N°7

COURRIER DES STATISTIQUES

Janvier 2022

Insee
Mesurer pour comprendre

STATISTIQUE PUBLIQUE

Rédaction en chef
Odile Rascol
Contribution
Insee : Mathias André, Benjamin Camus, Laura Castell, Frédéric Comte, Arnaud Degorre, Stéphane Legleye, Romain Lesur, Emmanuel L'Hour, Olivier Meslin, Benoît Rouppert, Éric Sigaud, Benoît Werquin
CGDD : Ronan Le Saout
Depp : Thierry Rocher
DEPSD : Amandine Schreiber
Santé publique France : François Beck
SSP : Hervé Le Grand
Directeur de la publication
Jean-Luc Tavernier
Directeur de la collection
Pascal Rivière

Rédaction
Pierre Glénat, Marine Le Roux, David Martineau, Odile Rascol, Pascal Rivière
Composition
Agence LATITUDE Nantes
5, rue Jacques Brel
« Les Reflets » Bâtiment A
44800 SAINT-HERBLAIN
0622/21 - 0623/21
02 51 25 06 06
www.agence-latitude.fr
Photo de couverture
Adobe Stock®

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

www.insee.fr

Courrier des statistiques N7

SOMMAIRE

PRÉSENTATION DU NUMÉRO

Musique Maestro !

Avec ce septième numéro, le Courrier des statistiques fête son troisième anniversaire. L'ambition de la revue est toujours d'aborder un large panel des grandes problématiques auxquelles se confronte la *statistique publique*. Elle propose une tonalité pédagogique, ouverte à la variété des sujets, des auteurs, des points de vue. Elle s'adresse au statisticien, débutant comme expert, et au citoyen, même si ce dernier pourra parfois trouver un peu aride sa lecture. Elle permet de témoigner de notre capacité collective à évoluer et innover tant sur les méthodes et outils, que sur des questions institutionnelles ou encore juridiques. La revue veille à rester attentive aux pratiques extérieures, en France comme à l'étranger, afin de se positionner vis-à-vis de notre communauté, de nourrir nos propres réflexions et de témoigner de nos travaux.

La revue, quoi de plus normal, s'intéresse particulièrement à la donnée et les différents moyens de la collecter ou de la produire. Elle traite, dans ce numéro, d'une évolution majeure de ces dernières années : l'intégration du multimode dans les enquêtes, à savoir la mise en musique, dans ces dernières, de modes complémentaires de collecte. Comme dans le numéro précédent, elle cherche aussi à savoir comment les statisticiens œuvrent de plus en plus pour tirer parti des gisements de données qui existent déjà mais qui restent insuffisamment valorisés. Enfin, on verra comment une grande opération statistique, le recensement agricole, se modernise.

La donnée est donc centrale dans ce numéro, elle qui constitue le cœur du métier du statisticien. Pour autant, ce dernier doit, aujourd'hui encore plus qu'hier, adopter une large gamme d'instruments. Au sein de son répertoire, la maîtrise des technologies de l'informatique dans les nuages (le *Cloud computing*) ou des développements informatiques les plus récents renforcent son autonomie, sa responsabilité et le champ des possibles avec lequel il pourra jouer et orchestrer ses traitements. La capacité à jouer de concert avec d'autres disciplines académiques est mise en évidence dans ce numéro, avec l'exemple de la psychométrie dans la mesure des compétences des élèves. Enfin, les statisticiens, issus d'organismes différents, doivent diffuser des données incontestables et cohérentes. Pour cela ils développent et adoptent un solfège commun pour ranger, classer et analyser les données : ce solfège, c'est la nomenclature. Le numéro N7 l'illustre avec la nouvelle nomenclature statistique des infractions, désormais commune à tous les acteurs de la statistique pénale.

En prélude de ce numéro, deux articles traitent de l'introduction d'internet et singulièrement des protocoles multimodes dans la collecte des enquêtes. ***François Beck, Laura Castell, Stéphane Legleye et Amandine Schreiber*** procèdent à une large revue de cette évolution : le multimode , mélange de collecte en face-à-face, par téléphone, papier ou internet, répond au contexte de difficulté croissante à contacter les ménages. Mais par ailleurs, il complexifie le processus tout entier, tant dans la définition du protocole de collecte que dans les traitements statistiques. ***Éric Sigaud et Benoît Werquin*** détaillent ensuite la nécessaire mise au diapason des différentes étapes de la collecte, pour les équipes en charge de la concevoir et de la mettre en œuvre. Cela requiert *ex-ante* de conceptualiser chacune de ses phases, depuis la conception et génération automatique des questionnaires jusqu'à leur traitement et la consolidation des données. Ils nous offrent, ce faisant, une nouvelle interprétation de la démarche de pilotage par les métadonnées actives.

Le recensement agricole de 2020 est à l'honneur de ce numéro dans un article de ***Hervé Le Grand***. Cette opération majeure de la statistique agricole est au cœur de son système d'information : elle en est en quelque sorte son métronome qui donne le rythme des autres enquêtes et garantit de produire des données cohérentes, au niveau national et européen. La dernière édition porte cinq innovations majeures, se répercutant sur les enquêtés, les enquêteurs et les statisticiens. Les données ont été collectées, pour la première fois, majoritairement par internet ou par téléphone. À l'heure où nous écrivons ces lignes, une sixième innovation, la *data visualisation*, apporte une touche finale à ce recensement décidément riche en évolutions.

Un trio de pirates, ***Frédéric Comte, Arnaud Degorre et Romain Lesur***, nous emmène voyager dans le SSPCloud. Environnement informatique d'aide à l'expérimentation sur les nouvelles méthodes de la *data science*, le SSPCloud est constitué d'un ensemble de ressources informatiques permettant de réaliser des prototypes, de tester des traitements statistiques et de s'approprier de nouvelles pratiques de travail. Avec le SSPCloud, le statisticien s'inscrit dans un courant d'inspiration de type FabLab lui permettant de valoriser les nouveaux gisements de données. Ici on compose à plusieurs et l'adoption de solutions *open source* garantit les possibilités de réutilisation. Le SSPCloud, c'est enfin un mélange fertile des deux univers professionnels : celui de la statistique et celui de l'informatique.

Dans la droite ligne de l'article sur le SSPCloud, ***Emmanuel L'Hour, Ronan Le Saout et Benoît Rouppert*** nous parlent du statisticien selfeur. Le métier de statisticien requiert aujourd'hui une bonne maîtrise des outils informatiques. Terminé le temps d'une interprétation *a cappella* de sa partition. Il doit coder selon les règles de l'art car si les programmes informatiques écrits doivent permettre de produire des résultats, ils sont, au-delà des livrables, des éléments de preuve de la qualité des traitements et doivent pouvoir être réutilisés pour d'autres travaux ou réinterprétés par d'autres statisticiens selfeurs.

Le sixième article de cette revue nous conduit dans l'univers de l'exploitation et de l'appariement de sources administratives exhaustives***. Mathias André et Olivier Meslin*** décrivent le projet qu'ils conduisent en duo : créer une nouvelle base statistique, afin de pouvoir étudier le patrimoine immobilier des ménages et le profil redistributif de la taxe foncière. Attachés à mobiliser les sources administratives disponibles, ils ont fait l'expérience des obstacles rencontrés avant d'accéder aux données ainsi que des chausses-trappes de l'appariement et du traitement statistique de fichiers venant d'univers différents, conçus pour d'autres usages. L'article décrit avec précision les étapes de l'épopée : il retient les bonnes pratiques qui leur ont permis d'aboutir à une base de production, laquelle vient désormais compléter le panorama de l'information statistique concernant le patrimoine des ménages. Il met également en évidence des enseignements de ce projet à destination du statisticien souhaitant mener des travaux sur des bases administratives.

Les travaux du statisticien prennent parfois un chemin singulier, du fait que l'objet que l'on souhaite mesurer ne pré-existe pas à l'opération de mesure elle-même. ***Thierry Rocher*** décrit les solutions déployées par le service statistique de l'Éducation, pour approcher une mesure des compétences des élèves. Il ouvre une fenêtre sur les concepts relevant du champ de la psychométrie. Il nous décrit les choix opérés – procédures et modélisations spécifiques – pour arriver à produire des évaluations standardisées des compétences des élèves. Ce faisant, il nous rappelle l'étendue des dispositifs, nationaux et internationaux, qui cherchent à définir des statistiques comparables, et utiles à tous les niveaux, depuis celui de l'enseignant et du chef d'établissement, jusqu'à celui d'un ministre.

Les dernières mesures de ce numéro nous entraînent du côté d'un défi assez rare dans la vie d'un statisticien : celui de l'élaboration d'une nomenclature statistique. Jusqu'alors, ministère de l'Intérieur et ministère de la Justice utilisaient des nomenclatures de diffusion différentes, ce qui empêchait de disposer de statistiques fines cohérentes tout au long de la filière pénale. ***Benjamin Camus*** décrit comment l'ONU a mis au point en 2015 une nomenclature internationale, s'affranchissant des différences de législations pénales, en choisissant une approche fondée sur le comportement de l'auteur de l'infraction. Celle-ci a donné le *tempo* et fourni l'occasion de lancer le chantier en France : un groupe de travail interministériel en a défini une déclinaison française ancrée sur une codification détaillée du droit pénal. En décembre 2021, la Nomenclature française des infractions a vu le jour : articulée avec la nomenclature internationale pour les grandes catégories, mais comprenant un niveau de détail plus pertinent dans le contexte français, elle porte en elle le germe de statistiques réconciliées.

Odile Rascol
Rédactrice en chef, Insee

LE MULTIMODE DANS LES ENQUÊTES AUPRÈS DES MÉNAGES

UNE COLLECTE MODERNISÉE, UN PROCESSUS COMPLEXIFIÉ

François Beck, Laura Castell**, Stéphane Legleye***, Amandine Schreiber*****

Avec l'introduction d'internet comme nouveau mode de collecte et les difficultés croissantes à contacter les ménages, l'évolution des enquêtes vers des protocoles multimodes est devenue une orientation stratégique forte pour les services statistiques publics. De nombreux protocoles multimodes sont envisageables, permettant de tirer profit des avantages de chaque mode de collecte en fonction des contraintes, de la thématique de l'enquête et des populations ciblées.

Cependant, le multimode complexifie le processus d'une enquête. Des adaptations sont nécessaires pour garantir la qualité des résultats : d'abord sur le questionnaire et sa durée, puis sur la définition du protocole de collecte et enfin sur les traitements statistiques d'agrégation des données en aval de la collecte. Par les efforts de standardisation et de simplification qu'elle impose avec notamment l'introduction de séquences auto-administrées, l'évolution vers le multimode constitue un véritable changement de paradigme pour les enquêtes ménages.

With the introduction of the Internet as a new data collection mode and the increasing difficulties in contacting households, the evolution of surveys towards mixed-mode protocols has become a strong strategic orientation for official statistical offices. Many mixed-mode protocols are possible, making it possible to take advantage of the benefits of each collection mode depending on the constraints, the survey topic and the target populations.

However, such protocols tend to make the survey process more complex. Adaptations are necessary to guarantee the quality of the results: firstly, the questionnaire and its duration, then the definition of the collection protocol and finally the statistical processing of the data after collection. Through the efforts of standardisation and simplification that it imposes, in particular with the introduction of self-administered sequences, the evolution towards mixed-mode surveys constitutes a real paradigm shift for household surveys.

* Directeur de la Prévention et de la promotion de la santé, Santé publique France, *francois.beck@santepubliquefrance.fr*

** Experte sur la collecte en ligne et la collecte multimode, Insee, *laura.castell@insee.fr*

*** Chef de la division Conditions de vie des ménages, Insee, *stephane.legleye@insee.fr*

**** Cheffe du Département des études, de la prospective, des statistiques et de la documentation, ministère de la Culture, *amandine.schreiber@culture.gouv.fr*

Les protocoles mobilisant plusieurs modes de collecte (face-à-face, téléphone, internet, etc.) ne sont pas une nouveauté dans les enquêtes de la statistique publique. Néanmoins, la généralisation de l'usage d'internet, l'évolution des pratiques des ménages en termes de communication et de participation, et la standardisation des outils de collecte[1] constituent des changements profonds, favorables au développement du multimode dans les enquêtes auprès des ménages. Cette évolution a été impulsée depuis les années deux-mille-dix dans les services statistiques publics, elle a connu tout récemment une accélération dans le contexte particulier de la crise sanitaire de la Covid-19. Plus que jamais, elle se dessine comme une orientation stratégique forte pour la statistique publique. Or le multimode complexifie les processus de collecte comme le traitement des données à l'aval, au risque d'entraîner des ruptures de séries. Au-delà des opportunités offertes par ce changement de paradigme pour les enquêtes auprès des ménages, il convient d'identifier les défis que cette transition pose aux statisticiens. Les choix effectués en France s'appuient sur une réflexion aux dimensions internationales[2].

LE MULTIMODE, UN CHANGEMENT DE PARADIGME POUR LES ENQUÊTES AUPRÈS DES MÉNAGES

Le caractère multimode d'une enquête concerne à la fois le protocole de contact et le protocole de collecte. Le premier contact consiste très souvent en une lettre postale, alors même que la collecte s'effectue rarement *via* un questionnaire papier[3].

Néanmoins, dans les faits, on parle d'enquête multimode lorsque le protocole de collecte implique le recours à plusieurs modes de recueil de l'information : en face-à-face avec un enquêteur, par téléphone, par internet, etc.

Différents paramètres entrent en jeu pour apprécier l'opportunité d'une collecte multimode : la population cible, les informations disponibles pour l'atteindre (base de sondage, données de contact) et la thématique. Plusieurs enquêtes de la statistique publique sont ainsi réalisées en multimode depuis de nombreuses années, par exemple :

- l'enquête auprès des ménages sur les *Technologies de l'information et de la communication* (TIC) est réalisée par internet, papier et téléphone depuis 2007 ;
- l'enquête *Emploi en continu* conjugue les modes face-à-face, téléphone (depuis 2003) et internet (depuis 2021[4]) selon les vagues d'interrogation ;
- l'enquête *Cadre de vie et sécurité* (CVS) réalisée par l'Insee en face-à-face de 2006 à 2021, collecte une partie de son questionnaire sous casque en auto-administré, etc.

1. Pour ce qui concerne l'impact du multimode sur l'organisation de la collecte et le travail des enquêteurs, voir l'article d'Éric Sigaud et Benoît Werquin sur « La mise en musique d'enquêtes multimodes » dans ce même numéro.
2. L'article s'appuie notamment sur les travaux séminaux de (Dillman *et alii*, 2014) aux États-Unis et sur ceux de (De Leeuw, 2018) en Europe, ainsi que sur un premier état des lieux réalisé dans le contexte de la statistique publique française (Razafindranovona, 2015).
3. De très rares enquêtes auprès des ménages continuent à utiliser le papier (*TIC*), voire commencent à le faire en complément d'internet (par exemple l'enquête Vécu et Ressenti en matière de Sécurité (*VRS*) conduite par le Service statistique ministériel de la sécurité intérieure (SSMSI)).
4. Une enquête sur les non-répondants de l'EEC a été menée à partir de 2007 par internet et papier et ses résultats ont été intégrés à la pondération de l'enquête en 2007. Voir l'article du Courrier des statistiques N6 consacré au nouveau protocole de collecte de l'enquête Emploi (Guillaumat-Tailliet et Tavan, 2021).

Depuis une dizaine d'années, l'Insee s'est lancé dans un plan d'expérimentations pour évaluer l'impact de l'introduction d'internet dans les protocoles d'enquêtes. Cette place croissante accordée au mode internet conduit à renouveler l'analyse sur les protocoles multimodes. De fait, dans les enquêtes de la statistique publique, internet est rarement envisagé comme unique mode de collecte, mais plutôt comme un mode complémentaire.

Trois raisons expliquent le passage à des enquêtes multimodes.

La première d'entre elles est la diminution tendancielle du taux de réponse aux enquêtes ménages[5], bien que la situation française s'avère moins problématique que dans d'autres pays européens (Luiten *et alii,* 2020). Or la diversification des moyens de contacter les personnes accroît les chances de les atteindre et augmente *a priori* le taux de collecte, même si le taux de réponse n'est pas en soi un indicateur définitif de qualité (Groves et Peytcheva, 2008). Ainsi, des travaux récents (Cornesse et Bosnjak, 2018) ont montré que les enquêtes multimodes sont de meilleure qualité en termes de représentativité et que les enquêtes exclusivement conduites sur internet accusent de faibles taux de participation et d'importants effets de sélection[6].

> *« La diversification des moyens de contacter les personnes accroît les chances de les atteindre. »*

De manière générale, le face-à-face apparaît supérieur au téléphone et à internet pour la représentativité de certains segments de la population, notamment les moins favorisés ou éduqués, les immigrés, les personnes peu à l'aise dans la langue, ou encore les personnes vivant en milieu urbain (Buelens et Van den Brakel, 2010).

L'évolution de l'équipement et des pratiques des ménages constitue également un contexte favorable au multimode. Au cours de la dernière décennie, le taux d'équipement des foyers en accès internet (matériel et abonnement) a connu une croissance spectaculaire en France : internet est désormais d'utilisation courante[7]. Une telle situation rapproche désormais la France des pays dans lesquels l'implantation d'internet a été plus ancienne, tels que les Pays-Bas, les pays nordiques, le Royaume-Uni et l'Allemagne. Les conditions sont donc de plus en plus favorables pour recourir à ce moyen pour enquêter les ménages. Les attentes des ménages eux-mêmes paraissent par ailleurs fortes : certains ménages enquêtés comprennent de moins en moins qu'un tel service ne soit pas offert alors que les progrès récents en matière de démarches administratives sur le *web* sont très visibles.

Certaines personnes n'accèdent par ailleurs à internet que par le biais d'un *smartphone :* plus de 60 % de la population européenne dispose d'un accès à internet *via* un téléphone mobile. Ceci souligne l'enjeu de proposer une interface et un questionnaire conçus d'emblée pour ce type de support *(Mobile first)* et dont le format des pages s'adapte de manière ergonomique (*Responsive design*) ***(encadré 1)***.

5. Par exemple, le taux de réponse a baissé pour les enquêtes en face-à-face Cadre de vie et sécurité (de 72 % à 66 % entre 2012 et 2021) et *SRCV* (de 85 % à 80 % entre 2010 et 2019).
6. C'est un tel constat qui avait conduit l'Insee à écarter la solution d'enquêtes monomodes *via* internet.
7. Entre 2007 et 2021, le taux d'équipement en accès à Internet à domicile est passé de 54 % à 91 % et le taux d'utilisation quotidienne ou quasi-quotidienne est passé de 36 % à 72 % (enquête sur les TIC auprès des ménages).

Les enquêtes téléphoniques, qui ont été très utilisées dans les années quatre-vingt-dix par les instituts de sondages et une partie de la statistique publique, se heurtent en revanche à des taux de participation plus faibles désormais, à cause de la lassitude engendrée par les incessantes démarches marketing entreprises par ce moyen[8]. Le téléphone fixe, qui était le socle de ces enquêtes jusqu'aux années 2010, est désormais un outil qui est très rarement décroché (voire activé) par les nouvelles générations[9].

> *« Les contraintes budgétaires conduisent à s'intéresser aux coûts relatifs des différents modes de collecte. »*

Les contraintes budgétaires conduisent par ailleurs à s'intéresser aux coûts relatifs des différents modes de collecte : les entretiens avec les enquêteurs (en face-à-face ou téléphone) étant les plus coûteux, on préférera les réserver aux thématiques ou aux populations pour lesquelles leur expérience est indispensable.

Chaque mode de collecte ayant ses forces et ses faiblesses en termes de biais de couverture et de sélection, il importe de tirer profit des avantages de chacun dans l'élaboration des protocoles des enquêtes.

Encadré 1. Vers une conception *Mobile First*

Avec les tablettes puis les téléphones mobiles, le besoin de générer des questionnaires « agnostiques », c'est-à-dire à même d'offrir la meilleure ergonomie possible quels que soient la situation d'usage et l'appareil utilisé, est devenu crucial, en particulier dans les enquêtes auprès des jeunes.

De fait, il n'est pas possible d'interdire aux enquêtés de répondre sur leur *smartphone*, mais cela impose souvent des efforts de zoom et de « défilement » (*scrolling*) pour le répondant. Or si l'ergonomie du questionnaire n'est pas adaptée à cet outil, le risque est fort d'une qualité de remplissage moindre : réponses systématiques, non-réponses partielles ou abandon en cours de questionnaire.

Le questionnaire et les questions doivent de ce fait être aussi courts que possible et minimiser l'effort de *scrolling* (Couper et Peterson, 2017). Il est par ailleurs recommandé d'utiliser les « boutons radio » (ou case d'option) en adaptant leur taille au petit écran et de rendre l'ensemble de la modalité sélectionnable (pas seulement le bouton radio). Ces auteurs incitent même à aller vers les questions ouvertes plutôt que vers de longues listes de modalités.

Les données les plus récentes indiquent que lorsqu'une enquête est conçue de manière appropriée pour les *smartphones*, la qualité des données peut être aussi bonne que celle des enquêtes en ligne réalisées sur ordinateur fixe (Antoun *et alii*, 2017). Il y a donc un enjeu fort à se situer dans une logique d'optimisation des questionnaires pour les supports mobiles, et notamment le *smartphone*.

Note : Mobile First *est un concept de* web design *optimisé pour le mobile qui va au-delà du* Responsive web design. *Il consiste à concevoir un site en mettant la priorité sur la version mobile et en adaptant progressivement la conception* web *pour les écrans plus larges, à rebours de l'approche la plus répandue précédemment qui consistait à dégrader progressivement un site* web *pour l'adapter à un affichage sur téléphone portable.*

8. Le taux de réponse à la première vague de l'enquête *Camme* est passé de 63 % à 53 % entre 2013 et 2020.
9. Selon TIC 2021, seuls 66 % des 15-29 ans disposent d'un téléphone fixe : 42 % y prennent tous les appels, 21 % ne décrochent jamais et 23 % ne décrochent que lorsqu'ils connaissent le numéro appelant. Par comparaison, 99 % sont équipés d'un portable : 65 % y prennent tous les appels, 2 % ne décrochent jamais, et 31 % décrochent seulement lorsque le numéro appelant est connu.

DES MODES DE COLLECTE QUI S'ADAPTENT AU CONTEXTE DE CHAQUE ENQUÊTE

L'analyse des biais de mesure – c'est-à-dire de l'impact du mode de collecte sur la réponse des enquêtés – oppose les modes auto-administrés (papier et internet) et les modes avec enquêteurs (face-à-face et téléphone). Les deux biais principaux induits par la présence ou l'absence d'un enquêteur sont le biais de désirabilité sociale et le *satisficing* :

- la **désirabilité sociale** désigne le fait que certains enquêtés peuvent être amenés à fournir une réponse donnant une image valorisante d'eux-mêmes ou dont ils croient qu'elle satisfait les attentes normatives. Elle survient plus fréquemment lors d'une interrogation menée par un enquêteur, et pour des questions d'opinion ou des questions sensibles ou intimes, telles que la consommation d'alcool ou de drogues illicites (Beck et Peretti-Watel, 2001) ou les comportements sexuels (Legleye et Charrance, 2021) ;
- le biais de ***satisficing***[10] désigne le fait que certains enquêtés renoncent à fournir l'effort nécessaire pour renseigner la réponse exacte (Krosnick, 1991). Ce phénomène survient plus fréquemment avec des questionnaires auto-administrés, difficiles, longs ou répétitifs pour lesquels le niveau de concentration et d'engagement sont moindres que face à un enquêteur. Il peut se manifester par des réponses approximatives, par une tendance à choisir plutôt les premières modalités proposées (*primacy effect*) ou les dernières (*recency effect*), à ne pas recourir à des documents, à arrondir les réponses chiffrées, à choisir des modalités médianes ou encore par de la non-réponse partielle, voire un abandon du questionnaire. Le moindre niveau de concentration associé aux modes auto-administrés impose une durée nettement raccourcie par rapport au téléphone, et encore plus comparé au face-à-face.

Par ailleurs, les modes auto-administrés souffrent de taux de réponse inférieurs et d'un risque d'auto-sélection par rapport à la thématique de l'enquête plus important, puisque la réponse suppose une démarche plus pro-active de la part des enquêtés.

Ces effets peuvent varier suivant les caractéristiques des répondants. Ainsi, les jeunes les plus diplômés, les personnes ayant des compétences et une appétence pour le numérique, pourront trouver naturelle la réponse par internet : leurs réponses sur ce mode seront peu affectées par le *satisficing*. À l'inverse, ces populations peuvent être plus difficiles à enquêter par téléphone ou en face-à-face et leurs réponses peuvent se trouver affectées par une forte désirabilité sociale en présence d'un enquêteur. De plus, les coordonnées de contact disponibles sont variables : si l'adresse postale est connue pour tous les ménages, ce n'est pas toujours le cas du numéro de téléphone fixe voire mobile et de l'adresse courriel. La disponibilité de ces différentes coordonnées de contact diffère dans la population, ainsi que la faculté à contacter effectivement les personnes, et leur appétence à répondre dans les différents modes de collecte.

Une étude conduite auprès d'une population de migrants a ainsi montré que le papier était le plus approprié pour les hommes, et le téléphone plus adapté pour les femmes – notamment pour limiter l'impact de la présence du conjoint (Kappelhof, 2015).

Chaque mode de collecte comporte donc ses avantages et ses inconvénients en fonction de la thématique, des populations ciblées et des modalités de contact disponibles ***(figure 1)***.

10. Ce mot-valise est formé de *satisfying* (satisfaisant) et *sufficing* (suffisant) : il apparaît en 1957 dans le discours du sociologue, économiste et psychologue Herbert Simon dans le cadre de ses recherches sur le comportement humain.

Figure 1. À chaque mode ses caractéristiques

	AVANTAGES	RISQUES OU INCONVÉNIENTS
Internet	Coûts modérés Remontée d'informations rapide Possibilité d'introduire des filtres complexes Disponibilité des enquêtés Recours à des contenus variés (vidéos, son, images) Adapté aux 15-30 ans et aux populations connectées	Non-réponse, auto-sélection Biais de *satisficing* Problèmes d'affichage Défaut de couverture Nécessite une bonne compréhension de la langue écrite Risque d'incompatibilité informatique
Papier	Disponibilité des enquêtés Couverture large Adapté aux 75 ans et plus	Remontée d'informations lente Aucune interactivité Impossibilité d'introduire des filtres complexes Non-réponse partielle Non-réponse, auto-sélection Nécessite une bonne compréhension de la langue écrite Biais de *satisficing*
Téléphone	Intermédiation permettant un meilleur respect des consignes et des concepts Possibilité d'introduire des filtres complexes	Pas de recours à des contenus variés (vidéos, son images) Défaut de couverture Peu adapté pour les 15-30 ans Biais de désirabilité sociale
Face-à-face	Taux de réponse élevé Intermédiation permettant un meilleur respect des consignes et des concepts Possibilité d'introduire des filtres complexes Recours à des contenus variés (vidéos, son, images) Couverture large	Coûts élevés Biais de désirabilité sociale Problème de disponibilité des enquêtés

UNE MULTITUDE DE MULTIMODES : CONCURRENTIEL, SÉQUENTIEL, MIXTE…

Les protocoles construits sur la base des différentes possibilités offertes par les modes de collecte prennent des formes très variées ***(figure 2)***.

Les différents modes de collecte peuvent être soit proposés en même temps aux enquêtés, les modes étant en quelque sorte mis en concurrence (on parle de **multimode concurrentiel**), soit proposés successivement, la proposition d'une alternative n'étant faite qu'aux non-répondants de la phase antérieure (on parle de **multimode séquentiel**). En général, le multimode séquentiel consiste à proposer d'abord les modes de collecte les moins coûteux.

La plupart des protocoles débouchent sur une période où plusieurs modes sont en concurrence : c'est le cas des séquentiels *web*/téléphone pour lesquels la possibilité de répondre sur internet reste offerte même lorsque la phase d'enquête téléphonique a commencé. On parle alors de **protocole mixte.**

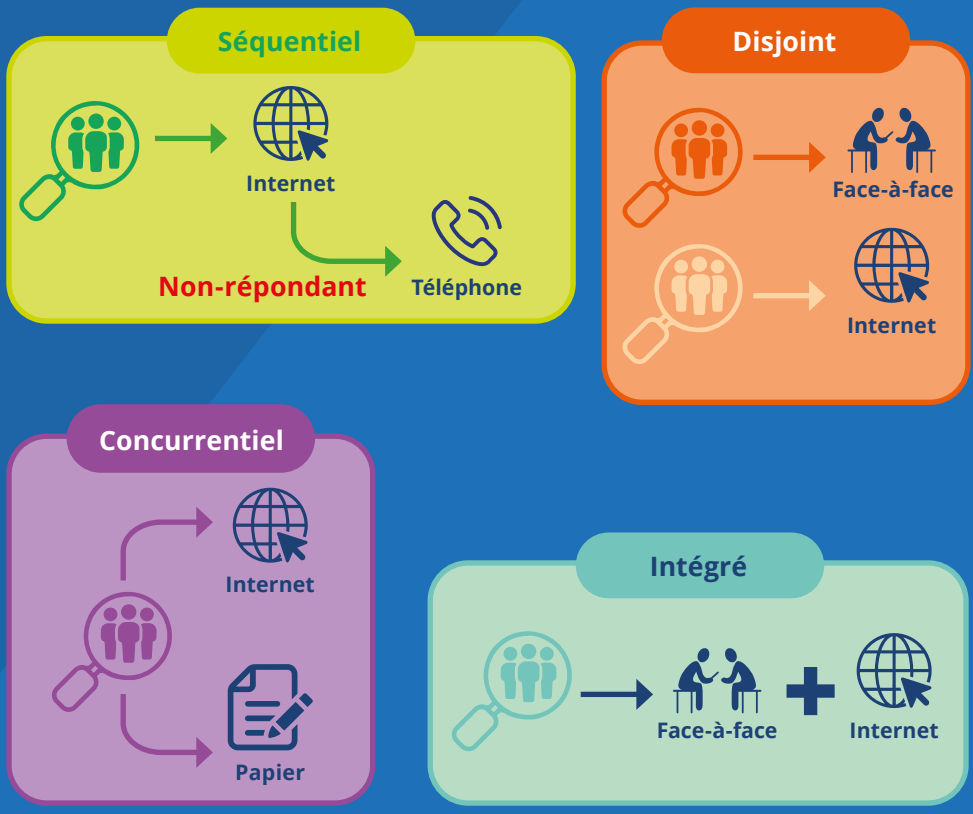
La limite principale du protocole séquentiel est qu'il incite certains enquêtés à répondre sur un mode qui n'est pas celui pour lequel ils ont le plus d'appétence et que cela pourrait dégrader la qualité de leurs réponses (sujet sensible, aversion pour le mode, etc.). Il entraîne aussi un risque que des enquêtés, qui auraient accepté de répondre s'ils avaient été sollicités par un enquêteur en premier lieu, se trouvent dans de moins bonnes dispositions lorsque ce contact intervient après qu'ils aient refusé de répondre par un premier mode de collecte auto-administré : persistance dans le refus initial, réticence marquée à fournir les efforts nécessaires à la qualité des réponses, etc. Par ailleurs, sauf à diminuer les chances de réussite de chacun des modes, il impose une durée de collecte minimale à chaque étape. Un temps de terrain court incite donc à proposer un multimode concurrentiel ou mixte plutôt que séquentiel.

« Le multimode séquentiel reste le protocole jugé le plus efficace. »

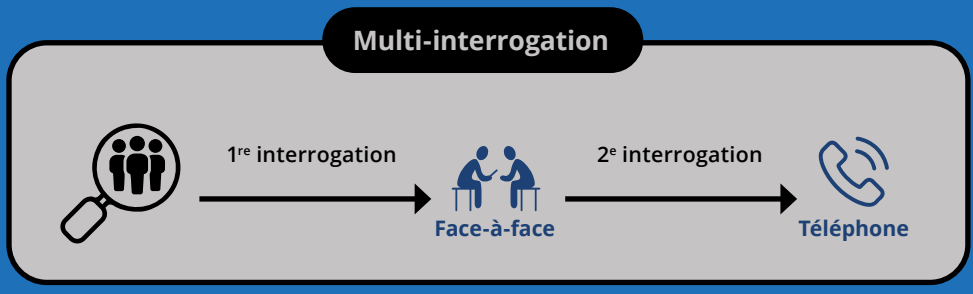
Le multimode séquentiel reste néanmoins, en général, le protocole jugé le plus efficace, en particulier sur le plan financier (Dillman *et alii*, 2014). Par ailleurs, le fait de laisser le choix du mode de réponse à l'enquêté peut être vécu comme une charge cognitive supplémentaire (Schwartz, 2005), à un moment où il attend souvent que l'enquêteur, ou à défaut le protocole, le guide pour s'acquitter au plus vite de sa tâche. Laisser le choix du mode s'avère surtout intéressant dans les enquêtes en panel ou avec des réinterrogations, l'enquêté pouvant exprimer sa préférence au terme de la première vague et renseigner des coordonnées de contact facilitant le recours à un mode particulier pour les vagues suivantes.

Figure 2. Avec le multimode, la collecte gagne en souplesse

Les différents modes (face-à-face, internet, téléphone, papier) peuvent se combiner pour des protocoles de collecte adaptés aux populations enquêtées…



… et adaptés au rythme et aux thématiques des enquêtes



Nota : les combinaisons ci-dessus sont des exemples de protocoles multimode, ce ne sont pas les seules possibles.

… OU INTÉGRÉ (POUR LES QUESTIONS SENSIBLES)

Le **multimode intégré** désigne le recours à un mode de collecte complémentaire en cours de collecte pour l'ensemble des enquêtés. C'est par exemple le cas lorsqu'un enquêteur présent au domicile laisse l'enquêté répondre seul à une partie du questionnaire sur un ordinateur, comme dans l'enquête Cadre de vie et sécurité (CVS) où il peut utiliser un casque et répondre seul *via* le clavier.

Ce type de protocole apparaît particulièrement bien adapté aux questions sensibles (Turner *et alii*, 1998). Il permet sans doute de tirer au mieux profit des différents modes sollicités, mais c'est en général le plus coûteux.

Dans les enquêtes par téléphone, il peut aussi être demandé de basculer complètement sur internet pour remplir le questionnaire (Legleye et Charrance, 2021). Pour une série de questions sensibles, l'enquêteur peut éventuellement laisser sa place à un serveur vocal interactif (SVI[11]) avec lequel on espère que le biais de désirabilité sociale sera diminué (Kreuter *et alii*, 2008). Cela amène à des abandons assez nombreux durant la collecte, car l'expérience n'est probablement pas jugée très agréable par les enquêtés qui perdent la motivation induite par la présence sonore de l'enquêteur et peuvent couper court à la poursuite du questionnaire. Il semble donc préférable de ne recourir à ce mode de collecte que pour une courte série de questions.

DES PROTOCOLES ADAPTABLES AU RYTHME DES ENQUÊTES…

Dans les enquêtes longitudinales, il est possible d'adapter les modes de collecte selon les vagues d'interrogation. La présence d'un enquêteur est souvent nécessaire pour initier le processus de collecte. En revanche, les vagues ultérieures (réinterrogations) peuvent souvent être réalisées sur des modes moins coûteux et moins invasifs, comme le téléphone ou internet. En France, l'enquête *Emploi en continu* et l'enquête *Loyers et charges* reposent depuis de nombreuses années sur ce type de protocole, enchaînant une première vague d'interrogation en face-à-face, puis des réinterrogations au téléphone avant de clore le processus en face-à-face. Depuis janvier 2021, un multimode séquentiel *web*/téléphone est proposé en réinterrogation dans l'enquête Emploi (Guillaumat-Tailliet et Tavan, 2021).

Il existe par ailleurs des protocoles en deux phases reposant sur une double enquête. La première consiste en une enquête courte, menée auprès d'un nombre très important d'individus, basée sur des questions visant à identifier une population d'intérêt spécifique (personnes en situation de handicap, victimes de violences, minorités sexuelles, etc.). Étant donné la taille de l'échantillon, il peut être avantageux de mener un multimode séquentiel, s'appuyant sur des modes de collecte moins coûteux : internet, puis papier, puis téléphone, éventuellement pour une partie des non-répondants. La seconde phase, qui peut également être multimode, est réalisée auprès d'une sélection de répondants de la première phase, dans laquelle certaines caractéristiques d'intérêt sont surreprésentées grâce aux informations de l'enquête filtre. Elle peut être multimode elle aussi. Un tel protocole a été retenu avec succès pour l'enquête *Vie quotidienne et santé (VQS)/Autonomie* du service statistique du ministère des Solidarités et de la Santé (Drees) et pour l'enquête *Genre et Sécurité (Genese)* du service statistique du ministère de l'Intérieur (SSMSI).

11. Le SVI est utilisé pour l'assistance aux enquêtés, les relances, ou en automate posant les questions et recueillant les réponses. Sur les sujets sensibles, il donne de meilleurs résultats que le téléphone, sans atteindre le niveau des questionnaires auto-administrés qui restent privilégiés pour diminuer le biais de désirabilité sociale.

Enfin, la pénétration croissante des *smartphones* offre désormais aux chercheurs la possibilité de collecter des données auprès de leurs utilisateurs par le biais d'une collecte de données passives *via* l'historique et l'utilisation des applications (géolocalisation, mouvements physiques, comportement en ligne, etc.). Une partie de la collecte peut même s'effectuer sur *smartphone* (carnet d'emploi du temps, interrogations courtes à intervalles réguliers par exemple). Cette opportunité ne doit pas cacher les difficultés que représente, pour certains enquêtés, le téléchargement d'une application sur leur *smartphone*, ni les enjeux soulevés en matière de consentement au recueil et à l'exploitation de telles données.

Au final, les contraintes de chaque projet d'enquête peuvent amener le concepteur à orienter ses choix parmi les possibles évoqués ci-dessus (voir les choix opérés sur certaines grandes enquêtes européennes en ***encadré 2***).

Encadré 2. Les choix de protocole effectués par les différents pays sur les enquêtes sociales européennes

Dans le cadre du projet européen *Mixed Mode Designs in Social Surveys* (MIMOD*) piloté par l'institut italien Istat de janvier 2018 à avril 2019, un questionnaire portant sur quelques grandes enquêtes sociales européennes a été réalisé pour établir un état des lieux des pratiques en matière de multimode.

Une majorité des pays mène désormais des enquêtes en multimode, ce qui n'était pas encore le cas il y a cinq ans. La variété des protocoles apparaît très grande et, pour l'instant, la majorité d'entre eux (57 %) n'inclut pas de mode internet.

Selon les experts de chaque pays, certaines enquêtes s'avèrent plus adaptées au mode internet. Pour la première interrogation de l'enquête Emploi, c'est l'absence d'enquêteur qui semble le plus préjudiciable, tandis que pour celle de SRCV**, c'est la longueur du questionnaire. Les vagues de réinterrogation sont très majoritairement identifiées comme compatibles avec un recueil sur internet, en particulier celles de l'enquête Emploi.

Bien sûr, les contextes nationaux sont parfois très structurants : la disponibilité de registres de population dans les pays nordiques par exemple, peut permettre d'alléger nettement le questionnement lorsque certaines informations y figurent, ce qui favorise le recours à internet. Les pays qui s'appuient beaucoup sur le téléphone ont des systèmes connectés qui permettent de récupérer les paradonnées (nombre de tentatives, date et heure de chaque tentative, etc.) même lorsque les enquêteurs appellent de chez eux, grâce à des systèmes de plateau téléphonique virtuel.

** L'ensemble des rapports des différents groupes de travail est disponible sur le site d'Istat (https://www.istat.it/en/archivio/226140).*

*** Statistiques sur les ressources et conditions de vie. Le système statistique EU-SILC a pour vocation de permettre la production d'indicateurs structurels sur la répartition des revenus, de la pauvreté et de l'exclusion comparables pour les pays de l'Union européenne.*

… ET ADAPTABLES AU PROFIL DES POPULATIONS ENQUÊTÉES

Ces différents types de protocoles multimodes peuvent ne pas s'appliquer de la même manière à toutes les unités échantillonnées d'une même enquête. Ainsi, depuis quelques années, des protocoles permettant d'adapter la stratégie de recueil des données au profil sociodémographique des enquêtés ont été conceptualisés et commencent à être mis en œuvre (Chun *et alii*, 2018). L'***Adaptive design*** utilise ainsi l'information disponible en amont de la collecte des données pour affecter le mode de collecte *a priori* le plus efficace à chaque unité enquêtée, que ce soit en termes de taux de réponse ou de limitation des biais de mesure.

Par exemple, dans le cas des enquêtes sur les usages d'alcool, de tabac et de drogues illicites, sexe et âge apparaissent comme des caractéristiques clivantes dans le choix du mode de collecte (auto-administré ou bien téléphone) (Beck *et alii*, 2014). Une des différences entre l'*Adaptive design* et les autres formes de multimode est que la préférence pour un mode est fixée d'emblée par le concepteur (contrairement au concurrentiel où c'est l'enquêté qui le choisit) et qu'elle est assumée (tel mode est plus adapté pour telle sous-population).

La longueur du questionnaire étant un facteur limitant l'intérêt du recours à internet, on peut aussi cibler une population ayant une bonne capacité à répondre sur internet sans que la qualité soit dégradée (les individus jeunes, issus de ménages de 1 ou 2 personnes, n'ayant pas de ce fait un ménage trop complexe à décrire), tandis que le reste de la population est interrogé par téléphone ou en face-à-face.

Il existe également des protocoles plus complexes consistant à mobiliser l'information disponible pendant la collecte (données recueillies par questionnaire ou grâce aux paradonnées[12]) afin d'ajuster la stratégie de collecte pour le reste de l'échantillon : il s'agit du ***Responsive design***. Il peut ainsi être décidé, en cours de collecte, de porter l'effort sur un profil particulier en le ciblant directement ou indirectement en recourant à un mode de collecte particulier. Avant d'être utilisé dans le cadre des enquêtes de la statistique publique française, ce type de protocole nécessiterait d'être testé pour en évaluer l'opportunité.

LE QUESTIONNAIRE : PIERRE ANGULAIRE D'UNE ENQUÊTE MULTIMODE RÉUSSIE

La mise en œuvre d'un protocole multimode nécessite d'adapter toutes les étapes de conception et de production d'une enquête afin de proposer une ergonomie qui limite au maximum les effets de mode. Cet effort est grandement facilité par les outils de conception et de génération des supports de collecte fondés sur les métadonnées actives[13].

Le premier défi est l'adaptation du questionnaire aux différents modes envisagés : doit-on chercher à tout prix à garder la même structure et les mêmes questions pour tous les modes ou au contraire chercher à faire au mieux au sein de chaque mode ? Les recommandations internationales penchent nettement en faveur de la première solution (De Leeuw, 2018), même si les travaux les plus récents invitent plutôt à une troisième voie cherchant à optimiser chacun des questionnaires tout en réduisant les écarts inter-mode au strict minimum.

12. Les informations, non visées par la collecte, qui peuvent être recueillies lors de celle-ci, par exemple : si une unité appartient à l'échantillon, si elle a répondu, le nombre de tentatives pour la joindre, le mode de collecte, etc.
13. La notion de métadonnées actives a été abordée dans le numéro N3 de la revue. Voir également l'article d'Éric Sigaud et Benoît Werquin dans ce même numéro.

Cette **approche omnimode (ou agnostique)**, dans laquelle s'inscrivent plusieurs instituts nationaux de statistique (Pays-Bas, Grande-Bretagne), consiste à concevoir le questionnaire indépendamment du mode de collecte et à opérer dans un second temps les adaptations nécessaires dans chacun des modes afin de limiter au maximum les effets de mode. Les grandes recommandations en la matière sont les suivantes :

- adopter une approche omnimode : les questions ne doivent pas être attachées à un mode particulier ;
- s'en tenir à une seule question par écran pour que le répondant reste concentré sur une question à la fois, à moins qu'il ne s'agisse de questions filtres ou de suivi ;
- répéter les textes des questions sur la page suivante pour donner à nouveau le stimulus complet ;
- inclure systématiquement toutes les options de réponse présentes à l'écran dans les questions à lire à haute voix par les enquêteurs ;
- éviter les tableaux et grilles pour privilégier des séries de questions dans tous les modes ;
- réduire au minimum les instructions et les explications, et les présenter de manière similaire dans les différents modes.

Les adaptations nécessaires portent notamment sur les choix de charte graphique, de taille et de position des boutons / cases à cocher, sur l'affichage systématique ou en seconde intention des options « ne sait pas » / « ne veut pas répondre » ou sur leur absence systématique.

De manière générale, le passage au multimode, et notamment à l'introduction de modes de collecte auto-administrés, constitue une bonne occasion de se re-questionner sur ce que l'on cherche à mesurer et ainsi revoir les complexités qui ont pu être introduites au fil des éditions d'enquêtes. Si la compréhension des concepts et des questions est endossée par l'enquêteur, elle ne peut être déportée sur les enquêtés eux-mêmes, au risque d'obtenir de la non-réponse ou des réponses incorrectes.

> *« Le passage au multimode constitue une bonne occasion de se re-questionner sur ce qu'on cherche à mesurer. »*

Le recours à l'auto-administré impose ainsi d'utiliser des formulations de questions simples et de fournir des exemples éclairants pour les enquêtés, en particulier en population générale, car aucune reformulation ni aide ne pourra être apportée par un enquêteur (Koumarianos et Schreiber, 2021). La présentation des consignes, aides et exemples doit faire l'objet d'une réflexion *ad hoc* afin qu'elle soit attractive et judicieuse. Les enquêteurs assurent une mission importante de contrôle des saisies et de respect des consignes durant la passation : lorsqu'ils sont absents, tout ce travail de vérification est transféré aux statisticiens en charge de l'apurement des données, avec un risque important d'erreur en raison de l'impossibilité de demander des précisions sur la saisie *a posteriori*.

AJUSTER LA DURÉE DE QUESTIONNEMENT AUX MODES LES PLUS CONTRAIGNANTS

Répondre à une enquête est toujours un exercice difficile, mais la participation à une enquête et la qualité des informations recueillies dépendent de la longueur et de la complexité du questionnaire. Si la présence d'un enquêteur, en particulier en face-à-face, ouvre des possibilités pour des questionnements longs et complexes, l'auto-administré en revanche renvoie très vite aux limites de ce que peut et souhaite accomplir la frange de la population la plus en difficulté de littératie ou la moins intéressée ou concernée par la thématique. Au-delà d'une certaine durée qui dépend du mode de collecte, le risque de *satisficing* s'accroît et avec lui de réponse aléatoire, d'abandon, de recours plus fréquent au « ne sait pas », de réponses plus courtes aux questions ouvertes et globalement d'une baisse de la qualité (Galesic et Bosnjak, 2009). Dans le cas du multimode, la longueur du questionnaire est fixée par le mode de collecte qui génère le plus de contraintes chez l'enquêté.

La durée maximale recommandée pour le téléphone est de l'ordre d'une demi-heure (Roberts *et alii*, 2010), mais la présence d'un enquêteur conduit fréquemment les concepteurs à dépasser cette limite, sans toutefois le plus souvent excéder trois quarts d'heure. Pour les enquêtes *web*, il est grandement recommandé de ne pas dépasser 20 minutes, l'idéal étant de proposer un questionnaire dont le temps médian de réponse est de 10 minutes (Revilla et Ochoa, 2017).

Pour parvenir à limiter le temps de questionnement, plusieurs solutions sont envisageables si la réduction du questionnaire n'est pas suffisante. Une première consiste à administrer le questionnaire en plusieurs séquences, ce qui revient à « panéliser » l'enquête. Cela n'est pas sans risque ni coût : attrition, gestion des déménagements ou des éclatements du ménage, étalement de la référence temporelle, complexification de l'aval statistique, etc. Les protocoles d'*Adaptive design* cherchent pour leur part à répondre à ces problématiques en ajustant le questionnaire de façon à proposer à chaque enquêté une durée de passation optimale du point de vue de la qualité des informations recueillies. Ainsi, il peut être envisagé de ne pas poser toutes les questions à tous les enquêtés, mais par exemple de déterminer un tronc commun de questions, puis d'affecter aléatoirement des blocs de questions différents à plusieurs sous-échantillons (Beck et Richard, 2013). La question des incitations financières, en cas de questionnaire plus difficile ou plus long, est évoquée dans l'***encadré 3.***

LE DÉFI DE L'AGRÉGATION DES DONNÉES…

Dans les enquêtes multimodes, l'agrégation de données issues de modes de collecte différents nécessite certaines précautions. De fait, ce qui est observé sur un mode peut ne pas être directement comparable à ce qui est observé sur un autre. On appelle cette différence **un effet de mode.** Il peut être décomposé en deux parties :

- **le biais de sélection :** un mode de collecte peut entraîner une sous-représentation de certaines catégories de population, car leur propension à répondre dépend du mode qui leur est proposé. Ainsi, la composition des répondants diffère d'un mode à l'autre, de sorte que la moyenne des variables d'intérêt s'en trouve affectée : on parle alors de biais de composition ou de biais de sélection ;

- **le biais de mesure :** un mode de collecte peut induire des réponses qui diffèrent de celles apportées sur d'autres modes, car la situation d'enquête va influer sur la façon de répondre des enquêtés (désirabilité sociale, *satisficing*, biais de mémoire, difficulté face à une longue liste d'items en auto-administré ou au téléphone, etc.). Le biais de mesure est la conséquence directe du mode de collecte sur la réponse d'un individu donné (ou d'individus semblables, puisqu'on observe rarement la réponse d'un même individu sur deux modes différents).

> *« Il est important de neutraliser tous les biais de composition pour évaluer proprement ce qui ne relève que du mode de collecte, c'est-à-dire d'évaluer le biais de mesure. »*

L'existence d'un biais de sélection découle d'un des objectifs du multimode qui est d'augmenter les possibilités de répondre des enquêtés. Ce n'est en soi pas un problème, à partir du moment où il n'y a pas d'effet de mesure et que l'ensemble des répondants, quel que soit leur mode de collecte, est représentatif de la population cible.

Il est donc important de neutraliser tous les biais de composition pour évaluer proprement ce qui ne relève que du mode de collecte, c'est-à-dire d'évaluer le biais de mesure.

Pour cela, on utilise des méthodes permettant de rendre parfaitement comparables les répondants à chacun des modes de collecte. Cependant, dans la pratique, il peut être difficile de distinguer les biais résultants de ces deux effets, sélection et mesure (Klausch, 2014). De fait, toutes ces méthodes reposent sur l'hypothèse forte qu'il n'existe pas de biais de sélection non-ignorable, c'est-à-dire de facteur non observé qui influe à la fois sur

Encadré 3. Des leviers pour accroître la participation aux enquêtes –

Suivant les travaux théoriques de (Groves *et alii*, 2000) sur les déterminants de la participation à une enquête, l'acceptation se trouve influencée par : la nature de l'organisme commanditaire de l'étude, l'habileté de l'enquêté avec un *smartphone*, la durée de la période de collecte des données, etc., mais avant tout par le recours à une **incitation**.

Les études menées au niveau international au cours des deux dernières décennies ont montré qu'une incitation financière était la solution la plus efficace pour accroître la participation aux enquêtes ménages (Singer et Ye, 2012 ; Edwards *et alii*, 2009), plus que les « cadeaux ». Elle est notamment utilisée par certains instituts statistiques lorsque la charge pesant sur les répondants est particulièrement lourde (par la complexité ou par la durée du questionnaire), par exemple lorsque l'individu ou le ménage est sollicité pour participer à un panel, ou lorsque la durée de questionnaire dépasse trois quarts d'heure. Elle permet de limiter l'attrition. Notons que les expérimentations de la sorte sont rares en France (Legleye *et alii*, 2014 ; 2016) et que leur mise en œuvre par un organisme de la statistique publique, pour des enquêtes la plupart du temps obligatoires, pose des questions de plusieurs ordres, notamment budgétaire (renchérissement) et déontologique (par exemple : faut-il monétiser la participation civique à des projets d'intérêt social ? Y a-t-il rupture d'égalité en cas de ciblage de sous-populations particulières ? Faut-il étendre l'indemnisation à toutes les enquêtes ? etc.). Un tel choix complexifie les procédures de production et nécessite une évaluation stricte (arbitrage coût-efficacité en termes de représentativité, gain en variance ou en biais). Enfin, des effets à long terme pourraient surgir : altération de l'image et de la notoriété des instituts ; l'appréciation de l'importance d'une enquête à l'aune de l'indemnisation prévue pour y répondre pourrait pousser à la concurrence inter-enquêtes et à une hausse tendancielle des indemnisations, etc.

la participation et sur la valeur des variables d'intérêt, et qui soit peu corrélé aux facteurs observés par lesquels on contrôle (Rubin et Little, 2002). Dans ce cas, surviennent alors des problèmes de sélection endogène particulièrement complexes à traiter (Lee, 2009 ; Castell et Sillard, 2021 ; Heckman, 1979).

Les travaux menés sur différents jeux de données issus d'enquêtes françaises ont permis de constater des écarts entre les valeurs obtenues par des enquêtes de référence en face-à-face et des enquêtes *web* posant les mêmes questions (Razafindranovona, 2015). Ces différences sont atténuées par la prise en compte de caractéristiques sociodémographiques, et plus encore si les variables de contrôle sont enrichies de variables de l'enquête en lien direct avec la thématique de l'enquête[14].

Mais il peut demeurer un écart résiduel traduisant à la fois un biais de mesure et un biais de composition sur des variables non observables, comme l'intérêt pour la thématique de l'enquête (biais de sélection endogène).

… ET DES MÉTHODES DE CORRECTION QUI NE LÈVENT PAS DÉFINITIVEMENT TOUTES LES DIFFICULTÉS

Les effets de mode sont sans doute le frein le plus puissant face au développement du multimode : ils complexifient les analyses, rendent difficile l'estimation des « vrais » niveaux des indicateurs d'intérêt de l'enquête et peuvent entraîner des ruptures de séries. L'estimation et la correction de ces biais nécessitent de faire des hypothèses. Pour s'assurer de leur plausibilité, il est nécessaire de se doter de moyens spécifiques, à penser en amont de la collecte ***(encadré 4)***.

Le projet MIMOD ***(encadré 2)*** a permis de recenser les méthodes concrètement utilisées dans les instituts nationaux statistiques pour agréger les données issues des différents modes et de mettre à jour les connaissances acquises sur le multimode. Si les expériences en matière de mesure des effets de mode commencent à être riches et partagées au sein des pays européens, celles en matière de corrections de ces effets restent en revanche rares, à l'image de la littérature scientifique sur ce thème.

Plusieurs méthodes de correction sont envisageables, mais elles restent sujettes à des limites fortes, d'où le choix le plus souvent fait de ne pas corriger les effets de mode.

Tout d'abord, le biais de mesure est conceptuellement un problème qui se situe au niveau de la réponse partielle (comparativement au biais de sélection qui est un problème au niveau de la réponse totale). Il peut dans ce sens s'apparenter à un problème de valeur manquante : il se traite donc efficacement par **imputation.** Cependant, cette méthode constitue un choix méthodologique, voire déontologique, fort puisqu'on modifie la réponse des enquêtés. On préférera alors des méthodes « parcimonieuses »[15] sélectionnant les observations les plus porteuses du biais afin de minimiser les modifications de la base de données (Legleye *et alii*, 2019).

14. Cela peut être le cas du sentiment d'insécurité déclaré pour évaluer l'effet de mode sur les taux de victimation, ou encore d'un score de bien-être pour évaluer l'effet de mode sur un certain nombre de conditions de travail, etc.
15. C'est-à-dire, des méthodes limitant le nombre de modifications de la base de données, en ne sélectionnant que les observations les plus porteuses du biais, par opposition à des méthodes qui modifieraient toutes les réponses fournies sur un mode.

Encadré 4. Se doter en amont de moyens de corriger les effets de mode à l'aval

L'analyse des effets de mesure nécessite de faire l'hypothèse qu'il n'existe pas de biais de composition inobservable entre les répondants aux différents modes et que chaque répondant d'un mode de collecte trouvera un individu comparable parmi les répondants à l'autre mode. Pour être au plus proche de ces hypothèses, des solutions existent, mais elles doivent être pensées en amont de la collecte.

- Une première méthode consiste à disposer d'un **échantillon de contrôle disjoint**, sur lequel l'enquête sera réalisée *via* l'un des modes de collecte de l'échantillon multimode. Dans l'idéal, le protocole dédié à cet échantillon devrait à la fois reposer sur le mode de référence pour la thématique de l'enquête – c'est-à-dire conduisant à un biais de mesure minimal – et maximiser le taux de collecte, ce qui est souvent difficile. En cas d'arbitrage entre réduction du biais de mesure et maximisation du taux de collecte, il faut privilégier le taux de collecte afin de disposer d'un échantillon comprenant tous les profils à comparer dans chacun des modes. Cette méthode peut donc être généralisée à plusieurs sous-échantillons donnant lieu à plusieurs combinaisons de modes différentes.
- Une autre méthode, complémentaire de la première, consiste à **ajouter des questions pertinentes** directement dans le questionnaire. Ces questions doivent permettre d'assurer la comparabilité entre les répondants à chacun des modes de collecte, de manière plus robuste qu'avec les informations disponibles dans la base de sondage ou les variables sociodémographiques classiques. Elles peuvent par exemple porter sur les usages des différents modes de collecte. Il est en revanche important que ces variables ne soient pas sujettes elles-mêmes à un effet de mesure.

Ces méthodes sont efficaces, mais il y a toujours un risque que des profils de répondants particuliers soient très peu représentés dans un des modes, rendant difficiles l'évaluation et la correction des effets de mesure. De plus, les effets peuvent varier d'une sous-population à l'autre, ce qui peut très largement compliquer l'analyse.

- Une autre méthode, particulièrement performante, consiste à **réinterroger les mêmes individus, avec un certain délai, avec deux modes différents**. Le test repose sur l'hypothèse forte que les réponses au second mode ne sont pas affectées par la participation antérieure au premier mode. Dans la mesure où la situation créée par ce protocole est très artificielle, il est nécessaire de soigner particulièrement la prise de contact avec l'enquêté. Il est également recommandé d'apporter des modifications au questionnaire en remplaçant les questions les moins importantes par des efforts didactiques afin de justifier cette nouvelle interrogation (Klausch, 2014).
- Une autre solution consiste à **affecter aléatoirement un mode de collecte après le recueil d'une acceptation de participer**, obtenue à la suite d'une première sollicitation (idéalement faite sur un autre mode de contact que les modes de collecte testés). L'estimation est alors sans biais (sur l'échantillon des répondants) si le taux de réponse post-acceptation atteint 100 % (ce qui est peu probable). Mais cette méthode permet en principe de contrôler les deux phases de sélection (l'une relative à l'enquête, l'autre au mode), ce qui augmente sa fiabilité relativement à un protocole multimode classique.

Une autre méthode consiste à recourir au **calage.** Celui-ci permet de neutraliser le biais de sélection mais aussi le biais de mesure à un niveau agrégé, en introduisant des contraintes sur les niveaux de la variable d'intérêt. Ainsi, au lieu de modifier les données, on modifie les poids de façon à obtenir le bon niveau dans tout ou partie de l'échantillon. En revanche, le biais de mesure n'est pas corrigé au niveau individuel. Là aussi, on privilégiera des marges sur les sous-populations les plus porteuses du biais et le recours à un échantillon de contrôle monomode est fortement recommandé.

Il est également possible de ne pas corriger le biais de mesure, mais de chercher à le contenir dans le temps : dans des enquêtes répétées, il est possible d'introduire dans le redressement une contrainte maintenant à valeurs constantes les parts des différents modes (Buelens et Van den Brakel, 2010). Ceci présente l'avantage de la simplicité mais pose également des questions de pérennité en cas de disparition d'un mode ou bien de changement de part des différents modes au cours du temps.

Au-delà du biais de mesure, on peut être amené à corriger un autre biais : un biais de sélection inobservable portant sur l'ensemble des répondants. Cette problématique n'est pas directement liée au multimode, qui au contraire peut améliorer la représentativité de l'enquête, en fonction du protocole. Mais le multimode s'associe le plus souvent à l'introduction d'internet et l'utilisation de modes de collecte avec un moindre taux de réponse et une moindre représentativité des répondants que le face-à-face. Dans ce cas, des méthodes de correction existent, fondées sur l'approche d'Heckman (Heckman, 1979). Toutefois, leur validité repose sur l'hypothèse forte qu'il n'existe pas de biais de mesure (Castell et Sillard, 2021), car aujourd'hui, il n'existe pas de méthode permettant de traiter simultanément les deux, si la réponse sur un mode résulte d'un choix de l'enquêté (Lee, 2009).

De manière générale, les méthodes de correction des effets de mesure nécessitent un travail spécifique. Ainsi, toutes les variables d'une enquête ne pourront pas être analysées, et encore moins corrigées. Il est donc important de sélectionner avec soin les variables d'intérêt principales. Par ailleurs, les méthodes nécessitent de faire l'hypothèse qu'un des modes de collecte constitue la référence, soit parce qu'il est le mode historique d'une série d'enquêtes, soit parce qu'il est réputé produire la mesure de la meilleure qualité. Un échantillon de contrôle de la qualité, collecté en monomode (ou avec multimode embarqué) s'avère donc souvent nécessaire.

Ainsi, la détection des effets de mode peut justifier d'adapter le protocole d'enquête et son plan de sondage, notamment pour que des sous-échantillons disjoints soient collectés selon des compositions de modes différents ***(encadré 4)***.

FAIRE DU MULTIMODE UN ATOUT POUR AMÉLIORER LA QUALITÉ DES ENQUÊTES

Le développement du multimode nous invite à repenser la conception des enquêtes ménages autour de bonnes pratiques fondamentales et structurantes, garantes d'une qualité minimum des réponses telles que le respect d'un temps de réponse maximum pour un questionnaire et le recours à des concepts et des formulations compréhensibles par une large majorité de la population.

Dans son « programme Multimode », l'Insee a ainsi fait le choix de ne proposer le *web* que dans certains contextes (enquêtes courtes ou séquencées, concepts simples, réinterrogations après un premier contact intermédié). À l'heure où certains instituts de sondage abandonnent les enquêtes en face-à-face et parfois même les enquêtes téléphoniques, la force de l'Insee est de disposer d'un réseau d'enquêteurs très expérimentés en face-à-face, et très bien répartis sur le territoire français.

Au final, l'évolution vers le multimode constitue certes une occasion d'améliorer les enquêtes sur de nombreux aspects, mais elle représente aussi un coût organisationnel et un défi pour la gestion des effets de mode : ceci doit nous inviter à apporter le plus grand soin dans les choix méthodologiques envisagés au moment de la conception de l'enquête. Face à la généralisation du multimode en Europe, il serait notamment nécessaire de faire évoluer les règlements européens vers des questionnaires omnimodes, donc forcément simplifiés par rapport à ceux en vigueur actuellement et qui ont été conçus dans un univers dominé par le face-à-face. À l'automne 2021, un groupe de travail *ad hoc* (Eurostat, 2021) a présenté des recommandations allant dans ce sens : un « *position paper* » (coordonné par la France) identifie les champs d'investigation des prochaines années – en termes d'enjeux méthodologiques et de collecte – pour adapter les enquêtes européennes auprès des ménages au contexte du multimode.

BIBLIOGRAPHIE

ANTOUN, Christopher, COUPER, Mick P. et CONRAD, Frederick G., 2017. Effects of Mobile versus PC Web on Survey Response Quality: A Crossover Experiment in a Probability Web Panel. In : *Public Opinion Quarterly*. [en ligne]. 28 mars 2017. Vol. 81, n°S1, 2017, pp. 280-306. [Consulté le 15 décembre 2021]. Disponible à l'adresse : *https://doi.org/10.1093/poq/nfw088*.

BECK, François et PERETTI-WATEL, Patrick, 2001. Les usages de drogues illicites déclarés par les adolescents selon le mode de collecte. In : *Population*. [en ligne]. 56e année, n°6, pp. 963-985. [Consulté le 15 décembre 2021]. Disponible à l'adresse : *https://www.persee.fr/doc/pop_0032-4663_2001_num_56_6_7214*.

BECK, François, et RICHARD, Jean-Baptiste, 2013. Le Baromètre santé de l'INPES, un outil d'observation et de compréhension des comportements de santé des jeunes. In : *Agora Débats Jeunesses*. [en ligne]. Presses de Sciences Po, 2013, n° 63, pp. 51-60. [Consulté le 15 décembre 2021]. Disponible à l'adresse : *https://doi.org/10.3917/agora.063.0051*.

BECK, François, GUIGNARD, Romain et LEGLEYE, Stéphane, 2014. *Does Computer Survey Technology Improve Reports on Alcohol and Illicit Drug Use in the General Population? A Comparison Between Two Surveys with Different Data Collection Modes In France*. [en ligne]. 22 janvier 2014. PLOS ONE. [Consulté le 15 décembre 2021]. Disponible à l'adresse : *https://doi.org/10.1371/journal.pone.0085810*.

BUELENS, Bart et VAN DEN BRAKEL, Jan, 2010. On the Necessity to Include Personal Interviewing in Mixed-Mode Surveys. In : *Survey Practice*. [en ligne]. 30 septembre 2010. Vol. 3, n°5. [Consulté le 15 décembre 2021]. Disponible à l'adresse : *https://doi.org/10.29115/SP-2010-0023*.

CASTELL, Laura, SILLARD, Patrick, 2021. *Le traitement du biais de sélection endogène dans les enquêtes auprès des ménages par modèle de Heckman*. [en ligne]. 24 mars 2021. Insee. Document de travail n°M2021/02. [Consulté le 15 décembre 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/5237610*.

CHUN, Asaph Young, HEERINGA, Steven G. et SCHOUTEN, Barry, 2018. Responsive and Adaptive Design for Survey Optimization. In : *Journal of Official Statistics*. [en ligne]. Septembre 2018. Vol. 34, n°3, pp. 581-597. [Consulté le 15 décembre 2021]. Disponible à l'adresse : *http://dx.doi.org/10.2478/JOS-2018-0028*.

CORNESSE, Carina et BOSNJAK, Michael, 2018. Is there an association between survey characteristics and representativeness? A meta-analysis. In : *Survey Research Methods*. [en ligne]. 12 avril 2018. Vol. 12, n°1, pp. 1-13. [Consulté le 15 décembre 2021]. Disponible à l'adresse : *https://dx.doi.org/10.18148/srm/2018.v12i1.7205*.

COUPER, Mick P, PETERSON, Gregg J, 2016. Why Do Web Surveys Take Longer on Smartphones? In : *Social sciences computer review*. [en ligne]. 11 février 2016. Vol. 35, n°3, pp. 355-377. [Consulté le 15 décembre 2021]. Disponible à l'adresse : *https://doi.org/10.1177/0894439316629932*.

DE LEEUW, Edith, 2018. Mixed-Mode: Past, Present, and Future. In : *Survey Research Methods*. [en ligne]. 13 août 2018. Vol. 12, n°2, pp. 75-89. [Consulté le 15 décembre 2021]. Disponible à l'adresse :
https://dx.doi.org/10.18148/srm/2018.v12i2.7402.

DILLMAN, Don A., SMYTH, Jolene D., CHRISTIAN, Leah Melani, 2014. *Internet, Phone, Mail and Mixed-Mode Surveys: The Tailored Design Method*. Août 2014. Éditions Wiley. ISBN 978-1-118-45614-9.

EDWARDS, Philip James, ROBERTS, Ian, CLARKE, Mike J., DIGUISEPPI, Carolyn, WENTZ, Reinhard, KWAN, Irene, COOPER, Rachel, FELIX, Lambert M. et PRATAP, Sarah, 2009. *Methods to increase response to postal and electronic questionnaires.* [en ligne]. 8 juillet 2009. Éditions John Wiley & Sons. Cochrane Database Systematic Reviews, n°3, article n°MR000008. [Consulté le 15 décembre 2021]. Disponible à l'adresse :
https://core.ac.uk/reader/13098130?utm_source=linkout.

EUROSTAT, 2021. *Position Paper on Mixed-Mode Data Collection in Household Surveys*. [en ligne]. 19 octobre 2021. Groupe de travail sur les enquêtes ménages en multimode. Projet présenté aux directeurs des statistiques sociales (DSS), aux directeurs de la méthodologie (DIME) et aux directeurs des systèmes d'information (IT). [Consulté le 15 décembre 2021]. Disponible à l'adresse :
https://inseefrlab.github.io/ESS-Multimode-PP/.

GALESIC, Mirta et BOSNJAK, Michael, 2009. Effects of Questionnaire Length on Participation and Indicators of Response Quality in a Web Survey. In : *Public Opinion Quarterly.* [en ligne]. 28 mai 2009. Oxford Academics Journal. Vol. 73, n°2, pp. 349-360. [Consulté le 15 décembre 2021]. Disponible à l'adresse :
https://doi.org/10.1093/poq/nfp031.

GROVES, Robert M. et PEYTCHEVA, Emilia, 2008. The Impact of Nonresponse Rates on Nonresponse Bias: A Meta-Analysis. In : *Public Opinion Quarterly.* [en ligne]. 7 mai 2008. Oxford Academics Journal. Vol. 72, n°2, pp. 167-89. [Consulté le 15 décembre 2021]. Disponible à l'adresse :
https://doi.org/10.1093/poq/nfn011.

GROVES, Robert M., SINGER, Eleanor et CORNING, Amy, 2000. Leverage-Saliency Theory of Survey Participation: Description and an Illustration. In : *Public Opinion Quarterly.* [en ligne]. 1er novembre 2000. Oxford Academics Journal. Vol. 64, n°3, pp. 299-308. [Consulté le 15 décembre 2021]. Disponible à l'adresse :
https://doi.org/10.1086/317990.

GUILLAUMAT-TAILLIET, François et TAVAN, Chloé, 2021. Une nouvelle enquête Emploi en 2021, entre impératif européen et volonté de modernisation. In : *Courrier des statistiques*. [en ligne]. 8 juillet 2021. Insee. N° N6, pp. 7-27. [Consulté le 15 décembre 2021]. Disponible à l'adresse :
https://www.insee.fr/fr/statistiques/fichier/5398681/courstat-6-art-1.pdf.

HECKMAN, James J., 1979. Sample Selection Bias as a Specification Error. In : *Econometrica, Journal of the econometric society.* [en ligne]. Janvier 1979. Vol. 47, n°1, pp. 153-161. [Consulté le 15 décembre 2021]. Disponible à l'adresse :
https://doi.org/10.2307/1912352.

KAPPELHOF, Johannes W. S., 2015. Face-to-Face or Sequential Mixed-Mode Surveys Among Non-Western Minorities in the Netherlands: The Effect of Different Survey Designs on the Possibility of Nonresponse Bias. In : *Journal of Official Statistics.* [en ligne]. Mars 2015. Vol. 31, n°1, pp. 1-30. [Consulté le 15 décembre 2021]. Disponible à l'adresse : *https://doi.org/10.1515/jos-2015-0001*.

KLAUSCH, Lars Thomas, 2014. *Informed Design of Mixed-Mode Surveys − Evaluating mode effects on measurement and selection error.* [en ligne]. 10 octobre 2014. Utrecht University − Department of Methodology and Statistics. ISBN 978-90-393-6192-4. [Consulté le 15 décembre 2021]. Disponible à l'adresse : *https://dspace.library.uu.nl/handle/1874/300673*.

KOUMARIANOS, Heïdi et SCHREIBER, Amandine, 2021. *Conception de questionnaires auto-administrés*. [en ligne]. 15 décembre 2021. Insee. Document de travail n° M2021/03. [Consulté le 15 décembre 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/6010788*.

KREUTER, Frauke, PRESSER, Stanley et TOURANGEAU, Roger, 2008. Social Desirability Bias in CATI, IVR, and Web Surveys: The Effects of Mode and Question Sensitivity. In : *Public Opinion Quarterly.* [en ligne]. Décembre 2008. Oxford Academics Journal. Vol. 72, n°5, pp. 847-865. [Consulté le 15 décembre 2021]. Disponible à l'adresse : *https://academic.oup.com/poq/article-pdf/72/5/847/5188667/nfn063.pdf*.

KROSNICK, Jon A., 1991. Response strategies for coping with the cognitive demands of attitude measures in surveys. In : *Applied Cognitive Psychology*. [en ligne]. Mai/Juin 1991. Vol. 5, n°3, pp. 213-236. [Consulté le 15 décembre 2021]. Disponible à l'adresse : *https://doi.org/10.1002/acp.2350050305*.

LEE, David S., 2009. Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects. In : *The Review of Economic Studies*. [en ligne]. Juillet 2009. Vol. 76, n°3, pp. 1071-1102. [Consulté le 15 décembre 2021]. Disponible à l'adresse : *https://doi.org/10.1111/j.1467-937X.2009.00536.x*.

LEGLEYE, S., BOHET, A., RAZAFINDRATSIMA, N., BAJOS, N. et MOREAU, C., 2014. A randomized trial of survey participation in a national random sample of general practitioners and gynecologists in France. In : *Revue d'Épidémiologie et de Santé Publique*. [en ligne]. Août 2014. Vol. 62, n°4, pp. 249-255. [Consulté le 15 décembre 2021]. Disponible à l'adresse : *https://doi.org/10.1016/j.respe.2014.04.007*.

LEGLEYE, Stéphane et CHARRANCE, Géraldine, 2021. *Sequential and Concurrent Internet-Telephone Mixed-Mode Designs in Sexual Health Behavior Research.* [en ligne]. 30 août 2021. Journal of Survey Statistics and Methodology. [Consulté le 15 décembre 2021]. Disponible à l'adresse : *https://doi.org/10.1093/jssam/smab026*.

LEGLEYE, Stéphane, RAZAFINDRANOVONA, Tiaray et DE PERETTI, Gaël, 2019. *Agregating mix-mode survey data: a practical approach to neutralize measurement bias.* [en ligne]. 19 juillet 2019. European Survey Research Association. Conférence internationale 2019, Zagreb. [Consulté le 15 décembre 2021]. Disponible à l'adresse : *https://www.europeansurveyresearch.org/conferences/programme?sess=143#583*.

LEGLEYE, Stéphane, RAZAKAMANANA, Nirintsoa, CORNILLEAU, Anne et COUSTEAUX, Anne-Sophie, 2016. *Intéressement financier, motivation initiale et caractéristiques des enquêtes : effets sur le recrutement et la participation à long terme dans le panel ELIPSS.* [en ligne]. 30 octobre 2016. Université du Québec en Outaouais. 9e Colloque francophone sur les sondages, Gatineau. [Consulté le 15 décembre 2021]. Disponible à l'adresse : *http://sondages2016.sfds.asso.fr/wp-content/uploads/2016/10/Session03-Legleye.pdf*.

LUITEN, Annemieke, HOX, Joop et DE LEEUW, Edith, 2020. Survey Nonresponse Trends and Fieldwork Effort in the 21st Century: Results of an International Study across Countries and Surveys. In : *Journal of Official Statistics*. [en ligne]. 24 juillet 2020. Vol. 36, n° 3, pp. 469-487. [Consulté le 15 décembre 2021]. Disponible à l'adresse : *https://doi.org/10.2478/jos-2020-0025*.

RAZAFINDRANOVONA, Tiaray, 2015. *La collecte multimode et le paradigme de l'erreur d'enquête totale.* [en ligne]. 27 mars 2015. Insee. Document de travail n°M2015/01. [Consulté le 15 décembre 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/1381054*.

REVILLA, Melanie et OCHOA, Carlos, 2017. Ideal and Maximum Length for a Web Survey. In : *International Journal of Market Research.* 24 octobre 2017. Vol. 59, n° 5, pp. 557-566.

ROBERTS, Caroline, EVA, Gillian, ALLUM, Nick et LYNN, Peter, 2010. *Data Quality in Telephone Surveys and the Effect of Questionnaire Length: a Cross-National Experiment.* [en ligne]. 9 novembre 2010. Institute for Social and Economic Research. Working Paper Series n°2010-36. [Consulté le 15 décembre 2021]. Disponible à l'adresse : *https://www.iser.essex.ac.uk/research/publications/working-papers/iser/2010-36.pdf*.

RUBIN, Donald B. et LITTLE, Roderick J. A., 2002. *Statistical Analysis with Missing Data*. 26 août 2002. Éditions John Wiley & Sons, Wiley Series in Probability and Statistics. ISBN 978-1119013563.

SCHWARTZ, Barry, 2005. *The Paradox of Choice - Why More is Less.* 18 janvier 2005. Harper Perennial. ISBN 978-0060005696.

SINGER, Eleanor et YE, Cong, 2012. The use and effects of incentives in surveys. In : *The Annals of the American Academy of Political and Social Science.* [en ligne]. 26 novembre 2012. Vol. 645, n°1, pp.112-141. [Consulté le 15 décembre 2021]. Disponible à l'adresse : *https://doi.org/10.1177/0002716212458082*.

TURNER, C. F., KU, L., ROGERS, S. M., LINDBERG, L. D., PLECK, J. H., et SONENSTEIN, F. L., 1998. Adolescent Sexual Behavior, Drug Use, and Violence: Increased Reporting with Computer Survey Technology. In : *Science*. [en ligne]. 8 mai 1998. Vol. 280, n°5365, pp. 867-873. [Consulté le 15 décembre 2021]. Disponible à l'adresse : *https://www.science.org/lookup/doi/10.1126/science.280.5365.867*.

LA MISE EN MUSIQUE D'ENQUÊTES MULTIMODES

Eric Sigaud et Benoît Werquin***

Pour l'Insee, le développement d'enquêtes multimodes relève d'une orchestration délicate et rigoureuse. Il s'agit, entre autres, de construire des questionnaires internet cohérents avec des questionnaires enquêteurs, de coordonner le travail des enquêteurs et gestionnaires mobilisés sur des collectes multiples dans différents modes, le tout en conservant une mise en œuvre harmonieuse et efficace prenant en considération les contraintes de ressources et de calendrier. Ces enjeux de construction et de transformation sont autant de tonalités de la démarche mise en œuvre par l'institut depuis une dizaine d'années. Celle-ci est rythmée par des implémentations en vraie grandeur sur des opérations d'enquêtes, d'abord auprès des entreprises puis des ménages, d'abord via internet puis avec les enquêteurs, d'abord sur des protocoles simples ne mobilisant qu'un seul mode de collecte puis sur des protocoles complexes, avec plusieurs collectes, dans des modes différents, simultanément.

Depuis le lancement de cette évolution structurante pour la statistique publique, le même schéma s'applique : penser d'abord le métier, l'exprimer, le conceptualiser, puis passer aux phases de mise en œuvre ou de passage à l'échelle. L'expérience acquise à chaque étape favorise une construction progressive de l'outillage conceptuel et technique des enquêtes « réellement » multimodes, dans toute leur complexité.

For INSEE, the development of multi-mode surveys is a delicate and rigorous orchestration. Among other things, it is a matter of constructing Internet questionnaires that are consistent with survey questionnaires, and coordinating the work of interviewers and managers mobilised for multiple collections in different modes, all while maintaining a harmonious and efficient implementation that takes into consideration resource and schedule constraints. These issues of construction and transformation are all part of the approach implemented by the Institute over the last ten years. This has been punctuated by full-scale implementations of survey operations, first with companies and then with households, first via the Internet and then with interviewers, first with simple protocols using a single collection mode and then with complex protocols, with several collections, in different modes, simultaneously.

Since the launch of this structuring evolution for official statistics, the same pattern has been applied: first think about the business, express it, conceptualise it, then move on to the implementation or scaling-up phases. The experience acquired at each stage favours the gradual construction of the conceptual and technical tools of "truly" multimode surveys, in all their complexity.

* Maître d'ouvrage délégué du programme Métallica, département des Prix à la consommation et des enquêtes ménages, Insee, *eric.sigaud@insee.fr*

** Chef de projet informatique de Métallica, service national du développement informatique, Insee Hauts de France, *benoit.werquin@insee.fr*

Pour nombre d'enquêtes de la statistique publique française, auprès des entreprises ou des ménages, la collecte par internet constitue une alternative à la collecte papier depuis quelques années déjà[1]. Elle devient également une alternative ou un complément pour la collecte en face-à-face ou par téléphone, créant, de fait, de nouveaux protocoles de collecte, plus complexes, qualifiés de « multimode ».

Au-delà des délicates questions statistiques ou méthodologiques posées par l'avènement de tels protocoles[2], il convient de se pencher, à la lumière de l'expérience acquise, sur la complexité opérationnelle induite par le multimode.

Comment construire un ensemble d'outils cohérents capable de rendre les services attendus ? Quels sont précisément les services attendus ? Comment les organiser ? Sur quel acteur repose quelle tâche ? Quelle offre de services mettre en œuvre ? Dans quel ordre ?

La démarche s'apparente en quelque sorte à l'écriture par l'Insee de sa partition et de ses arrangements pour arriver à une collecte multimode harmonieuse[3]. Des premiers solos, opérations mono-modes, ont d'abord permis à l'institut de rôder les concepts, les outils et les processus. Puis des déclinaisons « polyphoniques », où cette fois différents modes sont mis en œuvre indépendamment les uns des autres, ont permis de caractériser chaque mode de collecte, constituant une étape préalable qualifiée de « poly-mode ». Elle prépare à la mise en œuvre de protocoles multimodes à proprement parler, où les différents modes de collecte doivent cette fois agir de concert.

POUR PASSER D'UN SEUL MODE DE COLLECTE À PLUSIEURS

Le développement de la collecte par internet, en complément d'un autre mode, s'accompagne immanquablement de son lot de questions pratiques : comment construire un questionnaire *web* ? Doit-on avoir le même questionnement dans les différents modes ? Comment contacte-t-on les répondants ? Comment les relance-t-on ?

> *« Conceptualiser ce qui est déjà fait et ce qu'il y aurait à faire. »*

Si le multimode pose des questions complexes, sur la qualité et l'exploitabilité des données collectées, il faut auparavant résoudre la question de la faisabilité : comment fait-on pour conduire une collecte recourant à plusieurs modes, dont un nouveau ? Un premier réflexe est de s'appuyer sur les pratiques courantes pour définir les nouvelles, pour tenter de reproduire ce qui est déjà fait. Cependant, la complexité induite par le nouveau mode de collecte (développement d'un nouveau questionnaire, nécessité de gérer des identifiants/mots de passe, etc.) et ses particularités (pas toujours de phase d'approche par l'enquêteur, contexte de réponse moins contraint, etc.) conduisent à pousser la réflexion un cran plus loin : conceptualiser ce qui est déjà fait et ce qu'il y aurait à faire.

1. Au tout début des années deux-mille, le service statistique du ministère de l'Industrie lançait les premières enquêtes par internet de la statistique publique. En « régime » courant en 2021, 40 enquêtes auprès des entreprises utilisent le mode *web*, et 4 enquêtes auprès des ménages, dont l'enquête Emploi.
2. Voir sur ces questions l'article de François Beck, Laura Castell, Stéphane Legleye et Amandine Schreiber sur « *Le multimode dans les enquêtes auprès des ménages : une collecte modernisée, un processus complexifié* » dans ce même numéro.
3. [NDLR] Les métaphores musicales ne surprendront pas les lecteurs des numéros précédents sur les sujets connexes, comme (Cotton et Dubois, 2019 ; Haag et Husseini-Skalitz, 2019 ; Koumarianos et Sigaud, 2019).

LE «POLY-MODE» (PREMIÈRES VOCALISES)

La première étape de conceptualisation enclenchée par l'Insee, de loin la plus mûre aujourd'hui puisqu'en place depuis plusieurs années[4], est la **génération automatique d'instruments de collecte** (appelée parfois par abus de langage génération de questionnaires). Elle repose sur le principe des **métadonnées actives** (Bonnans, 2019) et part de l'idée simple qu'un processus répété, en l'occurrence le développement d'un questionnaire d'enquête, gagne souvent à être automatisé.

Pour conceptualiser un processus, celui de « fabrication » du questionnaire, le GSBPM[5] est le standard référence en la matière. Il suggère un premier découpage, entre une phase de conception et une phase de construction. Dans la première, on spécifie un questionnement, en dehors d'un mode donné. Dans la seconde, on construit des instruments de collecte[6] dédiés à un mode. À l'Insee, ce découpage structure les outils *Pogues* et *Eno* (Cotton et Dubois, 2019 ; Koumarianos et Sigaud, 2019) : *Pogues* permet la spécification d'un questionnaire dans une application tandis que *Eno* supporte la génération des instruments de collecte ***(figure 1)***.

Cette approche générative permet de capitaliser sur les bonnes pratiques entre questionnaires et de mettre en œuvre les fonctionnalités propres à chaque mode :

- des codes-barres pour les questionnaires papier ;
- une accessibilité et un fonctionnement sur différents supports (*smartphone*, tablette, etc.) pour les questionnaires *web* ;
- une navigation et une ergonomie adaptées pour le questionnaire déroulé par un enquêteur ;
- une contextualisation du questionnement au mode (liste déroulante en *web*/carte code en face-à-face vs champ libre sur papier), etc.

Ainsi, la production de questionnaires peut changer d'échelle, passant du développement *ad hoc* de questionnaires spécifiques à chaque mode et à chaque enquête, à une automatisation de la production d'instruments de collecte, pour plusieurs modes de collecte, s'appuyant sur un même outil de spécification et rationalisant les travaux de développement : formellement, on ne conduit plus de développement *ad hoc*.

La logique « **mono-mode** », où le processus et les outils sont pensés et mis en œuvre pour un mode donné, laisse la place à une logique « **poly-mode** », où processus et outils sont conceptualisés pour supporter plusieurs modes de collecte et plusieurs cas d'enquête (*web*, papier, téléphone, plusieurs séquences de collecte, etc.).

On comprend alors tout l'intérêt de l'approche par les métadonnées actives, qui permet la génération automatique des questionnaires, et plus largement des « instruments de collecte » (Koumarianos et Sigaud, 2019). L'approche ne se limite d'ailleurs pas aux seules phases de conception et de construction des instruments de collecte. Et après ces premières phases du GSBPM[7], c'est le processus de collecte dans son ensemble qui pourra faire l'objet

4. Les premiers travaux de génération de questionnaires à l'Insee remontent à 2013 et portaient sur le questionnaire de l'enquête sectorielle annuelle de la sphère entreprise.
5. *Generic Statistical Business Process*, ou modèle générique de processus de production statistique, mis au point dans le cadre de l'Unece. Voir (Unece, 2019) et (Erikson, 2020) pour un exemple de mise en œuvre opérationnelle.
6. Un instrument de collecte est l'instanciation d'un questionnement dans un contexte opérationnel (questionnaire internet, questionnaire adapté à un appel téléphonique d'un enquêteur, adapté à la visite en face-à-face d'un enquêteur, questionnaire papier).
7. Formellement, le GSBPM distingue 8 phases : la définition des besoins, la conception, la construction, la collecte, le traitement des données, l'analyse, la diffusion et l'évaluation.

Figure 1. Le poly-mode, première étape vers l'utilisation des métadonnées au service de la collecte

*DDI : *Data Documentation Initiative*

Pour plus d'information sur la génération automatique d'instruments de collecte, voir (Cotton et Dubois, 2019 ; Koumarianos et Sigaud, 2019).

d'une conceptualisation afin de permettre son instanciation selon l'enquête, son mode et son protocole. Ainsi, des métadonnées telles que des découpages en opérations (plusieurs séquences, vagues de collecte), des calendriers d'opérations de collecte (dates de début/fin de collecte, de relance), ou des caractéristiques de protocoles (concurrentiel ou séquentiel, avec ou sans phase de repérage, avec ou sans reprise enquêteurs) pourraient être également « activées » et permettre le pilotage des processus mis en œuvre.

Un exemple de la transformation apportée par la conceptualisation illustre parfaitement l'apport de cette approche par les métadonnées : celui du passage de la fiche-adresse (document contenant les coordonnées d'un ménage à enquêter) à l'unité enquêtée ***(encadré 1)***. Le concept d'unité enquêtée désigne les unités de compte d'une opération de collecte, qu'elle soit enquêtée par le *web*, le téléphone, le face-à-face, le papier. La notion permet notamment d'associer des données collectées par un questionnaire et les informations relatives au déroulement de la collecte (du type paradonnées), indépendamment du mode.

CONCEPTUALISER AVANT DE METTRE EN ŒUVRE

Définir précisément le processus à mettre en œuvre, s'appuyer sur le GSBPM pour le découper en « phases », ensemble d'actions métier cohérentes, sont des préalables à toute mise en œuvre dans les différents modes. Une réflexion décorrélée conduirait à construire des outils et processus incohérents les uns avec les autres ou incompatibles avec les besoins les plus exigeants, tels que des collectes *web* et enquêteurs concomitantes et concurrentes, ou encore un suivi du processus consolidé entre les modes.

> *« C'est ce découpage fin du processus qui va permettre une réutilisation et une mutualisation dans le contexte des enquêtes auprès des ménages. »*

Ceci n'empêche pas une approche incrémentale, tant dans les opérations que dans la complexité considérée. Ainsi à l'Insee, les premiers développements ont concerné les enquêtes auprès des entreprises, par nature plus homogènes que les enquêtes auprès des ménages.

LES ENQUÊTES AUPRÈS DES ENTREPRISES (PREMIÈRE SYMPHONIE)

Le projet *Coltrane*[8] de l'Insee a mis en place des services automatisés et partagés à destination des enquêtes auprès des entreprises, afin de permettre une collecte par internet et par papier. *Coltrane* portait un enjeu fort de rationalisation des coûts, notamment sur les travaux de spécification, de recette et de développements informatiques. La plateforme et les services associés instancient un processus, certes adapté aux enquêtes auprès des entreprises, mais s'inscrivant dans un cadre conceptuel plus global. Notamment un découpage fin du processus, s'appuyant sur le GSBPM, permet une séparation des phases d'approche ou d'accès du répondant, des phases de questionnement à proprement parler. Voit alors le jour un portail « Mes enquêtes » ***(figure 1)***, lequel propose un « tableau de bord » pour les personnes répondant en entreprise, interrogées plusieurs fois et souvent dans plusieurs enquêtes en même temps.

8. (Collecte Transversale d'Enquêtes) Aujourd'hui c'est une offre de service complète qui permet aux concepteurs d'enquête : de collecter les réponses par internet, de disposer d'un référentiel de contacts au sein des entreprises et de le gérer, d'envoyer des courriers, des mails et des questionnaires papier aux entreprises qui le souhaitent, le tout en disposant d'un système d'assistance dédié (Haag et Husseini-Skalitz, 2019).

De même sont mis en place une offre de courriers standardisés[9] ainsi que des processus d'envois automatiques et intégrés, pour l'ensemble des enquêtes auprès des entreprises[10].

C'est ce découpage fin du processus, notamment en phases d'accès aux enquêtes et de réponse à un questionnaire, qui va permettre une réutilisation et une mutualisation dans le contexte des enquêtes auprès des ménages.

Encadré 1. Un changement de partition: de la fiche-adresse à l'unité enquêtée

Les enquêtes auprès des ménages de la statistique publique française se sont longtemps appuyées sur une collecte en face-à-face, habitude qui se retrouve dans la dénomination usuelle, à l'Insee, de l'objet « enquêté ménage » sous le vocable de « fiche-adresse ».

Une fiche-adresse, dans le vocabulaire de l'Insee, désigne au départ la fiche papier sur laquelle sont indiquées les coordonnées d'un enquêté, individu, ménage ou logement qu'un enquêteur doit interroger.

Dans un contexte mono-mode, il y a une certaine logique conjointe entre la production de ces documents papier à destination des enquêteurs, le tirage de l'échantillon et la configuration des applications et des flux pour la mise en œuvre d'une collecte. On retrouve, à chaque fois, peu ou prou, les mêmes informations que celles contenues sur ces « fiches-adresses ». Ainsi, par un abus de langage, le répondant est également dénommé « fiche-adresse » et tout le monde se comprend.

Mais que devient cette fiche-adresse dans une collecte *web* ou téléphone quand aucun enquêteur ne se déplace ? Faut-il encore imprimer ce papier ? Que devient ce papier quand on a surtout besoin de coordonnées téléphoniques plutôt que d'une adresse physique ? Comment reporter sur ce papier les informations provenant d'une collecte *web* concomitante ?

Autant de questions qui nécessitent une réflexion préalable sur le processus (approche ou repérage, prise de contact/rendez-vous ou accès au questionnaire en ligne, récupération des données collectées, etc.) avant de pouvoir décliner l'enquête dans un nouveau mode.

La **fiche-adresse** devient alors **unité enquêtée** et celle-ci traverse les différentes phases du processus de collecte, de sa « naissance », dans les berceaux de l'échantillonnage, à son « éducation », qu'elle soit sous la vigilance bienveillante d'un enquêteur ou livrée aux interactions plus « chaotiques » d'une modernité connectée, jusqu'à son « passage à l'âge adulte », transformée par des traitements en variables statistiques. La conception de ce **« cycle de vie de l'unité enquêtée »** permet d'identifier **les concepts et les événements** nécessaires pour retracer l'histoire vécue par l'unité enquêtée. Ensuite on pourra identifier des objets qui peuvent être partagés au sein des différents systèmes : coordonnées, qualification du type de non-réponse, etc. Ces précisions sur les événements peuvent relever d'un mode en particulier – essai de contact ou de repérage pour les enquêtes mobilisant des enquêteurs téléphoniques ou se déplaçant en face-à-face, informations d'authentification pour les enquêtes mobilisant le *web*, notices papier pour celles offrant la possibilité de répondre par papier, etc. – ou se décliner dans chaque mode – non-répondant, date de début/fin de collecte, relance, etc.

9. Ces modèles de courriers mutualisés sont par ailleurs conformes aux préconisations du comité du Label de la statistique publique.
10. Avec les services spécialisés dans la production éditique.

LES ENQUÊTES *WEB* AUPRÈS DES MÉNAGES (VARIATION SUR LE MÊME THÈME)

Si les premières opérations ont concerné les enquêtes auprès des entreprises, l'Insee a depuis étendu la réflexion et les travaux aux enquêtes auprès des ménages.

Pour envisager de réutiliser dans un contexte « ménage » les services mis en œuvre dans un contexte « entreprise », même par « morceaux », c'est l'analyse fine du processus de collecte qui permet de proposer des scénarios. Les services rendus et les outils mis en œuvre, correctement découpés, peuvent être ensuite « recomposés ». En l'espèce, il a fallu opérer quelques variations par rapport à l'expérience *Coltrane*.

Le questionnement statistique a d'emblée été analysé comme un même besoin métier : il pouvait donc être mis en œuvre par les mêmes services techniques, moyennant un paramétrage pour offrir une contextualisation adaptée à chaque type de répondant ; cela concernait :

- les premières pages du questionnaire (contenant notamment le cadre juridique de l'enquête) ;
- la charte graphique du questionnaire (logo des services statistiques partenaires par exemple) ;
- les mécanismes d'authentification sur le site *web* (un thème et des consignes différentes) ;
- les formules d'assistance aux répondants.

En revanche, pour l'accès au questionnaire, les besoins divergent sensiblement entre les deux univers. Les répondants en entreprise sont en effet plus familiers des enquêtes de la statistique publique, car ils sont souvent interrogés par plusieurs d'entre elles. Dans un contexte « entreprise », un portail d'accès doit donc répondre à des besoins de visibilité sur l'ensemble des enquêtes pour lesquelles un répondant est sollicité ; il doit également prendre en compte des liens parfois complexes entre la personne « contact » et les entreprises pour lesquelles il est autorisé à répondre (par exemple, les comptables sont souvent répondants pour plusieurs entreprises différentes pour plusieurs enquêtes différentes). Ainsi, le choix a porté sur un portail « *Mes enquêtes* », un site où le répondant peut accéder aux différents questionnaires qui le concernent, avec des rappels sur les échéances. Sont également implémentées des fonctions de mise à jour des coordonnées personnelles, informations particulièrement importantes dans un contexte où les ré-interrogations d'une même entreprise sont nombreuses (à chaque vague de certaines enquêtes, ou par différentes enquêtes).

Le besoin des ménages est tout autre : les répondants ne sont en général sollicités que pour une seule enquête[11], ils ne sont pas forcément familiers des concepts d'enquête de statistique publique et sont souvent moins enclins à y consacrer du temps. Dans ce contexte « ménage », le besoin est davantage tourné vers l'information, voire la promotion de l'enquête auprès du répondant. Et ainsi c'est un portail dit de « *Promotion* » qui est proposé, affichant des informations relatives à une enquête ménage en particulier, ses résultats précédents, son cadre juridique et tout autre élément susceptible de favoriser une réponse par le répondant ménage.

11. Les services de tirage d'échantillon offerts par l'Insee à ses enquêtes ménage et à celles des services statistiques ministériels (SSM) garantissent la non-réinterrogation d'une enquête à l'autre d'un même ménage, afin de limiter la charge supportée individuellement par les répondants.

MUTUALISER SANS DOGMATISME (UNE PARTITION COMMUNE, DES RISQUES DE DISSONANCE)

Cette approche « découplée » est séduisante à plusieurs titres. D'abord, elle offre une modularité intéressante, chaque service pouvant être, au choix, ré-implémenté ou mutualisé. Elle permet également une mise en œuvre incrémentale où la cible métier est atteinte par paliers : plus d'application monolithique « à tout faire », couplée à ses inévitables projets de refonte, mais des ensembles de produits, qui évoluent à des rythmes différents.

C'est ce qui a permis par exemple d'offrir un portail mutualisé à l'enquête Emploi en continu, tout en conservant ses applications de collecte historiques[12] : ainsi, les concepteurs de l'enquête pouvaient mener les travaux méthodologiques nécessaires à la sécurisation du passage au multimode, et ne viser une « bascule » complète qu'avec de nouveaux outils matures. Ou encore, c'est ce qui permet d'assurer la collecte du volet *web* d'enquêtes multimodes, telles que Vie quotidienne et santé (*VQS*) ou Technologies de l'information et des communications (*TIC*), avec du suivi et de l'assistance autour de cette collecte *web*, tout en conservant là aussi les outils historiques pour le volet enquêteur de la collecte.

La cible conceptuelle parfaite ne saurait être complètement atteinte, les habituelles contraintes opérationnelles et budgétaires étant des réalités bien concrètes. Ainsi plus qu'une cible idéale, c'est un cadre, une cohérence globale que les différentes réalisations doivent respecter, des bonnes pratiques qui doivent régir tant la construction des outils que les pratiques métier.

Par exemple, pour les courriers et les différents supports de communication, le cadre initié avec les enquêtes auprès des entreprises a été étendu aux enquêtes auprès des ménages. Ou encore, les services d'assistance offerts aux répondants et leurs processus afférents s'assurent de respecter les mêmes critères de sécurité et de confidentialité des informations personnelles. Ce sont les mêmes outils de conception de questionnaires (*Pogues*, *Eno*) qui sont utilisés, lesquels intègrent déjà les bonnes pratiques méthodologiques.

> *« La mutualisation ou la « généricisation » à outrance ne doit pas devenir un dogme et les réutilisations doivent être étudiées au cas par cas. »*

Mais le pragmatisme implique aussi d'éviter de compliquer inutilement un système déjà inévitablement complexe. La mutualisation ou la « généricisation » à outrance ne doit pas devenir un dogme et les réutilisations doivent être étudiées au cas par cas.

L'existence des deux types de portail des enquêtes de la statistique publique française en est une bonne illustration. Il pouvait paraître séduisant d'imaginer un « méta-portail » proposant des contextes différents selon le type de répondant. Pourtant, une fois correctement découpées en briques fonctionnelles, seules les fonctions d'authentification et d'assistance présentaient des opportunités réelles de mutualisation. Les deux portails couvrent des besoins sensiblement différents et sont des applications simples qui gagnent à le rester, afin de faciliter leur maintenance et leur évolution. Aucun gain structurel significatif n'est à attendre (à ce jour) d'une mutualisation. Là encore, l'approche par les métadonnées offre un angle d'analyse intéressant, des natures de métadonnées différentes, ici de l'information générale sur une enquête contre des liens entreprise-contact et des multiples calendriers d'enquête, suggèrent souvent des besoins « éloignés ».

12. Voir (Guillaumat-Tailliet et Tavan, 2021).

POLY-MODE ET MULTI-MODE, MÊME MUSIQUE ?

Les premiers travaux de conceptualisation pour un passage à l'échelle ont permis de consolider le processus de collecte et de le ré-instancier dans plusieurs contextes : des services du domaine « entreprise » mobilisés dans le contexte ménage, un processus de collecte construit autour du concept d'unité enquêtée, indépendante du mode. On est en mesure de contextualiser les différents concepts, en fonction du mode de collecte.

Mais peut-on simplement mobiliser ce processus pour un mode *web* et pour un mode enquêteur et parler de collecte multimode ? A-t-on résolu pour autant la question du multimode avec cette vision « poly-mode » ?

On pourrait être tenté de répondre oui, mais le multimode ne peut être réduit uniquement à plusieurs processus indépendants. Car il est nécessaire que ces processus interagissent (par exemple, les réponses issues d'un mode doivent pouvoir être reversées dans un autre) et parfois même cohabitent (cas de collecte concurrentielle[13] entre enquêteur et *web*). Se contenter d'une vision poly-mode n'est pas suffisante et il convient de prendre en compte des complexités supplémentaires inhérentes au multimode.

Et complexités supplémentaires entraînent également concepts supplémentaires :

- un questionnement multimode (remobilisation de réponses d'un mode à l'autre, consolidation des réponses en aval, etc.) ;
- un processus multimode (fonctions supplémentaires de suivi multimode, contrôle de la qualité dans un contexte multimode) ;
- des rôles pour les différents acteurs d'une collecte multimode (mobilisation des enquêteurs pour la reprise de questionnaires *web* incomplets, mobilisation de gestionnaires pour la reprise de questionnaires *web* « en erreur », etc).

Il s'agit là encore d'une analyse préalable qui constitue une étape indispensable à toute mise en œuvre d'une collecte multimode. Dans ce cheminement vers un « vrai » multimode, il faut repartir de la question centrale des enquêtes : être en capacité d'exploiter les données collectées.

À CHAQUE TYPE DE DONNÉES SON OUTIL DE CAPTATION (LA FIN DE L'HOMME-ORCHESTRE)

Les données collectées sont l'enjeu principal du processus… de collecte. Celui-ci vise à capter ces données et à permettre leur exploitation. Il s'agit avant tout des données collectées auprès d'un répondant par le biais de questions, mais il existe de fait des données d'autres types :

- des données permettant le suivi et le **pilotage** de la collecte (actions enquêteurs/gestionnaires, plis papier retournés pour adresse invalide, relances envoyées) ;
- des données sur le déroulement de l'enquête : les **paradonnées**, informations techniques captées sur le comportement du répondant (les différents clics, le matériel utilisé, la durée de session, etc.).

13. Les différents types de multi-mode sont déclinés dans l'article déjà cité de François Beck, Laura Castell, Stéphane Legleye et Amandine Schreiber dans ce même numéro.

Les filières « historiques » des enquêtes à l'Insee n'ont pas attendu les réflexions sur le multimode pour mobiliser et collecter ces différents types de données : l'approche strictement mono-mode conduit souvent à produire un questionnaire, un instrument de collecte, avec le *data model*[14] qui sert « d'application à tout faire » (collecter des réponses statistiques, produire des indicateurs de gestions et de suivi, mesurer les temps techniques, qualifier les phases de contact ou la non-réponse, etc.).

Pourtant, les données manipulées ne sont pas destinées aux mêmes acteurs, ni aux mêmes usages. Et une approche découplée préconise qu'elles ne soient pas captées par les mêmes outils. Ainsi, au sein des nouveaux outils pour la mise en œuvre de la collecte multimode :

- les données collectées relèvent de l'instrument de collecte (le questionnaire) ;
- les données de suivi/gestion sont de la responsabilité des outils d'organisation de la collecte (poste de gestion ou poste de collecte enquêteur) ;
- quant aux paradonnées, elles reposent sur des services techniques dédiés et spécialisés.

Dès lors, on est en mesure de proposer une visualisation des données collectées dans un questionnaire ou un autre (reprise de données *web* par un enquêteur par exemple), de permettre de visualiser simplement ou bien de modifier un questionnaire dans un contexte de contrôle qualité sans interférer sur les données de suivi de l'activité enquêteur[15], ou de proposer aux concepteurs d'enquêtes des services de type « entrepôt de paradonnées » afin de supporter des approches expérimentales et des analyses méthodologiques[16].

LE QUESTIONNAIRE «OMNIMODE» (PLUSIEURS COUPLETS, UN MÊME ARRANGEMENT)

Une fois les différents types de données récoltés, il devient possible de traiter l'une des complexités majeures des enquêtes en multimode : leur exploitabilité statistique. Les « effets de mode » – où les réponses à ce que l'on croit être une même question varient en fonction du contexte de questionnement – posent en effet de difficiles questions d'exploitation statistique. Afin d'en limiter l'effet, un certain nombre de principes sont mis en œuvre au sein des protocoles de collecte multimodes[17], mais aussi dans la conception du ou des questionnaires de l'enquête.

> *« Les réflexions sur la portée du questionnaire multimode ne se limitent pas au seul questionnement. »*

Tout d'abord, le questionnement se doit d'être multimode. C'est-à-dire vu comme une consolidation de différents processus de collecte, déclinés opérationnellement selon les modes d'interaction mis en œuvre : les données collectées doivent rester sensiblement les mêmes, les adaptations limitées aux considérations opérationnelles ou « ergonomiques ».

14. En l'occurrence, le *data model* Blaise.
15. Cela évite par exemple d'être obligé de se faire passer pour un enquêteur quand on est gestionnaire, ou de développer une application *ad hoc*.
16. *Généric* est une application de reprise par des gestionnaires, s'appuyant sur des métadonnées actives et la génération automatique de questionnaires. La mise en œuvre d'un entrepôt de paradonnées et du processus de reprise de questionnaires issus du *web* est par ailleurs en cours.
17. Voir l'article de François Beck, Laura Castell, Stéphane Legleye et Amandine Schreiber dans ce même numéro.

Les réflexions sur la portée du questionnaire multimode ne se limitent pas au seul questionnement, ce sont les variables collectées en elles-mêmes et les processus « post-collecte » qui se doivent d'être consolidés.

Pour garantir leur exploitabilité, les données collectées par des outils différents devront mobiliser des chaînes techniques différentes avant d'être enfin mises à disposition dans des bases consolidées. Il y a là aussi une nécessité d'avoir une vision consolidée du processus pour garantir une cohérence : mobiliser les mêmes contrôles post-collecte ou encore les mêmes outils de contrôle de la qualité. Une telle consolidation impose une importante rationalisation du système d'information pour faire « coopérer » les différents modes de collecte.

Les outils doivent permettre la conception d'un questionnaire **omnimode**, en ce sens où il est conçu pour garantir des questionnements les plus similaires possibles d'un mode à l'autre, des valeurs de variables collectées les plus comparables et une exploitation rationalisée.

LE CHAMP DU QUESTIONNAIRE OMNIMODE (SUR TROIS OCTAVES)

Derrière le terme omnimode se cachent en fait trois dimensions :

- le questionnaire est « unique » et commun à tous les modes ;
- le questionnaire est adapté à tous les modes ;
- les réponses issues des différents modes doivent être réconciliées en une même variable statistique.

Les deux premières dimensions sont largement prises en compte dans les outils de conception de questionnaires déjà mis en œuvre à l'Insee : *Pogues* et *Eno* permettent d'adapter automatiquement les mêmes objets de questionnaire (questions, champ de réponse, etc.) aux différents modes de collecte et ils limitent les objets spécifiques à un mode de collecte donné.

La déclinaison d'objets du questionnaire aux différents modes de collecte a déjà été évoquée, quant aux éléments réservés à un seul mode, c'est une pratique que les outils permettent mais limitent. La contextualisation d'une consigne (destinée à un enquêteur ou destinée à un répondant *web* par exemple) est assez courante et usuelle, mais filtrer des questions ou blocs de question selon le mode doit être encadré pour préserver le principe de questionnaire omnimode avec les mêmes variables collectées. La bonne pratique consiste à limiter ces cas aux questions dont les réponses peuvent être imputées automatiquement, garantissant une certaine indépendance des traitements post-collecte (chaque réponse collectée passe par les mêmes chaînes de traitement quel que soit le mode).

L'exemple théorique de la question « Avez-vous, personnellement, la possibilité d'accéder à un internet ? »[18] en est une bonne illustration. Dans un contexte *web*, elle paraît superflue et la réponse pourrait être imputée automatiquement à « Oui ». On peut également citer les questions relatives à la description d'un logement (distinction entre appartement ou maison individuelle par exemple) qui peuvent relever des phases de repérage préalables par l'enquêteur et alléger d'autant le questionnement statistique en face-à-face (tout en restant pertinentes dans les autres modes de collecte).

18. Au-delà des considérations de formulation, elle est vue comme un cas d'école exemple de question « absurde » dans le contexte *web*.

Pour ce qui est de la troisième dimension du questionnaire omnimode, des variables statistiques consolidées, elle constitue un champ d'évolution à considérer. Il semblerait efficace de pouvoir spécifier tant la variable statistique, que son mode de collecte (la question, son libellé, etc.), que les contraintes de format (les différents contrôles dans les questionnaires) et aller jusqu'aux règles de consolidation ou d'imputation automatique dont elle doit faire l'objet post-collecte.

Ainsi, en reprenant l'exemple didactique précédent, en même temps que le libellé de la question « Avez-vous personnellement accès à internet ? », une règle d'imputation automatique à « Oui » pour les répondants issus du mode *web* pourrait être spécifiée. Ou encore la spécification d'une variable à valeur unique prise dans une liste de modalité (une « liste déroulante » sur internet) pourrait être enrichie des règles de redressement et de consolidation multimode : on pense à l'exemple classique de la réponse exclusive « oui/non » pour laquelle un répondant facétieux peut cocher les deux réponses sur un questionnaire papier ; à l'occasion de la conception de la question, une règle de redressement automatique pourrait également être spécifiée (mettre à blanc ou privilégier le « Oui » en cas de réponses multiples non souhaitées).

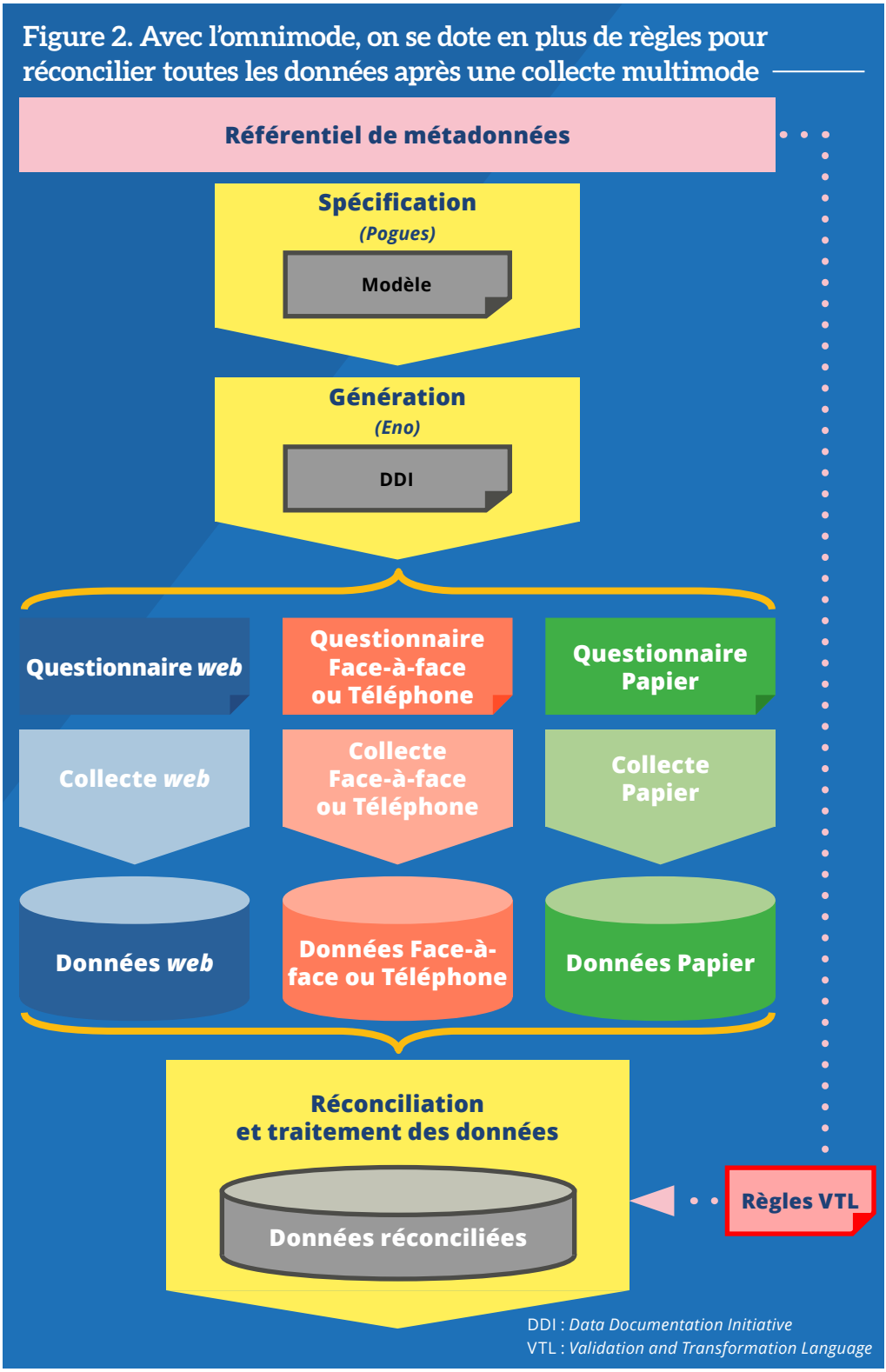
Ce type de traitement est actuellement réalisé *a posteriori* dans le traitement en aval de la collecte, et de manière spécifique à chaque enquête. Une règle d'imputation de ce type nécessite un certain formalisme d'écriture, et viendrait enrichir les éléments de spécification du questionnaire pour décrire ces traitements tout au long du processus. L'Insee a engagé des travaux autour de VTL, méta-langage de spécification pour les règles algorithmiques, promu par Eurostat : on spécifie dans le même langage, les filtres, les contrôles en cours de collecte ou post-collecte, les variables calculées au sein du questionnaire, ou les variables statistiques à livrer[19]. Et on « active » au passage de nouvelles métadonnées.

> *« Plus qu'un simple questionnement, il s'agit de concevoir la captation de variables statistiques, un « arbre de variables » dont on souhaite disposer. »*

À partir d'une conceptualisation omnimode de ce que doit être le questionnement statistique, on limite les effets de mode, on améliore l'exploitabilité statistique des résultats et on rationalise les outils mobilisés dans la consolidation des variables collectées. Ainsi, plus qu'un simple questionnement, il s'agit de concevoir la captation de variables statistiques, un « arbre de variables » dont on souhaite disposer, demandant l'instanciation de processus de collecte multiples, et l'ensemble des réponses collectées devant se « réconcilier » dans une base consolidée ***(figure 2)***.

19. Cette spécification consolidée peut se décliner ensuite automatiquement dans les différents outils et traitements bien que dans des technologies informatiques différentes (Java, JavaScript, etc.). Voir (SDMX, 2020 ; Banca d'Italia, European Central Bank et Insee, 2021).

Figure 2. Avec l'omnimode, on se dote en plus de règles pour réconcilier toutes les données après une collecte multimode



DDI : *Data Documentation Initiative*
VTL : *Validation and Transformation Language*

LES ENJEUX DU PROCESSUS MULTIMODE (DES ACTEURS QUI DOIVENT AGIR DE CONCERT)

Une collecte multimode, s'appuyant sur un questionnaire omnimode pose encore un certain nombre de problématiques opérationnelles et organisationnelles.

> *« Découpler les enjeux méthodologiques et statistiques d'un concepteur et ceux plus opérationnels d'un maître d'œuvre du processus de collecte. »*

Quelle charge représente pour un enquêteur un questionnaire collecté par internet ? Un questionnaire repris d'internet ? Comment correctement coordonner l'action d'un enquêteur quand il y a des relances papier de masse ? Comment répercuter des événements complexes tels que les éclatements de ménage, de logement ou de budget sur les différents processus ? Comment organiser correctement le travail des enquêteurs quand de tels évènements proviennent de réponses collectées par internet ?

La qualité statistique obtenue dépend tout autant de la conception rigoureuse du questionnement que de la mise en œuvre du processus de collecte où chaque acteur a son rôle à jouer. Avec l'arrivée du multimode, c'est tout un processus et les rôles et responsabilités afférents des différents acteurs qui est à penser. Et là encore, découpler les enjeux méthodologiques et statistiques d'un concepteur de questionnement (questions correctement comprises, variables d'intérêt identifiées, protocole maîtrisé) et ceux plus opérationnels d'un maître d'œuvre du processus de collecte (suivi d'activité des enquêteurs, gestion du réseau enquêteur, priorisation fine d'opérations) relève d'une étape préalable. Leurs objectifs, sans être divergents, gagnent à être respectivement identifiés et consolidés.

UN NOUVEAU CONCEPT ? L'UNITÉ D'INTÉRÊT D'ENQUÊTE (UNE NOUVELLE PORTÉE)

La maîtrise d'ouvrage de l'enquête, qui conçoit une enquête, son protocole, son questionnaire, son calendrier, etc. s'intéresse avant tout à l'unité statistique « initiale » ou « finale ».

C'est le concept d'**unité d'intérêt d'enquête**[20] qui est ici introduit : il correspond à l'unité de compte statistique issue de l'échantillonnage initial. Il garantit que les données liées à cette unité seront correctement collectées (pas de doublon, pas de « trous de collecte ») et correctement livrées (une réconciliation technique sous forme de donnée statistique avant la correction d'effet de modes éventuels).

Le maître d'œuvre statistique doit pour sa part organiser le travail des différents acteurs, et mettre en œuvre les outils nécessaires, afin de permettre le bon déroulement des différents processus de collecte, parfois concomitants, sur plusieurs enquêtes. Il a besoin, lui, d'une unité de collecte, l'**unité enquêtée**, qui va lui garantir que les événements, les coûts et charges induits, liés au déroulement du processus de collecte d'une unité seront correctement captés pendant les opérations.

Sans cette dualité de concept, les informations orientées « contrôle de gestion » sont mélangées avec celles plus orientées « statistique » et les objectifs de chacune de ces deux approches sont dévoyés.

20. Qu'elle fasse l'objet d'une collecte multimode concurrentielle ou séquentielle, elle reste une unité de compte unique, à laquelle peuvent être rattachées plusieurs unités enquêtées, dans des modes différents, à des moments différents. Elle peut formellement être rapprochée du concept d'unité d'intérêt statistique du GSIM (*General Statistical Information Model*).

Un exemple caractéristique est celui de la filière historique Insee où étaient produits des « codes résultats ». Ces agrégats étaient calculés en cours de collecte afin d'offrir tant des éléments pour le calcul de la performance des enquêteurs que ceux nécessaires au redressement, imputation et autres traitements statistiques dit « aval ». La liste des « codes résultats » existants était devenue complexe, difficile à maintenir, difficile à consolider entre les enquêtes et engendrait parfois des compromis imposés quant au besoin statistique ou opérationnel couvert. L'ambiguïté sur le terme de « hors-champ » en est une bonne illustration :

- il peut être entendu au sens opérationnel du terme – un logement détruit par exemple – et n'être comptabilisé ni dans le calcul de performance des enquêteurs, ni dans les résultats statistiques ;
- ou entendu au sens statistique exclusif – un ménage sans fonctionnaire pour une enquête sur la fonction publique d'état comme autre exemple – et pourtant devant être comptabilisé comme une enquête réussie pour l'enquêteur.

Pour répondre aux deux besoins, il convient de découpler les concepts, et surtout de déporter le calcul des agrégats dans les processus qui en sont consommateurs : l'un pour charger un infocentre utile au contrôle de gestion sur l'activité des enquêteurs, l'autre pour alimenter les chaînes de traitements avec la qualification de la réponse, ou plus souvent de la non-réponse, statistique. À l'instar des deux portails adaptés aux deux types de répondants aux enquêtes, ce découplage permet de limiter la complexité induite et de garantir que chacun voit son besoin couvert, chacun étant responsable de ses propres agrégats.

Aujourd'hui, les outils que l'Insee met à disposition des enquêteurs implémentent ces principes. Par exemple, les fonctions d'organisation et de suivi de la collecte sont dans des « briques » applicatives séparées des fonctions liées à la conduite d'un questionnement statistique.

SAVOIR ET POUVOIR S'ENTENDRE ENTRE MODES (POUR ÉVITER LA CACOPHONIE)

La dernière étape de conceptualisation, non la moindre, pour être en mesure de mettre en œuvre correctement une collecte multimode concerne la « communication multimode ».

Un processus multimode permet de mettre en œuvre une collecte quel que soit le mode. Une consolidation au niveau du questionnement statistique garantit une exploitabilité statistique. Une consolidation au niveau de l'unité de compte statistique garantit un pilotage cohérent et efficace des collectes successives d'une enquête. Mais une dimension supplémentaire n'est pas encore prise en compte : le besoin des processus de communiquer « en direct » les uns avec les autres.

Cette communication est indispensable pour permettre une collecte multimode concurrentielle ou à séquences rapprochées (par exemple, les processus « collecte *web* » et « collecte enquêteur » devant s'avertir l'un l'autre), pour permettre un suivi consolidé des différentes collectes (le suivi des différentes unités enquêtées, quel que soit leur mode de collecte, doit être cohérent et consolidé dans une même interface), ou encore la mise en œuvre de collectes hybrides (comme celles à base de carnet de collecte, où le démarrage de la collecte *web* du carnet doit être synchronisé avec le passage d'un enquêteur).

Dans une collecte multimode « totale », on peut facilement imaginer trois à quatre processus de collecte simultanés devant communiquer les uns avec les autres. Ces différents processus en forment-ils un seul complexe, ou plusieurs se « parlant » ? Comment s'assurer que cette communication ne va pas devenir effectivement cacophonique ? En clair, comment doivent être conçues ces collectes parallèles ?

LES DIFFÉRENTS MODES DE COLLECTE À L'UNISSON

À l'Insee cette conception a été assez « organique ». Partant d'une conceptualisation initiale poly-mode, où un processus de collecte peut être mis en œuvre plusieurs fois dans des modes différents, de premières opérations d'enquêtes ont pu être conduites. Ainsi quand la délicate question de faire communiquer ces processus s'est posée – par exemple quand une collecte internet doit « prévenir » un enquêteur qu'une réponse a été reçue, ou encore quand un enquêteur doit « prévenir » un carnet *web* de collecte qu'un nouveau répondant va arriver – l'enjeu était de préserver la stabilité des systèmes existants, comme de pérenniser les opérations de collecte déjà mises en œuvre.

Dans une approche « mono-processus », un processus complexe avec plusieurs instruments de collecte, plusieurs règles de gestion et de synchronisation, permet de gérer toutes les collectes, de la plus simple ne mobilisant qu'un mode de collecte, jusqu'aux multimodes. Plutôt que d'opter pour cette approche, le choix s'est porté sur une approche « poly-processus », où plusieurs processus cohabitent indépendamment les uns des autres et communiquent avec un système central ***(figure 3)***. Ce système est alors le chef d'orchestre de la collecte multimode, seul porteur de la complexité des règles de gestion liées aux protocoles (quels modes faut-il mobiliser, quelles actions déclencher, etc.).

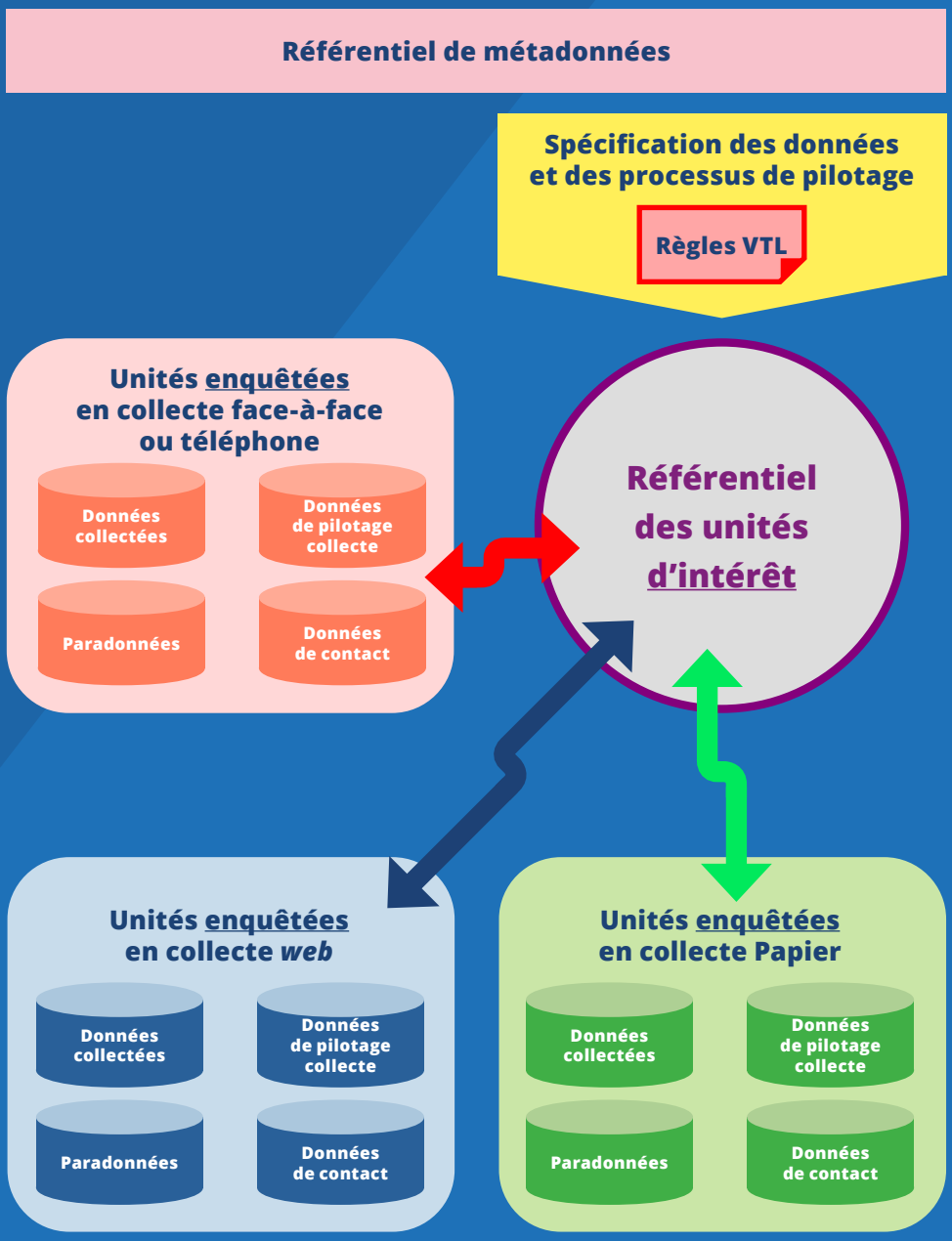
Ainsi, la montée en charge et la complexité croissante des protocoles considérés ont des impacts limités sur la complexité de chaque processus. D'une part, seuls les protocoles de collecte multimode les plus complexes requièrent une réelle communication inter-mode (collecte concurrentielle pour les chiffres de l'emploi, ou collecte s'appuyant sur des carnets d'activité ou de budget notamment). La grande majorité des collectes, mêmes multimodes, sont séquentielles et un modèle « poly-processus » convient parfaitement. D'autre part, c'est le système central, en charge de la communication entre chacun des processus, qui est le plus impacté par la complexité à mettre en œuvre. Les évolutions à porter dans les systèmes existants pour supporter les protocoles les plus complexes sont ainsi limitées (schématiquement des échanges de « messages » avec un système central).

PREMIÈRES OPÉRATIONS MONO-MODES (AVEC DES TONALITÉS MULTIMODES)

Les premières étapes du passage à l'échelle, conforme avec le cadre conceptuel défini, visent à sécuriser des systèmes de collecte indépendants. Une première collecte par internet dans un contexte ménage, une première collecte par téléphone, une première collecte par face-à-face.

Si ces premières opérations ne mobilisent pas de « communication » multimode, à proprement parler, la dimension multimode est déjà présente dans le processus et le travail des différents acteurs. Par exemple, la première collecte par téléphone se déroule dans une opération multimode séquentielle, le pilote 2022 de la nouvelle enquête Logement.

Figure 3. Le pilotage d'une collecte multimode génère un fort besoin de synchronisation



VTL : *Validation and Transformation Language.*

L'unité enquêtée est l'unité de compte d'une opération de collecte.

L'unité d'intérêt est l'unité de compte statistique.

À une unité d'intérêt peut correspondre plusieurs unités enquêtées (par exemple si un questionnement web s'enchaîne avec une collecte par téléphone pour un même ménage ou une même entreprise.).

Dans des collectes séquentielles, où le premier mode est le *web*, des stratégies d'optimisation des relances et de bascule entre les modes sont mises en œuvre. Elles nécessitent que les acteurs maîtrisent les différents processus et puissent manipuler les différents types de données (données collectées, paradonnées, données de suivi) même si les outils ne communiquent pas formellement directement.

Dans ces mêmes collectes, mobiliser des réponses antérieures issues d'internet dans une collecte par enquêteurs, nécessite que soit précisé le cadre méthodologique. Attend-on de l'enquêteur une reprise exhaustive du questionnaire, les réponses internet servant de « fil rouge » à un entretien classique ? Ou au contraire une réponse efficace, les réponses internet servant de « pis-aller » pour permettre un entretien raccourci ? La formation des enquêteurs, la prise en compte de ce nouveau métier dans leurs tâches fait partie intégrante des impacts du multimode, et ce dès les premières opérations réputées « mono-modes ».

Plus généralement, décliner et formaliser les protocoles multimodes en consignes, tâches ou rôles des acteurs intervenants dans la collecte (de l'enquêteur, à la maîtrise d'ouvrage d'enquête ou la maîtrise d'œuvre statistique) est l'un des enjeux de ces premières opérations.

GAGNER EN VIRTUOSITÉ EN ABORDANT UN MULTIMODE DE PLUS EN PLUS COMPLEXE

Avec la mise en œuvre de collectes multimodes plus complexes, ce ne sont plus seulement les applications, les outils ou les systèmes techniques qui doivent évoluer, mais aussi les métiers.

La méthode d'analyse préalable des processus cible est encore à l'œuvre, avec son lot de questions :

- Qu'est-ce que le suivi multimode ? Sur quels indicateurs s'appuyer pour suivre une collecte consolidée ?
- Qu'est-ce que le contrôle qualité des questionnaires internet dans la sphère ménage ? Faut-il se calquer sur un processus type « entreprise », s'appuyant sur des gestionnaires ? Ou faire intervenir l'expertise de l'enquêteur ?
- Combien de temps prend effectivement la reprise d'un questionnaire internet par un enquêteur ? Plus longtemps, car il y a un temps de préparation préalable et une phase d'accompagnement du répondant un peu plus longue ? Ou moins longtemps, car l'entretien est raccourci ?

Autant de questions complexes qui ne sauraient accepter des réponses *a priori*. L'expérimentation sera donc encore de mise, pour pouvoir tester et constater dans les faits. Cela nécessitera que les processus conçus et que les outils proposés soient suffisamment souples : des services standardisés faciles à mettre en œuvre, des acteurs formés et sensibilisés au caractère expérimental. Car dans un système de collecte multimode, la complexité impose de la souplesse, pour que s'approfondisse la connaissance de ce nouveau métier et que les interprètes des prochaines collectes s'approprient progressivement les techniques de demain.

BIBLIOGRAPHIE

BANCA D'ITALIA, EUROPEAN CENTRAL BANK et INSEE, 2021. *VTL Community. Towards a community of VTL developers.* [en ligne]. GitHub. [Consulté le 6 décembre 2021]. Disponible à l'adresse : *https://github.com/VTL-Community/VTL-Community*.

BONNANS, Dominique, 2019. RMéS, le référentiel de métadonnées statistiques de l'Insee. In : *Courrier des statistiques*. [en ligne]. 27 juin 2019. Insee. N° N2, pp. 46-57. [Consulté le 6 décembre 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/4168396/courstat-2-6.pdf*.

COTTON, Franck et DUBOIS, Thomas, 2019. Pogues, un outil de conception de questionnaires. In : *Courrier des statistiques*. [en ligne]. 19 décembre 2019. N° N3, pp. 17-28. [Consulté le 6 décembre 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/4254216/courstat-3-3.pdf*.

COTTON, Franck, DUBOIS, Thomas, SIGAUD, Éric et WERQUIN, Benoît, 2021. *A fully metadata-driven platform for the conception of survey questionnaires and the management of multimode data collection.* [en ligne]. 27-30 septembre 2021. Unece, Conférence des statisticiens européens, Expert Meeting on Statistical Data Collection. [Consulté le 6 décembre 2021]. Disponible à l'adresse : *https://unece.org/sites/default/files/2021-09/DC2021_S1_France_Werquin%20Cotton_AD.pdf*.

ERIKSON, Johan, 2020. Le modèle de processus statistique en Suède – Mise en œuvre, expériences et enseignements. In : *Courrier des statistiques.* [en ligne]. 29 juin 2020. Insee. N° N4, pp. 122-141. [Consulté le 6 décembre 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/4497085/courstat-4-8.pdf*.

GUILLAUMAT-TAILLIET, François et TAVAN, Chloé, 2021. Une nouvelle enquête Emploi en 2021, entre impératif européen et volonté de modernisation. In : *Courrier des statistiques*. [en ligne]. 8 juillet 2021. Insee. N° N6, pp. 7-27. [Consulté le 6 décembre 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/5398681/courstat-6-art-1.pdf*.

HAAG, Olivier et HUSSEINI-SKALITZ, Anne, 2019. Collecte par internet des enquêtes auprès des entreprises : l'Insee entre dans l'ère Coltrane. In : *Courrier des statistiques*. [en ligne]. 19 décembre 2019. N° N3, pp. 45-60. [Consulté le 6 décembre 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/4254223/courstat-3-5.pdf*.

KOUMARIANOS, Heïdi et SIGAUD, Éric, 2019. Eno, un générateur d'instruments de collecte. In : *Courrier des statistiques*. [en ligne]. 19 décembre 2019. N° N3, pp. 29-44. [Consulté le 6 décembre 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/4254218/courstat-3-4.pdf*.

SDMX, 2020. *Validation and Transformation Language (VTL).* [en ligne]. Mise à jour du 4 août 2020. The official site for the SDMX community. A global initiative to improve Statistical Data and Metadata eXchange. [Consulté le 6 décembre 2021]. Disponible à l'adresse : *https://sdmx.org/?page_id=5096*.

UNECE, 2019. *Generic Statistical Business Process Model GSBPM.* [en ligne]. Janvier 2019. Version 5.1. [Consulté le 6 décembre 2021]. Disponible à l'adresse : *https://statswiki.unece.org/display/GSBPM/GSBPM+v5.1*.

LE RECENSEMENT AGRICOLE DE 2020

CINQ INNOVATIONS QUI FERONT DATE

*Hervé Le Grand**

Opération internationale et décennale, le recensement agricole ambitionne de produire une photographie complète du monde agricole. Cette opération majeure de la statistique agricole est au cœur de son système d'information. L'édition de 2020 a été conduite entre octobre 2020 et mai 2021, malgré le contexte sanitaire, en s'appuyant sur cinq innovations majeures impactant les enquêtés, les enquêteurs et les statisticiens.

Pour la première fois, le recensement a été collecté majoritairement par internet ou par téléphone. Le terrain reste l'apanage des enquêteurs du réseau agricole, notamment dans les départements d'Outre-mer. Mais le recours à des prestataires spécialisés permet de développer des alternatives au face-à-face, plus adaptées aux contraintes des enquêtés. Le multimode a été doublé d'un recours massif aux données administratives, afin d'alléger la charge de réponse et d'améliorer la qualité des données. Pour faciliter l'adhésion à ces nouveaux modes de collecte, une démarche inspirée du* Nudge *a conduit à repenser les supports de communication. Enquêteurs et prestataires ont par ailleurs bénéficié d'une formation à distance : en soi c'est une petite révolution pour une opération plus que centenaire. Enfin, le plan de sondage des modules complémentaires sur l'élevage et la main-d'œuvre a été optimisé.

The agricultural census is an international, ten-yearly operation designed to provide a complete picture of the agricultural world. This major operation of agricultural statistics is at the heart of its information system. The 2020 edition was conducted between October 2020 and May 2021, despite the health context, based on five major innovations impacting the respondents, the interviewers and the statisticians.

For the first time, the census was collected mainly by internet or telephone. Fieldwork remains the prerogative of the agricultural network's interviewers, particularly in the overseas departments. However, the use of specialised service providers makes it possible to develop alternatives to face-to-face interviews that are better adapted to the constraints of the respondents. The multi-mode approach has been coupled with extensive use of administrative data, in order to lighten the response burden and improve the quality of the data. To facilitate adherence to these new collection methods, an approach inspired by Nudge led to a rethinking of the communication media. Interviewers and service providers also benefited from distance training: in itself a small revolution for an operation that is more than a century old. Finally, the survey plan for the complementary modules on livestock and labour was optimised.

* À la date de rédaction de l'article, chef du bureau des Statistiques structurelles, environnementales et forestières, Service de la statistique et de la prospective du ministère de l'Agriculture et de l'Alimentation,
herve.le-grand@insee.fr

Tous les dix ans, le recensement agricole permet de collecter de nombreuses données sur l'ensemble des exploitations françaises. Opération centrale de la statistique agricole, le recensement est aussi une des plus importantes opérations statistiques réalisées en France.

Le recensement concerne l'ensemble des exploitations agricoles, qu'elles soient grandes ou petites, que l'agriculture corresponde à l'activité principale ou secondaire de l'exploitant. Afin d'assurer une couverture complète du monde agricole, ce sont ainsi plus de 500 000 unités qui sont interrogées.

Pour chaque exploitation, plus de 900 données sont recueillies sur les superficies cultivées, les cheptels, la main-d'œuvre, les modes de production et de commercialisation ainsi que les activités de diversification et de transformation des produits à la ferme. Le recensement fournit ainsi une photographie précise et exhaustive du monde agricole et de sa diversité, en France métropolitaine mais aussi dans les départements d'Outre-mer.

L'agriculture est en constante adaptation, faisant écho aux évolutions de la société. Le secteur utilise les technologies informatiques, diversifie ses modes de commercialisation et ses débouchés, revoit ses pratiques pour préserver l'environnement, développe des labels de qualité, etc. C'est toute la réalité des professionnels du monde agricole qui change. De même, le recensement agricole se devait d'évoluer et la collecte 2020 a ouvert la voie à plusieurs innovations structurantes.

Le recueil des données est désormais réalisé en majorité par internet ou par téléphone. L'ensemble de la communication adressée aux enquêtés a été révisée. Le volet de données complémentaires sur la main-d'œuvre et les bâtiments d'élevage est collecté sur un échantillon pour lequel le plan de sondage a été optimisé. Les enquêteurs ont été formés en ligne. Autant d'innovations qui ont facilité la poursuite de la collecte malgré le contexte sanitaire imposé par la Covid-19.

LE RECENSEMENT AGRICOLE, ESSENTIEL POUR LE PILOTAGE DES POLITIQUES PUBLIQUES…

Du fait de leur exhaustivité, les données des recensements constituent des références importantes pour tous les acteurs du monde agricole : exploitants, organisations professionnelles, syndicats, chercheurs, pouvoirs publics, etc. Elles permettent d'analyser l'agriculture française et ses évolutions, et d'en mesurer le poids au sein de l'Union européenne. L'agriculture est certes un secteur stratégique en tant que secteur économique, mais aussi pour son rôle dans l'indépendance alimentaire, face à l'augmentation de la population et au changement climatique. Dans ce contexte, les résultats du recensement agricole vont contribuer au pilotage et à l'évaluation des politiques publiques agricoles et alimentaires mises en œuvre aux échelons régional, national, et communautaire.

Les données du recensement permettent également de quantifier des mesures liées au développement durable et aux politiques agro-environnementales et d'évaluer les politiques de soutien économique aux exploitations. Les résultats du précédent recensement ont par exemple été utilisés pour simuler des modifications apportées dans les règles de classement des territoires éligibles à l'indemnité compensatoire de handicaps naturels (ICHN). Cette aide vient soutenir les agriculteurs installés dans des territoires où les conditions de production sont plus difficiles qu'ailleurs, du fait de contraintes naturelles ou spécifiques.

Le recensement agricole permet d'évaluer l'état de l'agriculture, mais aussi sa position et son évolution, en comparant les résultats à ceux des précédents recensements, ou à ceux des autres pays européens. Ce besoin d'information sur le secteur agricole dépasse de loin le seul cadre national.

...QUI S'INSCRIT DANS UN CADRE INTERNATIONAL

La FAO, l'organisation des Nations Unies pour l'alimentation et l'agriculture[1], a établi un programme mondial et élaboré des concepts et des méthodes pour appuyer la réalisation coordonnée des recensements agricoles dans le monde (FAO, 2020).

Au sein de l'Union européenne, le règlement sur les statistiques intégrées sur les exploitations agricoles (IFS – *Integrated Farm Statistics*[2]) organise la réalisation du recensement dans les 27 États membres (Eurostat, 2020).

Les périodes de collecte sont harmonisées. Si un socle commun d'informations doit être transmis à Eurostat par l'ensemble des pays, chacun peut également décider d'ajouter des questions pour répondre à des besoins nationaux.

Entre deux recensements décennaux, sont intercalées des **enquêtes sur la Structure des exploitations agricoles (Esea)** pour actualiser les données ***(figure 1)***. Les prochaines Esea seront menées en 2023 et 2026.

Le règlement européen distingue par ailleurs deux types de données : un tronc commun devant être renseigné exhaustivement pour chaque exploitation et huit modules pouvant être collectés par échantillon. En 2020, les trois modules à collecter portaient sur la main-d'œuvre, le développement rural et le logement des animaux. En 2023, le module sur le logement des animaux sera remplacé par quatre modules : irrigation, gestion du sol, machines et équipements, vergers.

Les États membres doivent être en mesure de garantir la qualité des données collectées, mais restent libres de choisir la façon dont ils obtiennent les informations demandées : questionnement par internet, par téléphone, en face-à-face, mobilisation de sources administratives, etc. La collecte par internet a été développée dans de nombreux pays (Autriche, Allemagne, Espagne, etc.).

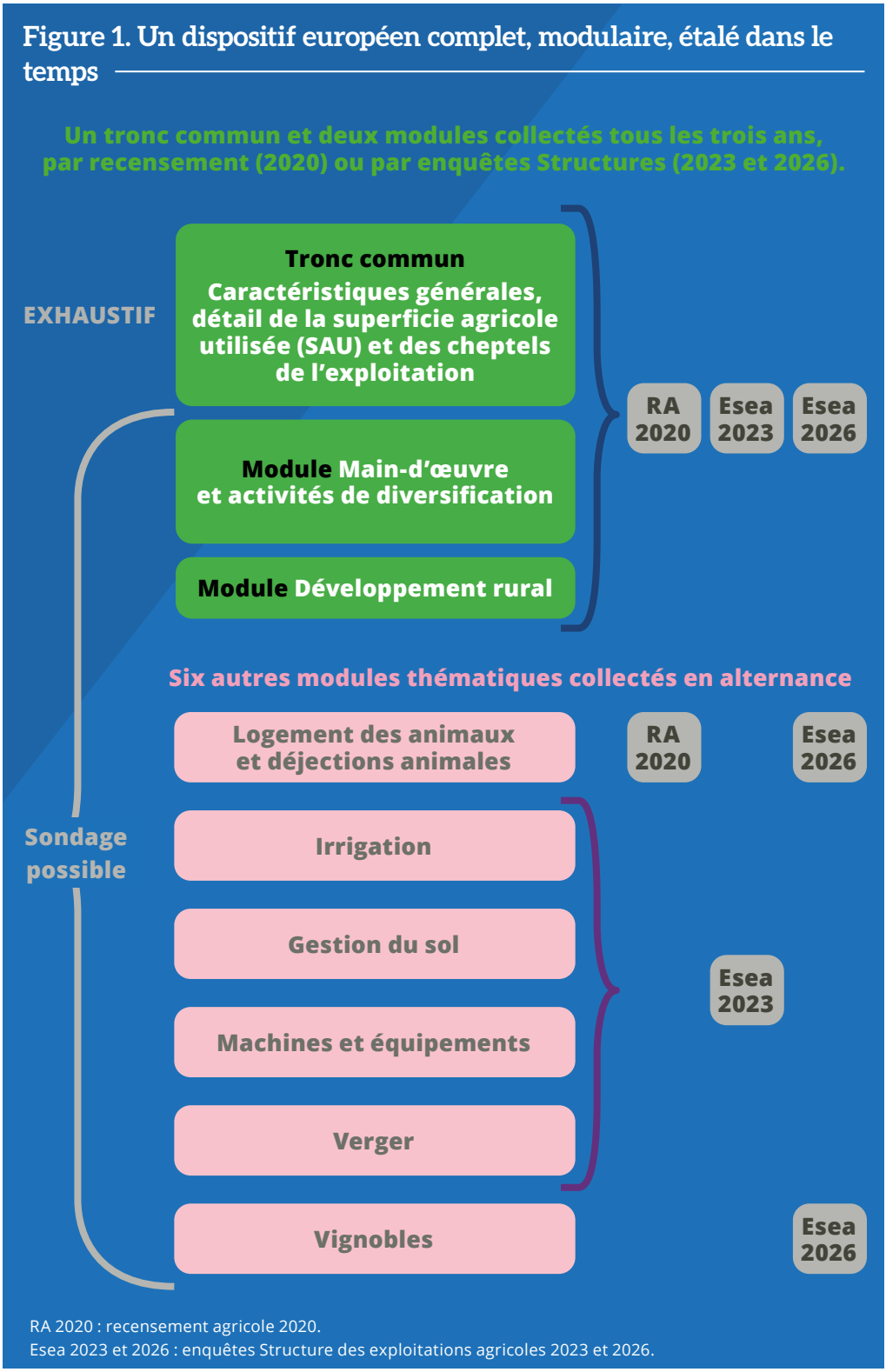
> *« Le contexte sanitaire a aussi ouvert la voie à une modernisation du processus de collecte et des méthodes de travail. »*

L'organisation du recensement agricole de l'année 2020 a été touchée par la pandémie de Covid-19 dans la plupart des pays de l'Union européenne. Celle-ci a retardé plusieurs activités, notamment la préparation des outils de collecte, la diffusion d'informations aux personnes interrogées, le recrutement des enquêteurs, la formation du personnel et la collecte des données elle-même. Toutefois, le contexte sanitaire a aussi ouvert la voie à une modernisation du processus de collecte et des méthodes de travail. Plusieurs pays, dont la France, ont ainsi davantage eu recours aux sources de données administratives et à la collecte des questionnaires par internet et par téléphone.

Mais quand on parle de recensement, on pense d'abord à s'assurer de son exhaustivité.

1. FAO (*Food and Agriculture Organization*).
2. Voir les références juridiques en fin d'article.

Figure 1. Un dispositif européen complet, modulaire, étalé dans le temps



RA 2020 : recensement agricole 2020.
Esea 2023 et 2026 : enquêtes Structure des exploitations agricoles 2023 et 2026.

DÉFINIR LES UNITÉS STATISTIQUES DANS LE CHAMP DU RECENSEMENT

Un recensement se doit d'être exhaustif. C'est la raison pour laquelle, la première question à régler est celle de l'univers statistique de cette opération. Toutes les exploitations agricoles sont interrogées dans le recensement agricole, aussi bien en France métropolitaine que dans les départements d'Outre-mer. Mais une seule personne sur chaque exploitation est invitée à répondre à l'enquête : c'est en général le chef d'exploitation, même dans les cas où l'agriculture ne constitue pas son activité principale.

Certaines unités interrogées s'avéreront *ex-post* comme hors champ[3] du recensement agricole : pour ne retenir dans les résultats que les exploitations dépassant une certaine dimension économique, des règles sont appliquées à partir des caractéristiques de l'exploitation relevées lors du recensement (superficie cultivée, nature de la production, etc.) ***(encadré 1)***. Ces règles s'appuient sur des seuils, définis aux niveaux européen et national. Les seuils retenus au niveau national sont généralement plus bas qu'au niveau européen. Ils sont calqués pour la plupart sur ceux en vigueur lors du précédent recensement (2010) afin de permettre une continuité dans les séries statistiques. Le règlement européen sur les statistiques agricoles (IFS, voir *infra*) fixe ainsi un seuil de surface agricole utilisée (*SAU*) minimale de 5 ha, contre 1 ha pour le seuil français.

Le respect de ces conditions assure en théorie que l'exploitation retenue est un acteur économique en capacité à participer à une transaction commerciale à partir de sa production.

Mais comment établir *ex-ante* la liste des unités statistiques qui constituent le champ de la collecte ?

POUR CONSTITUER L'UNIVERS STATISTIQUE DU RECENSEMENT, ON A FAIT FEU DE TOUT BOIS

Lors du recensement de 2010, le service de la Statistique et de la prospective (SSP) avait fait appel aux 36 000 communes qui devaient vérifier des listes d'exploitants agricoles sur leur territoire, établies par les services régionaux (Srise)[4]. Ce travail, certes utile, était cependant sujet à erreurs d'interprétation sur le concept d'exploitation agricole, omissions, bref de qualité très hétérogène.

Pour éviter ce travers, l'univers statistique du recensement de 2020 a été délibérément établi sur une large base d'informations déjà disponibles. Partant du répertoire des entreprises (*Sirene*), géré en continu à partir des créations et cessations déclarées, il a été complété avec les fichiers administratifs du monde agricole : déclarations dans le cadre de la politique agricole commune (PAC), casier viticole informatisé[5], fichiers des mouvements des animaux, mutualité sociale agricole (MSA) et à l'aide des enquêtes thématiques réalisées par le service de la statistique et de la prospective (SSP).

3. *In fine*, 13 % des exploitations agricoles ont été classées hors champ du recensement agricole, essentiellement en raison de cessation d'activité ou d'activité interrompue.
4. Le service statistique ministériel de l'Agriculture est constitué du Service de la statistique et de la prospective (SSP) et des Services régionaux de l'information statistique et économique (Srise).
5. Le casier viticole informatisé (CVI) est un outil que les États membres de l'Union européenne doivent tenir obligatoirement. Il contient notamment toutes les informations relatives aux entreprises viti-vinicoles, aux parcelles plantées ou arrachées, les niveaux de production et de stock. En France, ce répertoire est géré par les services des douanes.

Encadré 1. Des règles en cascade pour entrer dans le champ du recensement agricole

Pour ne retenir dans les résultats que les exploitations dépassant une certaine dimension économique, deux types de seuils sont définis, aux niveaux européen et national. Cette question s'avère complexe, car les seuils sont nombreux et leur application obéit à une série de règles en cascade, appliquées après la collecte.

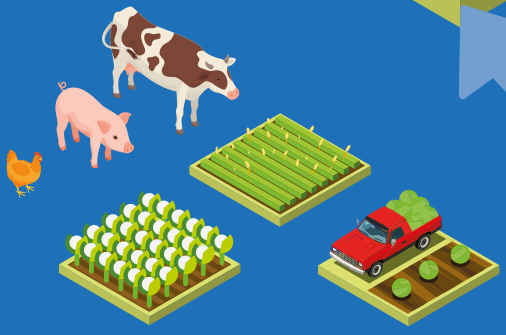
Superficie agricole utilisée ≥ 1 ha

Superficie en cultures spécialiées (houblon, semences, maraîchage, vignes, etc.) ≥ 0,2 ha

Règle n°1

Règle n°2

Règle n°3



Production agricole (en nombre d'animaux; surface de production, volume de production) supérieure à un seuil ou production spécialisée à forte valeur ajoutée

La 3e règle concerne en particulier les exploitations orientées vers l'élevage et dépourvues de surface comme c'est le cas pour les élevages hors-sol. Elle vise aussi à conserver dans le champ du recensement certaines productions spécialisées à forte valeur ajoutée pour lesquelles la surface minimale de 20 ares s'avère trop élevée. Ainsi, 5 ares de vigne en champagne figurent parmi les seuils imposés par la règle n° 3 et permettront de retenir dans le champ des petites exploitations.

Un important travail a été ensuite réalisé pour actualiser les coordonnées des exploitations, en particulier les courriels et numéros de téléphone. Ainsi a été constitué un référentiel, la base de sondage des exploitations agricoles (Balsa), en rénovation dès 2021 pour qu'à terme, les données du recensement agricole permettent de valider l'univers des exploitations et que l'intégration de données administratives soit annualisée et automatisée. Ceci fait du recensement un élément central pour les autres opérations de la statistique agricole.

LE MULTIMODE : UNE AVANCÉE MAJEURE POUR LE RECENSEMENT DE 2020 EN FRANCE

En France, le recensement agricole a depuis son origine été collecté en face-à-face par des enquêteurs[6]. Lors de l'édition de 2010, les enquêteurs avaient pour la première fois été dotés d'ordinateurs pour contrôler en direct les données saisies. Mais c'est la mise en œuvre d'une véritable collecte multimode qui aura constitué la principale innovation du recensement agricole de 2020.

> *« La mise en œuvre d'une véritable collecte multimode aura constitué la principale innovation du recensement agricole de 2020. »*

La refonte du mode de collecte poursuivait deux objectifs. Tout d'abord, la collecte par internet devait faciliter la remontée des données : le service est ouvert 24 h/24 et 7 j/7, les exploitants peuvent donc renseigner le questionnaire au moment qui leur convient le mieux, en fonction du planning de travail imposé par la conduite de leur exploitation. Elle devait également permettre d'alléger la charge du service statistique[7].

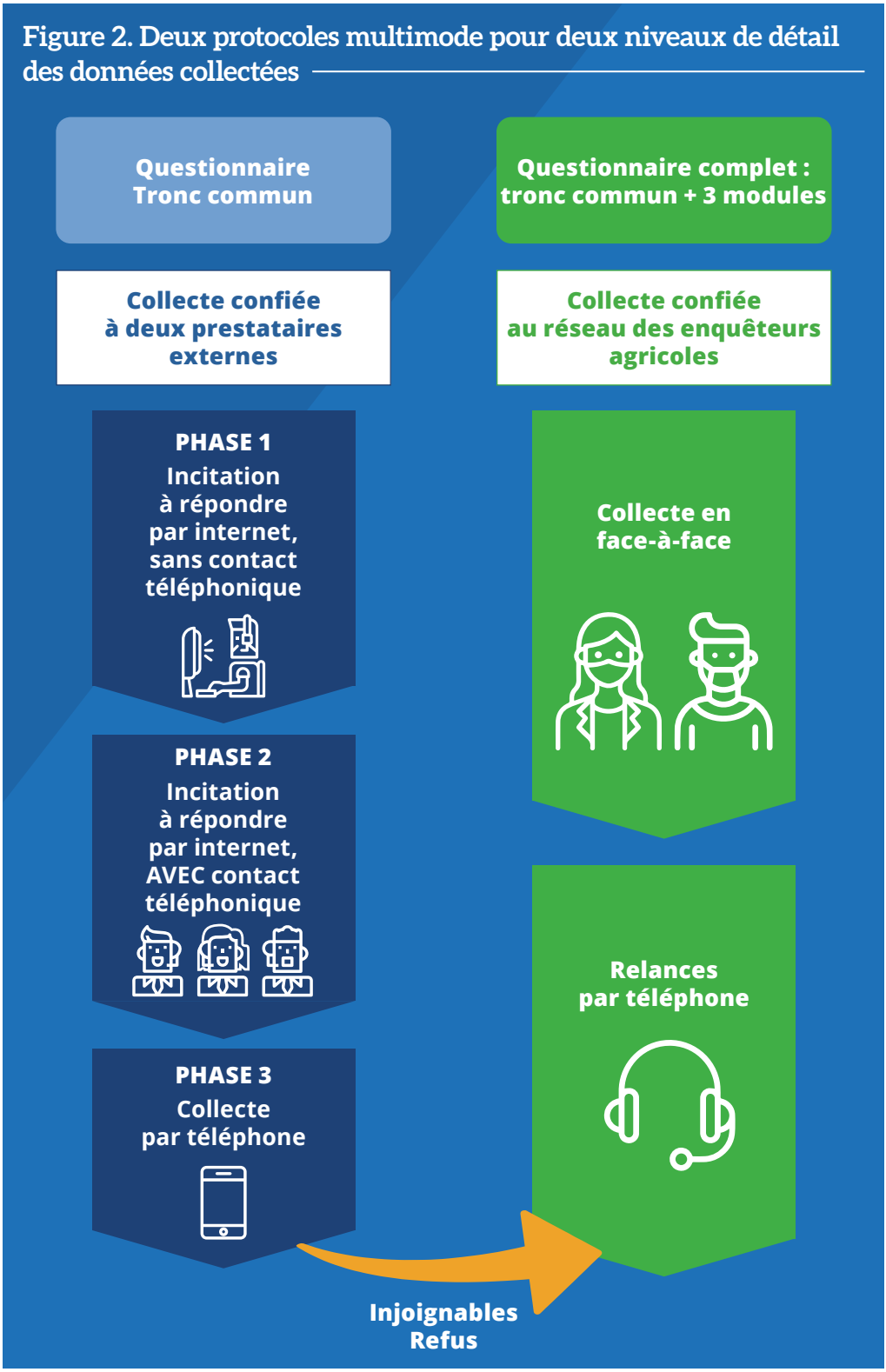
Concrètement, en 2020 et pour la métropole[8], le service statistique agricole a donc choisi de mobiliser les trois modes[9], internet, téléphone et face-à-face ***(figure 2)*** :

- le tronc commun était dans la plupart des cas proposé *via* un questionnaire internet ;
- pour les personnes sans accès à internet ou rencontrant des problèmes informatiques, il était possible de basculer sur la collecte par téléphone, faisant appel à des prestataires externes ;
- les exploitations de l'échantillon concerné par les modules ont été interrogées en face-à-face par des enquêteurs du réseau de la statistique agricole, pour les modules comme pour le tronc commun ;
- les non-répondants à la collecte par internet ou par téléphone ont été *in fine* relancés par des enquêteurs.

La collecte s'est déroulée entre octobre 2020 et mai 2021.

6. La statistique agricole mobilise de longue date un réseau d'enquêteurs, pour la plupart issus du monde agricole.
7. Voir l'article de François Beck, Laura Castell, Stéphane Legleye et Amandine Schreiber sur la méthodologie de la collecte multimode dans ce même numéro.
8. Voir les adaptations apportées à la collecte dans les départements d'Outre-mer dans l'***encadré 2***.
9. En toute rigueur on devrait parler de quatre modes, puisqu'un peu moins de 600 questionnaires ont été collectés sur support papier. La plupart n'ont d'ailleurs pas pu être saisis sans le recours à des enquêteurs qui ont dû rappeler les exploitants pour compléter leur réponse.

Figure 2. Deux protocoles multimode pour deux niveaux de détail des données collectées



PREMIER RETOUR D'EXPÉRIENCE SUR LE RECOURS À DES PRESTATAIRES EXTERNES

Près de 400 000 unités ont été invitées à répondre au questionnaire par internet, de manière sécurisée. Pour ce faire, elles ont reçu au préalable un courrier sur lequel figuraient identifiant, mot de passe et lien vers le site de collecte, ainsi que les coordonnées de l'assistance : numéro vert gratuit, et adresse mail de contact. Après un certain temps, en l'absence de connexion, des relances par mail d'abord, puis par téléphone, ont été effectuées pour inciter à répondre au questionnaire en ligne. Enfin, dans une troisième phase, la collecte était réalisée par téléphone.

Deux instituts de sondage[10] ont été retenus pour assurer la collecte par téléphone. Chaque prestataire était chargé de six régions de France métropolitaine[11], et devait dérouler le protocole par vagues successives d'une durée maximale de trois mois, afin de lisser l'activité de leur plateau téléphonique. Les prestataires ont également relancé par SMS les non-répondants dont on connaissait le numéro de téléphone portable.

Les plateaux téléphoniques étaient associés à un système d'information dédié, permettant d'optimiser le rythme des appels et d'automatiser certaines tâches, par exemple la prise d'appel. Ces outils ont permis au SSP de réaliser une économie de coût pour la collecte des unités du tronc commun.

La durée cumulée de passation du questionnaire internet et téléphone, telle que constatée par les prestataires, a été de 20 à 25 minutes en moyenne, contre 50 minutes pour la collecte en face-à-face. Ceci s'explique aisément par la différence de longueur du questionnaire : en face-à-face il intégrait, en plus du tronc commun, des questions sur les trois modules.

Les prestataires ont obtenu un taux de retour de 91 %. Ce taux varie principalement en fonction du degré de connaissance des coordonnées au moment de la collecte. L'ensemble des unités interrogées par les prestataires pour la collecte par internet et par téléphone avait été classé en cinq groupes, allant du groupe 1 où toutes les coordonnées étaient renseignées (téléphone fixe, téléphone mobile, courriel) au groupe 5 où seule l'adresse postale était connue.

LE FACE-À-FACE AVEC LES ENQUÊTEURS AGRICOLES, POUR APPROFONDIR CERTAINES THÉMATIQUES

Près de 70 000 exploitations dont le siège est situé en France métropolitaine ont été tirées dans un échantillon représentatif et ont ensuite été interrogées sur la base du questionnaire dit « complet », plus détaillé, car comportant, outre le tronc commun, les trois modules de 2020. L'échantillonnage a conduit à généraliser ce type de collecte à toutes les exploitations de Corse et des départements d'Outre-mer.

La collecte en face-à-face, ainsi que la relance des non-répondants au tronc commun, a été réalisée par le réseau des enquêteurs du service statistique ministériel de l'Agriculture. Ce réseau existe depuis de nombreuses années, il est composé d'environ 850 personnes, en grande partie issues du monde agricole (agriculteurs ou anciens agriculteurs, conjoints

10. Il s'agit de BVA et Ipsos.
11. Hors la Corse qui a été recensée exclusivement en face-à-face, par les enquêteurs du Srise (voir *infra*).

d'exploitants notamment), ce qui lui confère une forte compétence et une certaine légitimité lors de la conduite des enquêtes en face-à-face. Il a été renforcé par près de 300 personnes recrutées pour les besoins du recensement.

Malgré le contexte lié à la pandémie de Covid-19, les enquêteurs ont souligné la qualité de l'accueil que leur ont réservé les agriculteurs, « *des gens passionnés par leur travail et qui aiment en parler* ».

Accompagnés par un gestionnaire du service régional statistique, notamment lors des premiers entretiens, les enquêteurs ont bénéficié d'une formation d'un nouveau genre, qui constitue la deuxième innovation majeure de ce recensement de 2020.

LA FORMATION EN LIGNE: UNE IDÉE QUI TOMBE À PIC EN PÉRIODE DE CRISE SANITAIRE…

Jusqu'en 2019, sur l'ensemble du dispositif d'enquêtes agricoles, toutes les formations des enquêteurs étaient assurées par les services régionaux de statistique agricole (Srise), eux-mêmes formés par les responsables de l'enquête au niveau national (SSP).

Plusieurs problèmes étaient posés par ces modalités de formation. Des déperditions étaient parfois constatées dans la transmission des consignes aux formateurs régionaux puis aux enquêteurs. Par ailleurs, elles généraient une lourde charge logistique pour trouver près de 80 lieux de formation, et gérer les déplacements des enquêteurs. *A posteriori*, on imagine que ces difficultés auraient été particulièrement accrues avec les restrictions de circulation et de regroupement imposées par la crise sanitaire de 2020.

Une première expérimentation de formation des enquêteurs en ligne avait été conduite dès 2019 avec l'enquête Teruti[12]. Elle s'est traduite par la bonne participation du réseau d'enquêteurs, dont une majorité a d'ailleurs fait part de sa satisfaction, jugeant positivement les outils développés et appréciant de ne pas avoir à se déplacer loin de leur domicile. La principale difficulté provenait de la mauvaise voire de l'absence de connexion à internet.

Riche de ces enseignements, l'équipe en charge du recensement agricole a conçu une formation en ligne en collaboration avec des membres du réseau de formateurs régionaux. La collaboration des Srise, dans le cadre d'un groupe de travail chargé de la conception des supports, a été particulièrement importante pour la réussite de cette opération. La mise en place d'une formation en ligne constitue en effet un grand changement dans les relations entre les gestionnaires d'enquête en Srise et le réseau d'enquêteurs.

Cette formation a été montée avec une structure calquée sur celle du questionnaire du recensement. La plate-forme de formation[13] reprend l'identité graphique du questionnaire et des applications de saisie afin que l'enquêteur puisse faire le lien facilement entre la collecte et les concepts exposés en formation. Pour chaque partie, une vidéo introductive décrit les points qui sont abordés dans les modules qui suivent. Chaque module est constitué d'un diaporama animé et commenté, présentant les concepts à maîtriser. Enfin des quiz sont proposés afin que l'enquêteur puisse vérifier sa compréhension et sa mémorisation des consignes.

Afin de faciliter la réutilisation ultérieure des modules de formation, le choix a été fait de ne pas citer dans les commentaires les caractéristiques propres au recensement agricole, telles que la formulation littérale de chaque question ou le millésime de la campagne de

12. Enquête statistique annuelle permettant un suivi longitudinal de l'occupation et de l'usage du sol au niveau national, régional et départemental (mode de consommation des terres agricoles et des espaces naturels, artificialisation et imperméabilisation des sols) et la quantification des principaux flux entre grands types d'occupation. Voir *https://www.cnis.fr/enquetes/occupation-et-lutilisation-du-territoire-teruti-enquete-sur-l-2020a063ag/*.
13. La plate-forme utilisée est en *open source* (Moodle, 2021).

cultures par exemple. Ainsi, certains modules pourront être réutilisés dans le cadre des futures enquêtes menées par le SSP, assurant ainsi une cohérence dans les concepts manipulés dans l'ensemble de la statistique agricole.

Le programme de formation est adapté à chaque type de questionnaire. Des modules spécifiques ont ainsi été développés pour les départements d'Outre-mer, dont les modalités de questionnement ont été adaptées (voir ***encadré 2***)[14].

Un compte individuel a été créé pour chaque enquêteur. Ainsi, les gestionnaires d'enquête des Srise, qui constituaient auparavant le réseau des formateurs en région, pouvaient suivre l'activité de formation de tous les enquêteurs qui leur sont rattachés. Si 1 123 comptes enquêteurs ont été créés, seulement 943 enquêteurs ont de fait été formés en ligne. Une solution hors connexion a donc été proposée à ceux qui rencontraient des problèmes d'accès internet. Les supports ont également été utilisés par les prestataires pour former leurs propres opérateurs des plateformes téléphoniques.

Encadré 2. Une collecte adaptée dans les départements d'Outre-mer

Dans les DOM, l'intégralité de la collecte a été réalisée par les enquêteurs du service statistique agricole. La collecte est réalisée **le plus souvent en face-à-face**, quelquefois par téléphone. Ce mode de collecte a été privilégié pour plusieurs raisons.

Hormis en Guyane, le territoire de collecte est moins étendu, facilitant ainsi les déplacements des enquêteurs dans les exploitations. En Guyane, les réseaux de télécommunication ne permettent pas de collecter les informations par téléphone et internet dans les zones éloignées des centres urbains. La barrière de la langue en particulier à Mayotte et en Guyane nécessite parfois la médiation d'un enquêteur local pour s'assurer de la bonne traduction des questionnaires. Enfin, la faiblesse des données administratives dans les DOM conduit à mettre hors champ une part significative des unités identifiées dans l'univers au lancement du recensement.

Le questionnaire utilisé pour les DOM est très proche de celui retenu pour la France métropolitaine. Les principaux aménagements portent sur :

- les cultures (ajout de canne à sucre, banane, café, cacao, etc. et retrait de betteraves, artichaut, etc.) ;
- les cheptels (composition, logement, absence de pâturages collectifs) ;
- les signes et démarches qualité ;
- l'origine de la propriété des terres (indivision, colonages, etc.) ;
- l'irrigation (origine de l'eau) ;
- la diversification (absence de transformation de céréales mais présence de transformation de tubercules ou de canne à sucre).

Pour la description des surfaces cultivées en légumes, un questionnement spécifique a été prévu pour décrire le plus précisément la complexité de la composition des abattis en Guyane et des jardins mahorais. Les abattis sont des cultures sur brûlis où l'agriculteur implante successivement des espèces à cycle court (gombo, pastèque, aubergine) ou une graminée (maïs), puis des tubercules (dachine, igname), ensuite des espèces à cycle moyen (patates douces) qui céderont la place aux plantes majeures telles que le manioc. De même, dans le jardin mahorais, différentes cultures compatibles forment plusieurs étages de productions. On peut par exemple trouver de la patate douce, des plants d'ananas, un pied de bananier, le tout entouré de cocotiers, de manguiers ou de jacquiers.

Jardin mahorais avec manioc et citrouille au 1er plan, bananier et manguier au 2e plan.

Les adresses des exploitations sont souvent moins précises dans les DOM qu'en Métropole. Le questionnaire a été adapté afin de **recueillir pour chaque exploitation un lieu-dit**, dont le nom a été normalisé et affecté de coordonnées spatiales précises : dans des communes de Guyane qui peuvent avoir la taille d'un département métropolitain, on en percevra tout l'intérêt lors de la diffusion des résultats.

14. Ces modules ne sont visibles que pour les enquêteurs des DOM.

UNE COMMUNICATION VERS LES ENQUÊTÉS REPENSÉE SELON LES PRINCIPES DU *NUDGE*

Pour garantir de bons taux de réponse par internet, plusieurs leviers d'action sont disponibles, notamment en perfectionnant les courriers et les courriels de relance, l'assistance téléphonique et le site *web* dédié à la collecte.

Cette démarche s'inscrit dans les actions animées par la direction interministérielle de la transformation publique sur la simplification des documents administratifs (DITP, 2021). Elles reposent sur les enseignements des sciences comportementales développées notamment par D. Kahneman, D. Ariely, C. Sunstein et R. Thaler[15]. Ces travaux ont montré d'une part l'existence de biais cognitifs dans la prise de décision et d'autre part le caractère prédictif de ces biais de décision : un individu ne prend pas forcément les décisions les plus rationnelles, mais sa décision n'est pas surprenante. Il fait, en effet, appel le plus souvent à un « système » automatique, rapide, lié à l'émotion plutôt qu'à un mode réflexif plus lent. Une démarche d'analyse est alors possible pour utiliser ces biais cognitifs dans le but de faire adopter le comportement souhaité. C'est le principe du *Nudge* qui va tenter de donner un coup de pouce pour favoriser la réponse par internet.

En se basant sur cette approche, les courriers adressés aux enquêtés ont été revus tant sur le fond que sur la forme, suite à un groupe de travail sollicitant l'avis d'une sélection d'agriculteurs[16].

L'expérience a d'abord montré l'intérêt de développer la personnalisation des courriers, et de privilégier la formulation à la première personne. Le courrier est donc adressé individuellement et il est signé par la cheffe du service statistique du ministère, car identifier le messager est également important. Le caractère officiel du courrier rassure, une remise en forme a donc été nécessaire pour respecter la charte gouvernementale de communication ***(figure 3)***.

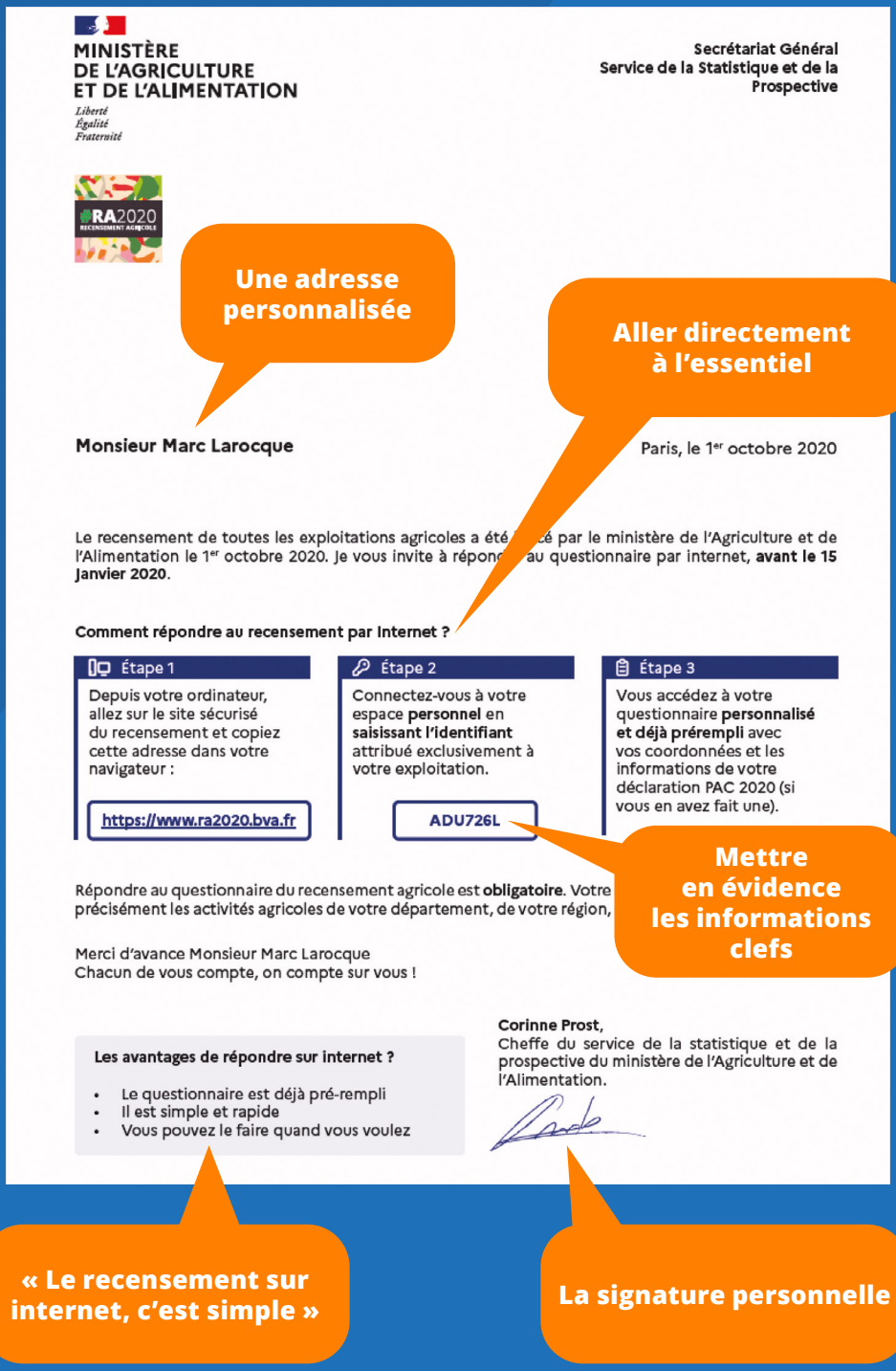
> *« Le recensement agricole, c'est simple et rapide. »*

Des formulations plus incitatives ont également été recherchées. L'enquêté adopte le « bon » comportement lorsqu'il a un minimum de cadrage. La saillance a été travaillée pour attirer l'attention sur les éléments importants, qui devaient être facilement identifiés, en particulier sur les trois étapes de la connexion au questionnaire. Le visuel utilisé (trois étapes, trois icônes) contribuait à mettre en évidence le message souhaité : « Le recensement agricole, c'est simple et rapide ». La nouvelle version du courrier se distingue donc des versions utilisées antérieurement par un visuel plus graphique et aéré et une formulation plus didactique.

Enfin, des éléments ont été ajoutés en cours d'enquête sur le site pour montrer que nombreux étaient les autres exploitants qui avaient déjà répondu : ce faisant, la norme sociale contribuait alors à déclencher une réponse chez ceux qui s'étaient jusque-là abstenus. D'autres améliorations ont été apportées pour rendre plus visible le numéro vert, ajouter une adresse postale de contact, certains exploitants souhaitant envoyer un courrier, bref rendre l'opération la plus simple possible pour les enquêtés.

15. Voir (Ariely, 2010), (Kahneman, 2012) et (Thaler et Sunstein, 2012).
16. Ce groupe de travail a été mis en place par un des prestataires sélectionné pour la collecte par téléphone, car il avait en la matière une certaine expérience.

Figure 3. Une application des principes du *Nudge* aux courriers du recensement agricole



UN NOUVEAU PLAN DE SONDAGE POUR LES MODULES THÉMATIQUES

En théorie, un recensement, par son caractère exhaustif, ne fait pas appel à un plan de sondage. Mais pour conserver la possibilité de collecter la majorité des exploitations par internet, il était important que le questionnaire ne soit pas trop long et trop complexe. En outre, le nouveau règlement européen permettait de basculer une partie de la collecte sur des enquêtes par sondage. C'est pourquoi, il a été décidé que les questions prévues dans les modules complémentaires (portant sur les bâtiments d'élevage et sur la main-d'œuvre) ne soient pas posées à toutes les exploitations agricoles mais seulement à un échantillon. Ce premier échantillon pour un recensement agricole a été construit pour produire des résultats représentatifs à l'échelle départementale.

En termes de volume de données collectées et de nombre d'exploitations interrogées, l'échantillon du recensement de 2020 est comparable à une enquête Structure (Esea), enquête intercalaire réalisée deux fois entre chaque recensement (voir *supra*), dont la dernière édition était relativement récente (2016). Plutôt que de reproduire l'échantillonnage des Esea, il a été décidé de conduire, dans le cadre du recensement de 2020, une révision intégrale du plan de sondage, selon deux axes :

- l'**optimisation de la stratification**, en s'appuyant principalement sur deux variables fortement corrélées aux variables collectées, à savoir l'orientation technique de l'exploitation (*Otex*) et la production brute standard (PBS), une variable estimant la taille de l'exploitation. Les strates ont été déterminées à l'aide d'un outil proposé par la direction de la Méthodologie de l'Insee[17]. Ce travail a abouti à la création de près de 2 000 strates croisant Otex, taille et département géographique (lorsque le nombre d'unités était suffisant) ;
- la **définition de la strate exhaustive** dans laquelle sont placées les unités qui ne doivent pas faire l'objet d'une pondération lors de la production des résultats. L'objectif était, pour le recensement, d'élargir le périmètre des enquêtes Structure précédentes, où seules les très grosses exploitations figuraient. En effet, lors de l'enquête réalisée en 2016, la strate exhaustive était principalement constituée de 2 000 exploitations, celles dont la production brute standard dépassait 1,5 millions d'euros par an et celles employant au moins 50 salariés permanents. Toutes les autres exploitations agricoles faisaient l'objet d'une pondération, pouvant occasionner des problèmes de robustesse des résultats produits aux niveaux les plus fins. Les travaux réalisés pour le recensement de 2020 ont abouti à retenir deux seuils beaucoup plus bas : un seuil fixé à 250 000 € pour les exploitations en maraîchage, horticulture ou aviculture et un autre fixé à 500 000 € pour toutes les autres orientations. En complément, toutes les exploitations qui emploient au moins 10 salariés en emploi permanent sont également introduites dans la strate exhaustive. Au final, près de 25 000 unités ont été placées dans la strate exhaustive en raison de leur taille, soit plus de dix fois plus qu'en 2010. Les départements d'Outre-mer et de Corse ont également été placés dans la strate exhaustive. Ce choix a été motivé à la fois par la qualité des sources administratives, qui ne permet pas de constituer un univers complet avec des variables de stratification bien renseignées et par des effectifs limités qui auraient généré des taux de sondage très élevés. Enfin, les unités du réseau d'information comptable agricole (Rica), enquête statistique qui assure le suivi des revenus et des activités des exploitations, ont également été ajoutées à la strate exhaustive.

> *« Au final, près de 25 000 unités ont été placées dans la strate exhaustive. »*

17. Il s'agit de la fonction R *Palourde*, qui automatise les calculs de l'algorithme Koubi-Mathern afin de déterminer le nombre d'unités à interroger dans chaque strate. Les variables de stratification ont ensuite été découpées de manière optimisée en appliquant la méthode Lavallée-Hidiroglou pour le calcul des bornes de stratification.

LA COLLECTE DES 900 VARIABLES S'APPUIE LARGEMENT SUR LES DONNÉES ADMINISTRATIVES…

L'objectif du recensement agricole 2020 est de décrire les productions des exploitations, avec les superficies cultivées et les cheptels, ainsi que les principaux facteurs de production mobilisés en agriculture, en particulier la main-d'œuvre occupée et le mode de faire-valoir du foncier. Des questions portent également sur l'engagement dans des démarches spécifiques (démarches de qualité et ou environnementales), sur la diversification des activités et les modalités de commercialisation des produits.

Le questionnaire de 2020 s'appuie sur des informations déjà connues par ailleurs, par exemple avec le pré-remplissage des surfaces des cultures à partir des déclarations « PAC » (voir *infra*). En effet, pour réussir la collecte par internet, il était primordial d'alléger la charge de réponse pour les exploitants, et de diminuer la taille du questionnaire. Deux sources ont été notamment mobilisées :

- la **base de données nationale d'identification animale** (BDNI) sert de support à l'identification des animaux dans l'objectif d'assurer la traçabilité et le contrôle des aides européennes ; cette source permet de décrire le cheptel bovin élevé dans l'exploitation. Elle a été mobilisée pour les modules traitant de cette thématique ;
- les **déclarations en vue de bénéficier d'une subvention européenne de la politique agricole commune** (PAC), qui concernent trois quarts des exploitations françaises, sont également mobilisées pour renseigner les surfaces des cultures et l'intégralité du module « Développement rural », ce qui a simplifié considérablement le questionnaire pour ces unités[18].

… AVEC PARFOIS QUELQUES DIFFICULTÉS

L'appariement entre les unités recensées et les sources administratives a été réalisé sur le numéro d'établissement, le *Siret*. Cet identifiant est parfois mal renseigné dans les sources mobilisées, ce qui nécessite un contrôle approfondi lors de la constitution du fichier de lancement du recensement.

> *« Une réflexion à moyen terme visant à mieux articuler la nomenclature PAC avec le questionnaire du recensement (qui est lui aussi issu d'un règlement européen) devrait aplanir les différences. »*

L'imputation des données PAC en lieu et place de la déclaration de l'agriculteur a par ailleurs amené trois types de problèmes. En premier lieu, des différences entre les nomenclatures de culture des données de la PAC et celle du recensement agricole obligent à poser des questions complémentaires pour certaines cultures. Par exemple, les cultures déclarées à la PAC en « *Pomme de terre de consommation* » doivent être éclatées entre « *Pomme de terre primeur ou nouvelle* », « *Pomme de terre de conservation ou demi-saison* » et « *Plants de pomme de terre* ». Les divergences de nomenclature sont particulièrement marquées pour les prairies, les légumineuses fourragères et les fruits. Une réflexion à moyen terme visant à mieux articuler la nomenclature PAC avec le questionnaire du recensement (qui est lui aussi issu d'un règlement européen) devrait aplanir les différences.

18. Trois surfaces sont enregistrées à la PAC et figureront dans le fichier de diffusion : la surface graphique qui correspond à la réalité du terrain, *i.e.* la surface de la culture réellement en place, la surface admissible qui est une surface recalculée intégrant les bordures considérées comme cultivées, et la surface constatée, qui sert d'assiette au calcul de la subvention.

Deuxième difficulté, le taux de couverture par la PAC varie fortement selon les cultures, ce qui impose de compenser par des enquêtes supplémentaires.

Enfin, les données déclarées à la PAC varient parfois fortement avec celles déclarées par l'exploitant. Une comparaison entre les données PAC et les données déclarées dans l'Esea 2016 montre qu'entre 10 et 20 % des exploitations selon les cultures ont un écart en valeur absolue de plus de 5 % entre la déclaration PAC et l'enquête. Toutefois, ce constat ne préjuge pas de la qualité d'une source par rapport à une autre. Dans l'enquête, l'exploitant a tendance à arrondir les surfaces et il confond parfois les campagnes de cultures, donnant ainsi les surfaces de la campagne précédente ou de la campagne suivante.

LE RECENSEMENT AGRICOLE C'EST (VRAIMENT) SIMPLE ET RAPIDE

Sur les 510 000 unités interrogées, 494 000 questionnaires ont été collectés, soit 97 % de retour sur l'ensemble du territoire national, en France métropolitaine comme dans les DOM. La collecte du recensement agricole 2020 est donc un franc succès ***(encadré 3)***. Elle a permis de porter cinq innovations qui touchent tous les acteurs de la collecte :

- les exploitants agricoles eux-mêmes (recours au multimode et aux données administratives, amélioration de la communication) ;
- les enquêteurs agricoles et les opérateurs des plateformes externes (formation en ligne, collecte par internet) ;
- et *in fine* les statisticiens (refonte du plan de sondage).

Ce résultat conforte donc le SSP dans la plupart des choix qui ont été opérés en amont. En premier lieu, la flexibilité de la solution internet s'est révélée bien adaptée à l'activité des exploitants agricoles. Au final, seulement une réponse sur cinq a été obtenue par téléphone ou par courrier.

> *« La flexibilité de la solution internet s'est révélée bien adaptée à l'activité des exploitants agricoles. Au final, seulement une réponse sur cinq a été obtenue par téléphone ou par courrier. »*

L'assistance par des opérateurs spécialement formés a été indispensable à la bonne participation par internet. En outre, la diversité des supports de relance (courrier, mail, SMS, téléphone) a permis une animation efficace de la collecte.

La relance par téléphone du réseau des enquêteurs agricoles, à l'issue du travail des prestataires, a été payante puisque 63 % des 35 000 non-répondants ont été récupérés. L'articulation entre les deux modes de collecte est donc efficace.

Les phases de relance par téléphone et d'interviews par téléphone sont complémentaires et peuvent se dérouler simultanément, en gérant l'orientation dans chacune des phases avec des priorités. En phase de relance, un exploitant qui ne peut pas ou ne veut pas répondre par internet doit pouvoir être interrogé immédiatement par téléphone. En phase d'interrogation, une personne qui affirme vouloir répondre par internet ne répondra pas par téléphone, même en insistant.

La durée du questionnaire varie selon les caractéristiques de l'exploitation agricole et selon le mode de réponse, mais il reste toujours dans des normes acceptables :

- les mises en hors champs sont très rapides (5 minutes) ;
- à l'inverse, les exploitations pour lesquelles aucune déclaration PAC n'a pu être retrouvée mettent plus de temps pour répondre : elles doivent passer en revue toutes les productions, aucune information ne pouvant être chargée sur les surfaces ;
- le questionnaire est plus rapide en passation par téléphone (22 minutes) qu'en durée de connexion sur internet (28 minutes) ;
- la durée varie également selon l'orientation de l'exploitation, allant, pour la collecte par téléphone, de 18 minutes pour les grandes cultures ou les éleveurs de bovins à 24 minutes pour les éleveurs d'ovins-caprins.

De cette édition 2020 ressort également un ensemble de points de vigilance ou d'améliorations à envisager pour les prochaines collectes.

Un élément déterminant pour assurer un bon taux de réponse est de disposer du maximum de coordonnées sur l'unité interrogée. La recherche manuelle de coordonnées complémentaires permet de gagner quelques points de participation. Elle est certes coûteuse, car réalisée par un gestionnaire d'enquête, mais devrait rester moins onéreuse que le face-à-face.

Encadré 3. Le recensement agricole 2020 en chiffres

- 7e recensement général s'adressant à toutes les exploitations agricoles depuis 1955
- 1 150 enquêteurs mobilisés dans le réseau de la statistique agricole
- 510 000 unités interrogées, 392 000 avec le questionnaire tronc commun, 118 000 avec le questionnaire complet dont 42 100 dans les DOM
- 490 000 questionnaires collectés avec 900 variables renseignées pour chaque exploitation agricole
- Collecte d'octobre 2020 à mai 2021
- Taux de réponse global (tronc commun et questionnaire complet) : 97 %

Activité des prestataires pour collecter le tronc commun

- 392 000 unités interrogées, 358 000 réponses collectées
- 145 000 connexions aux portails internet des prestataires
- 62 000 appels à l'assistance
- 800 000 courriers envoyés
- 950 000 courriels envoyés
- 480 000 SMS envoyés

Taux de retour sur le tronc commun

- 47 % de réponses par internet avec uniquement des relances automatiques
- 73 % de réponses après les relances téléphoniques
- 91 % après la collecte par téléphone par les prestataires
- 97 % après la collecte complémentaire par les Srise

En mode auto-administré, plusieurs questions courtes et simples sont parfois préférables à une seule question longue et complexe, surtout *via* internet où la lecture sur l'écran se fait rapidement. Les contrôles embarqués permettent de garantir une bonne qualité des données, mais ils doivent être bien calibrés compte-tenu de la très grande diversité des situations rencontrées.

Certains exploitants ne se sont pas sentis concernés par le recensement agricole et n'ont donc pas répondu spontanément au questionnaire. C'est le cas par exemple de certains viticulteurs ou apiculteurs. Une communication les incluant plus spécifiquement, serait à développer pour le prochain recensement agricole.

En complément des solutions d'assistance proposées en cours de collecte, un répondeur vocal pourrait raconter les étapes à venir et expliquer les solutions possibles pour répondre.

L'organisation retenue a mis en œuvre trois applications de saisie (deux pour les prestataires et une pour le SSP). Cela a généré une charge très importante pour spécifier, suivre le développement et valider le fonctionnement de chacune d'entre elles. La réalisation d'une application unique pour tous les acteurs intervenant dans la collecte paraît aujourd'hui plus optimale, et mériterait en tout cas d'être étudiée.

Le contexte de tous les travaux réalisés en trois ans, depuis la préparation jusqu'à la fin de la collecte, a été évidemment marqué par la crise sanitaire. Cela n'a pas empêché le bon déroulement du protocole : le recensement de 2020, avec son cortège de nouveautés, débouche fin 2021 sur une couverture très efficace de l'univers agricole[19].

La diffusion s'accompagne d'une sixième innovation. Une *data visualisation* présente les premiers résultats sur tous supports, en particulier sur *smartphone*. Elle s'appuie sur des techniques qui mélangent graphiques et explications : les parties significatives du visuel sont mises en évidence ou zoomées au gré du défilement de la page commandé par l'utilisateur (SSP, 2021). L'objectif de ce mode de diffusion est de toucher tous les publics, et en tout premier lieu, les agriculteurs.

19. Les premiers résultats ont été publiés en décembre 2021. Voir par exemple (Barry et Polvêche, 2021).

BIBLIOGRAPHIE

ARIELY, Dan, 2010. *Predictably Irrational : The Hidden Forces That Shape Our Decisions*. 27 avril 2010. Éditions Harper Perennial. ISBN 978-0061353246.

BARRY, Catherine et POLVÊCHE, Vincent, 2021. *Recensement agricole 2020. Surface moyenne des exploitations agricoles en 2020 : 69 hectares en France métropolitaine et 5 hectares dans les DOM*. [en ligne]. Décembre 2021, SSP, Agreste n° 5. [Consulté le 10 décembre 2021]. Disponible à l'adresse : *https://agreste.agriculture.gouv.fr/agreste-web/disaron/Pri2105/detail/.*

CNIS, 2019. *Recensement agricole 2020.* [en ligne]. 17 octobre 2019. Avis d'opportunité N°146/H030 modifié. [Consulté le 10 décembre 2021]. Disponible à l'adresse : *https://www.cnis.fr/wp-content/uploads/2020/06/AO_2019_SSP_RA2020_def_modif.pdf.*

CNIS, 2020. *Recensement agricole 2020.* [en ligne]. 28 mai 2020. Comité du label de la Statistique publique. N°2020_11698_DG75-L002. [Consulté le 10 décembre 2021]. Disponible à l'adresse : *https://www.cnis.fr/wp-content/uploads/2020/06/AC_2020_SSP_RA_2020.pdf.*

DITP, 2021. Simplifier les documents administratifs. In : *site de la DITP.* [en ligne]. 22 février 2021. Ministère de la Transformation et de la Fonction Publiques. Équipe sciences comportementales. [Consulté le 10 décembre 2021]. Disponible à l'adresse : *https://www.modernisation.gouv.fr/outils-et-methodes-pour-transformer/simplifier-les-documents-administratifs.*

EUROSTAT, 2020. Le recensement agricole 2020. In : *site d'Eurostat*. [en ligne]. [Consulté le 10 décembre 2021]. Disponible à l'adresse : *https://ec.europa.eu/eurostat/fr/web/agriculture/census-2020.*

FAO, 2020. Programme du recensement mondial de l'agriculture 2020. In : *site de la FAO.* [en ligne]. Organisation des Nations Unies pour l'alimentation et l'agriculture. [Consulté le 10 décembre 2021]. Disponible à l'adresse : *http://www.fao.org/world-census-agriculture/wcarounds/wca2020/fr/.*

KAHNEMAN, Daniel, 2012. *Système 1 / Système 2 – Les deux vitesses de la pensée.* Paris, éditions Flammarion, Collection Essais. Traduction de Raymond Clarinard. ISBN 2-08-121147-5.

MOODLE, 2021. Documentation de Moodle 3.x. In : *site de Moodle*. [en ligne]. Mis à jour le 14 juillet 2021. [Consulté le 10 décembre 2021]. Disponible à l'adresse : *https://moodle.org.*

SSP, 2010. Recensement agricole 2010 – Premières tendances. In : *Agreste Primeur – France métropolitaine*. [en ligne]. Septembre 2011. Ministère de l'Agriculture, de l'Alimentation, de la Pêche, de la Ruralité et de l'Aménagement du territoire. N° 266. [Consulté le 10 décembre 2021]. Disponible à l'adresse : *https://gallica.bnf.fr/ark:/12148/bc6p06znqmf/f1.pdf.*

SSP, 2021. VIZagreste, les résultats de la statistique agricole en datavisualisation. [en ligne]. Décembre 2021. Ministère de l'Agriculture et de l'Alimentation. [Consulté le 10 décembre 2021]. Disponible à l'adresse : *https://vizagreste.agriculture.gouv.fr/*.

THALER, Richard et SUNSTEIN, Cass, 2010. *Nudge : la méthode douce pour inspirer la bonne décision.* Mars 2010. Paris, éditions Vuibert, collection Signature. Traduction de Marie-France Pavillet. ISBN 978-2-311-00105-1.

FONDEMENTS JURIDIQUES

Règlement (UE) 2018/1091 du Parlement européen et du Conseil du 18 juillet 2018 concernant les statistiques intégrées sur les exploitations agricoles. In : *site EUR-Lex*. [en ligne]. [Consulté le 10 décembre 2021]. Disponible à l'adresse : *https://eur-lex.europa.eu/legal-content/FR/TXT/PDF/?uri=CELEX:32018R1091&from=FR*.

LE SSPCLOUD : UNE FABRIQUE CRÉATIVE

POUR ACCOMPAGNER LES EXPÉRIMENTATIONS DES STATISTICIENS PUBLICS

Frédéric Comte, Arnaud Degorre**, Romain Lesur****

Environnement d'aide à l'expérimentation sur les nouvelles méthodes de la* data science*, le SSPCloud pour le système statistique public est un ensemble de ressources informatiques permettant de réaliser des prototypes, de tester des traitements statistiques et de s'approprier de nouvelles pratiques de travail. Inscrit dans un courant d'inspiration de type FabLab, il apporte les conditions (im)matérielles pour favoriser la créativité du statisticien et l'aider à valoriser les nouveaux gisements de données. Il s'appuie sur des technologies de l'informatique dans les nuages (le* Cloud computing*) qui renforcent l'autonomie – et la responsabilité – des utilisateurs dans l'orchestration de leurs traitements.

Construit autour d'une communauté ouverte à l'ensemble des statisticiens publics, le SSPCloud se veut un atelier d'apprentissage, où le geste statistique se réinvente à plusieurs. La collaboration s'y trouve facilitée par l'adoption de solutions* open source*, garantissant les possibilités de réutilisation. Le SSPCloud propose un mélange fertile des deux univers professionnels de la statistique et de l'informatique, pour progresser plus particulièrement dans la mise en place de processus répondant aux standards de la reproductibilité appliqués aux traitements de la donnée et aux travaux d'études.

The SSPCloud for the official statistical system is an environment for experimenting with new data science methods. It is a set of computer resources for creating prototypes, testing statistical processing and adopting new work practices. Inscribed in a Fab Lab type of inspiration, it provides the (im)material conditions to encourage the statistician's creativity and help him or her make the most of new data sources. It is based on technologies of Cloud computing which reinforce the autonomy – and the responsibility – of users in the orchestration of their processing.

Built around a community open to all public statisticians, the SSPCloud is intended to be a learning workshop, where the statistical gesture is reinvented by many. Collaboration is facilitated by the adoption of open source solutions, guaranteeing the possibility of reuse. The SSPCloud offers a fertile mix of the two professional worlds of statistics and computer science, to progress more particularly in the implementation of processes that meet the standards of reproducibility applied to data processing and research work.

* Chef de projet, division Innovation instruction technique, DSI, Insee, *frederic.comte@insee.fr*
** Ancien chef de l'unité Innovation et stratégie du système d'information, DSI, Insee, *arnaud.degorre@insee.fr*
*** Chef de la division Innovation instruction technique, DSI, Insee, *romain.lesur@insee.fr*

UNE REFONTE DES APPAREILS DE PRODUCTION DE LA STATISTIQUE PUBLIQUE…

À l'issue de la conférence *The European path towards Trusted Smart Statistics* dédiée à l'émergence d'une société des données, les instances représentatives du système statistique européen adoptaient le 12 octobre 2018 le Mémorandum de Bucarest, posant les principes d'une refonte majeure des appareils de production de la statistique publique dans les différentes nations européennes. Ces principes visent plus particulièrement à doter le système statistique européen des capacités nécessaires pour prendre en compte les nouveaux gisements de données et de méthodes alternatives de traitement du chiffre, et cela dans de multiples dimensions, incluant le cadre juridique, les compétences techniques et les solutions informatiques de mise en œuvre (Eurostat, 2018).

La notion de *Trusted Smart Statistics*[1] embrasse ainsi un vaste ensemble d'évolutions liées à la démultiplication des sources d'information sur la société au sens large, dont les illustrations ne manquent pas :

- appréhender la tension sur le marché du travail *via* l'étude des offres d'emploi en ligne ;
- cartographier finement les mouvements de population au jour le jour, heure par heure, à partir des données de téléphonie mobile ;
- mesurer la vulnérabilité énergétique à partir des données des compteurs connectés d'électricité et de gaz, etc.

Autant d'exemples où la statistique publique est amenée à investir pour valoriser des informations de forme et de nature bien différentes des données d'enquête (Unece, 2013a ; 2013b).

Ces investissements s'accompagnent également d'innovations dans les processus statistiques, pour être en mesure de valoriser la richesse de ces nouvelles sources, mais aussi pour faire face à leur complexité, ou à leurs imperfections, qui impliquent des traitements conséquents de mise en conformité statistique. Au premier rang de ces innovations, figurent les méthodes dites d'apprentissage (*machine learning*) et leurs cas d'usage prometteurs dans les domaines de la codification et de la classification, de l'édition et de l'imputation des données. Le *Machine Learning Project* mis en place par la Commission économique pour l'Europe des Nations Unies (Unece) a ainsi pris la mesure, dans son rapport conclusif, des investissements réalisés à ce jour dans la communauté de la statistique publique : « *De nombreux instituts nationaux de statistique étudient la manière dont le machine learning peut être utilisé pour accroître la pertinence et la qualité des statistiques officielles dans un environnement caractérisé par des demandes croissantes d'informations fiables, des technologies accessibles et en développement rapide et de nombreux concurrents.* » (Unece, 2021).

… RENDUE POSSIBLE PAR DE NOUVELLES CAPACITÉS INFORMATIQUES

Les gisements de données massives, au cœur des *Trusted Smart Statistics*, présentent des caractéristiques qui, du fait de leur volumétrie, de leur vélocité (avec leur vitesse de constitution et de renouvellement) ou de leur variété (données structurées mais aussi non structurées, comme des images), en rendent la manipulation particulièrement complexe.

1. Qu'on pourrait traduire par « statistiques intelligentes et fiables ».

Sont ainsi considérées comme des données massives ou *big data*[2] les informations dont les caractéristiques sont telles qu'elles ne peuvent être facilement collectées, stockées, manipulées ou analysées par des équipements informatiques traditionnels.

Il est alors nécessaire de mettre en place des infrastructures d'un nouveau type, permettant d'anticiper sur les cas d'usage de la statistique publique, voire de les inspirer, dans une approche portée par les opportunités rendues possibles par l'innovation technologique[3].

Derrière la notion d'infrastructure informatique, il faut entendre de multiples strates, qui sont chacune des variables clefs dans la composition des services attendus : les capacités de stockage de la donnée selon différentes méthodes (fichier, objet, base de données, etc.), la puissance de traitement (mémoire vive, CPU[4], GPU[5]) ou encore les services de traitement (logiciels permettant d'exécuter des langages comme *R*, *Python*, etc.), mais aussi des architectures techniques d'**orchestration** dédiées à la mise en relation de ces différentes strates. Par exemple, le calcul distribué est une méthode de traitement de la donnée fondée sur la division d'un problème unique en une multitude de problèmes plus petits, afin de résoudre en parallèle chacun de ces problèmes sur un même centre de calcul (*via* le *multithreading*) ou en les répartissant sur plusieurs centres de calcul liés entre eux (appelés alors un *cluster*). La mise en œuvre effective d'un calcul distribué nécessite de mobiliser plusieurs éléments logiciels, appelés également *framework*, pour gérer l'accès aux données, les répartir entre des nœuds de traitement, d'effectuer toutes les opérations d'analyse nécessaires, puis de livrer les résultats.

LE RAPPORT DU STATISTICIEN À L'INFORMATIQUE ÉVOLUE, AVEC PLUS DE POUVOIRS… ET DE RESPONSABILITÉS

Ces infrastructures se destinent à des professionnels du traitement de la donnée qui sont appelés à développer, pour leurs besoins propres, des briques techniques et à les assembler en un processus intégré (ou *pipeline*). L'appellation de *data scientist*[6] traduit, parmi ses multiples acceptions, cette implication accrue du statisticien dans l'élaboration puis l'orchestration informatique de son traitement, au-delà des seules phases de conception ou de recette. Les nouvelles infrastructures de *data science* prennent en compte ce rôle étendu de ses utilisateurs, en leur accordant des possibilités d'action plus large qu'une infrastructure conventionnelle.

> *« L'appellation de* data scientist *traduit, parmi ses multiples acceptions, cette implication accrue du statisticien dans l'élaboration puis l'orchestration informatique de son traitement. »*

2. Pour une caractérisation des *big data*, le lecteur pourra par exemple consulter les publications du *National Institute of Standards and Technology* (NIST) *Big Data Public Working Group* (NIST, 2017).
3. Voir les deux modèles d'alignement stratégique d'une organisation (Anderson et Ventrakaman, 1990) : un alignement « descendant » des solutions informatiques sur la base des besoins portés par les processus métiers et un alignement « ascendant » des processus métiers qui saisissent des opportunités de transformation rendues possibles par les évolutions informatiques.
4. *Central Processing Unit,* désigne la plupart du temps le processeur d'un ordinateur. On peut le traduire en français par unité centrale de traitement (UCT) ou unité centrale de calculs.
5. *Graphics Processing Unit*, ou processeur graphique en français.
6. « *Le data scientist* […] *effectue des tâches complexes dans le traitement des données. Il est capable de traiter des données variées et de mettre en place des algorithmes optimisés de classification, de prédiction sur des données numériques, textuelles ou d'images.* […] *Le data scientist utilise des outils de programmation et doit savoir optimiser ses calculs pour les faire tourner rapidement en exploitant au mieux les capacités informatiques (serveur local, cloud, CPU, GPU, etc.)* » (Dinum et Insee, 2021).

À cet élargissement du domaine d'intervention, correspond également un renforcement de la place des travaux de prototypage dans l'activité des statisticiens publics, à travers des démarches de va-et-vient entre la conception d'un traitement statistique et sa mise en œuvre. Certes, les traitements statistiques ont toujours été des processus évolutifs, appelés à connaître des adaptations à l'épreuve des données. Cette dimension est toutefois accentuée s'agissant de données qui, contrairement à des fichiers d'enquête, ne présentent pas nativement les qualités attendues en termes de stabilité des concepts et de précision des mesures. Par exemple, les données de téléphonie mobile peuvent comporter des évolutions dans les informations de localisation, liées aux méthodes de triangulation à partir de la position des antennes-relais, dont l'implantation évolue au fil des ans. Ou encore, l'interprétation des images satellitaires ou aériennes pour établir des indicateurs statistiques comme la tâche urbaine se fonde sur des clichés dont la qualité dépend des saisons ou des conditions météorologiques. Enfin, l'analyse des offres d'emploi en ligne peut être conduite pour en déduire des éléments sur les compétences requises, sur des champs textuels dont le contenu est hétérogène d'un recruteur à un autre, pour décrire un même métier.

Ce type de données appelle en outre à conduire des traitements statistiques de nature évolutive, fondés sur des algorithmes apprenants, plutôt que des opérations déterministes. Le comportement de l'algorithme est ainsi conçu pour évoluer au fur et à mesure qu'il « apprend » des données qui lui sont transmises, et son orchestration nécessite de disposer, au sein de l'infrastructure informatique, de services et de briques techniques pensées en conséquence. Le statisticien doit non seulement concevoir l'algorithme de traitement, mais aussi s'intéresser à sa mise en œuvre, prenant ainsi à son compte des éléments qui relèvent habituellement de l'exploitation informatique.

SORTIR DU CADRE POUR MIEUX ACCUEILLIR LES DÉMARCHES EXPLORATOIRES

La forte évolutivité des nouveaux gisements de données appelle à engager, en complément des projets informatiques structurants, des approches plus agiles, conduites dans des calendriers courts : fondées sur du tâtonnement, des essais et erreurs, des intuitions, elles privilégient une opportunité pour laquelle il s'agit de faire le point sur les possibilités offertes, plutôt que d'expertiser la solution optimale pour un objectif défini à l'avance. Ces démarches de prospection privilégient l'expérimentation avec une mise en place accélérée d'un premier prototype, afin de faire la preuve du concept[7] de façon empirique, sans forcément traiter toutes les facettes d'un sujet.

Les expérimentations peuvent toutefois buter sur une difficulté pour accéder aux ressources nécessaires à leur réalisation… naturellement plus faciles à accorder à un projet qui présente, par ses études préalables, toutes les garanties souhaitées par des instances d'arbitrage. L'essor des « laboratoires » dans les grandes organisations publiques comme privées est une réponse pour contrebalancer la place prédominante des investissements *via* les projets structurants, en proposant des espaces de créativité laissant la part belle au prototypage et à l'expérimentation. Un laboratoire, ou « lab », a pour principale caractéristique de proposer une « autre façon d'inventer »[8] par rapport au processus de R&D dominant. Pour cela, il est, au moins en partie, en dehors des règles habituelles de fonctionnement de l'organisation, donc en dehors des procédures classiques de décision.

7. *Proof of concept* (POC) en anglais.
8. « *Les Open labs constituent un lieu et une démarche portés par des acteurs divers, en vue de renouveler les modalités d'innovation et de création par la mise en œuvre de processus collaboratifs et itératifs, ouverts et donnant lieu à une matérialisation physique ou virtuelle* » (Mérindol, Bouquin, Versailles *et alii*, 2016).

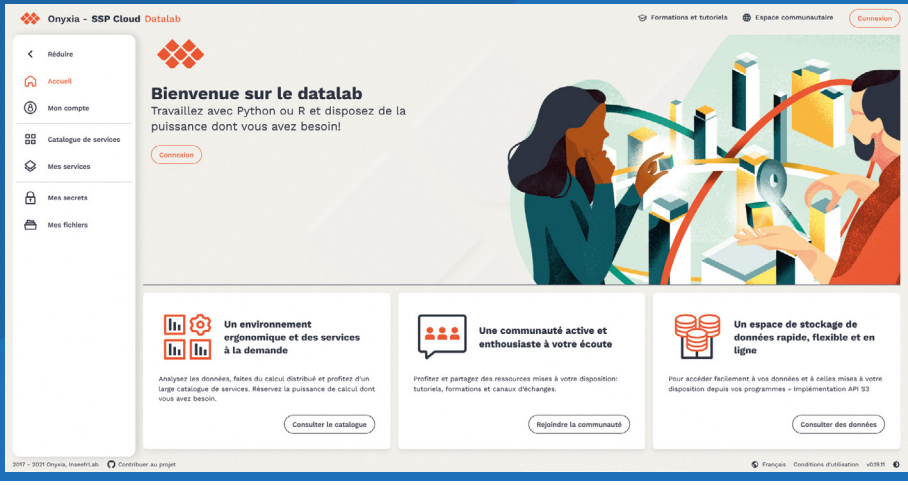
DE L'OMBRE À LA LUMIÈRE : L'ÉMERGENCE D'UNE NOUVELLE INFRASTRUCTURE POUR LE SSP

Conçu pour répondre aux besoins d'outillage en *data science*, le SSPCloud ***(figure 1)*** est une nouvelle infrastructure mise en place par l'Insee pour offrir, sur le versant informatique, un environnement « lab » propice à l'engagement d'expérimentations : sur les nouveaux gisements de données et les nouvelles méthodes de traitement, plus particulièrement sur des calculs distribués sur données massives, comme les données de téléphonie mobile pour calculer des populations présentes à des échelles géographiques fines (Suarez Castillo *et alii*, 2020) ; sur des méthodes d'apprentissage, comme des systèmes apprenants de codification automatique de libellés de professions dans des nomenclatures.

D'abord créé « en dehors de l'organisation » ***(encadré 1)***, le SSPCloud trouve ses origines en 2017, avec la participation d'une équipe de l'Insee au *hackathon New Techniques and Technologies for Statistics* (NTTS). Celle-ci joue le rôle de déclencheur : l'équipe parvient à proposer un traitement de données massives intégrant à la volée des restitutions sous forme de *data visualisation*, mais rapporte dans son retour d'expérience d'importantes limitations techniques rencontrées dans les infrastructures alors à leur disposition. Répondre à ce besoin sera le fil conducteur des travaux engagés par un collectif d'agents de la direction du système d'information. L'opportunité offerte par la disponibilité d'équipements informatiques décommissionnés de leur usage initial permettra, quelques mois plus tard, de créer le premier prototype fonctionnel d'une plateforme d'expérimentation en *data science*, qui évoluera ensuite pour devenir le SSPCloud.

Le dispositif a continué d'évoluer et de s'enrichir, jusqu'à recevoir plusieurs soutiens institutionnels internes[9] et externes[10], et s'ouvrir à l'ensemble du système statistique public en octobre 2020. Pour mettre en exergue cette volonté d'ouverture et faire référence au paradigme technologique utilisé, l'appellation de SSPCloud a été retenue pour ce nouveau service mutualisé entre l'Insee et les services statistiques ministériels.

Figure 1. Pour faire un tour sur le SSPCloud (https://datalab.sspcloud.fr)



9. La démarche a reçu le soutien progressif de l'institut, avec la création des labs statistique (SSP Lab) et informatique (division Innovation et instruction technique) de l'Insee, comptant en leur sein les agents ayant œuvré à la constitution de cet écosystème expérimental.
10. Le SSPCloud a également bénéficié d'un financement *via* le Fonds de transformation ministériel de Bercy en 2019, et de l'apport de trois entrepreneurs d'intérêt général, suite à un appel à projet remporté auprès de la direction du Numérique en 2020.

UNE FABRIQUE CRÉATIVE, OUVERTE EN LIBRE SERVICE

Le SSPCloud se présente comme une « fabrique créative », ou FabLab[11], de la statistique publique. Ouvert en libre service, il propose un accès direct à l'ensemble des matériels physiques (comme des cartes graphiques), logiques (comme des *framework* et briques logicielles), utiles à une activité de *data science* ***(figure 2)***. Le statisticien public y dispose des composants techniques nécessaires pour prototyper et tester un processus de traitement, de bout en bout.

L'accès à ces ressources est immédiat : il n'est pas nécessaire, par exemple, de demander le provisionnement préalable d'un espace de stockage, ou d'un environnement de recette, les technologies utilisées rendent autonome l'utilisateur pour en disposer.

Les réalisations conduites sur le SSPCloud ne font pas l'objet d'une régulation ou d'un contrôle[12] : l'utilisateur a le libre choix des éléments qu'il mobilise, sans être contraint par un cadre technique de cohérence qui viendrait délimiter le champ des possibles.

Pour rester une fabrique créative, le SSPCloud évolue par ailleurs en permanence, avec l'installation de nouveaux logiciels et leur mise à jour en continu[13]. L'utilisateur peut contribuer à élargir le catalogue de services et introduire de nouvelles briques techniques dont ses pairs pourront bénéficier. Le SSPCloud place ainsi le statisticien au cœur de la conception et du développement de ses futurs processus statistiques.

Encadré 1. « C'est plus marrant d'être un pirate que de s'engager dans la marine » (S. Jobs)

L'informatique et l'innovation de l'ombre

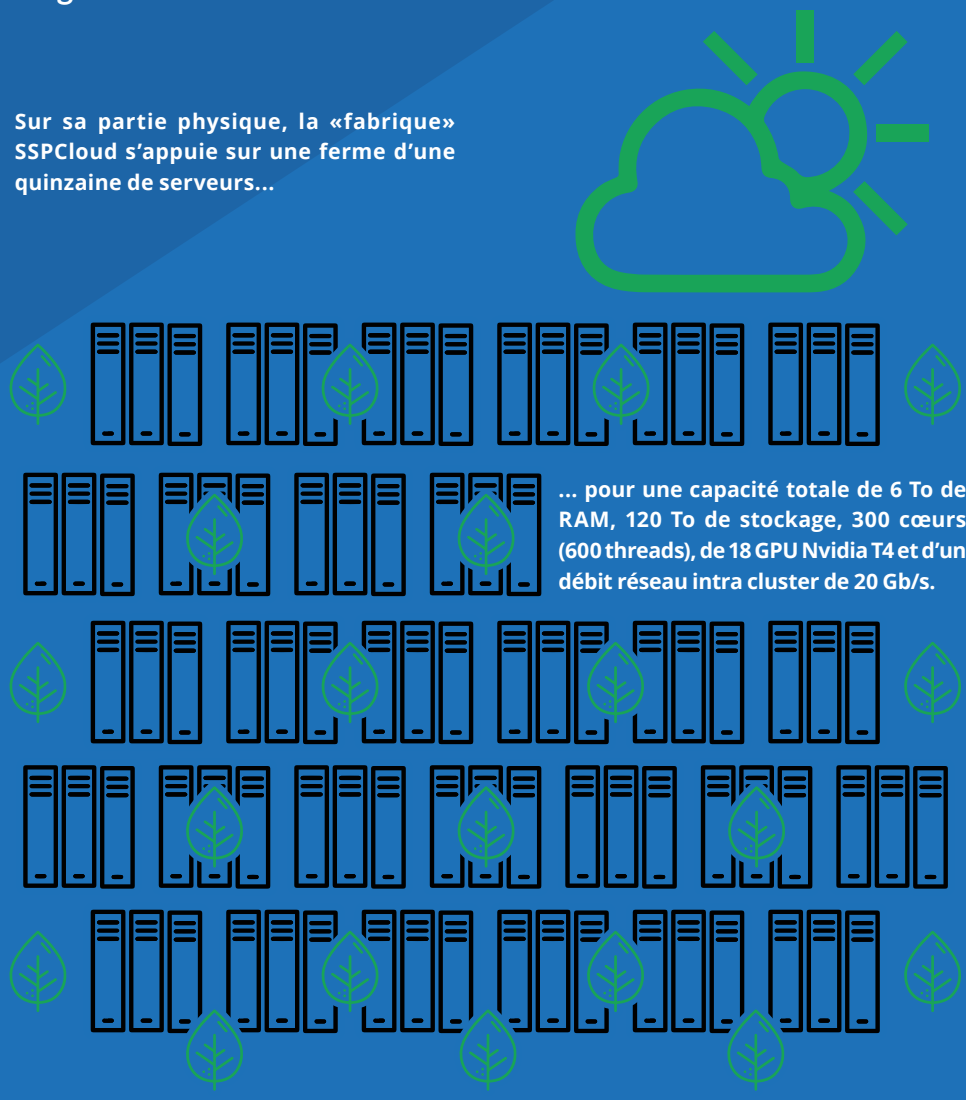
Ces mots de Steve Jobs permettent d'appréhender l'une des motivations amenant à l'apparition de démarches d'innovation conduites « en secret » au sein d'une organisation. S'interrogeant sur les ressorts de l'innovation dans les entreprises et les administrations, Donald A. Schon a introduit la notion de contrebande d'innovation (*bootlegging*) (Schon, 1963). Le *bootlegging* est défini comme une recherche dans laquelle des individus motivés organisent secrètement un processus d'innovation, sans la permission officielle des instances de direction, mais pour le bénéfice de l'entreprise. Dans le domaine des systèmes d'information, le même phénomène a pu être observé avec la *Shadow IT* – terme désignant les systèmes d'information réalisés et mis en œuvre au sein d'organisations sans approbation de la direction du système d'information.

Prenant conscience du potentiel de créativité des collaborateurs et de la nécessité de les accueillir, les grandes organisations visent désormais à accompagner les démarches « masquées » d'innovation, en apportant des ressources techniques d'usage libre sans avoir à rendre de compte (Robinson et Stern, 1997). Cette mise en lumière de « l'innovation de l'ombre » permet alors de capter l'esprit créatif des agents, de façon plus horizontale.

11. Dans l'imaginaire collectif, les laboratoires sont peuplés d'éprouvettes, de réactifs en tout genre, de machines sophistiquées, dans des assemblages parfois déconcertants, aux réactions inattendues… L'appellation FabLab permet de rappeler la place prise par ce matériel d'expérimentation dans le SSPCloud, et les créations qui peuvent découler de leur libre usage.
12. Tant qu'elles s'inscrivent dans les conditions générales d'usage. Ces dernières portent sur le type de données pouvant être traitées sur le SSPCloud. Pour en savoir plus, voir les Conditions générales d'utilisation du SSPCloud (*https://www.sspcloud.fr/tos_fr.md*).
13. Le dispositif doit aider à imaginer, demain, ce que pourraient être les futures chaînes de production de la statistique. Pour reprendre l'image du laboratoire de chimie, il s'agit de travailler non seulement sur la découverte d'une nouvelle molécule, mais aussi sur le procédé permettant de l'obtenir de façon industrielle.

Figure 2. Le SSPCloud vu de l'intérieur : dans la mécanique du nuage

Sur sa partie physique, la «fabrique» SSPCloud s'appuie sur une ferme d'une quinzaine de serveurs…



… pour une capacité totale de 6 To de RAM, 120 To de stockage, 300 cœurs (600 threads), de 18 GPU Nvidia T4 et d'un débit réseau intra cluster de 20 Gb/s.

Sur sa partie logique, elle propose un libre assemblage de services :

- environnements de développement (*Visual Studio Code*), dont des environnements spécialisés dans les langages de traitement de données (*RStudio, Jupyter*) ;
- systèmes de gestion de base de données (*PostgreSQL*, *MongoDB*) ;
- *frameworks* et composants dédiés à des traitements avancés de *data science* (*Apache Spark* pour les données massives, *TensorFlow* pour le *machine learning*, *fast.ai* pour le *deep learning*, *RAPIDS AI* pour l'utilisation de la puissance de GPU) ;
- moteurs avancés de recherche (*Elasticsearch*) ;
- gestion d'espaces sécurisés *via* la solution de gestion de secrets *Vault*.

DES CHOIX TECHNOLOGIQUES EN FAVEUR DE LA SCALABILITÉ

Dans ses partis pris technologiques, le SSPCloud s'appuie sur une architecture répondant à plusieurs sources d'influence de l'informatique contemporaine.

La capacité de dimensionner des traitements – ce que désigne la notion de scalabilité[14] – est une attente clef du *data scientist*, pour stocker et manipuler des données massives, mais aussi pour disposer de ressources computationnelles adaptées aux méthodes de *machine learning*.

Ce besoin trouve sa réponse dans la distribution du traitement sur plusieurs centres de calcul, par exemple avec une ferme de serveurs – une approche au cœur des technologies du *Cloud computing* (l'informatique en nuage). Le fournisseur de service, également appelé *Cloud provider*, met à disposition à la demande des ressources physiques (CPU, mémoire, disques, GPU) fournies par des serveurs distants, partagés entre plusieurs clients, de façon à ce que le service puisse être redimensionné au fur et à mesure du besoin exprimé, en mutualisant de vastes ensemble de serveurs.

Pour être mise en œuvre, cette approche nécessite également d'adapter l'orchestration des traitements : ce fut en particulier l'objectif poursuivi dans le développement du *framework Hadoop*[15]. L'idée principale est la colocalisation du traitement et de la donnée : si le fichier source se trouve réparti sur plusieurs serveurs (approche horizontale), chaque section du fichier source est directement traitée par un processus de la machine hébergeant cette section pour éviter les transits réseaux entre les serveurs.

Le SSPCloud a adopté une solution qui s'inscrit dans le sillage de ces travaux, en retenant le *framework Apache Spark* conçu comme une méthode pour accélérer le traitement des systèmes *Hadoop* ***(encadré 2)***. Il a d'ailleurs été utilisé à cette fin, au cours du premier semestre 2021, pour prototyper une architecture alternative à celle pré-existante, s'agissant du processus de traitement des données de caisse dans le cadre de l'élaboration de l'indice des prix à la consommation[16]. Une accélération des traitements pouvant atteindre jusqu'à un facteur 10 a ainsi été obtenue, pour des opérations qui prenaient jusqu'alors plusieurs heures d'exécution.

UN ENVIRONNEMENT ET DES RESSOURCES ACCESSIBLES EN TOUT LIEU

Dans une infrastructure *Cloud*, l'ordinateur de l'utilisateur devient un simple point d'accès pour exécuter des applications ou des traitements sur une infrastructure centrale. Le SSPCloud a été conçu sur ce modèle, en permettant aux statisticiens publics de s'y connecter depuis n'importe quel poste, dès lors que ce dernier accède à Internet. Indépendant des infrastructures propres de l'Insee ou des services statistiques ministériels, il ne nécessite donc pas d'être relié à un réseau local de l'administration d'appartenance, ce qui lui permet en outre d'accueillir des utilisateurs venus d'autres horizons (par exemple, des membres d'instituts statistiques européens, des universitaires, etc.).

14. En informatique matérielle et logicielle et en télécommunications, l'extensibilité ou scalabilité désigne la capacité d'un produit à s'adapter à un changement d'ordre de grandeur de la demande, en particulier sa capacité à maintenir ses fonctionnalités et ses performances en cas de forte demande.
15. Le *framework Hadoop* a été publié par Doug Cutting et l'entreprise *Yahoo* sous la forme d'un projet *open source* en 2008, en s'inspirant des travaux de *Google* (Dean et Ghemawat, 2004).
16. Voir (Leclair, 2019).

L'infrastructure du SSPCloud apporte ainsi un service dit ubiquitaire, accessible en tout lieu, depuis n'importe quel terminal doté d'un navigateur internet (ordinateur, tablette, téléphone, etc.).

Encadré 2. Comment l'écosystème de référence peut être détrôné en quelques années

Il faut imaginer, plutôt qu'un logiciel monolithique, tout un écosystème d'applications et de composants, en évolution permanente… jusqu'à ce qu'un nouvel écosystème l'emporte.

Début des années 2000

Les solutions de traitement de données massives voient leurs origines dans les travaux conduits pour les moteurs de recherche sur le *web*, avec les premières investigations de la fondation *Apache* pour des systèmes d'indexation de contenus massifs (projets *Lucene* et *Nutch*).

2004

Jeffrey Dean et Sanjay Ghemawat, employés chez Google, créent l'algorithme *MapReduce* pour paralléliser les traitements de grands volumes de données sur plusieurs serveurs.

2006

Doug Cutting, qui a mené les projets *Lucene* et *Nutch* avant de rejoindre *Yahoo*, crée un nouveau système de fichier distribué qu'il combine avec *MapReduce*, et nomme ce *framework Hadoop*…

Le *framework Hadoop* est destiné à créer des applications distribuées, au niveau du stockage des données et de leur traitement.

2009

…*Hadoop* est légué en 2009 à la fondation *Apache*…

…émergence d'*Apache Spark*, conçu par Matei Zaharia au sein de l'université de Californie à Berkeley…

Là où le *MapReduce* travaille par étape, *Spark* peut travailler sur la totalité des données et exécute toutes les opérations d'analyse en temps réel.

D'abord pensé en complémentarité à *Hadoop*, *Spark* peut aussi être utilisé avec d'autres méthodes de stockage, comme *Amazon S3*… et constituer ainsi le cœur d'un nouvel écosystème.

Voir (Dean et Ghemawat, 2004).

Cette logique visant à concilier scalabilité et ubiquité a également influencé le choix des technologies de stockage de la donnée. Le SSPCloud s'appuie sur S3 (*Simple Storage Service*)[17], un stockage dit « objet » – un objet se composant d'un fichier, d'un identifiant et de métadonnées, le tout de taille arbitraire, non bornée. Chaque dépôt de données (ici appelé un *bucket*) se trouve consultable directement avec un adressage unique (une URL par dépôt[18]) et des services d'accès par API[19]. Pensé pour offrir un dimensionnement adaptable, optimisé pour lancer des calculs intensifs[20], le stockage S3 apporte également des propriétés intéressantes pour faciliter l'accès aux données : ainsi, il comporte une API de sélection pour n'appeler dans une requête qu'un sous-ensemble d'un fichier, même si celui-ci est compressé ou chiffré. Surtout, il est le complément naturel à des architectures fondées sur des environnements dits conteneurisés, pour lesquels il apporte une couche de persistance, des modalités de connexion facilitées sans compromettre la sécurité, voire en la renforçant par rapport à un système de stockage traditionnel.

LA CONTENEURISATION POUR MAÎTRISER LES CONDITIONS D'EXÉCUTION

Dans le monde de l'informatique, un conteneur[21] est un regroupement logique de ressources, qui permet d'encapsuler et d'exécuter des ensembles logiciels, par exemple une application, des bibliothèques et d'autres dépendances réunies en un seul package. La conteneurisation ***(encadré 3)*** propose un système logique d'isolation, apportant une réponse à deux problématiques majeures des environnements de traitement de données :

- elle permet de gérer la concurrence d'accès aux ressources physiques (CPU, mémoire) entre les différents utilisateurs, en organisant la répartition des capacités entre les conteneurs ;
- elle assure une complète indépendance du contenu de chaque conteneur, ce qui permet de construire une large palettes de services et logiciels sans exposer les utilisateurs à des problèmes de compatibilité entre librairies.

Dans le cas d'un environnement dédié à l'expérimentation, le mécanisme des conteneurs permet de gérer une forte évolutivité de l'offre de services : en effet, contrairement à une plateforme centrale monolithique qui implique pour chaque utilisateur d'adapter son code à la montée de version logicielle du socle, un dispositif de conteneur permet à chacun de maîtriser les composants qu'il utilise. Ce dispositif permet de rejouer, dans des conditions techniques maîtrisées, l'intégralité d'un traitement.

17. En 2006, *Amazon web service* ouvre un service de stockage en ligne fondé sur une nouvelle technologie, désigné par l'appellation *Amazon S3*. Le logiciel *open source Minio* permet de construire et déployer un service S3 ne dépendant pas d'*Amazon*.
18. Une URL (*Uniform Resource Locator*, littéralement « localisateur uniforme de ressource ») est une chaîne de caractères qui permet d'identifier une ressource du *World Wide Web* par son emplacement.
19. Une API (*Application Programming Interface*) est un ensemble standardisé de méthodes par lequel un logiciel offre des services à d'autres logiciels. On parle d'API à partir du moment où une entité informatique cherche à agir avec un système tiers, et que cette interaction se fait en respectant les contraintes d'accès définies par le système tiers.
20. Contrairement au stockage HDFS qui s'appuie sur la colocalisation de la donnée et des unités de traitement, S3 permet de relier librement des conteneurs avec des données qui n'y sont pas physiquement adossées. Cette logique est pertinente pour des traitements qui nécessitent d'allouer une grande puissance de calcul de façon dynamique, comme pour le *machine learning*.
21. Voir *https://lwn.net/Articles/256389/*.

Encadré 3. Les techniques de conteneurisation renforcent l'autonomie du *data scientist*

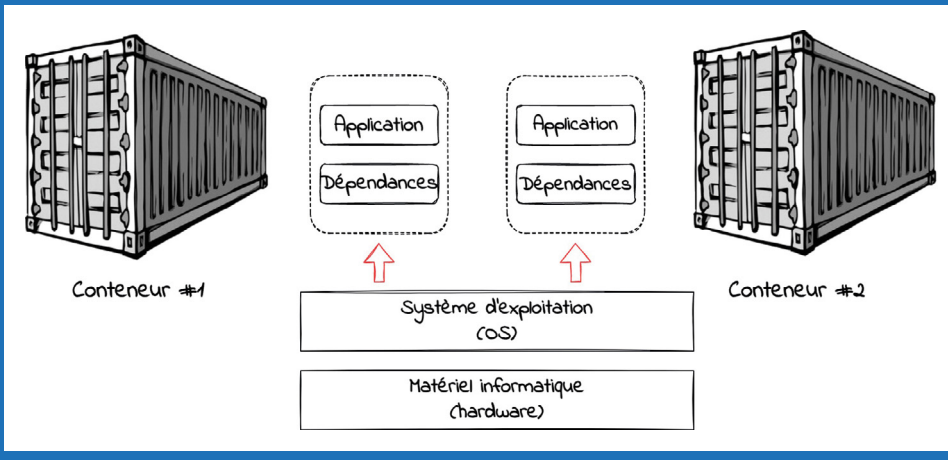
Un système d'information se compose de **services** (serveur *web*, base de données, etc.) qui s'exécutent en continu et de **tâches** dont l'exécution est planifiée de manière régulière ou ponctuelle. Le code nécessaire à l'exécution des services et des tâches s'appuie sur un **système d'exploitation**, qui organise l'accès aux ressources physiques (CPU, RAM, disques, etc.). En pratique, l'intervention d'acteurs spécialisés dans l'exploitation des infrastructures informatiques est nécessaire pour préparer et mettre à jour les systèmes sur lesquels s'exécute le code. Ainsi, dans l'organisation conventionnelle d'un SI, la mise en production et la maintenance d'un code (application complète ou script de traitement) nécessitent des interactions fortes entre les développeurs du code et les exploitants des infrastructures informatiques.

La **conteneurisation apporte une alternative** en créant des « bulles » propres à chaque service, tout en ayant recours à un même système d'exploitation support. Cette isolation assure la portabilité du code d'un environnement de développement à l'environnement d'exploitation, en maîtrisant l'ensemble des dépendances. La conteneurisation peut être harmonisée sur un ensemble de serveurs.

Pour qu'un utilisateur puisse demander, en toute autonomie, l'exécution d'un code, des ensembles logiciels assurent alors une fonction d'**orchestrateur** : *Kubernetes* optimise ainsi l'allocation des ressources physiques pour un ensemble de conteneurs et facilite leur mise en relation. Ce dispositif gère la **scalabilité** d'un traitement, par exemple en dupliquant un conteneur si nécessaire pour tenir la charge. Il gère aussi la **portabilité**, en déplaçant au besoin un conteneur dans un autre groupe de ressources (d'autres outils permettent même de les déplacer d'un *cluster* de calcul à un autre). Ces deux caractéristiques permettent de gérer des calculs intensifs avec les technologies *big data*.

Le SSPCloud correspond à un paradigme dit d'infrastructure codée (*infrastructure as a code*).

L'environnement conteneurisé est créé uniquement *via* les spécifications des scripts : le *data scientist* peut lui-même définir son environnement de travail, les ressources à y allouer, les logiciels à inclure (par exemple, *R*) et toutes les bibliothèques utiles à son traitement (par exemple, des *packages R*). La logique du SSPCloud est ainsi de proposer au statisticien d'assurer à la fois les fonctions de conception (écriture des traitements), de déploiement (écriture du conteneur qui encapsule le traitement) et d'exploitation (écriture de l'orchestration du conteneur).



Le SSPCloud est fondé sur l'utilisation de conteneurs, en s'appuyant sur l'environnement *open source Kubernetes*[22] pour déployer et gérer les services conteneurisés. Ce choix technologique, parfaitement en phase avec les environnements *Cloud*, répond en outre naturellement aux besoins de scalabilité : la ressource affectée à un conteneur est en effet paramétrable, et des traitements complexes peuvent être répartis au sein d'un *cluster* et pris en charge par plusieurs conteneurs en parallèle. En conséquence, les environnements dédiés aux données massives sont désormais de plus en plus souvent organisés *via* le recours à un mariage du *Cloud computing* et de l'orchestration de conteneurs. Les *frameworks big data* comme *Apache Spark* ou *Dask* fonctionnent en parfaite synergie avec les conteneurs : le traitement massif s'y trouve découpé dans une multitude de petites tâches et le conteneur est sans aucun doute la meilleure façon de déployer ces petites opérations unitaires, en minimisant la ressource requise[23].

ACCOMPAGNER LE GESTE INFORMATIQUE DU STATISTICIEN

L'ensemble des technologies réunies au sein du SSPCloud amène à revoir, en profondeur, le geste professionnel du statisticien dans son usage de l'environnement informatique. Au tournant des années deux-mille, la micro-informatique connaissant son apogée, une grande partie des ressources techniques d'un agent de l'Insee étaient locales, ou du moins dans une logique d'accès local : le statisticien avait, sur sa machine, son code et son logiciel de traitement, ainsi qu'un accès aux données *via* un système de partage de fichiers. Cet environnement a conduit, dans une certaine mesure, à l'essor d'une gestion manuelle des traitements réalisés « en self » par le statisticien.

Lorsque, dans une démarche de rationalisation des infrastructures informatiques, des systèmes de traitement mutualisés ont été de nouveau privilégiés, ces derniers ont cherché à préserver l'expérience utilisateur en proposant d'accéder à un « bureau distant ». Par exemple, l'Architecture Unifiée Statistique (AUS), un centre de calcul interne à l'Insee, propose une forme de transition entre une informatique locale et une informatique centralisée : elle concentre toutes les ressources sur des serveurs centraux, mais recrée un poste virtuel par agent, ce qui l'amène à garder la même pratique, largement manuelle, dans l'élaboration de ses scripts et dans l'exécution de ses traitements.

> *« Ce cadre permet d'adopter des pratiques vertueuses, pour mieux séparer le code, les données et le processus. »*

L'expérience utilisateur au sein du SSPCloud est radicalement différente, et rend impérative l'appropriation d'autres gestes. L'utilisateur ne dispose pas d'un bureau distant, et doit apprendre à composer avec des ressources qui, dans leur conception, sont évanescentes et n'existent qu'au moment de leur mobilisation effective. Ce cadre permet d'adopter des pratiques vertueuses, pour mieux séparer le code, les données et le processus, dont la persistance sera gérée selon des technologies différentes. Il lui faut également apprendre à organiser leur orchestration, dans un cadre qui concilie autonomie et automatisation. Le statisticien voit ainsi sa pratique se rapprocher de celle d'un développeur.

22. L'environnement *Kubernetes* est né dans *Google Cloud* : il y a été développé et publié en *open source* en 2014.
23. Alors qu'un dispositif de machines virtuelles (VM) nécessite d'installer un système d'exploitation complet pour chaque VM, les conteneurs sont des unités bien plus légères, partageant un même noyau de système d'exploitation.

RENDRE PÉRENNE CE QUI N'EST PAS PERSISTANT

La première transformation du geste du statisticien porte sur sa capacité à organiser la persistance des éléments qu'il mobilise dans son traitement. Dans le SSPCloud, l'ensemble des services sont dits « non-persistants » : ils sont conçus pour être désactivés lorsque l'utilisateur n'en a plus l'usage. Tout ce qui a été développé au sein de ce service disparaît à cette occasion – plus précisément, toutes les ressources conçues au sein d'un conteneur s'effacent avec l'extinction de ce dernier – contrairement à la pratique sur un poste local où l'utilisateur garde une trace de ses fichiers sur son espace de stockage, de même qu'avec un poste distant adossé à un répertoire de fichiers.

L'utilisateur du SSPCloud doit veiller à organiser la pérennité des ressources qu'il crée – à commencer par son programme informatique. Les fonctionnalités d'un outil de contrôle de version comme *Git*[24], utilisé avec une forge collaborative[25], permet au statisticien de gérer son code en dehors de l'environnement dans lequel il travaille, et d'y recourir depuis n'importe quel terminal. L'utilisation de *Git* permet en outre d'accroître la traçabilité des traitements et de les archiver au fur et à mesure qu'ils sont conçus, participant ainsi à l'amélioration continue des processus statistiques. D'autres services peuvent alors être adossés au dépôt du code source, pour « construire » le *pipeline* statistique, tester son intégrité, produire ses modules – c'est la fonction première d'une forge logicielle, un incontournable du génie logiciel qui devient également un élément pivot dans l'environnement de travail du statisticien. Le SSPCloud est prévu pour fonctionner de pair avec des forges logicielles – il comporte en son sein une instance privée du logiciel *GitLab*, et permet également d'appeler les services des instances publiques de *GitHub* et *GitLab*.

ÊTRE AUTONOME DANS L'ORCHESTRATION DE SON TRAITEMENT

La deuxième transformation apportée par le SSPCloud a trait à la méthode d'orchestration des traitements. L'utilisateur de cette infrastructure est appelé à construire de lui-même l'environnement technique adapté à son besoin, sans avoir à solliciter des équipes informatiques. À cette fin, il lui faudra apprendre à spécifier, de façon programmatique, les différents paramètres qui vont composer son environnement d'exécution.

> *« L'utilisateur de cette infrastructure est appelé à construire de lui-même l'environnement technique adapté à son besoin, sans avoir à solliciter des équipes informatiques. »*

Cette approche est possible grâce à l'utilisation des technologies *Cloud* et de la conteneurisation. En effet, les technologies *Cloud* sont agnostiques quant aux services qui y sont déployés et les conteneurs quant à eux sont façonnables à loisir par l'utilisateur, qui en définit lui-même le contenu et les ressources allouées. La possibilité offerte au statisticien de déployer en quelques secondes des bases de données, un *cluster* de calcul et un environnement de traitement statistique, qu'il a lui-même choisis voire conçus lui permet la plus grande créativité[26].

24. *Git* est un logiciel de gestion de versions, permettant de garder trace de toutes les modifications apportées à un code, pour l'ensemble des contributeurs. Il fonctionne de façon décentralisée : le code informatique développé est stocké sur l'ordinateur de chaque contributeur du projet, et le cas échéant sur des serveurs dédiés.
25. Une forge est un ensemble d'outils de travail, initialement conçus pour le développement de logiciels, utiles plus généralement pour d'autres types de projets comme l'écriture de codes statistiques. Pensée pour le travail à plusieurs mains, une forge collaborative permet d'organiser des processus de contribution impliquant plusieurs acteurs.
26. Le SSPCloud offre, au travers de son catalogue de services, des outils standards qui répondent à la majorité des besoins. Pour autant, l'utilisateur peut concevoir ses propres outils et les partager. Il est ainsi libre de tester un nouveau logiciel ou une nouvelle librairie *open source.*

Cette approche permet également de concevoir l'orchestration d'une suite d'opérations élémentaires. Ainsi, le statisticien peut concevoir un ordonnancement de tâches telles que par exemple, récupérer toutes les nuits des données issues du *web*, en extraire les informations pertinentes, les stocker dans des tables de données et alimenter une application interactive qui sera mise à jour et déployée automatiquement.

Cette logique, poussée jusqu'à son terme, a par exemple permis à une équipe de statisticiens de l'Insee de déployer un service de distancier (calcul de distances routières) utilisable par la communauté SSPCloud pour outiller des analyses spatiales. Ils ont ainsi à la fois transformé des données, déployé en toute autonomie un serveur offrant une API de calcul des distances routières et enfin, développé et déployé une application interactive permettant à un utilisateur de calculer des distances entre différents lieux. Fait remarquable, aucune intervention d'un exploitant informatique n'a été requise pour réaliser l'ensemble de ces tâches[27].

PROGRESSER DANS LA MISE EN PLACE D'UNE STATISTIQUE REPRODUCTIBLE

L'enjeu d'une pleine maîtrise du contexte d'exécution vaut tout autant pour les activités d'études que pour les activités de production statistique. Dans le cas de travaux académiques[28], il est par exemple exigé, pour des publications se fondant sur l'exploitation de données, de répondre à un nouveau critère de scientificité[29] : la possibilité pour un tiers de reproduire à l'identique l'ensemble des résultats publiés.

L'atteinte de cette reproductibilité demande donc de concevoir des solutions de traitement du chiffre qui puissent être rejouées d'une part, partagées avec des tiers d'autre part. Les technologies de conteneurs répondent parfaitement à cette attente, en décrivant de façon normée et datée toutes les ressources d'un ensemble de traitements, description qui peut être publiée sur des espaces ouverts (par exemple, des registres de conteneurs) et reliées à la publication du code source d'un traitement (par exemple, sur une forge logicielle publique comme *GitHub* ou *GitLab*).

> *« Le data scientist procède désormais d'une documentation active, où la description même des tâches à conduire va être menée de façon à automatiser leur mise en œuvre. »*

Cette exigence se diffuse progressivement au monde de la statistique publique, déjà attachée au fait de pouvoir rendre compte, par la documentation méthodologique et les rapports qualité, du processus d'élaboration d'un indicateur. Là où le statisticien public était amené à constituer une documentation passive des étapes de son processus, dans l'optique de rendre compte d'une conformité, le *data scientist* procède désormais d'une documentation active, où la description même des tâches à conduire va être menée de façon à automatiser leur mise en œuvre. La capacité à décrire les environnements de traitement de façon programmatique, avec des « contrats » de conteneurisation, est au cœur de cette démarche.

27. Voir également ***encadré 3***.
28. La détection d'erreurs majeures dans les calculs d'un article des économistes de renom Carmen Reinhart et Kenneth Rogoff, publié en 2010 et conduisant à un *erratum* en 2013, a appelé à rehausser les exigences académiques en matière de scientificité des publications économiques et statistiques.
29. En écho aux critères de scientificité de Karl Popper, ce dernier insistant sur la nécessité de pouvoir confronter des théories à des éléments de vérifications, sous forme d'expérience. Par extension, il s'agit d'avoir des éléments vérifiables pour s'assurer de l'exactitude des calculs effectués par des auteurs en démonstration de leur thèse.

S'OUVRIR POUR FACILITER LE PARTAGE DES SAVOIRS ET SAVOIR-FAIRE

En plus d'être une infrastructure informatique, le SSPCloud a été pensé comme un « lieu »[30] facilitant les rencontres entre des acteurs forts de compétences et d'expériences différentes. L'accès au SSPCloud depuis Internet rend possible son accès à l'ensemble du système statistique public – et plus largement, à l'ensemble des acteurs souhaitant s'inscrire dans un courant de partage de leur connaissance.

Date symbolique de cette volonté collaborative, la pré-ouverture du SSPCloud a été conduite le 20 mars 2020, au tout début du premier confinement lié à la crise pandémique, une période où beaucoup d'agents publics n'avaient plus d'accès à leur environnement administratif de travail. La version de test du SSPCloud a alors été dédiée à l'accueil de l'ensemble des agents publics, quelle que soit leur administration d'appartenance, pour leur apporter un espace de formation en distanciel aux logiciels de *data science*.

Couvrant tout autant des étapes de découverte (initiation à des langages) que des cas d'usages plus avancés (*machine learning*, calcul distribué), l'espace pédagogique du SSPCloud a immédiatement rencontré son public, en outillant par la même occasion une communauté d'agents dédiés à l'apprentissage de *R* et *Python*, la communauté *Spyrales*[31]. Le dispositif des conteneurs permet en particulier de proposer, aisément, des systèmes de tutoriels interactifs, en associant dans des environnements sur mesure les logiciels, les exercices et jeux de données, les exemples de programmes, sous la forme par exemple d'un *notebook*. Le SSPCloud est devenu, par la même occasion, la terre virtuelle d'accueil d'évènements à visée pédagogique, organisés en distanciel en 2020 et en 2021, à l'instar d'un jeu sérieux de formation en *R*[32].

BÉNÉFICIER DE LA MISE EN COMMUN EN PRIVILÉGIANT L'*OPEN SOURCE*

L'élargissement des publics bénéficie plus généralement à l'innovation, en facilitant la coopération entre des équipes qui, autrement, évoluent dans des systèmes d'information étanches.

L'essor de l'*open source* est venu apporter un cadre juridique et une méthode de travail permettant d'accompagner cette volonté de partage. La statistique a été marquée par l'essor des logiciels libres durant ces 20 dernières années : c'est ainsi que *R* et maintenant *Python* sont devenus les langages de référence pour la statistique en lieu et place, par exemple, de *SAS*®.

Au-delà de l'usage de langages et logiciels *open source*, les statisticiens ont plus généralement pris l'habitude de partager leurs programmes : ainsi, la publication d'un article proposant une nouvelle méthode de traitement des données est désormais quasi systématiquement

30. La notion d'infrastructure renvoie à l'idée d'un axe de passage (l'autoroute) où les utilisateurs se suivent ; la notion de lieu renvoie à un espace de jonction (la place publique) où les utilisateurs se rencontrent.
31. La communauté *Spyrales* a été constituée en mars 2020 pour aider les agents publics à se former aux langages *R* et *Python*, en facilitant la mise en relation de personnes désireuses de se former et de tuteurs/formateurs et en cataloguant des ressources pédagogiques.
32. Conçu sur les principes des « jeux sérieux », le *FuncampR* propose une approche ludique de l'apprentissage de *R*. Au cours d'un jeu vidéo, le stagiaire est amené à résoudre des énigmes dont la réponse est apportée dans des tutoriels *R*. Le *FuncampR* est une formation déployée en ligne sur le SSPCloud en utilisant la technologie des conteneurs.

accompagnée par la publication, par les mêmes auteurs, d'une librairie *R* ou *Python* implémentant leurs méthodes. Le partage du savoir statistique s'effectue donc de façon duale, à la fois scientifique mais aussi technique.

Le SSPCloud a fait le choix de fonder ses services, exclusivement, sur des briques *open source* – logiciels statistiques, systèmes de gestion de base de données, outils de développement, etc. Il s'agit en effet de s'assurer de la portabilité des travaux menés sur le SSPCloud vers n'importe quel autre environnement de *data science*, sans barrière de propriété, là où l'introduction de logiciels commerciaux conduit inéluctablement à limiter les possibilités de réutilisation.

De façon emblématique, l'interface graphique du SSPCloud est elle-même le fruit d'un projet collaboratif *open source*[33] développé à cette occasion par l'Insee. Ce projet logiciel, intitulé *Onyxia*, permet à l'utilisateur de lancer ses services, de gérer ses dépôts de données, de paramétrer ses ressources ***(figure 1)***. Elle est conçue pour pouvoir être déployée sur d'autres infrastructures utilisant des technologies de conteneurs, et peut ainsi être réutilisée par d'autres directions informatiques désireuses de constituer des services de même nature.

NOUER DES PARTENARIATS TECHNIQUES, POUR INSPIRER ET S'INSPIRER

Fidèle à sa vocation de fabrique créative, le SSPCloud se veut une pierre de touche pour aider le statisticien à découvrir de nouvelles technologies, mais aussi pour permettre aux ingénieurs informatiques d'imaginer de nouvelles architectures de traitement du chiffre.

Bac-à-sable pour ses utilisateurs finaux, le SSPCloud l'est également pour ses concepteurs, dans une posture de veille sur les nouvelles opportunités techniques à prendre en compte. La démarche du SSPCloud vise ainsi à nouer des partenariats technologiques avec des tiers, tant au niveau du système statistique européen qu'auprès d'autres acteurs spécialisés dans la manipulation et l'exploration de données (centres de recherche, observatoires). L'ensemble du projet, fondé sur l'*open source*, permet d'en imaginer une réutilisation par de futurs partenaires, créant des instances propres tout en contribuant à améliorer, avec leurs retours d'expérience, la conception d'ensemble du SSPCloud. C'est ainsi qu'Eurostat a mis en place en 2021, sur la base du code source partagé, un environnement expérimental de traitement de la donnée, engageant la dynamique d'échanges au cœur de l'intention portée par les équipes du SSPCloud.

33. Voir *https://github.com/InseeFrLab/onyxia-ui*.

BIBLIOGRAPHIE

ANDERSON, John C. et VENKATRAMAN, N., 1990. *Strategic alignment: a model for organizational transformation through information technology.* [en ligne]. Novembre 1990. Center for Information Systems Research, Massachusetts Institute of Technology. CISR WP N°217. [Consulté le 16 décembre 2021]. Disponible à l'adresse : *https://dspace.mit.edu/bitstream/handle/1721.1/49184/strategicalignme90hend.pdf*.

DEAN, Jeffrey et GHEMAWAT, Sanjay, 2004. *MapReduce: Simplified Data Processing on Large Clusters.* [en ligne]. OSDI'04, Sixth Symposium on Operating System Design and Implementation, pp. 137-150. [Consulté le 16 décembre 2021]. Disponible à l'adresse : *https://research.google/pubs/pub62.pdf*.

DINUM et INSEE, 2021. *Évaluation des besoins de l'État en compétences et expertises en matière de donnée.* [en ligne]. Juin 2021. Insee, Direction interministérielle du numérique. [Consulté le 16 décembre 2021]. Disponible à l'adresse : *https://www.numerique.gouv.fr/uploads/RAPPORT-besoins-competences-donnee.pdf*.

EUROSTAT, 2018. Bucharest Memorandum on Official Statistics in a Datafied Society (Trusted Smart Statistics). In : *site d'Eurostat*. [en ligne]. 12 octobre 2018. 104e conférence des directeurs généraux des instituts nationaux statistiques (DGINS). [Consulté le 16 décembre 2021]. Disponible à l'adresse : *https://ec.europa.eu/eurostat/fr/web/european-statistical-system/-/dgins2018-bucharest-memorandum-adopted*.

MÉRINDOL, Valérie, BOUQUIN, Nadège, VERSAILLES, David W. et alii, 2016. *Le Livre blanc des Open Labs. Quelles pratiques ? Quels changements en France ?* [en ligne]. Mars 2016. Futuris, PSB. [Consulté le 16 décembre 2021]. Disponible à l'adresse : *https://www.researchgate.net/profile/Ignasi-Capdevila/publication/301356114_Les_open_labs_de_la_recherche_et_de_l'enseignement_superieur/links/571501de08aedc7cbcc99e6d/Les-open-labs-de-la-recherche-et-de-lenseignement-superieur.pdf*.

NIST, 2017. *NIST Big Data Program*. [en ligne]. Mis à jour le 11 janvier 2017. The National Institute of Standards and Technology (NIST). U.S. Department of Commerce. [Consulté le 16 décembre 2021]. Disponible à l'adresse : *https://bigdatawg.nist.gov*.

ROBINSON, Alan G. et STERN, Sam, 1997. *Corporate creativity: How innovation and improvement actually happen*. Éditions Berrett-Koehler.

SCHON, Donald A., 1963. *Champions for Radical New Inventions*. Mars-avril 1963. Harvard Business Review.

SUAREZ CASTILLO, Milena, SEMECURBE, François, LINO, Galiana, COUDIN, Élise et POULHES, Mathilde, 2020. *Que peut faire l'Insee à partir des données de téléphonie mobile ? Mesure de population présente en temps de confinement et statistiques expérimentales.* [en ligne]. 15 avril 2020. Blog Insee. [Consulté le 16 décembre 2021]. Disponible à l'adresse : *https://blog.insee.fr/que-peut-faire-linsee-a-partir-des-donnees-de-telephonie-mobile-mesure-de-population-presente-en-temps-de-confinement-et-statistiques-experimentales/*.

UNECE, 2013a. *Utilisation des « données massives » dans les statistiques officielles. Note du secrétariat.* [en ligne]. 18 mars 2013. Groupe de haut niveau sur la modernisation de la production et des services statistiques. Conférence des statisticiens européens des 10-12 juin 2013, Genève. [Consulté le 16 décembre 2021]. Disponible à l'adresse : *https://statswiki.unece.org/download/attachments/77170614/Big%20Data%20Published%20version%20FR.pdf?version=1&modificationDate=1370507699015&api=v2*.

UNECE, 2013b. *Big data and modernization of statistical systems, Report of the Secretary-General.* [en ligne]. 20 décembre 2013. 45th Statistical Commission. [Consulté le 16 décembre 2021]. Disponible à l'adresse : *https://unstats.un.org/unsd/statcom/doc14/2014-11-bigdata-e.pdf*.

UNECE, 2021. *HLG-MOS Machine Learning Project.* [en ligne]. Mis à jour le 13 octobre 2021. [Consulté le 16 décembre 2021]. Disponible à l'adresse : *https://statswiki.unece.org/display/ML/HLG-MOS+Machine+Learning+Project*.

QUELQUES BONNES PRATIQUES DE DÉVELOPPEMENT LOGICIEL À L'USAGE DU STATISTICIEN SELFEUR

(OU « SAVOIR COMPTER, SAVOIR CODER »)

Emmanuel L'Hour, Ronan Le Saout** et Benoît Rouppert****

En plus de compétences en méthodologie statistique et d'une bonne connaissance des sources de données disponibles, le métier de statisticien nécessite une bonne maîtrise des outils informatiques. Les programmes informatiques écrits permettent non seulement de produire des résultats, mais ils peuvent aussi devenir des livrables du travail réalisé, que ce soit en tant qu'élément de preuve ou en tant qu'outils réutilisables pour d'autres travaux. Dans cette optique, le statisticien se doit d'acquérir les bonnes pratiques de développement logiciel qui lui permettront de garantir une appropriation facile de ses programmes par d'autres utilisateurs, ou une ré-appropriation par lui-même, au-delà de la période pendant laquelle le développement a eu lieu.

Ces bonnes pratiques couvrent tous les aspects du cycle de développement logiciel : la définition des exigences, l'architecture du programme, les styles de programmation, les choix techniques, les outils de développement et les tests. Elles permettront de garantir une bonne réponse au besoin des utilisateurs, après une étape de questionnement de ces besoins, et une fiabilité des résultats obtenus tout en maîtrisant le coût de programmation. Et plus important encore : en rendant les programmes facilement lisibles, elles aideront le producteur de statistiques publiques à communiquer sur ses choix méthodologiques et sur la façon d'utiliser les données, renforçant ainsi la confiance que lui accordent ses utilisateurs.

In addition to skills in statistical methodology and a good knowledge of available data sources, the profession of statistician requires to be comfortable with IT tools. Programs not only produce the statistical results, they can also become deliverables, either as evidence or as reusable tools for future work. Having this in mind, the statistician must acquire software development best practices that will allow him to guarantee an easy understanding of his programs by other users, or re-appropriation by himself in future developments.

These best practices cover all aspects of the software development cycle: definition of requirements, program architecture, programming styles, technical choices, development tools and testing. They will make it possible to guarantee an appropriate answer to users' needs, after a step of questioning these needs, and quality of the results obtained while controlling the development costs. And more important, by making the programs easily readable, they will help the producer of official statistics to communicate on his methodological choices and how to use the data, and thus strengthen the confidence of his users.

* Chef du Service statistique, direction interrégionale Insee La Réunion-Mayotte, *emmanuel.lhour@insee.fr*

** Expert en méthodologie statistique, Commissariat général au développement durable, *ronan.lesaout@developpement-durable.gouv.fr*

*** Ancien chef du département des Production et infrastructure informatiques, Insee

DE LA PROGRAMMATION À TOUS LES ÉTAGES

Parmi les statisticiens apparaissent de plus en plus des profils qui combinent des compétences en méthodologie statistique, des connaissances sur le contenu métier des sources de données, et une maîtrise des outils informatiques. Ce troisième élément prend une importance croissante, notamment par la capacité à « savoir coder »[1], c'est-à-dire savoir écrire un programme informatique. Mais derrière cette formule facile à retenir, se cache une diversité de pratiques de programmation, telles que :

- celle du responsable des retraitements post-collecte d'une enquête ;
- celle du chargé d'études s'appuyant sur des méthodes statistiques plus ou moins sophistiquées pour établir un résultat ;
- celle du chercheur désirant rendre ses travaux reproductibles ;
- ou celle du « diffuseur » qui veut mettre à la disposition d'un large public une application de visualisation des données.

Le point commun entre ces exemples est qu'ils mobilisent des compétences souvent associées à l'informatique et parfois ignorées du statisticien. Elles relèvent du dialogue avec l'utilisateur, de la conceptualisation de son processus de travail, des méthodes de travail en équipe et de la connaissance des méthodes et des outils du jour. Cet article vise donc à présenter ici quelques outils conceptuels à l'usage du statisticien pour se construire une aisance rédactionnelle avec ses programmes.

LE SELFEUR, UN «DÉVELOPPEUR» EXPERT DE SON DOMAINE MÉTIER

La profession de statisticien public n'est pas homogène. Elle recouvre en fait une diversité de métiers.

> *« Le terme de « selfeur » est un néologisme qui renvoie à la partie de l'activité du statisticien qui mobilise de la programmation informatique à travers une expertise « métier ». »*

L'Insee en identifie très précisément 38, regroupés en 6 familles professionnelles (production, études, action régionale, informatique, fonctions support, stratégie et pilotage)[2]. Les termes « développement » et « programmation » (au sens informatique) n'y apparaissent que pour l'analyste-programmeur. Ceci ne signifie évidemment pas qu'il y a absence de programmation informatique dans les autres activités de la statistique publique. En particulier, les descriptifs des métiers de production, d'études ou de l'action régionale[3] font mention de l'utilisation de logiciels statistiques. Le terme de « selfeur » est un néologisme qui renvoie à la partie de l'activité du statisticien qui mobilise de la programmation informatique à travers une expertise « métier »[4].

1. L'article s'appuie sur un document de travail des mêmes auteurs (L'Hour, Le Saout et Rouppert, 2016), dont le titre est librement inspiré par la série d'articles du Courrier des statistiques autour des techniques rédactionnelles, « Savoir compter, savoir conter » (Cotis *et alii*, 2009).
2. [NDLR] La grille des métiers de l'Insee à laquelle les auteurs se réfèrent est une note interne du département RH de l'Insee, en date du 16 janvier 2017. Les pages du site internet qui traitent des *métiers et des formations* reprennent cette typologie (hors stratégie et pilotage).
3. Dans l'organisation de l'Insee, l'action régionale recouvre une palette de travaux à destination des décideurs régionaux ou locaux (études, diffusion de données). C'est une des particularités de l'institut statistique français.
4. Le terme fait notamment référence au travail de programmation en « libre-service », en « self ».

Mais au-delà de l'expertise métier, la comparaison avec les activités de développement informatique dans un cadre « classique » laisse entrevoir une autre notion : celle de la responsabilité vis-à-vis du code. Lorsque la direction du Système d'Information (DSI) est mise à contribution dans un projet d'investissement, il est d'usage de distinguer deux natures de responsabilités : d'une part la responsabilité technique de l'équipe informatique, et d'autre part la responsabilité fonctionnelle de l'équipe « métier » (assurée par le chef de projet statistique ou l'administrateur d'application dans le cadre d'une maintenance). Le développeur informatique est alors essentiellement garant de la bonne exécution du code, tandis que la validité statistique du résultat est plutôt assumée par un responsable d'application.

> *« Le selfeur porte en effet la responsabilité de la bonne exécution du programme, mais aussi de la valeur statistique de ce que le programme produit. »*

Dans le cas d'un développement en autonomie (ou self, ou libre-service), ces responsabilités sont souvent cumulées, même si la responsabilité technique ne fait généralement l'objet que de peu de contrôles et que, de ce fait, on a tendance à l'oublier un peu. Le selfeur porte en effet la responsabilité de la bonne exécution du programme, mais aussi de la valeur statistique de ce que le programme produit (indicateur, base de données retraitée, étude économique, outil de visualisation des résultats, etc.). Il se distingue de l'informaticien par la mobilisation de compétences spécifiques au métier statistique. Le chargé d'enquête, le chargé de l'exploitation de fichiers administratifs ou le chargé d'études intègre ainsi au travail de programmation, des savoirs spécifiques sur les concepts à mesurer : il est capable de mobiliser des références bibliographiques sur la thématique, de prêter attention à des différences de champ dans les sources mobilisées, ou de repérer les difficultés de mesure, etc .

Le selfeur programme au premier abord essentiellement pour lui-même, pendant que le développeur informaticien travaille sur des applications ou des logiciels de large utilisation. Les contraintes ne sont donc *a priori* pas identiques.

L'expérience montre qu'en réalité, informaticien et selfeur portent un même type de responsabilité. Pourquoi donc porter attention à la lisibilité des programmes de retraitements si c'est pour une enquête unique ? Pour la simple raison qu'il est fort probable que ces programmes n'auront pas un usage unique et qu'il est difficile de prévoir les usages futurs : élément de preuve associé à la reproductibilité[5] des études économiques, transmission à des collègues d'éléments méthodologiques, ré-utilisation dans un autre cadre ou pour corriger une erreur, etc. Ce faisant, un selfeur est un développeur comme les autres.

Or, si les techniques rédactionnelles font partie de la formation de base d'un chargé d'études, les « techniques rédactionnelles du code » en sont absentes. Pourtant, les compétences de génie logiciel gagneraient à se propager hors des sphères des projets informatiques. Le statisticien selfeur pourrait s'inspirer des bonnes pratiques bien connues des développeurs informaticiens, pour sécuriser son code et faciliter sa réutilisation ou la reproductibilité de son étude. Ces pratiques s'appuient sur un cadre conceptuel simple et qui a fait la preuve de son efficacité dans les développements « classiques » : un cycle de développement et des étapes-clefs à ne pas négliger.

5. Pour aller plus loin sur la conception orientée reproductibilité (*reproducibility by design*), voir par exemple (Langlais et Eprist, 2020). Sur la reproductibilité de traitements sur des données confidentielles, voir par exemple (Gadouche, 2019).

UN CYCLE DE DÉVELOPPEMENT À TROIS TEMPS

Le cycle de développement d'un logiciel met en jeu trois[6] activités principales ***(figure 1)*** :

- la première a pour objectif de déterminer ce à quoi doit répondre le programme ; c'est ce que l'on appelle les **exigences** ; elle comporte des phases d'élucidation, de spécification puis de validation de la spécification[7] ;
- la deuxième consiste à **concevoir et écrire le programme** ;
- et la troisième à **tester le programme**, c'est-à-dire vérifier qu'il fait bien ce qui est attendu de lui.

Ce processus n'est jamais linéaire. Chacune de ces activités peut faire l'objet d'itérations, à trois niveaux :

- au sein d'une même activité : on peut avoir besoin de réécrire les exigences pour être plus clair, ou corriger un *bug* dans le code (niveau 1) ;
- des allers-retours entre les activités sont également à prévoir : en codant, des exigences peuvent apparaître ambiguës, les tests peuvent identifier des *bugs* et conduire à revoir le code, voire les spécifications des exigences (niveau 2) ;
- enfin, on peut augmenter progressivement le nombre ou le périmètre des livrables du programme, par itérations de cycles (niveau 3). Par exemple, on va d'abord éditer un tableau de données sur une année, puis ajouter des années, puis faire des graphiques, puis les rendre interactifs, etc.

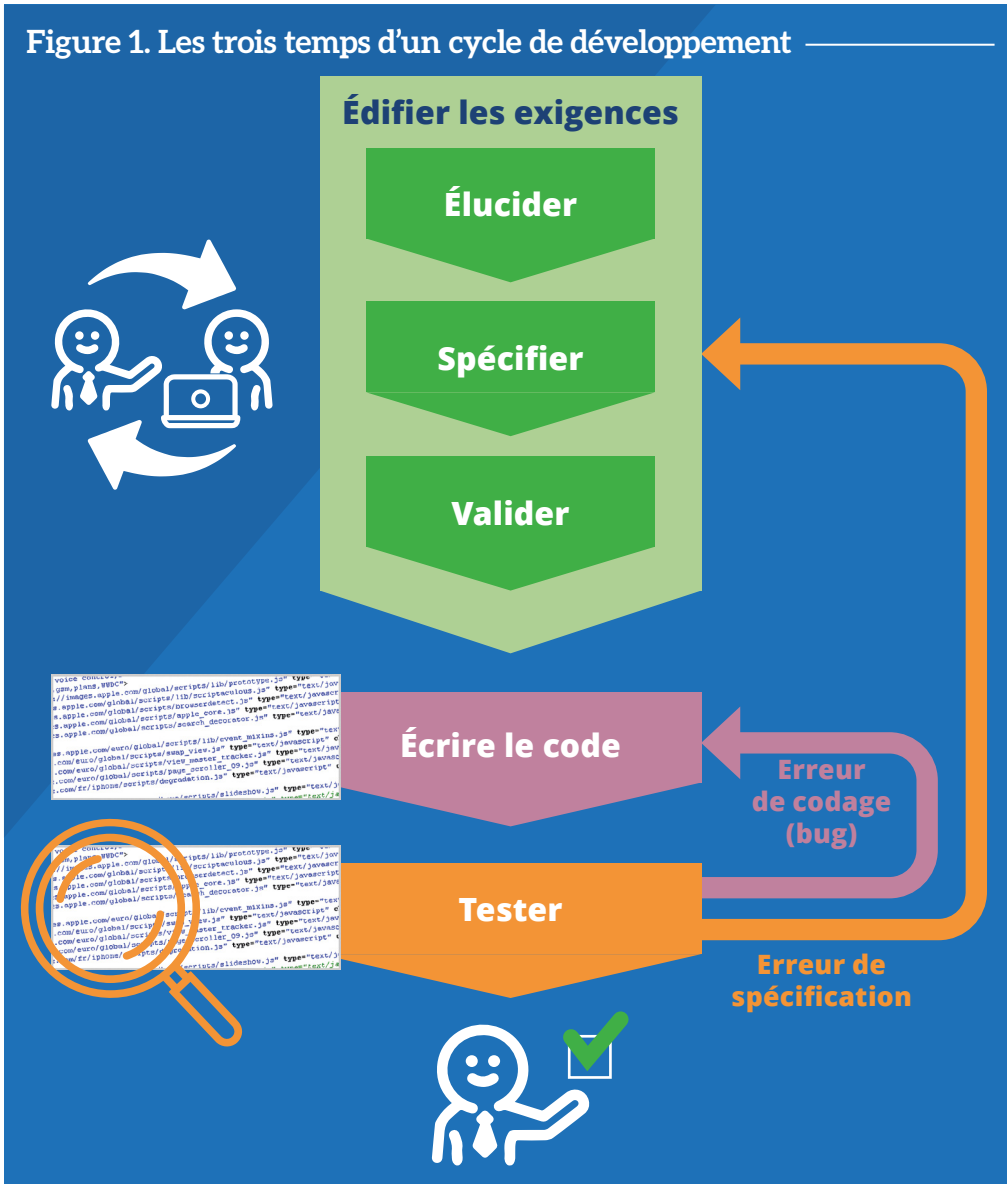
Néanmoins, le plus efficace (McConnell, 2005, chapitre 3) consiste à prévenir les itérations de deuxième niveau, signes de spécifications mal définies ou d'erreurs de codages ***(figure 1)***, qui se révèlent très coûteuses. On pourra ainsi privilégier les itérations au sein de chaque activité (premier niveau), et les itérations de cycles (troisième niveau), où l'on produit toujours quelque chose de fonctionnel en augmentant progressivement la qualité, c'est-à-dire l'adaptation du résultat à la demande.

> *« Le plus efficace consiste à prévenir les itérations de deuxième niveau, signes de spécifications mal définies ou d'erreurs de codages. »*

On conçoit que le statisticien qui écrit un code à usage unique n'éprouvera peut-être pas spontanément le besoin de conceptualiser autant le cadre de l'écriture de son programme. Mais il est probable que la méthode et les outils se révéleront tout particulièrement utiles pour le selfeur isolé, en particulier pour l'aider à gérer ses « conflits » de responsabilité. Dans les projets informatiques, l'organisation du travail collectif porte en soi la méthode, notamment les espaces de concertation en amont et en aval du développement. Si le développement en self s'insère dans un ensemble plus vaste, par exemple entre un processus de collecte doté d'applicatifs spécifiques et un processus de diffusion des données doté de ses contraintes propres, il devient également très intéressant de se pencher sur ce que les informaticiens de métier ont noté comme pièges à éviter et pratiques à préconiser. En tout premier lieu, il convient d'analyser le besoin qui justifie l'écriture d'un programme, ce qu'on nomme les exigences.

6. La littérature autour du développement piloté par les tests *(tests driven development)* utilise une typologie identique : design – code – test.
7. La littérature sur l'ingénierie des exigences *(requirement management)* est plutôt anglophone. Voir par exemple (Robertson et Robertson, 2013) ou (Cockburn, 2000). En français, on peut lire aussi (Constantinidis, 2018).

Figure 1. Les trois temps d'un cycle de développement



ÉDIFIER LES EXIGENCES…

En génie logiciel, une exigence est « *l'expression d'une condition ou d'une fonctionnalité à laquelle doit répondre un système ou un logiciel* »[8].

Établir les exigences, c'est donc identifier les besoins[9], mais également les contraintes, puis décrire ce qu'on est censé faire. Bien poser le problème, savoir quoi exiger, est crucial pour ne pas rater son objectif et éviter d'avoir à tout recommencer (voire abandonner). Un problème mal spécifié expose en général à des difficultés bien plus grandes qu'une mauvaise ligne de code (McConnell, 2005, chapitre 3).

Ne pas se précipiter vers le code et prendre le temps d'établir des exigences a de multiples vertus. Pour un informaticien, cela garantit la satisfaction de l'utilisateur et du donneur d'ordre et réduit aussi les risques d'échec du projet et la durée d'écriture du programme.

L'IREB[10] retient une typologie en trois activités majeures pour établir des exigences : les élucider, les spécifier et les valider. Pour les mettre en œuvre, les outils sont essentiellement sémantiques (distinguer le besoin du superflu ou de la contrainte, se mettre d'accord sur les termes métier utilisés), linguistiques (formuler des exigences), et visuels (faire des schémas). Le souci de sobriété doit prévaloir, car la complexité est source de fragilité (Volle, 2001a ; McConnell, 2005, chapitre 27).

… CE QUI SUPPOSE QUE L'ON CONNAISSE LES UTILISATEURS

Dans un projet informatique classique, l'utilisateur et le donneur d'ordre (appelé aussi client) sont généralement distincts. Le client donne les moyens pour que le projet se fasse (la hiérarchie qui donne l'aval du projet, le partenaire qui finance, etc.). L'utilisateur manipulera le produit du projet. C'est le rôle du développeur de satisfaire la demande du client, mais en s'assurant qu'il fournit un produit qui convient à l'utilisateur[11].

> *« C'est le rôle du développeur de satisfaire la demande du client, mais en s'assurant qu'il fournit un produit qui convient à l'utilisateur. »*

Qui sont ces utilisateurs ? Si aucun utilisateur n'est identifié *a priori*, on peut s'interroger sur l'opportunité du projet. La question n'est pas simple : elle amène une confusion possible entre les utilisateurs des données produites à l'aide du programme développé et les utilisateurs directs du programme.

Dans le premier cas, on parlera d'utilisateur final, par essence assez éloigné de l'informaticien et parfois même du selfeur. La statistique publique dispose généralement d'une représentation des utilisateurs finaux de ses productions, *via* le Cnis[12], un Cries[13], un comité d'utilisateurs d'une enquête, le comité de pilotage d'une étude en partenariat dans une région, etc. Mais ils se situent par construction très en amont du développement.

8. Voir également le site de la Specief (Société pour la promotion et la certification de l'ingénierie des exigences en langue française) *http://specief.org/index.php/ingenierie-des-exigences/*.
9. *« La définition des besoins […] est complexe et délicate, en raison du nombre et de la diversité des parties prenantes, des demandes souvent divergentes, des contraintes variées et, […] du facteur humain »* (Constantinidis, 2018).
10. L'IREB (*International Requirements Engineering Board*), organisation à but non lucratif, est le fournisseur du schéma de certification en ingénierie des exigences (CPRE pour *Certified Professional for Requirements Engineering*).
11. Sur la différence entre la satisfaction de la demande et celle des besoins, voir par exemple (Volle, 2001b).
12. Conseil national de l'information statistique, voir (Anxionnaz et Maurel, 2021).
13. Comité régional pour l'information économique et sociale.

L'utilisateur qui intéresse le développeur, c'est celui à qui il va livrer son programme, ou le résultat de son programme. Il faut donc qu'il soit en capacité d'accéder directement à cet utilisateur. Certes, à défaut, on pourra à certains moments demander à un collègue de jouer le rôle de cobaye. Mais l'idéal reste néanmoins de pouvoir observer l'utilisateur face au résultat produit : le graphique ou le tableau est-il facilement intelligible ? Quels passages de l'étude a-t-il du mal à comprendre ? Le modèle est-il clair et convaincant ? Cet outil interactif est-il aisé à manipuler ?

DANS LE CAS D'UN DÉVELOPPEMENT EN SELF, QUI SONT CES UTILISATEURS?

À première vue, on aura tendance à répondre : le selfeur lui-même. En effet, on associe spontanément au terme de « self » l'image d'un programme développé pour l'usage d'une seule personne, voire un développement à usage unique, par exemple en statistique exploratoire.

Or en pratique, la notion de self recouvre une grande diversité de situations. Rien qu'à l'Insee, sur le périmètre de la production statistique, on a récemment recensé 360 opérations statistiques outillées par des développements applicatifs en self : il s'agit d'une pratique généralisée dans toutes les activités de l'institut, qu'on retrouve dans toutes les phases des processus, de la conception à la diffusion. Le statisticien selfeur qui a écrit le code de ce type de traitements, sous-processus d'un processus plus large, est donc dans une situation assez voisine de l'informaticien dans un projet « classique ». Cependant, il ne pourra pas toujours identifier avec précision l'utilisateur de son programme et sera relativement éloigné des utilisateurs finaux du processus englobant.

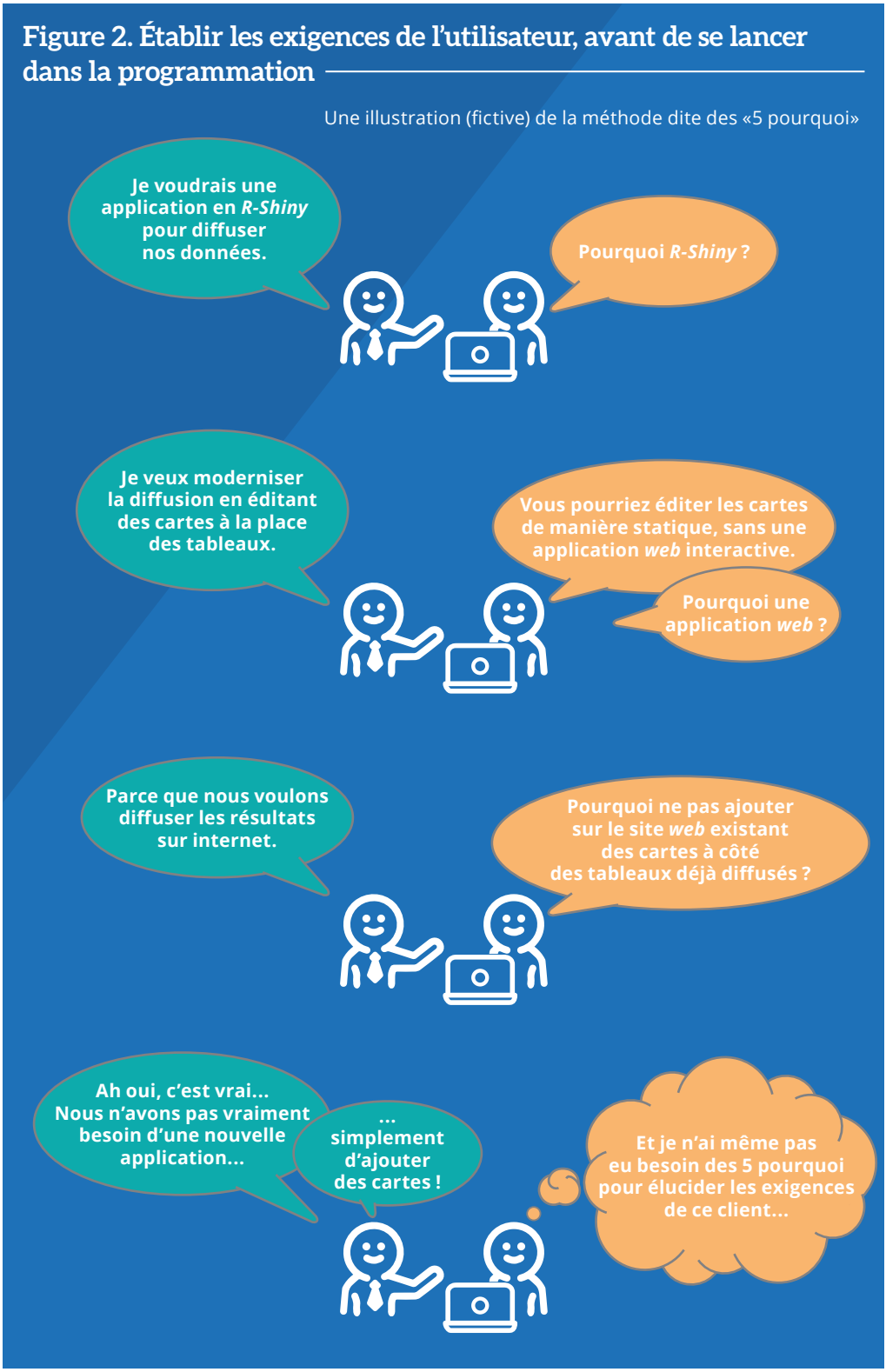
S'il est expert « métier », le statisticien selfeur se sentira particulièrement performant pour comprendre le besoin de l'utilisateur final. Pour autant, il lui faudra éviter de tomber dans le double piège qui consiste à croire, d'une part que l'utilisateur de son programme est le même que l'utilisateur final ; et d'autre part qu'il saurait mieux que son utilisateur ce dont ce dernier a besoin. Par exemple, pour développer un outil de visualisation des adresses à enquêter sur une carte, des échanges avec les enquêteurs aboutiront au choix de cartes interactives numériques plutôt que de cartes imprimées, facilitant ainsi l'usage du GPS sur le terrain.

FAIRE LA CHASSE À L'IMPLICITE, AU NON-DIT

En définitive, élucider les exigences, c'est aller au bout de la démarche de compréhension des besoins de l'utilisateur pour bien distinguer le besoin (ce qui est utile) et la contrainte (légale, matérielle, technique…) du superflu. Sans cela, des exigences que l'utilisateur ne formule pas spontanément, basées sur des besoins implicites ou des contraintes inamovibles seront découvertes tardivement, au risque de devoir abandonner ou tout refaire. Une méthode simple permet d'ailleurs de rendre explicite ce que son interlocuteur tait, souvent parce qu'il le considère comme évident : la série des « cinq pourquoi ». Celle-ci consiste très simplement à demander cinq fois à la suite à son interlocuteur pourquoi il cherche à réaliser ce qu'il demande. À l'usage, les cinq « pourquoi ? » ne seront pas systématiquement tous nécessaires pour aller au fond des choses ***(figure 2)***.

Figure 2. Établir les exigences de l'utilisateur, avant de se lancer dans la programmation

Une illustration (fictive) de la méthode dite des «5 pourquoi»



Dans un autre ordre d'idée, mais pour autant tout aussi primordial, le développeur, comme le selfeur, devront lister les besoins purement informatiques, qu'on appelle « non fonctionnels », pour s'interroger sur la faisabilité du projet avant de se lancer dans la programmation : les exigences de disponibilité ou de sécurité, ont-elles été correctement instruites ? Quel sera le nombre d'utilisateurs en simultané ? A-t-on seulement besoin d'un accès en lecture des fichiers produits, ou aussi en modification ? À quelle fréquence doit-on mettre à jour les données ? Par exemple, pour le suivi de la conjoncture économique, il est nécessaire de pouvoir actualiser les prévisions très fréquemment pendant une période précise du trimestre. Il sera donc très probablement nécessaire d'automatiser totalement l'outil développé pour atteindre l'objectif de réactivité.

MODÉLISER À L'AIDE D'OUTILS SÉMANTIQUES ET VISUELS

Le développeur qui a analysé les besoins fonctionnels auxquels son code doit répondre, et apprécié les contraintes techniques qui s'imposent à lui, doit s'assurer que sa vision est partagée avec celle de son client. Un premier pas dans cette voie consiste à formaliser **le cadre conceptuel** de son travail. Car l'usage de concepts partagés évite les incompréhensions (Evans, 2003).

> *« L'usage de concepts partagés évite les incompréhensions. »*

Pour le statisticien selfeur, une manière d'appliquer cette bonne pratique est de se poser la question des concepts qu'il manipule à travers son code. S'il travaille par exemple sur des données d'entreprises, il devra prendre en compte des concepts statistiques, d'identification ou d'analyse (Siren, Siret, entreprise profilée, taille d'entreprise…) et des concepts comptables (chiffres d'affaires, valeur ajoutée). La liste des concepts servira de référentiel pour nommer les entités du code et des données (tables, variables et leurs modalités, fonctions). À l'Insee, le référentiel des métadonnées statistiques (RMéS) centralise la définition des concepts et des sources de données de la statistique publique[14]. Il constitue donc un outil commode pour le selfeur, comme pour l'utilisateur.

Les schémas sont en outre de précieux outils de communication et de conception. Représenter graphiquement les exigences permet de renouveler la discussion avec l'utilisateur. De surcroît, si les schémas sont confus, traduisant des dépendances multiples et erratiques, cela reflète bien qu'on n'est pas prêt à se mettre à coder. C'est aussi un bon support de documentation.

La modélisation du *Generic Statistical Business Process Model* (GSBPM) répertorie et catégorise les phases d'un processus de production statistique[15]. Cette schématisation d'un processus est générique : elle s'applique quel que soit le processus et permet ainsi de ne pas oublier d'étapes importantes.

14. Pour plus d'information sur le référentiel RMéS, voir (Bonnans, 2019).
15. On trouvera plus d'information sur le GSBPM et sur un modèle qui s'en inspire dans (Erikson, 2020).

Les exemples sont indispensables pour parer au risque d'une modélisation trop abstraite. En particulier, on n'omettra pas la description des exigences en matière de validation de données ou de traitements. La validation d'un traitement peut consister à vérifier les données, individuelles ou agrégées, en cherchant à repérer des valeurs atypiques. Par exemple, un responsable d'enquête auprès d'entreprises identifiera des réponses dont le niveau ou dont l'évolution paraît suspecte, en définissant des seuils, des intervalles, etc. Un chargé d'études, avant de se lancer dans les calculs, construira un ensemble de données de référence à partir d'une bibliographie thématique, ensemble auquel il confrontera ensuite les résultats de ses calculs.

VALIDER LES EXIGENCES AVEC TOUTES LES PARTIES PRENANTES

Dans un projet informatique classique, outre l'utilisateur et le développeur, il convient que le donneur d'ordre valide également les exigences, à l'aide des **spécifications**.

Les spécifications ont ainsi un double usage : elles préparent la phase de codage, mais elles servent aussi à solliciter les utilisateurs. Elles explicitent en quoi le processus proposé répond au besoin de l'utilisateur. Cela peut passer par la description de tests fonctionnels (cf. *infra*) ou la réalisation de prototype. L'écriture des spécifications peut ainsi réduire la frontière entre les cycles de développement, avec une anticipation sur l'écriture du programme et des tests. Si les exigences ne sont pas validées, il faut alors déterminer en quoi le besoin a mal été compris et ce qui doit être revu dans les spécifications.

Notre statisticien selfeur aura ici encore peut-être du mal à identifier son donneur d'ordre, mais le terme de parties prenantes aura très certainement une vraie signification. Ira-t-il jusqu'à faire écrire puis valider des spécifications à celui qu'il aura identifié comme son donneur d'ordre ? Il faut l'essayer pour s'en convaincre : même s'il est son propre donneur d'ordre, son propre utilisateur, le selfeur a tout intérêt à décrire ce qu'il anticipe être le résultat de son programme.

ENCORE UN PEU DE PATIENCE : AVANT DE CODER, DÉFINIR L'ARCHITECTURE

Pour implémenter les exigences, il sera nécessaire de réaliser une succession de traitements (automatisés ou non). Ces traitements sont autant de processus (informatiques) : ils exigent des données en entrée et donnent en sortie leurs résultats.

En pratique, décomposer son processus de travail en enchaînements de modules « entrée-traitement-sortie » se traduit par exemple par quelques pratiques saines :

- isoler les paramètres immuables des traitements dans un fichier spécifique (variables d'environnement) ;
- éviter de mettre les statistiques calculées dans le code. On préférera éditer un fichier dédié ;
- identifier les tables en entrée, ce qui permettra de revenir facilement aux données brutes, par exemple pour modifier les imputations ;
- identifier les tables intermédiaires, ce qui facilitera les analyses explicatives (approfondir *a posteriori* un chiffre calculé, creuser des évolutions étonnantes relevées par les utilisateurs) et la recherche de *bugs*.

L'identification des différents traitements utilisés permettra de définir la **structure du programme**. Un traitement pourra lui-même être décomposé en sous-processus.

Pour définir une architecture qui favorise la lecture, l'évolution future et l'exécution du programme, il est important de découper le traitement global en un ensemble de sous-processus, qui communiquent les uns avec les autres, mais qui restent indépendants pour leurs fonctionnements internes : « *L'objectif principal de l'architecture est de soutenir le cycle de vie du système. Une bonne architecture rend le système facile à comprendre, facile à développer, facile à maintenir et facile à déployer. Le but ultime est de minimiser le coût du système pendant sa durée de vie et de maximiser la productivité des programmeurs* » (Martin, 2017).

On retrouve dans cette organisation la notion de « **barrière d'abstraction** » (Abelson, Sussman et Sussman, 1996) : pour utiliser une sous-partie du programme, il est uniquement nécessaire de savoir quelles données sont utilisées en entrée et quelles données seront produites en sortie, par contre il n'est pas nécessaire de savoir comment le traitement a été implémenté. Cette barrière permet qu'une partie d'un processus ne soit pas affectée par des modifications sans lien avec les traitements qu'il implémente. Les différentes parties peuvent alors être construites, remplacées et corrigées séparément. Une partie de programme construite suivant ce principe est appelé un **module**, et un processus construit ainsi est dit modulaire.

Une approche simple pour évaluer la modularité d'un traitement est de vérifier si le traitement est bien circonscrit (arrive-t-on à le nommer simplement ?) et si les données en entrée et en sortie sont bien identifiées. On dispose ainsi d'une « façade », un patron de conception, qui fait abstraction de la manière dont le traitement est réalisé : manuellement ou automatiquement ? Avec quel langage de programmation ?

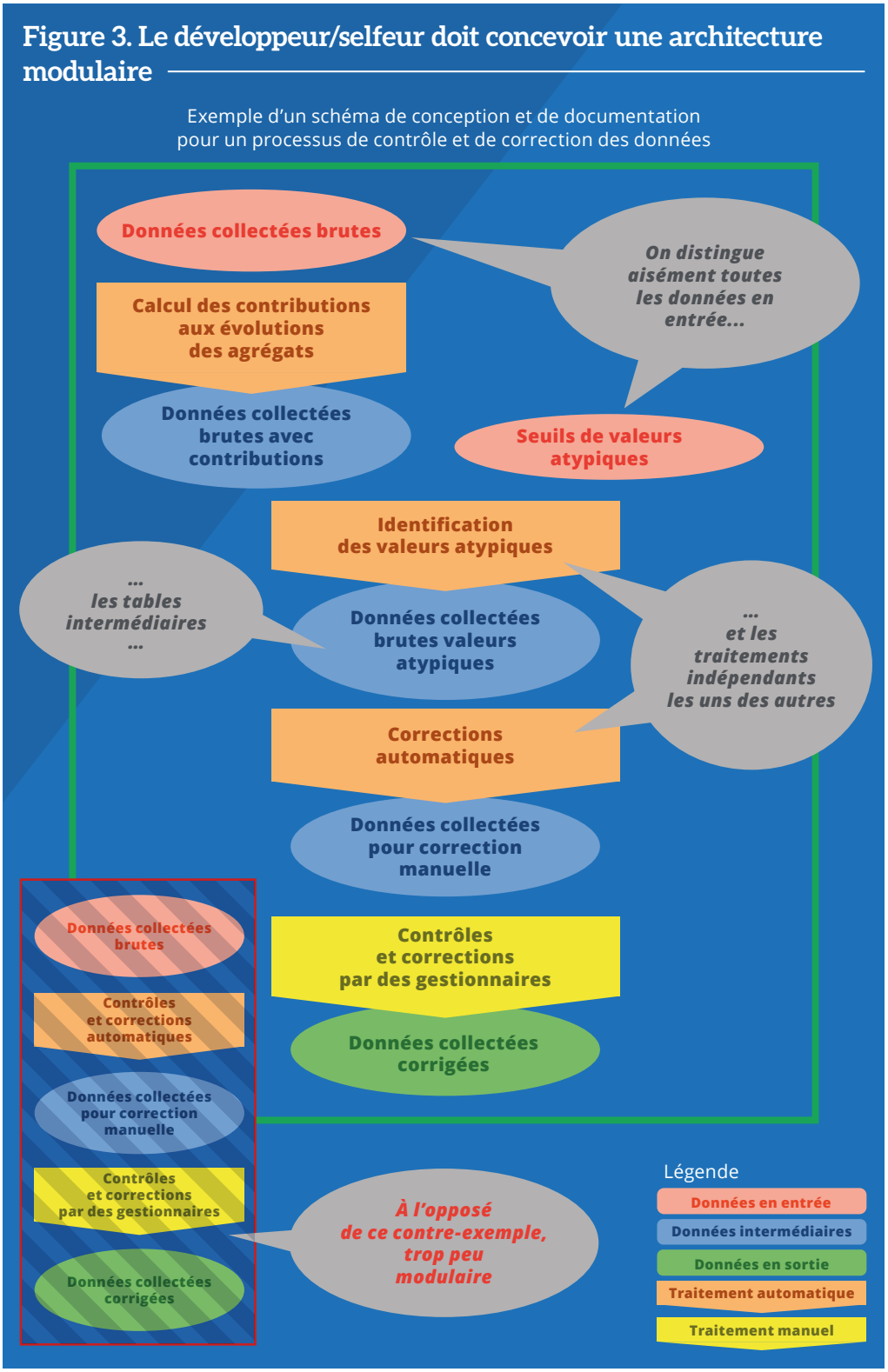
Grâce à la modularité, on se laisse la liberté de répondre à ce type de questions plus tard, et de pouvoir modifier ces réponses au cours du temps. Par exemple, pour passer des données individuelles aux données finales, une étape d'agrégation puis une étape de désaisonnalisation sont souvent nécessaires. Les identifier comme deux parties distinctes est important car cela permettra de faire évoluer séparément aussi bien les méthodologies que les outils informatiques utilisés[16].

C'est lors de l'exécution d'un programme qu'apparaît de la façon la plus flagrante l'intérêt d'un découpage en modules. Dans des projets informatiques, il est fréquent qu'un traitement complet nécessite plusieurs heures ou jours pour être réalisé de bout en bout. Si le traitement est implémenté sous la forme d'un unique programme monolithique, toute erreur en cours de route nécessitera de repartir à zéro. La seule façon efficace de corriger ce défaut est de découper le traitement en modules dont les durées d'exécution sont brèves. Si le découpage suit la logique décrite ci-dessus, alors il sera facile de corriger les *bugs* dans le code et de reprendre l'exécution des programmes à un point intermédiaire. La ***figure 3*** illustre ce que peut être la schématisation d'un processus. Ici, la modularité permet d'éviter certains écueils : tel qu'il est structuré, le programme ouvre la possibilité de connaître voire de modifier les paramètres des contrôles et des corrections automatiques, et d'exécuter séparément les calculs de contributions, les contrôles et les corrections automatiques.

16. Par exemple, l'agrégation pourra se faire avec le langage SQL alors que la désaisonnalisation utilisera un outil spécialisé comme JDemetra+.

Figure 3. Le développeur/selfeur doit concevoir une architecture modulaire

Exemple d'un schéma de conception et de documentation pour un processus de contrôle et de correction des données



ADOPTER UN STYLE D'ÉCRITURE

Écrire un programme, c'est automatiser des tâches qui devraient sinon être réalisées manuellement. Si un programme permet d'éviter les erreurs qui apparaissent régulièrement ou aléatoirement dans les opérations manuelles, il rend systématiques les erreurs qui résultent des défauts d'écriture.

Pour réduire le risque de *bug*, il faut s'assurer de la lisibilité du programme tout au long de l'écriture de celui-ci. Cela passe par une attention particulière au nommage et au style, ainsi que par la revue et le partage du code.

Quand on cherche à automatiser des traitements, les concepts métiers doivent être « traduits » en objets informatiques. Cette action a un impact direct sur la qualité d'un programme, en particulier sur sa lisibilité. Nommer correctement les données et variables permet de partager le programme avec d'autres ou de le comprendre lors d'une relecture future[17].

Le rédacteur du code, gagne, comme le rédacteur d'une étude, à travailler la simplicité de son style, à écrire pour le plus grand nombre. Il existe des styles de programmation qui se distinguent assez simplement par la manière de :

- nommer les variables (par exemple `camelCase`, majuscule pour la première lettre des mots, sauf pour le premier, et `snake_case`, mots en minuscules séparés par des tirets bas, etc.) ;
- indenter les lignes de code (généralement par deux, trois ou quatre espaces) ;
- ou d'utiliser les commentaires.

> *« Un style efficace est surtout constant, homogène d'un programme à un autre. »*

Un style efficace est surtout constant, homogène d'un programme à un autre[18]. Un langage informatique respecte un ensemble de règles formelles strictes qui réduisent fortement la liberté d'écriture du développeur. Pour autant, le développeur peut organiser le programme suffisamment à sa guise pour développer un véritable style. Et ce style comprendra toujours une dimension visuelle, contrairement à la plupart des styles littéraires. En effet l'utilisation des passages à la ligne et des espaces en début de ligne, les indentations, jouent un rôle primordial dans la mise en valeur des structures logiques.

Quand on écrit un code, il faut se relire et se faire relire. La revue de programme (*code review*) est une pratique courante de gestion de la qualité du code, en complément des tests. Tant la forme que le fond du code y sont mis à l'épreuve. S'il ne dispose pas de collègues coopératifs, le selfeur gagnera simplement à prendre de la distance avec son code, en le laissant de côté quelques jours, quelques heures, quelques instants, et à le relire ensuite avec un regard neuf. S'il peine à rentrer dedans, c'est sans doute qu'il doit rendre son code plus accessible.

17. Pour de bonnes pratiques de nommage voir par exemple (McConnell, 2005, chapitre 11 ; Martin, 2009, chapitre 2).
18. La plupart des outils de développement proposent des fonctions de mise en forme automatique. Ces fonctions permettent de respecter un style défini *a priori*, soit par la communauté informatique dans son ensemble, soit par une équipe projet. Cependant les éditeurs de logiciels statistiques sont actuellement plus pauvres de ce point de vue (SAS® et R notamment) et contraignent à suivre manuellement les règles de style choisies.

Par ailleurs, on peut coder à plusieurs. Cela peut paraître évident mais c'est assez peu pratiqué. Deux selfeurs en binôme (pratique désignée par le terme de *pair programming*) feront très souvent beaucoup plus ensemble que s'ils étaient restés chacun de leur côté (Hannay *et alii*, 2009 ; Kessler et Williams, 2002). De surcroît, cela revient à une revue de code en continu. C'est un formidable outil de formation et de tutorat pour apprendre à supprimer le code inutile et à éviter les répétitions.

Au-delà de l'équipe, il est possible de s'inscrire dans une communauté plus grande, *via* le partage de programmes en *open source*. Dans ce cadre, on développe *de facto* pour les autres, la lisibilité du code est donc primordiale. Il est indispensable de respecter les normes retenues par la communauté du logiciel auquel on contribue, ces normes étant dans la plupart des cas des standards. S'inscrire dans une démarche *open source* est un moyen très puissant de progresser dans l'appropriation des bonnes pratiques. La communauté fait bénéficier de retours d'expérience plus nombreux et plus variés, elle devient un moyen de se former par tutorats réciproques. Depuis la mise à disposition du *modèle Inès* sur son site internet, l'Insee diffuse de plus en plus de programmes en *open source* (notamment les modèles, comme *Avionic*, *Destinie*, *Mésange*, *Mélèze* et *Omphale*[19]). En cela, l'institut s'inscrit pleinement dans une pratique devenue courante parmi les *data scientists*.

Dans un contexte qui lui est donc favorable, le statisticien public selfeur qui a su définir la finalité de son développement, qui sait structurer convenablement son programme et qui saura vérifier avec l'aide d'un tiers la lisibilité de son code, a encore un dernier choix à faire.

CÔTÉ TECHNIQUE…

Le terme de « technique » est source d'un débat toujours renouvelé : doit-on maîtriser la technique pour savoir définir une cible et construire le chemin qui y mène ?

Le statisticien selfeur doit-il se transformer en informaticien pour écrire un programme dans les règles de l'art ? Si l'on y regarde de plus près, toutes les activités étudiées jusqu'à présent relèvent d'une technicité particulière, que ce soit définir comment mesurer un concept, estimer un modèle, concevoir un enchaînement de traitements, écrire un programme, le tester, etc. Ce qui se cache derrière cette dernière question c'est la réalité de l'importance des outils utilisés pour réaliser les objectifs fixés.

> *« Pas de « techniques miracles » pour diminuer la quantité de bugs dans les programmes. »*

Dans les années quatre-vingt, l'ingénieur logiciel américain Frederick P. Brooks (Brooks, 1986) a utilisé une expression restée fameuse : « *No Silver Bullet »,* autrement dit il n'y a « pas de baguette magique », ou pas de « techniques miracles » pour augmenter la productivité des programmeurs et diminuer la quantité de *bugs* dans les programmes. Brooks estime que les difficultés de réalisation des logiciels se divisent en difficultés accidentelles (langages de programmation et systèmes laborieux et malaisés à utiliser) et en difficultés essentielles (inhérentes à la production de logiciels). Or, selon lui, les difficultés accidentelles ont déjà été en grande partie éliminées, par exemple par l'adoption de langages de haut niveau ; il n'y aura donc pas dans l'avenir de nouveaux progrès techniques permettant des gains importants de productivité. Les gains (en délai de réalisation, en qualité) doivent donc d'abord être cherchés dans le travail de conception, et ensuite seulement, dans le choix des outils.

19. Les codes-sources de ces modèles sont disponibles à l'adresse suivante : *https://github.com/InseeFr*.

Tout en gardant à l'esprit leurs limites, parmi les nombreux choix techniques à faire (*web* ou pas, format de la base de données, algorithmie, méthodes *big data* ou classiques, etc.), certains se révèlent plus importants que d'autres.

LE CHOIX LE PLUS STRUCTURANT EST CELUI DU LANGAGE…

Il n'est pour autant pas nécessaire de chercher la perfection : le langage à la syntaxe la plus subtile, permettant les syntaxes les plus courtes pour implémenter les traitements les plus complexes, ou autorisant de surprenantes opérations est rarement celui qu'il faut retenir. Les langages à haut niveau d'abstraction sont nombreux, et souvent suffisamment puissants pour la nature des travaux à réaliser. Le choix du langage se fera plutôt en prenant en compte son écosystème : évolue-t-il régulièrement ? Sa documentation est-elle abondante et facilement disponible ? Est-il porté par une communauté d'utilisateurs dynamique ? Existe-t-il des librairies tierces permettant de ne pas avoir à tout redévelopper ? Peut-il facilement s'interfacer avec d'autres langages ? Existe-t-il une offre suffisante d'outils de développement et de tests pour ce langage ?

Aujourd'hui, pour un statisticien développant lui-même ses programmes, le choix se portera en premier lieu sur les langages *R* et *Python*. Ils bénéficient tous deux d'un écosystème de qualité, et répondront à la plus grande partie des besoins en statistique[20].

… DES BIBLIOTHÈQUES DE PROGRAMMES SANS BOGUES…

On s'appuie généralement sur des programmes écrits par d'autres et mis à disposition sous forme de bibliothèques (*libraries* en anglais) réutilisables. En premier lieu, il faut s'assurer qu'elles ne présentent pas de *bugs*[21]. Ce point est particulièrement sensible dès que l'on touche au domaine de la méthodologie. Les traitements à implémenter peuvent être complexes et les erreurs générées par une mauvaise implémentation peuvent avoir un impact considérable sur le résultat, tout en étant difficiles à détecter. D'autres critères sont (sans chercher à être exhaustif) la performance, la pérennité, la disponibilité d'une documentation, la facilité d'utilisation, éventuellement le coût de la licence d'utilisation. Afin de guider le selfeur vers les librairies donnant les meilleures garanties, des initiatives de certification ont vu le jour. L'Insee a créé début 2019 un Comité de certification des packages R afin d'accompagner l'utilisation de plus en plus étendue de ce langage. Toujours au sein de la communauté R, on peut citer l'initiative rOpenSci[22] qui s'inscrit dans la logique de la science reproductible (voir *infra*).

… ET DES ÉDITEURS STANDARDS (ÉVITER LES PROPRIÉTAIRES)

Les interfaces entre les modules sont un endroit stratégique où il faut faire les bons choix. Les modules d'un même traitement doivent rester indépendants d'un point de vue technologique, sinon leurs cycles de vie ne seront pas indépendants. Pour cela il faut que les technologies choisies (type de fichiers de données utilisés par exemple) introduisent

20. Une liste d'outils utilisés dans la statistique publique est disponible à l'adresse suivante : *https://github.com/SNStatComp/awesome-official-statistics-software*.
21. Une telle vérification ne va pas de soi. Le mieux est de pouvoir s'appuyer sur un tiers qui a la compétence pour le faire. À défaut il faut disposer de jeux de tests permettant de valider la librairie.
22. Voir le site *https://ropensci.org/*.

peu de contraintes. Le choix devra donc se porter sur des standards faciles à produire et à manipuler, par les machines mais si possible aussi par les humains. Concrètement on s'efforcera d'échanger les informations sous forme de fichier textes dans des formats standards reconnus et non propriétaires (par exemple des fichiers CSV, JSON ou XML). L'usage de fichiers Excel millésimés (97, 2003, etc.) est source d'erreurs en cas de changements de version. Le format des fichiers de données SAS® l'est également avec d'autres logiciels statistiques.

L'utilisation des tableurs pour faire de la statistique est un sujet qui ne fait pas consensus. Leurs défenseurs mettent en avant leur facilité d'utilisation, aussi bien pour manipuler la donnée que pour produire des graphiques. Leurs détracteurs, dont font partie les auteurs de cet article, conseillent de ne pas utiliser ces outils pour des ensembles de données qu'il n'est pas possible d'afficher en entier sur un écran. En effet, un tableur rend impossible d'appliquer le conseil qui veut qu'un programme soit fait pour être lu avant d'être exécuté. Dans un tableur tout est mélangé : les données en entrée, les règles de calcul, l'architecture générale du traitement et le résultat. Utiliser des tableurs expose à des risques importants (Powell, Baker et Lawson, 2009) : difficulté à comprendre les traitements implémentés pour celui qui ne les a pas écrits (ou qui les a oubliés), quasi-impossibilité de construire des traitements modulaires et à mettre en œuvre des tests, corruption des traitements lorsque l'on remplace par erreur la formule contenue dans une cellule par une valeur.

> *« Un tableur rend impossible d'appliquer le conseil qui veut qu'un programme soit fait pour être lu avant d'être exécuté. »*

Pour répondre aux exigences de performance, il sera parfois nécessaire de choisir des outils spécifiques. La portée de ce choix devra être restreinte au maximum et il ne devra être fait qu'après la conception du traitement, une erreur classique étant de construire le programme de façon à mettre en valeur toute la puissance de l'outil retenu. Les choix d'outils pour des raisons de performance sont les plus à même d'être caduques au fur et à mesure que les performances globales de l'informatique évoluent. Par exemple SAS® et R ont souvent été opposés sur la question de la capacité à traiter des fichiers de grandes tailles, R nécessitant de les charger en mémoire vive ou de mobiliser un serveur de base de données, SAS® étant capable de les traiter séquentiellement à partir du disque dur[23]. Mais avec l'évolution des capacités matérielles et des possibilités logicielles le débat a perdu de sa pertinence. L'opposition se focalise maintenant plus sur la capacité offerte par les deux logiciels de pouvoir partager ses programmes avec d'autres. À nouveau, un traitement modulaire facilitera le travail : des modules bien définis permettront de ne pas faire porter le choix de la technologie au-delà du périmètre pour lequel cela se justifie, et si un jour le problème de la performance ne justifie plus une technologie spécifique, le retour vers des solutions standards sera facilité.

23. Ainsi, dans un article publié par SAS®, il apparaît que le chargement des données en mémoire vive pratiqué par *R* ne lui permet pas de manipuler des jeux de données aussi volumineux que SAS® (Ames, Abbey et Thompson, 2013).

GÉRER (AUTOMATIQUEMENT) SES VERSIONS, VIRTUALISER SON ENVIRONNEMENT

Pour suivre les modifications apportées à un programme, on peut ajouter à son nom une date ou un numéro de version. Manuellement, cela devient très vite fastidieux et source d'erreur. Il existe des outils pour enregistrer les versions d'un programme, accéder à l'historique, examiner les différences entre plusieurs versions, développer plusieurs versions en parallèles, etc. L'outil de référence en la matière est *Git*. Il nécessite un endroit où déposer et partager l'ensemble des versions du programme (tels que github.com et gitlab.com) mais facilite d'autant le travail en équipe.

L'utilisation fréquente de librairies et logiciels tierces et l'évolution rapide des langages informatiques nécessitent de référencer également les versions des outils utilisés et non pas seulement les versions du programme. Un programme pourra ne pas fonctionner avec une version plus ancienne ou plus récente que celle utilisée lors de son développement. C'est ce qu'on appelle **la gestion des dépendances**[24].

Afin de faciliter l'utilisation d'un ensemble d'outils de développement informatique, des environnements complets prêts à l'emploi existent[25]. Les environnements de développement « statistique » se sont perfectionnés et diversifiés (Besse, Guillouet et Laurent, 2018). Ils améliorent l'ergonomie de développement, la reproductibilité des résultats et aussi l'apprentissage des langages de programmation. En particulier, les carnets de code (*notebooks*[26]) permettent d'intégrer le code exécutable (et modifiable) au sein d'un rapport ou d'une présentation.

Enfin, la virtualisation consiste à créer une représentation virtuelle, basée logicielle, d'un objet ou d'une ressource telle qu'un système d'exploitation, un serveur, un système de stockage ou un réseau. Ces ressources simulées ou émulées sont en tous points identiques à leur version physique. La virtualisation permet même d'isoler un environnement d'exécution (*containers*) pour un projet[27]. Le statisticien public selfeur dispose maintenant d'un tel environnement avec le SSPCloud[28].

TESTER TOUT AU LONG DU DÉVELOPPEMENT, ET MÊME APRÈS

La démarche exposée précédemment (exigences, architecture, développement) permet de maîtriser la qualité du traitement *a priori*, lors de la conception. Les tests permettent de la maîtriser *a posteriori*, en fonctionnement réel. Il en existe de plusieurs sortes ayant des objectifs divers.

Une première catégorie, les **tests fonctionnels**, a pour but de s'assurer que les traitements donnent bien les résultats attendus du point de vue du métier. Au niveau le plus fin, il s'agit de tester que chaque fonction du programme est bien implémentée, ce sont les tests unitaires[29]. On testera par exemple la création, à partir de valeurs prédéfinies, d'un relevé de prix dans la base de données. Au niveau du programme dans son ensemble, il peut

24. Pour le langage R, *Renv* fait référence actuellement.
25. En anglais, on parle d'IDE pour *Integrated Development Environment*.
26. Les *notebooks Jupyter* ont inspiré au-delà de Python, en particulier Rmarkdown pour R.
27. Par exemple avec d*ocker* : *https://www.docker.com/*.
28. Voir à ce sujet l'article de Frédéric Comte, Arnaud Degorre et Romain Lesur sur le SSPCloud, dans ce même numéro.
29. Sur les tests unitaires, se reporter par exemple à (Martin, 2009).

s'agir de scénario reproduisant les comportements types d'un utilisateur, ou bien de calculs utilisant des jeux de données de test de grande taille et présentant des cas de figure variés. On testera par exemple le calcul d'un indice de prix à partir de données détaillées pour lesquelles le résultat est connu. Lorsque l'exécution de tests de cette première catégorie est automatisée, cela permet de s'assurer qu'il n'y a pas de régression fonctionnelle lors de l'ajout de nouvelles fonctions. Par exemple pour les retraitements d'une enquête répétée dans le temps dont les méthodes d'imputation seraient modifiées.

> *« Lorsque l'exécution de tests de cette première catégorie est automatisée, cela permet de s'assurer qu'il n'y a pas de régression fonctionnelle lors de l'ajout de nouvelles fonctions. »*

Une deuxième catégorie de tests permet de s'assurer de la qualité des aspects non fonctionnels, principalement la **performance** et la **sécurité**. Pour les programmes qui connaîtront un grand nombre d'évolutions au cours de leur vie, il est préférable d'automatiser ce type de tests.

A priori, les tests fonctionnels intéresseront plus particulièrement le statisticien, les aspects non fonctionnels étant vus comme l'affaire des informaticiens. Cependant il existe des cas où la maîtrise des exigences non fonctionnelles est un enjeu du métier statistique, que ce soit pour des raisons immuables, comme le respect du secret statistique, ou pour des raisons conjoncturelles, comme le besoin actuel de pouvoir traiter en des temps acceptables des données massives. Cette notion de performance a différentes dimensions. Il peut s'agir du temps de traitement, du volume maximal de données pouvant être pris en charge par le programme, du nombre d'utilisateurs simultanés (par exemple pour un outil *web* interactif). Aucun de ces objectifs de performance n'est un absolu en soi. C'est lors de la définition des besoins que l'on doit préciser lesquels doivent être atteints.

La plupart des tests peuvent, et doivent, être menés tout au long de la réalisation du programme. De ce point de vue, les tests de performance présentent une particularité. Les modifications apportées à un programme afin d'en améliorer les performances ont parfois pour effet d'en dégrader la structure interne et la lisibilité. De telles modifications ne doivent donc être apportées que si les performances du programme d'origine ne permettent pas de réaliser le traitement dans des conditions acceptables. Ainsi, les tests de performance doivent intervenir vers la fin du processus d'écriture, lorsque sa conception aura été menée jusqu'au bout. Les méthodes pour améliorer la performance sont diverses et leur impact sur la qualité du programme varie grandement (McConnell, 2005, chapitre 26). Il faut éviter de dégrader la structure et la lisibilité du programme pour atteindre des objectifs de performance qui ne sont pas nécessaires aux utilisateurs.

LE STATISTICIEN SELFEUR: UN DÉVELOPPEUR « AGILE », AU SERVICE DE LA QUALITÉ STATISTIQUE

Les grands principes énoncés jusque-là doivent bien évidemment s'adapter à la complexité du développement en self, et surtout à l'enjeu pour l'utilisateur du code produit. Ainsi, parmi les grands courants qui structurent aujourd'hui les méthodes de développement, celui de l'agilité est probablement le plus intéressant pour le selfeur : l'agilité a pour objectif d'orienter les efforts vers ce qui a le plus de valeur pour l'utilisateur, en s'adaptant aux changements à moindre coût.

Qu'il soit son propre utilisateur, ou qu'il insère son code dans un processus statistique complexe, le statisticien selfeur gagnera à s'inspirer des outils de pilotage d'un projet agile, si ce n'est à la lettre, du moins dans l'esprit.

Il gagnera aussi à s'inspirer des enjeux de la reproductibilité des études, ou des exploitations statistiques : pour cela, il s'assurera de livrer le résultat dans un environnement technique qui permet à tout un chacun de le reproduire à l'identique et de façon automatisée.

In fine, ce qui importe, c'est que le programme développé serve à produire une statistique publique de qualité : répondant à un besoin, dans les règles de l'art (statistique et informatique), documentée, reproductible, etc. Quel que soit le critère que l'on s'efforcera de respecter, « savoir coder » contribue à démontrer qu'on « sait compter », et ne peut que renforcer la confiance dans la donnée produite.

BIBLIOGRAPHIE

ABELSON, Harold, SUSSMAN, Gerald Jay, et SUSSMAN, Julie, 1996. *Structure and Interpretation of Computer Programs.* Juillet 1996. The MIT Press. Deuxième édition. ISBN 978-0262011532.

AMES, Allison J., ABBEY, Ralph et THOMPSON, Wayne, 2013. *Big Data Analytics. Benchmarking SAS®, R, and Mahout.* [en ligne]. Actualisé le 6 mai 2013. SAS Institute Inc., Cary, NC. Technical Paper. [Consulté le 13 décembre 2021]. Disponible à l'adresse : *https://support.sas.com/resources/papers/Benchmark_R_Mahout_SAS.pdf*.

ANXIONNAZ, Isabelle et MAUREL, Françoise, 2021. Le Conseil national de l'information statistique – La qualité des statistiques passe aussi par la concertation. In : *Courrier des statistiques.* [en ligne]. 8 juillet 2021. Insee. N° N6, pp. 123-142. [Consulté le 13 décembre 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/5398693/courstat-6-art-7.pdf*.

BESSE, Philippe, GUILLOUET, Brendan et LAURENT, Béatrice, 2018. Wikistat 2.0 : Ressources pédagogiques pour l'Intelligence Artificielle. In : *Statistique et Enseignement.* [en ligne]. 6 novembre 2018. Société française de statistique (SFdS). Volume 9, pp. 43-61. [Consulté le 13 décembre 2021]. Disponible à l'adresse : *http://statistique-et-enseignement.fr/article/view/694*.

BONNANS, Dominique, 2019. RMéS, le référentiel de métadonnées statistiques de l'Insee. In : *Courrier des statistiques*. [en ligne]. 27 juin 2019. Insee. N° N2, pp. 46-57. [Consulté le 13 décembre 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/4168396/courstat-2-6.pdf*.

BROOKS, Frederick P., 1986. No Silver Bullet – Essence and Accident in Software Engineering. In : *KUGLER, H.-J., 1986. Proceedings of the IFIP Tenth World Computing Conference.* [en ligne]. Elsevier Science BV, Amsterdam, pp. 1069-1076. [Consulté le 13 décembre 2021]. Disponible à l'adresse : *http://worrydream.com/refs/Brooks-NoSilverBullet.pdf*.

COCKBURN, Alistair, 2000. *Writing Effective Use Cases*. Octobre 2000. Addison-Wesley. ISBN 978-0201702255.

CONSTANTINIDIS, Yves, 2018. *Expression des besoins pour le SI. Guide d'élaboration du cahier des charges*. 11 janvier 2018. Eyrolles. 4e édition. ISBN 978-2212675771.

COTIS, Jean-Philippe, TEMAM, Daniel, BENVENISTE, Corinne, ANGEL, Jean-William, DARRINÉ, Serge, ROUMIGUIÈRES, Eve et GÉLY, Alain, 2009. Savoir compter, savoir conter. In : *Courrier des statistiques*. [en ligne]. Décembre 2009. Insee. Hors Série. [Consulté le 13 décembre 2021]. *https://www.bnsp.insee.fr/ark:/12148/bc6p06xt18x*.

ERIKSON, Johan, 2020. Le modèle de processus statistique en Suède – Mise en œuvre, expériences et enseignements. In : *Courrier des statistiques.* [en ligne]. 29 juin 2020. Insee. N° N4, pp. 122-141. [Consulté le 13 décembre 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/4497085/courstat-4-8.pdf*.

EVANS, Eric, 2003. *Domain-Driven Design: Tackling Complexity in the Heart of Software.* 20 août 2003. Éditions Addison-Wesley. ISBN 978-0321125217.

GADOUCHE, Kamel, 2019. Le Centre d'accès sécurisé aux données (CASD), un service pour la data science et la recherche scientifique. In : *Courrier des statistiques.* [en ligne]. 19 décembre 2019. Insee. N° N3, pp. 76-92. [Consulté le 13 décembre 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/4254227/courstat-3-7.pdf*.

HANNAY, Jo E., DYBA, Tore, ARISHOLM, Erik et SJOBERG, Dag I. K., 2009. The Effectiveness of Pair Programming: A Meta-Analysis. In : *Information and Software Technology.* Juillet 2009. Elsevier. Volume 51, n° 7, pp. 1110-1122.

KESSLER, Robert et WILLIAMS, Laurie, 2002. *Pair programming illuminated.* 19 juillet 2002. Éditions Addison-Wesley. ISBN 978-0201745764.

L'HOUR, Emmanuel, LE SAOUT, Ronan et ROUPPERT, Benoît, 2016. *Savoir compter, savoir coder.* [en ligne]. Juin 2016. Insee. Document de travail, Méthodologie statistique, n° M2016/04. [Consulté le 13 décembre 2021]. Disponible à l'adresse : *https://www.bnsp.insee.fr/ark:/12148/bc6p06zrjbz*.

LANGLAIS, Pierre Carl et EPRIST, 2020. *La recherche en crise de reproductibilité ?* [en ligne]. Avril 2020. EPRIST Analyse I/IST n°30. [Consulté le 13 décembre 2021]. Disponible à l'adresse : *https://www.eprist.fr/wp-content/uploads/2020/04/EPRIST_I-IST_Recherche-en-crise-de-reproductibilite_Avril2020.pdf*.

MARTIN, Robert C., 2009. *Coder proprement.* Février 2009. Pearson France, collection Campuspress. ISBN 978-2744023279.

MARTIN, Robert C., 2017. *Clean Architecture: A Craftsman's Guide to Software Structure and Design.* Addison-Wesley. ISBN 978-0132911221.

MCCONNELL, Steve, 2005. *Tout sur le code – Pour concevoir du logiciel de qualité, dans tous les langages*. 14 février 2005. Microsoft Press. 2e édition. ISBN 978-2100487530.

POWELL, Stephen G., BAKER, Kenneth R. et LAWSON, Barry, 2009. Errors in Operational spreadsheets. In : *Journal of Organizational and End User Computing*. [en ligne]. Juillet-septembre 2009. pp. 24-36. [Consulté le 13 décembre 2021]. Disponible à l'adresse : *http://faculty.tuck.dartmouth.edu/images/uploads/faculty/serp/Errors.pdf*.

ROBERTSON, Suzanne et ROBERTSON, James, 2013. *Mastering the Requirements Process: Getting Requirements Right*. Éditions Addison-Wesley Professional. Troisième édition. ISBN 978-0321815743.

VOLLE, Michel, 2001a. Pour une esthétique de la sobriété. In : *site de Michel Volle*. [en ligne]. 24 mars 2001. [Consulté le 13 décembre 2021]. Disponible à l'adresse : *http://www.volle.com/opinion/sobriete.htm*.

VOLLE, Michel, 2001b. L'expression des besoins et le système d'information. In : *site de Michel Volle*. [en ligne]. 31 décembre 2001. [Consulté le 13 décembre 2021]. Disponible à l'adresse : *http://www.volle.com/travaux/besoins.htm*.

PATRIMOINE IMMOBILIER DES MÉNAGES

ENSEIGNEMENTS D'UNE EXPLOITATION DE SOURCES ADMINISTRATIVES EXHAUSTIVES

*Mathias André et Olivier Meslin**

Les effets redistributifs de la taxe foncière selon le niveau de vie sont méconnus, notamment en raison d'un manque de données adaptées. L'étude de ces effets a donc nécessité la constitution d'une base exhaustive sur le patrimoine immobilier des ménages à partir de multiples sources administratives.

Ce type de démarche est appelé à se développer pour le statisticien, qu'il évolue dans le monde universitaire ou dans celui de la statistique publique : la période récente a vu en effet se développer sensiblement l'accessibilité des données administratives, tout comme leur nombre et leur variété. Pour autant, leur exploitation s'avère souvent complexe et exige de résoudre de multiples questions qui tiennent aux caractéristiques mêmes des sources administratives : taille importante, formats variés, informations manquantes ou de fiabilité variable.

Ce projet de rapprochement de données administratives exhaustives, dont certaines accessibles depuis peu, s'est déroulé en quatre grandes phases, depuis la récupération des données jusqu'à la structuration de la base statistique. Ce faisant, il permet d'en tirer des enseignements plus généraux : ici, le statisticien n'est plus maître du processus de production de l'information qu'il mobilise. Il doit donc acquérir de nouveaux réflexes et affronter des enjeux renouvelés.

The redistributive effects of the property tax according to the standard of living are poorly understood, in particular because of a lack of appropriate data. The analysis of these effects therefore required the creation of an exhaustive database on household property assets, based on multiple administrative sources.

This type of approach is likely to develop for the statistician, whether he or she works in the academic or official statistics world: the recent period has seen a significant increase in the accessibility of administrative data, as well as their number and variety. However, their use is often complex and requires solving many questions due to the very characteristics of the files: large size, varied formats, missing information or information of variable reliability.

To reconcile exhaustive administrative data, some of which has only recently become available, the project was carried out in four major phases, from data recovery to the structuring of the statistical database. In so doing, it allows us to draw some more general lessons: here, the statistician is no longer in control of the process of producing the information he uses. He must therefore acquire new reflexes and face new challenges.

* Chargés d'études, département des Études économiques, Insee,
mathias.andre@insee.fr
olivier.meslin@insee.fr

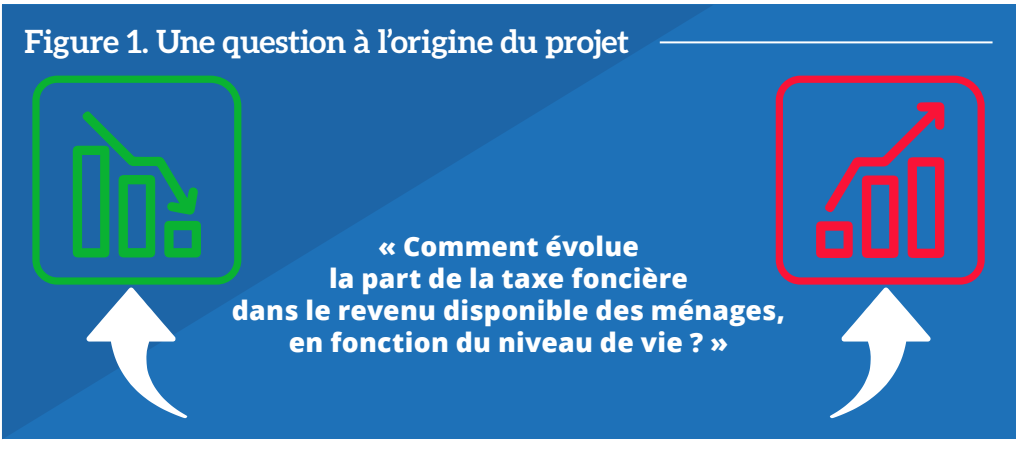
Deux questions sont à l'origine des travaux[1] présentés dans cet article. Une question économique d'abord ***(figure 1)*** : quel est le profil redistributif de la taxe foncière sur les locaux d'habitation ? Porte-t-il davantage sur les ménages aisés, médians ou modestes ? Peu d'études ont abordé cette thématique, à l'exception d'une publication récente (Carbonnier, 2019) qui s'intéresse aux résidences principales des ménages. Ces travaux peuvent également permettre de compléter l'étude du système socio-fiscal, comme le modèle Ines par exemple.

Une question statistique ensuite : est-il possible d'estimer précisément le patrimoine immobilier des ménages à partir de données administratives, de façon à compléter les données d'enquêtes ? En effet, les données des enquêtes de l'Insee ne permettent pas de répondre à toutes les questions, pourtant parfois centrales, notamment la répartition géographique fine du patrimoine immobilier, la concentration aux extrémités de la distribution ou le rôle des outils fiscaux comme les sociétés civiles immobilières (SCI).

Pour y répondre, il s'est avéré nécessaire de construire une nouvelle base statistique, regroupant des informations sur les ménages et l'ensemble des propriétés immobilières qu'ils possèdent, en s'appuyant sur des sources administratives. Une fois établi le lien entre les logements et les ménages qui en sont propriétaires, l'étape suivante consistera à estimer la valeur de marché de chaque logement, afin d'obtenir une mesure du patrimoine immobilier brut des ménages. Enfin, une fois le travail d'étude achevé se posera la question de la pérennisation de cette nouvelle source sous la forme d'une production statistique récurrente.

Mais le chemin est long pour passer d'une information administrative brute à un résultat statistique pertinent et robuste. C'est l'histoire de ce projet que cet article raconte. Ce travail de reconstitution des patrimoines immobiliers s'est heurté à de multiples défis qui touchent à la nature des données administratives : le statisticien qui exploite de telles sources se trouve constamment placé dans la situation paradoxale de chercher à répondre à des questions statistiques à l'aide de données qui n'ont pas été conçues pour cet usage. De ce point de vue, ce projet s'inscrit dans le cadre plus large de l'utilisation des données administratives. La démarche suivie est d'autant plus riche d'enseignements, qu'elle a abouti à une base de donnée exploitée par les divisions de production afin de déboucher sur une production d'indicateurs statistiques réguliers. Elle illustre en quelque sorte les embûches et les bonnes pratiques relatives à l'utilisation des données administratives par la statistique publique.

Figure 1. Une question à l'origine du projet



1. Les auteurs de l'article travaillent au département des Études économiques et ce projet a été mené en lien direct avec la division Logement de la direction des Statistiques démographiques et sociales, productrice du fichier Fidéli.

DONNÉES ADMINISTRATIVES: UN PAYSAGE QUI A RÉCEMMENT ÉVOLUÉ

L'histoire des statistiques, que ce soient les usages comme les méthodes, est intrinsèquement liée aux données à disposition des statisticiens (Rivière, 2020). Les sources mobilisées pour élaborer des statistiques et conduire des études sont d'une grande diversité :

- résultats d'enquêtes statistiques ;
- données administratives ;
- données issues des activités du secteur privé, comme les données de caisse de la grande distribution (Leclair, 2019), ou les données de téléphonie mobile (Cousin et Hillaireau, 2019) ;
- paramètres de barèmes législatifs dans les modèles de microsimulation comme c'est le cas pour le modèle Ines (Fredon et Sicsic, 2020), etc.

Des productions statistiques centrales comme le PIB, le taux de pauvreté ou l'inflation reposent d'ailleurs sur des combinaisons de ces différents types de sources.

L'usage des données administratives est ancien au sein de la statistique publique, et répond à des objectifs variés. Aujourd'hui, l'Insee produit ainsi des statistiques sur les revenus des ménages et sur les salaires en exploitant les données fiscales et les *DADS*[2]. Les données administratives servent également à constituer des bases de sondages pour le tirage des échantillons d'enquête : c'est le cas des fichiers de l'impôt sur le revenu et de la taxe d'habitation qui, une fois retraités et intégrés dans le fichier Fidéli, constitue la base de sondage pour les enquêtes ménages (Sillard *et alii*, 2020). Elles permettent en outre de compléter les données d'enquête voire de les remplacer : c'est par exemple le cas pour les enquêtes sur les Revenus fiscaux et sociaux (*ERFS*) ou sur les Ressources des jeunes (*ENRJ*), dans lesquelles les informations sur les revenus et les prestations sont recueillies grâce à un appariement avec des sources fiscales et sociales. Les enquêteurs se concentrent ainsi sur la collecte d'informations statistiquement pertinentes et *a priori* absentes des fichiers administratifs, comme les professions et catégories socioprofessionnelles (*PCS*) ou le statut d'activité.

UN ACCÈS AUX DONNÉES, PROGRESSIVEMENT ÉLARGI ET FACILITÉ

Cependant, avec l'augmentation des capacités informatiques et l'informatisation croissante des politiques publiques, la période récente se caractérise par une augmentation du nombre de bases administratives, mais aussi par une plus grande disponibilité de ces bases ; ce qui se traduit par une plus grande diversité d'usages. La diffusion de données administratives en *open data* a constitué une transformation majeure dans l'accès aux données publiques, notamment sous l'impulsion de la loi pour une République numérique[3]. Un nombre croissant de fichiers de l'administration sont ainsi mis à disposition des citoyens et des acteurs publics et privés[4],

> *« La diffusion de données administratives en* open data *a constitué une transformation majeure dans l'accès aux données publiques. »*

2. Déclaration annuelle de données sociales, voir par exemple (Lagarde, 2008).
3. Voir référence en fin d'article.
4. Il suffit pour s'en convaincre de consulter régulièrement la plateforme ouverte des données publiques françaises (*https://www.data.gouv.fr/fr/*) et de regarder le nombre de jeux de données proposés (près de 40 000 à la date de rédaction de l'article).

sans contrepartie. Cette diffusion se fait généralement par des formats ouverts comme le CSV (*comma separated value*), ou *via* des API (*application programming interface*). L'accessibilité croissante ne constitue cependant pas un gage de qualité : le statisticien, comme le citoyen, peut vite se trouver désorienté face à des données, certes nombreuses, mais non hiérarchisées.

En parallèle, pour les travaux de recherche, certaines sources soumises au secret statistique ou au secret fiscal ont été mises à disposition de façon plus systématique, notamment *via* le Centre d'accès sécurisé aux données (Gadouche, 2019). Les évolutions juridiques et techniques complètent ainsi les dispositifs d'échanges de données opérés depuis plusieurs années par l'entremise de conventions (par exemple, avec la direction générale des finances publiques (voir *infra*) ou la Banque de France). Elles accentuent, voire accélèrent un mouvement enclenché depuis 1986 (et l'article 7 bis de la loi de 1951) par lequel la statistique publique enrichit ses productions, ses études et ses publications à l'aide des données administratives : par exemple, la construction du panel « Tous salariés », l'exploitation de la déclaration sociale nominative (DSN[5]) comme celle dans un proche avenir du dispositif PASRAU[6] relèvent d'une longue tradition à l'Insee d'exploitation des données sociales.

Ainsi, en quelques années, nous sommes passés d'un monde où l'exploitation des données administratives était certes bien ancrée mais ciblée sur quelques utilisations, à un monde dans lequel le recours aux bases administratives se généralise et se diversifie.

UNE UTILISATION PAR LES STATISTICIENS, FACILITÉE PAR LES ÉVOLUTIONS TECHNIQUES

Au-delà de l'accessibilité croissante des données administratives, la grande nouveauté du point de vue du statisticien est apportée par les évolutions techniques qui accompagnent ce mouvement.

D'abord, il est désormais plus simple d'apparier ces sources entre elles et de les croiser longitudinalement. La présence d'identifiants individuels dans les bases facilite les rapprochements de fichiers, et ce malgré leur taille parfois importante[7]. Et lorsque les identifiants ne couvrent pas l'intégralité de la base, ou lorsqu'il n'y en a pas, le caractère exhaustif des fichiers permet quand même des appariements, dits « sur traits d'identité » (cf. *infra*).

Le statisticien a donc à portée de main des sources plus nombreuses, plus détaillées, plus accessibles, qu'il peut « facilement » croiser. Il peut donc s'en emparer plus aisément et il ne s'en prive pas : ainsi, en fédérant des informations administratives, de nouvelles bases statistiques sont créées pour de nouveaux usages, et répondant à de nouveaux enjeux.

5. La déclaration sociale nominative a constitué ces dernières années une simplification majeure des procédures déclaratives concernant les salaires et les revenus versés par un employeur. Voir (Humbert-Bottin, 2018).
6. Le dispositif PASRAU (Passage des revenus autres) prolonge la démarche de simplification et de rationalisation des déclarations sociales entamée avec la DSN, qu'il complète pour les « revenus de remplacement ».
7. À ce titre, les projets Résil (répertoire statistique individus et locaux d'habitation) et CSNS (code statistique non signifiant) menés à l'Insee illustrent le rôle central des appariements dans les travaux statistiques.

UNE OFFRE QUI RÉPOND À UNE DEMANDE (OU QUI LA GÉNÈRE ?)

Ce rôle croissant joué par les bases administratives s'inscrit par ailleurs dans un contexte où la statistique publique fait face à de nouvelles exigences, auxquelles les enquêtes ne répondent qu'imparfaitement. Par exemple, la taille usuelle des échantillons d'enquêtes ne permet pas de réaliser des analyses croisées à des niveaux géographiques fins, ou encore d'étudier précisément les extrémités de distributions concentrées, comme les revenus ou le patrimoine. Par ailleurs, les organismes de la statistique publique sont en recherche permanente de gains d'efficience : dans ce contexte augmenter la taille des échantillons ou la fréquence des enquêtes représenterait une charge excessive pour les ménages, et nécessiterait des moyens très conséquents.

> *« La demande sociale invite donc plus qu'auparavant à l'exploitation de données exhaustives. »*

La demande sociale, exprimée notamment au travers du Conseil national de l'information statistique (Anxionnaz et Maurel, 2021), invite donc plus qu'auparavant à l'exploitation de données exhaustives, entre autres parce que ces données sont plus accessibles. Mais elles présentent par ailleurs d'autres avantages.

Par exemple, les données administratives peuvent répondre à de nouveaux besoins, comme la publication d'informations plus rapides ou infra-annuelles. Au cours de l'année 2020, le contexte inédit de crise sanitaire a ainsi amené l'Insee à recourir à des nouvelles sources[8] afin de répondre à l'exigence de prévisions infra-trimestrielles. Les méthodes de prédiction pour le présent (*nowcasting*) peuvent s'appuyer sur des données d'enquête, comme c'est le cas avec le modèle Ines et les données ERFS. Mais le délai de recueil est *a priori* plus bref pour les bases administratives, et peut constituer un avantage pour produire des résultats statistiques de façon plus rapide.

Enfin, la possibilité de croiser les informations issues de différentes sources contribue à accroître la diversité des usages que la statistique publique peut en faire. Ce n'est pas aussi simple avec les données d'enquêtes auprès des ménages, dont la production se fait en « silos » parfois assez étanches. Les grandes enquêtes de l'Insee sont en effet thématiques, au sens où elles visent à recueillir des informations pertinentes sur un grand sujet : l'emploi, le patrimoine, le logement, etc. Si cette approche thématique permet de répondre précisément à de nombreuses questions dans le champ considéré et de limiter la charge de réponse des ménages enquêtés, elle présente néanmoins l'inconvénient de limiter le croisement des informations individuelles collectées entre différentes enquêtes.

De façon complémentaire, un rapprochement de sources administratives peut permettre de mener des travaux sur des thèmes variés[9]. C'est par exemple le cas des informations sur les ménages d'une part et des informations sur les entreprises d'autre part. Il serait précieux de pouvoir « mettre en transparence » les entreprises, c'est-à-dire de reconstituer le lien entre les entreprises et les ménages qui les possèdent. C'est notamment ce qui est fait pour les personnes morales que sont les SCI dans le projet dont la présentation détaillée suit ***(encadré 2)***.

8. Par exemple un panel anonymisé de clients de certaines banques, voir (Bonnet, Loisel et Olivia, 2021).
9. Par exemple les données fiscales et sociales pour la mesure des niveaux de vie.

APPARIER DE MULTIPLES SOURCES ADMINISTRATIVES : À LA CROISÉE DE DIFFÉRENTS MONDES

Afin de répondre aux deux questions qui ont été le point de départ de ce projet, de multiples sources administratives ont été mobilisées, avec un objectif : constituer une base de données exhaustive sur le patrimoine immobilier des ménages. La démarche n'a pas consisté à suivre une méthodologie préétablie, mais plutôt à construire une méthode pour répondre à des besoins et à des questions qui ont émergé à mesure que les travaux avançaient. Mener un tel projet s'est apparenté en pratique à une longue suite de problèmes à résoudre et de difficultés techniques à surmonter. Les paragraphes qui suivent décrivent les solutions retenues, en suivant un ordre chronologique et programmatique. *In fine*, ce projet aura connu quatre grandes phases ***(figure 2)*** et autant de défis à relever.

(PHASE 1) RÉCUPÉRER LES DONNÉES PERTINENTES

Le prérequis à la construction d'une base statistique nouvelle consiste à avoir connaissance de l'existence de données pertinentes et à rassembler celles qui seront finalement utiles à sa réalisation. Dans le cas présent, la colonne vertébrale du projet est constituée de données de revenus et de localisation des ménages, qui étaient déjà disponibles à l'Insee et bien documentées au sein des fichiers démographiques sur les logements et les individus (*Fidéli*)[10]. Les données cadastrales[11] sont la deuxième source essentielle du projet, indispensable pour reconstituer le lien entre les biens immobiliers et les ménages. Enfin, les autres sources que sont le registre national du commerce et des sociétés, les données sur les éléments d'imposition à la fiscalité directe locale (REI[12]) et les données sur les transactions immobilières (DVF[13]) ont été intégrées au projet au fur et à mesure, suite à des échanges avec des chercheurs et des statisticiens spécialistes du sujet ***(encadré 1)***.

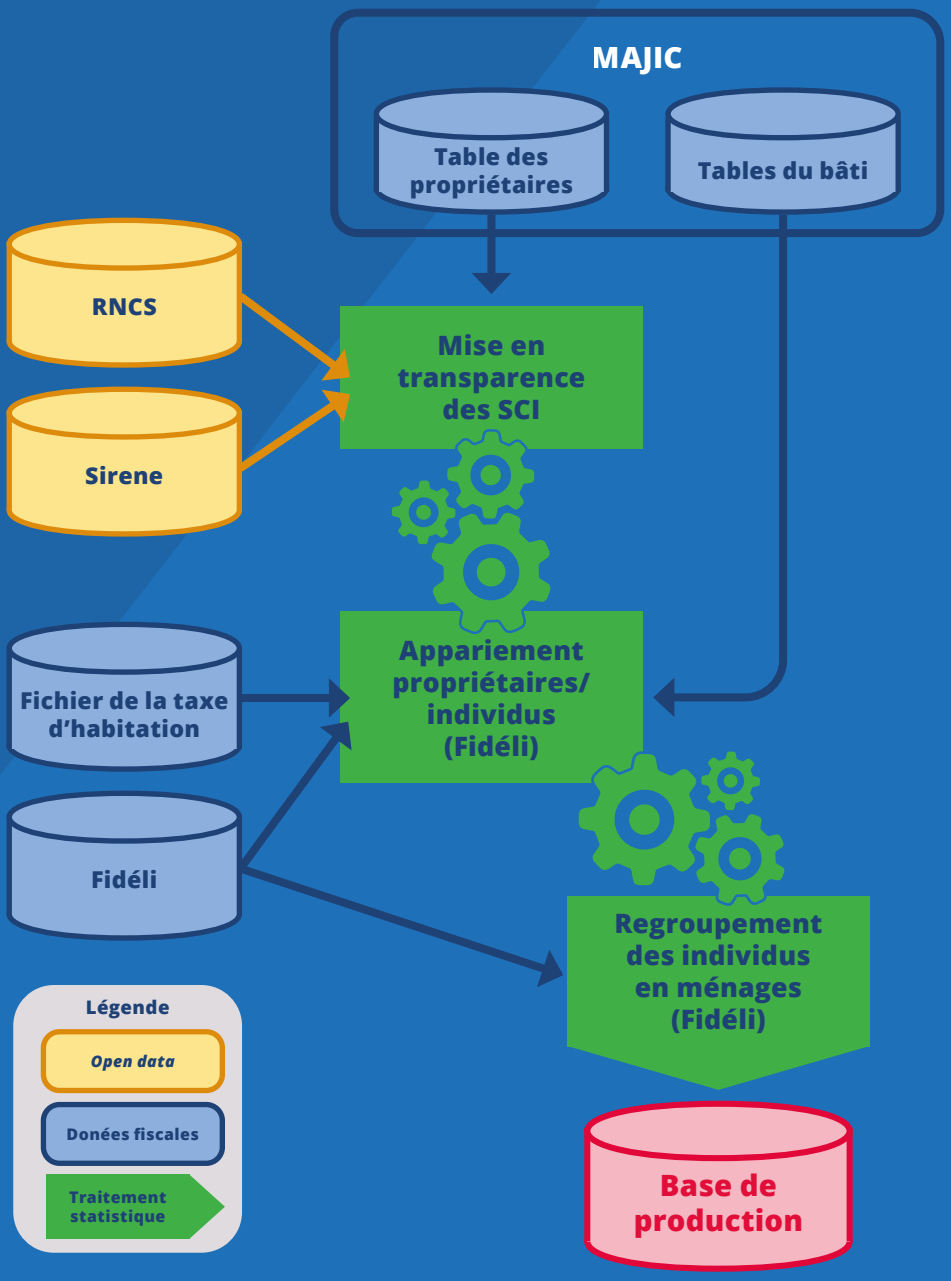
> *« Les difficultés ont plutôt consisté à avoir connaissance de l'existence de données potentiellement utiles. »*

S'il peut sembler aller de soi, ce travail de repérage et de centralisation des sources s'est avéré souvent complexe. En effet, il ne s'agit pas uniquement, ni même principalement, de suivre une procédure juridique précise auprès d'un guichet institutionnel déterminé. Les difficultés ont plutôt consisté à avoir connaissance de l'existence de données administratives potentiellement utiles au projet, puis à s'assurer qu'elles correspondaient bien aux besoins à partir de leur documentation, quand elle existe, pour enfin déterminer le meilleur moyen de les obtenir. Une hypothèse qui a guidé la recherche des sources était que les déclarations effectuées auprès de l'administration fiscale devaient « forcément » avoir été intégrées dans le système d'information : il suffisait donc de trouver la base centralisée correspondante.

Plus concrètement, cette recherche des sources a pris différentes formes, qui reflètent la diversité des situations rencontrées par le statisticien. Tout d'abord, pour les données déjà disponibles à l'Insee, comme les données cadastrales et les fichiers Fidéli, il a été nécessaire d'identifier l'unité disposant des données, puis de souligner la pertinence du projet pour pouvoir justifier la demande d'accès. De façon similaire, pour les données DVF dont l'Insee

10. Voir (Lamarche et Lollivier, 2021) pour plus d'informations sur Fidéli.
11. En l'occurrence, les fichiers *Majic* de la direction générale des finances publiques (DGFiP).
12. Recensement des éléments d'imposition à la fiscalité directe locale.
13. Demandes de valeurs foncières, source notariale et cadastrale.

Figure 2. Apparier, homogénéiser, unifier... pour créer une nouvelle source statistique



Fidéli : fichiers démographiques des logements et des individus
Majic : fichier des données cadastrales
RNCS : registre national du commerce et des sociétés
SCI : société civile immobilière
Sirene : répertoire administratif d'entreprises

ne disposait pas encore[14], la principale difficulté a consisté à identifier l'équipe dépositaire des données au sein de la direction générale des finances publiques (DGFiP), puis à établir une convention de transmission de données. Ensuite, c'est au cours d'une conférence universitaire que les chargés d'études ont appris que les données du registre national du commerce et des sociétés (RNCS) étaient mises à disposition par l'Institut national de la propriété intellectuelle (Inpi). Enfin, les données du recensement des éléments d'imposition (REI) relatives aux taux de taxe foncière votés par les collectivités locales sont disponibles sur le site de la DGFiP.

L'étape finale avant de passer aux traitements a consisté à préciser le cadre juridique du traitement des données personnelles, en conformité avec le Règlement général sur la protection des données[15]. Pour ce faire, il a fallu rédiger une déclaration de traitement et établir un dossier de conformité à la protection des données personnelles (DC-POD), avec l'aide de l'unité juridique de l'Insee.

(PHASE 2) RÉPONDRE AU DÉFI DU VOLUME ET DE L'HÉTÉROGÉNÉITÉ DES DONNÉES

Une fois les données brutes obtenues, le deuxième défi a été de mettre en place un environnement adapté à leur traitement. Les outils informatiques standards du statisticien ont été mis à l'épreuve et des échanges avec les responsables des infrastructures informatiques ont été nécessaires, en amont puis tout au long du processus. Tant la confidentialité des données que leur volume (300 Go de données brutes) ont imposé l'usage des serveurs sécurisés de l'Insee, avec un espace de stockage conséquent. Pour mener les traitements, le logiciel SAS® s'est d'abord imposé comme la solution offerte aux chargés d'études de l'Insee. Cependant au final, le projet combine l'utilisation de SAS® pour les grandes étapes de production statistique et le recours à R pour l'exploitation de chaque source.

Encadré 1. Les cinq sources du projet

- Les fichiers fonciers standards, dits fichiers *Majic*, constitués par la direction générale des Finances publiques à partir des informations du cadastre. Ils identifient les propriétés bâties et non bâties ainsi que leurs propriétaires (52 millions de locaux en 2017).
- Fidéli (fichiers démographiques sur les logements et les individus) décrit les logements et leurs occupants. Il est constitué par l'Insee à partir des données fiscales sur les individus (72 millions d'individus sur plusieurs années). Voir (Lamarche et Lollivier, 2021).
- Le registre national du commerce et des sociétés (RNCS), constitué par les greffes des tribunaux de commerce, contient des informations sur les sociétés et les personnes physiques qui les représentent (9 millions de représentants de 4,5 millions de sociétés).
- Les données DVF (demandes de valeurs foncières) décrivent les transactions immobilières (3,5 millions de transactions sur la période 2015-2019). Il s'agit d'une version enrichie par rapport à celle disponible en *open data* (elle contient notamment un identifiant cadastral).
- Les données REI (recensement des éléments d'imposition) décrivent la fiscalité locale au niveau de chaque commune (environ 36 000 d'observations en 2017).

14. L'Insee a récupéré les données en 2019 au cours du projet. Cette version est plus riche que celle disponible en *open data*, puisqu'elle contient en particulier un identifiant cadastral.
15. Voir référence en fin d'article.

La difficulté à traiter les données administratives ne tient pas seulement à leur volume, mais aussi aux formats dans lesquels elles sont fournies. En effet, elles peuvent prendre la forme de fichiers plats, parfois nombreux, souvent volumineux et difficiles à exploiter directement. Il a donc été nécessaire de les structurer dans un format adapté aux traitements envisagés. Par exemple, les données cadastrales se présentent sous la forme de 216 fichiers plats, recensant toutes les propriétés bâties situées en France ainsi que leurs propriétaires. La première étape du traitement de ces données a consisté à les structurer en sept fichiers nationaux homogènes sous forme de tables SAS®.

> *« La difficulté à traiter les données administratives ne tient pas qu'à leur volume, mais aussi aux formats sous lesquels elles sont fournies. »*

Cette structuration des données a également dû surmonter des difficultés liées à l'hétérogénéité des formats de données, au sein même de sources dont le contenu est censé être similaire. Ainsi, les données brutes du RNCS sont issues de fichiers plats (268 fichiers), à l'exception de celles portant sur l'Alsace-Moselle et l'outre-mer, qui sont disponibles en format XML pour des raisons historiques. Il a donc fallu distinguer deux traitements différents, de façon à reconstituer un fichier national unique et cohérent.

C'est à l'issue de cette phase que les données étaient organisées selon un format homogène ; il restait alors à en retraiter le contenu.

(PHASE 3) UNIFIER LES RÉFÉRENTIELS, LES DÉFINITIONS ET LES CONCEPTS (SI C'EST POSSIBLE)

Le troisième défi a porté sur le contenu des bases mobilisées : en effet, les informations qu'elles contiennent sont adaptées aux usages qu'en font les administrations, mais ne le sont pas toujours pour un usage statistique. Le statisticien doit donc procéder à de multiples vérifications et retraitements afin d'homogénéiser le plus possible les concepts portés par les données. Dans le cas d'espèce, les problèmes conceptuels (normalisation et définitions des variables, différences de champ, etc.) ont été plus fréquents que les enjeux de qualité.

Tout d'abord, les formats de données de même nature n'étaient pas toujours cohérents d'une source à l'autre. Il a donc été nécessaire de définir une procédure de normalisation visant à recoder ces variables dans un référentiel adapté au travail du statisticien. Ainsi, dans le RNCS, l'adresse des propriétaires de SCI figure dans un champ textuel non normalisé (« 12, allée des Acacias 06 000 Nice ») : il a fallu la retraiter pour codifier l'adresse et la commune de résidence à l'aide des nomenclatures usuelles[16]. De même, dans les données cadastrales, la commune de naissance des propriétaires est renseignée dans un champ textuel (« Vierzon 18 100 », « Paris 01 ») et a été normalisée selon le code officiel géographique (*COG*).

Par la suite, il s'est avéré indispensable de définir de nouvelles variables, ou de retraiter des nomenclatures existantes. Par exemple, la forme juridique du propriétaire dans les données cadastrales est codifiée selon une nomenclature détaillée en plusieurs centaines de catégories. Celle-ci a été retravaillée pour regrouper les propriétaires en trois grandes catégories : personnes physiques, SCI, autres personnes morales. De même, les données

16. Référentiel Fantoir (Fichier annuaire topographique initialisé réduit, anciennement fichier RIVOLI, géré par la direction générale des Finances publiques (DGFiP)), pour la voie, code officiel géographique (Insee) pour la commune.

du RNCS utilisées pour la mise en transparence des SCI ***(encadré 2)*** ne contenaient pas le genre des propriétaires, qu'il a fallu déduire à partir de leurs prénoms, en s'appuyant sur le *fichier des prénoms* publié par l'Insee. C'est encore à cette étape que la taxe foncière associée à chaque bien immobilier a été calculée, à partir des informations figurant dans le cadastre et des taux de taxe foncière votés par les collectivités locales, disponibles dans le REI.

La mise en transparence des sociétés civiles immobilières (SCI) s'est appuyée elle aussi sur différentes sources administratives. En effet, les données cadastrales ne contiennent pas les mêmes informations sur les individus, selon qu'ils possèdent un local en leur nom propre ou par l'intermédiaire d'une SCI : lorsqu'un bien immobilier est possédé *via* une SCI, ces données comportent uniquement la dénomination et l'identifiant Siren de la société, mais pas l'état civil des propriétaires de cette société. Compréhensible au regard des usages administratifs, cette différence entre personnes physiques et personnes morales est un obstacle pour le statisticien qui souhaite connaître les propriétaires finaux des biens immobiliers, indépendamment de l'intermédiation par les SCI. Le fichier des propriétaires a donc été apparié avec le registre national du commerce et des sociétés, de façon à obtenir l'état civil des propriétaires de SCI.

Enfin, les sources peuvent contenir des informations erronées ou obsolètes, qu'il importe de repérer et de redresser en amont des traitements statistiques. Par exemple, le fichier des propriétaires du cadastre contient plusieurs centaines de milliers de lignes correspondant à des individus décédés récemment (en raison des délais de mise à jour de ce fichier).

Encadré 2. Deux bonnes raisons de «mettre en transparence» les sociétés civiles immobilières (SCI)

La prise en compte des biens immobiliers possédés par l'intermédiaire des SCI revêt une importance particulière pour l'étude du patrimoine immobilier, pour deux raisons.

D'une part, le recours aux SCI est nettement plus fréquent chez les ménages possédant un nombre élevé de logements : 7 % des ménages propriétaires de 2 à 4 logements possèdent au moins un logement *via* une SCI, contre 31 % pour les ménages possédant 5 logements ou plus et 66 % pour ceux détenant 20 logements ou plus.

D'autre part, les SCI sont fréquemment utilisées pour partager la propriété de biens immobiliers entre plusieurs personnes physiques ou morales : 50 % des logements détenus *via* une SCI sont possédés par deux ménages ou plus, contre 13 % des logements détenus en nom propre.

Pour ces deux raisons, il est essentiel de prendre en compte les SCI pour mesurer correctement la concentration de la propriété immobilière et les phénomènes de copropriété. Or cette prise en compte s'avère complexe en raison même de l'intermédiation induite par le recours à la SCI. En effet, lorsqu'un bien immobilier est possédé par des personnes physiques au travers d'une SCI, les données cadastrales ne contiennent que le nom et l'adresse de la SCI (qui est le propriétaire légal du bien immobilier), mais pas l'identité des personnes physiques associées de cette SCI.

Si on veut étudier le patrimoine immobilier des ménages, il est donc nécessaire de mettre en transparence les SCI, c'est-à-dire de retrouver l'état-civil des personnes physiques associées des SCI. Cette opération est menée en rapprochant les données cadastrales du registre national du commerce et des sociétés, qui contient des informations sur les sociétés et les personnes physiques qui les représentent.

Ces enregistrements ont été repérés en rapprochant ce fichier du *fichier des personnes décédées* publié par l'Insee. De même, dans les données cadastrales, l'identifiant Siren et la forme juridique des personnes morales sont parfois erronés. Un travail de redressement de ces variables a été mené à l'aide du répertoire *Sirene*[17], notamment en vue de repérer précisément les sociétés civiles immobilières.

Cette phase a ainsi permis d'homogénéiser les données administratives tant dans leur format que dans leur contenu afin de préparer la dernière phase, celle qui fonde l'analyse statistique.

(PHASE 4) CHANGER D'UNITÉ D'ANALYSE POUR CRÉER UNE NOUVELLE SOURCE STATISTIQUE

Pour finir, il fallait restructurer les données pour en faire une base statistique à proprement parler. En effet, les données administratives sont construites en fonction des besoins de l'administration qui les collecte, et l'unité d'observation répond rarement au besoin du statisticien.

Dans le cas du patrimoine immobilier, l'unité de gestion administrative est le bien immobilier, également appelé local. Les données cadastrales sont donc structurées de telle façon que les informations sur chaque local puissent être mobilisées aisément : chaque local est repéré par un identifiant unique. Il est ainsi rapide de connaître la liste des propriétaires d'un bien donné, ou de calculer la taxe foncière due sur celui-ci. Inversement, les données cadastrales ne se prêtent pas aisément à une approche par individu : elles ne contiennent pas d'identifiant national unique des individus propriétaires, mais seulement leur état civil. Par conséquent, déterminer la liste des biens immobiliers dont un individu est propriétaire est une tâche qui s'avère complexe. Or, du point de vue du statisticien, reconstituer la propriété immobilière implique justement une approche dans laquelle l'unité pertinente n'est pas le bien immobilier, mais l'individu ou le ménage. Ce passage de l'unité de gestion administrative (ici, le local) à l'unité d'intérêt statistique (ici, le ménage) constitue le point nodal à partir duquel les sources administratives se transforment en sources statistiques. Ce changement d'unité d'analyse implique un important retraitement mené à l'aide de Fidéli, qui a constitué la source de référence sur les individus et les ménages.

> *« Ce passage de l'unité de gestion administrative à l'unité d'intérêt statistique constitue le point nodal à partir duquel les sources administratives se transforment en sources statistiques. »*

Il a d'abord été nécessaire d'ajouter dans le fichier des propriétaires un identifiant unique de chaque personne physique, de façon à déterminer la liste des locaux dont chaque individu est propriétaire. Pour ce faire, un appariement sur traits d'identité a été mené entre le fichier des propriétaires et la table des individus du répertoire Fidéli, qui identifie tous les individus majeurs connus dans les sources fiscales. Cet appariement a consisté à rechercher dans Fidéli un individu dont l'état civil et l'adresse sont identiques, ou très similaires, à des occurrences dans le fichier des propriétaires du cadastre. Il a permis d'identifier dans 94 % des cas, tous les propriétaires des locaux possédés par des personnes physiques, et dans 98 % des cas au moins l'un des propriétaires.

17. Système informatique pour le répertoire des entreprises et des établissements, répertoire administratif des entreprises géré par l'Insee.

Deuxièmement, les individus ont été regroupés en ménages, de façon à établir la liste des locaux dont chaque ménage est propriétaire. Ce regroupement a été opéré grâce au répertoire Fidéli, qui détermine la localisation des personnes : par convention, les individus partageant une même résidence principale appartiennent au même ménage.

C'est à l'issue de cette phase de changement d'unité d'analyse, et uniquement de celle-ci, que les données retraitées constituent (enfin) une source statistique. Il devient alors possible de construire des indicateurs statistiques et de mener des études. C'est notamment à ce moment que nous définissons le champ de l'étude sur la taxe foncière : les logements et dépendances situés sur le territoire national et possédés par des ménages résidents en pleine propriété ou en usufruit, soit en leur nom propre soit par l'intermédiaire d'une société civile immobilière. C'est grâce aux traitements menés au cours des quatre phases successives, et notamment grâce aux appariements entre sources, qu'il est possible de retenir finalement une définition aussi précise et couvrant un champ aussi large.

La nouvelle source ainsi créée permet de répondre aux questions initiales, et, par exemple, d'étudier les effets redistributifs de la taxe foncière ou la distribution de la propriété immobilière (André, Arnold et Meslin, 2021). De multiples exploitations sont envisageables et illustrent la richesse des travaux rendus possibles par l'exploitation des données administratives ***(figure 3)***.

QUELS ENJEUX SE DESSINENT À TRAVERS CE CAS D'USAGE DE DONNÉES ADMINISTRATIVES ?

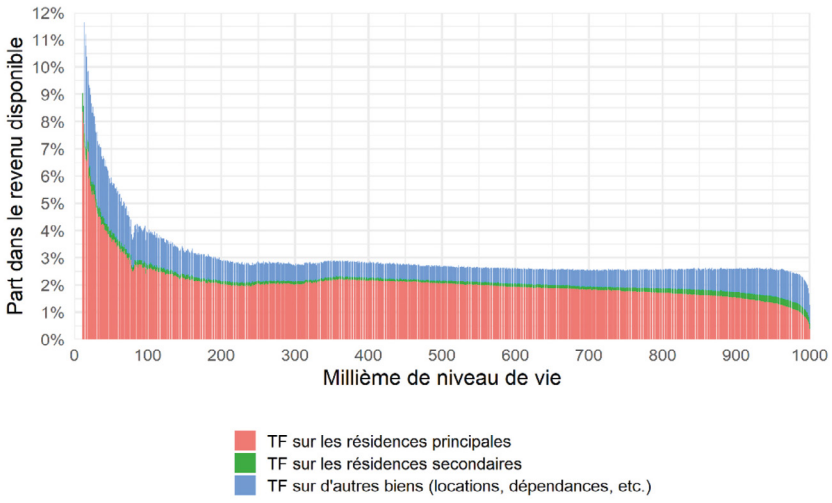
Forts de cette expérience, nous avons souhaité en tirer des enseignements pour le statisticien qui souhaiterait se lancer dans une entreprise similaire. Ces enseignements peuvent se résumer en trois caractéristiques distinctives des sources administratives, et quatre grandes questions que l'on doit se poser quand on les exploite à des fins statistiques.

Une source administrative présente, selon nous, trois caractéristiques distinctives :

- elle est **exhaustive… mais sur un certain champ**, dont la définition est pertinente au regard des objectifs poursuivis par l'administration qui la produit. Par exemple, le fichier de la taxe d'habitation comprend l'ensemble des foyers fiscaux assujettis à cet impôt, mais ne comprend pas les individus vivant en logement collectif non soumis à la taxe d'habitation ;
- son contenu reflète le travail de gestion accompli par les administrations et est le résultat de l'application de procédures administratives (déclaration des administrés, émissions d'avis d'imposition, etc.). Ce **contenu** est donc **en évolution constante** (création et suppression d'enregistrements, mise à jour d'informations) ;
- son contenu **n'est pas le produit d'un processus de collecte formalisé** au sens statistique. Par conséquent, des informations pertinentes pour l'analyse statistique peuvent être absentes des données administratives, et les métadonnées disponibles sur ces bases peuvent être incomplètes, voire inexistantes.

En définitive, les données administratives sont « subies » par le statisticien et non « construites » par lui. Ce constat nous conduit à proposer quatre sujets de réflexion préalable, indispensables pour réussir à construire une base à partir de ces sources particulières.

Figure 3. La nouvelle base répond à des questions jusque-là sans réponse

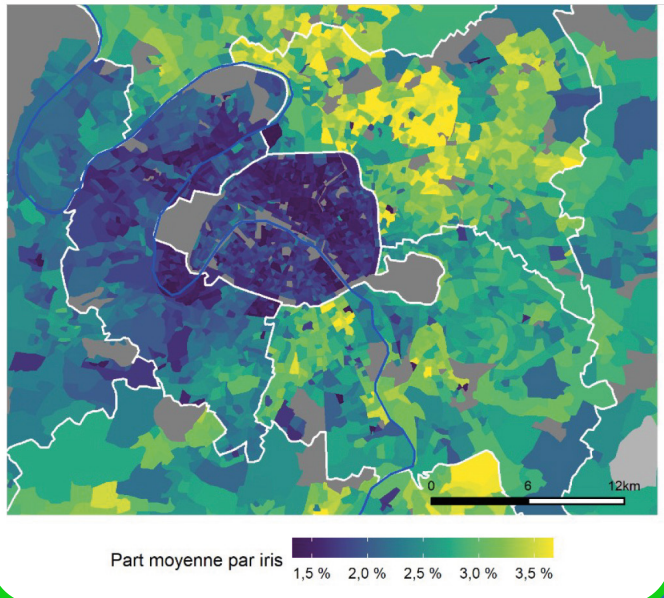


« Le graphique met en évidence que la part de la taxe foncière dans le revenu disponible des ménages imposables à cet impôt décroît en fonction du niveau de vie pour la première moitié de la distribution, puis est stable à environ 2,5 % du revenu disponible pour la seconde moitié de la distribution, à l'exception du top 5 % pour lequel elle décroît à nouveau. »

(André, Arnold et Meslin, 2021)

« Le taux d'effort de la taxe foncière, défini comme la part moyenne de la taxe foncière dans le revenu disponible des ménages imposables à cet impôt, varie notablement en fonction du lieu de résidence des ménages. [...] On] constate une différence nette entre les départements de la métropole parisienne : le taux d'effort moyen est le plus faible dans Paris et dans les Hauts-de-Seine (respectivement 1,6 % et 1,9 %), présente une valeur intermédiaire dans le Val-de-Marne (2,7 % en moyenne) et est la plus élevée en Seine-Saint-Denis (3,0 %). »

(André et Meslin, 2021)



À QUEL MOMENT OBSERVE-T-ON LES INFORMATIONS?

La première de ces questions touche à la nécessité de « récolter » les données à une certaine date, afin de pouvoir les exploiter : à quelle date figer les données ? Faut-il actualiser les sources, pour tenir compte du fait que l'administration peut mettre à jour ses données avec retard ?

Le statisticien est alors confronté aux difficultés d'exploitation des données « bi-datées », c'est-à-dire qui comprennent une date d'événement (par exemple un mariage) et une date de prise en compte dans les données administratives (l'enregistrement de ce mariage dans les fichiers). En effet, les bases administratives peuvent évoluer tous les jours, au gré de leur processus de récolte et d'actualisation. Certains fichiers peuvent être annuels mais avec des montants corrigés en plusieurs vagues, comme c'est le cas pour l'impôt sur le revenu ou les bases de la Cnaf[18], tenant compte des retards et indus uniquement quelques mois plus tard. En pratique, les millésimes de données administratives exploitées par le statisticien sont plus fréquemment définis par la date de traitement par l'administration que par la date des événements socio-économiques qu'elles décrivent : ainsi, le statisticien peut exploiter la base des déclarations de revenus à l'issue de la troisième émission de l'impôt sur le revenu (« POTE 2017 troisième émission »), le fichier à 6 mois pour les données sociales, l'extraction du registre national du commerce et des sociétés à la date du 6 mai 2017, etc.

DE QUELLE EXHAUSTIVITÉ PARLE-T-ON?

En second lieu, les bases administratives sont mises en avant pour l'exhaustivité de leurs informations. Pour autant, l'exhaustivité d'une source de données est une notion complexe à appréhender sur le plan statistique, car elle ne peut être définie que relativement au champ que cette source entend couvrir, qui est lui-même tributaire de l'usage qui en est fait : ainsi, certains fichiers sociaux sont exhaustifs sur les prestations ou minima sociaux, en ceci que tous les récipiendaires y figurent, mais une personne qui n'en perçoit pas n'apparaît pas dans une telle base. L'exhaustivité est donc relative au champ de la source administrative. Elle sera liée :

- aux limites de la base que l'administration fixe : France métropolitaine ? Ménages résidents ?
- au processus de recueil et à la finalité de gestion : revenus pour payer des impôts, carrières pour recevoir des retraites ou du chômage, etc.

Rappelons-le : les informations ne sont présentes que dans une finalité de gestion. Les définitions utilisées peuvent donc diverger des définitions statistiques usuelles. N'ayant pas d'objectif statistique, une administration ne collecte le plus souvent pas d'autres informations que celles dont elle a besoin dans ses missions : prélever des impôts, verser des prestations, etc. Ainsi, des informations essentielles pour l'analyse statistique peuvent être absentes. C'est par exemple le cas du revenu disponible qui est absent des sources fiscales. Par conséquent, une base administrative seule n'est pas toujours suffisante pour les besoins de la statistique, et il devient nécessaire d'apparier plusieurs bases avant d'obtenir une source rassemblant les informations pertinentes.

18. Caisse nationale des allocations familiales.

À cet égard, l'existence de référentiels communs à plusieurs bases, souvent indispensables au travail des administrations, peut faciliter le travail statistique de croisement des informations, en rendant possible des jointures exactes entre tout ou partie de bases différentes. C'est par exemple le cas de l'identifiant fiscal pour les différentes sources de la DGFiP ou le numéro d'inscription au répertoire ou numéro de sécurité sociale (*NIR*) pour la Cnaf. Il est également possible de mener des appariements sur traits d'identité en l'absence d'un référentiel commun, comme cela a été le cas dans le présent projet. Le processus de collecte d'une base administrative diffère fortement de celui d'une enquête statistique. C'est pour cette raison qu'un troisième point doit interpeller le statisticien qui les exploite : celui de la fiabilité des informations.

QUELLE EST LA FIABILITÉ DES INFORMATIONS ADMINISTRATIVES?

Une différence essentielle entre les sources administratives et les données d'enquête est que les premières ne sont pas pondérées. Chaque unité de gestion (ménage ou entreprise par exemple) ne représente qu'elle-même et correspond en quelque sorte à la « vraie » observation, et non plus à un estimateur d'une population par un échantillon représentatif. Dès lors, la question de la représentativité ne porte plus tant sur la correction de la non-réponse que sur le champ de la source administrative (cf. *supra*) et sur la fiabilité des informations qu'elle contient.

Or, la fiabilité d'une variable dans une source administrative dépend souvent de l'importance de cette information dans le processus de gestion : une variable de revenu sur laquelle s'appuie le calcul d'un impôt sera généralement très fiable car susceptible de recours de la part du contribuable et recueillie et vérifiée avec attention par l'administration. De même, dans les données cadastrales, l'adresse du propriétaire auquel l'avis de taxe foncière est envoyé est probablement plus fiable que l'adresse de ses éventuels copropriétaires, car seule cette adresse a un effet direct sur le recouvrement de l'impôt foncier. Inversement, une variable de moindre importance, comme l'âge d'une personne ou l'état général d'un bien immobilier peut s'avérer de qualité moindre, ou être moins souvent à jour, en raison de son importance réduite dans le processus de gestion. De la même manière, des variables déclaratives – comme pour le patrimoine immobilier dans l'impôt sur la fortune immobilière (IFI) – ou pré-remplies – le revenu déclaré par l'employeur – présenteront des degrés d'exactitude différents.

Ce travail d'expertise de la fiabilité implique des questions essentielles pour le statisticien : quelle est la définition administrative de cette variable ? Qui fournit l'information ? Est-elle à jour ? L'administration en a-t-elle contrôlé la qualité ? Y répondre requiert une compréhension fine du processus de gestion qui est à l'origine des données et un dialogue fréquent avec les administrations productrices de données. Enfin, une dernière question va se poser au statisticien qui rapproche ou s'appuie sur des bases administratives.

QUELS SONT LE CHAMP ET L'UNITÉ D'ANALYSE PERTINENTS POUR L'EXPLOITATION STATISTIQUE ?

Il s'agit là de l'étape à laquelle le statisticien n'échappe jamais, car c'est en répondant à cette question qu'il transforme les bases administratives en bases à usage proprement statistique. Dans un processus d'enquête, cette réflexion se situe en amont, au moment de la construction du questionnaire et de la constitution de l'échantillon[19]. Pour l'exploitation des sources administratives, elle est présente à chaque étape, tant au moment de la recherche des sources qu'en filigrane à toutes les étapes de production de la base finale. Dans le présent projet, il a fallu notamment répondre aux interrogations suivantes :

- quels types de biens immobiliers étudier ? Uniquement les logements, ou aussi les dépendances (garages, caves, etc.), voire les locaux industriels et commerciaux ?
- quelles modalités de détention retenir ? Uniquement les biens possédés en nom propre, ou également ceux possédés *via* une société (SCI, SA, SARL, SAS, etc.[20]) ?
- quels droits de propriété considérer ? Uniquement la pleine-propriété ? Que faire des usufruitiers et des nus-propriétaires[21] ?
- quelle unité de patrimoine immobilier étudier ? Au niveau des individus, des foyers fiscaux ou des ménages ?

Ainsi, les différentes « nuances d'exhaustivité », l'absence de production à visée statistique et les différences de définitions des unités d'observation portent à souligner l'importance de mener avec précaution les appariements entre les données d'enquête et les bases administratives ou entre bases administratives[22].

QUELS ENSEIGNEMENTS RETENIR DU PROJET ?

Trois principaux enseignements peuvent être tirés de ces travaux :

- le simple fait d'**avoir accès aux sources administratives n'est en aucun cas suffisant pour en faire des exploitations statistiques rigoureuses.** À partir d'une idée initiale, de nombreuses étapes sont nécessaires pour passer de bases initiales, riches mais hétérogènes et éparses, à une base statistique cohérente. Comme les autres travaux statistiques, les différentes difficultés qu'il a fallu surmonter ont correspondu à un cheminement vers toujours plus d'homogénéité, car les données administratives peuvent adopter des conventions, des définitions et des nomenclatures distinctes. Il revient alors au statisticien de les rapprocher et de les harmoniser avec patience et rigueur ;
- **l'exploitation systématique de données administratives correspond bien à un mode de collecte à part entière pour la statistique publique.** Celui-ci induit néanmoins un changement majeur par rapport aux autres modes : le statisticien ne maîtrise plus la production des données. Cette évolution implique le développement d'une expertise propre : comprendre le fonctionnement des administrations qui produisent les données, échanger avec les producteurs, anticiper les changements. Mais cela souligne également

19. Dans le modèle générique de processus de production statistique (GSBPM), ces aspects sont traités dans les phases de conception, au tout début du process.
20. Société civile immobilière, société anonyme, société à responsabilité limitée ou société par actions simplifiée.
21. La nue-propriété est le droit donnant à son titulaire, appelé nu-propriétaire, la faculté de disposer d'une chose mobilière ou immobilière (en la vendant, la donnant, la léguant, etc.) alors que l'usufruitier dispose seulement du droit d'en avoir l'usage.
22. Voir également l'expérience d'appariement relatée dans (Midy, 2021).

les limites imposées par la nature même de ces données, auxquelles il peut manquer des variables importantes pour l'étude, ou dont le contenu peut être d'une fiabilité relative et fonction de son importance dans les activités quotidiennes des administrations ;

- l'exploitation des données administratives ouvre des perspectives nouvelles à la statistique publique pour au moins deux raisons. D'une part, **l'exhaustivité rend possible l'étude de phénomènes rares**, tels que les patrimoines immobiliers les plus importants ou le croisement fin de plusieurs variables, **et la réalisation d'analyses à un niveau géographique très fin**, ce qui est difficilement réalisable avec des enquêtes. D'autre part, **des productions innovantes deviennent envisageables grâce aux appariements :** l'appariement entre les données cadastrales et le répertoire Fidéli permet de reconstituer la distribution conjointe des revenus et de la propriété immobilière, qui n'existait dans aucune source avec ce niveau de détail. De même, la mise en transparence des SCI permet de dépasser la distinction habituelle entre données sur les ménages et données sur les entreprises et donc d'étudier des phénomènes économiques peu documentés, comme le comportement de recours aux SCI en fonction de la composition du patrimoine et du niveau de revenu des ménages. Ces travaux pourraient aboutir à l'introduction d'un module sur les patrimoines immobiliers dans les fichiers Fidéli.

L'exploitation des données administratives est à la fois un défi et une opportunité. Un défi en ceci qu'elle exige des statisticiens qu'ils acquièrent de nouvelles compétences et renouvellent leurs méthodes. Une opportunité parce que la statistique publique élargit ainsi les informations qu'elle mobilise, source d'une grande richesse pour accomplir sa mission : mesurer et comprendre les phénomènes économiques et sociaux.

BIBLIOGRAPHIE

ANDRÉ, Mathias, ARNOLD, Céline et MESLIN, Olivier, 2021. 24 % des ménages détiennent 68 % des logements possédés par des particuliers. In : *France, Portrait Social.* [en ligne]. 25 novembre 2021. Insee références, édition 2021, pp. 91-104. [Consulté le 15 décembre 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/5432517?sommaire=5435421*.

ANXIONNAZ, Isabelle et MAUREL, Françoise, 2021. Le Conseil national de l'information statistique – La qualité des statistiques passe aussi par la concertation. In : *Courrier des statistiques.* [en ligne]. 8 juillet 2021. Insee. N° N6, pp. 123-142. [Consulté le 15 décembre 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/5398693/courstat-6-art-7.pdf*.

BONNET, Odran, LOISEL, Tristan et OLIVIA, Tom, 2021. *Impact de la crise sanitaire sur un panel anonymisé de clients de La Banque Postale. Les revenus de la plupart des clients ont été affectés de manière limitée et temporaire.* [en ligne]. 3 novembre 2021. Insee Analyses n° 69. [Consulté le 15 décembre 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/5760458*.

CARBONNIER, Clément, 2019. L'impact distributif de la fiscalité locale sur les ménages en France. In : *Économie et Statistique / Economics and Statistics.* [en ligne]. 11 juillet 2019. Insee. N° 507-508, pp. 31–52. [Consulté le 15 décembre 2021]. Disponible à l'adresse : *https://doi.org/10.24187/ecostat.2019.507d.1977*.

COUSIN, Guillaume et HILLAIREAU, Fabrice, 2019. Les données de téléphonie mobile peuvent-elles améliorer la mesure du tourisme international en France ? In : *Économie et Statistique / Economics and Statistics.* [en ligne]. 11 avril 2019. Insee. N° 505-506, pp. 89–107. [Consulté le 15 décembre 2021]. Disponible à l'adresse : *https://www.insee.fr/en/statistiques/fichier/3706269/ES_505-506_EN.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. 42-61. [Consulté le 15 décembre 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/4497070/courstat-4-4.pdf*.

GADOUCHE, Kamel, 2019. Le Centre d'accès sécurisé aux données (CASD), un service pour la data science et la recherche scientifique. In : *Courrier des statistiques.* [en ligne]. 19 décembre 2019. Insee. N° N3, pp. 76-92. [Consulté le 15 décembre 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/4254227/courstat-3-7.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 15 décembre 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/3647025/courstat-1-6.pdf*.

LAGARDE, Sylvie, 2008. La nouvelle exploitation exhaustive des DADS. In : *Courrier des statistiques*. [en ligne]. Juin 2008. N° 85-86, pp. 65-69. [Consulté le 15 décembre 2021]. Disponible à l'adresse : *https://gallica.bnf.fr/ark:/12148/bc6p06z99f7/f1.pdf*.

LAMARCHE, Pierre et LOLLIVIER, Stéfan, 2021. Fidéli, l'intégration des sources fiscales dans les données sociales. In : *Courrier des statistiques.* [en ligne]. 8 juillet 2021. Insee. N° N6, pp. 28-46. [Consulté le 15 décembre 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/5398683/courstat-6-art-2.pdf*.

LECLAIR, Marie, 2019. Utiliser les données de caisses pour le calcul de l'indice des prix à la consommation. In : *Courrier des statistiques.* [en ligne]. 19 décembre 2019. Insee. N° N3, pp. 61-75. [Consulté le 15 décembre 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/4254225/courstat-3-6.pdf*.

MIDY, Loïc, 2021. Un outil d'appariement sur identifiants indirects : l'exemple sur le système d'information des jeunes. In : *Courrier des statistiques.* [en ligne]. 8 juillet 2021. Insee. N° N6, pp. 82-99. [Consulté le 15 décembre 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/5398689/courstat-6-art-5.pdf*.

RIVIÈRE, Pascal, 2020. Qu'est-ce qu'une donnée ? Impact des données externes sur la statistique publique. In : *Courrier des statistiques.* [en ligne]. 31 décembre 2020. Insee. N° N5, pp. 114-131. [Consulté le 15 décembre 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/5008707/courstat-5-8.pdf*.

SILLARD, Patrick, FAIVRE, Sébastien, PALIOD, Nicolas et VINCENT, Ludovic, 2020. Pour les enquêtes auprès des ménages, l'Insee rénove ses échantillons. In : *Courrier des statistiques*. [en ligne]. 29 juin 2020. Insee. N° N4, pp. 81-100. [Consulté le 15 décembre 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/4497081/courstat-4-6.pdf*.

FONDEMENTS JURIDIQUES

Loi n° 2016-1321 du 7 octobre 2016 pour une République numérique. In : *site de Légifrance*. [en ligne]. Modifié le 23 août 2017. [Consulté le 15 décembre 2021]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/dossierlegislatif/JORFDOLE000031589829/*.

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. In : *site EUR-Lex*. [en ligne]. [Consulté le 15 décembre 2021]. Disponible à l'adresse : *https://eur-lex.europa.eu/legal-content/FR/TXT/HTML/?uri=CELEX:32016R0679&from=FR*.

L'ÉVALUATION DES COMPÉTENCES DES ÉLÈVES : UN PROCESSUS DE MESURE SINGULIER

*Thierry Rocher**

Les évaluations standardisées des compétences élèves se sont aujourd'hui imposées dans le débat public sur l'éducation. Leurs résultats sont utilisés à la fois pour éclairer les politiques éducatives mais également pour construire des indicateurs statistiques qui alimentent des outils de suivi mobilisés par les acteurs du système.

La mise au point de ces instruments suit une démarche assez singulière, du fait que l'objet de mesure – la compétence – est lui-même une construction, qui ne pré-existe pas à l'opération de mesure elle-même. Il en découle des procédures et des modélisations spécifiques, relevant de la psychométrie, domaine assez méconnu en France, alors qu'il y trouve une part importante de ses origines.

Cet article vise à donner un tour d'horizon des méthodes utilisées pour mesurer les compétences des élèves, en insistant sur leurs spécificités et en pointant quelques perspectives enrichissantes.

Standardised assessments of students' competences have now become an integral part of the public debate on education. Their results are used both to inform educational policies and to construct statistical indicators that feed into the monitoring tools used by educational stakeholders.

Building these instruments follows a rather singular approach, in that the object of measurement – the competence – is itself a construction, which does not pre-exist the measurement operation itself. This results in specific procedures and modelling, which are part of psychometrics, a field that is relatively unknown in France, even though it has a large part of its origins there.

This article aims to provide an overview of the methods used to measure students' competences, emphasising their specific features and pointing out some enriching perspectives.

* Adjoint au sous-directeur des Évaluations et de la performance scolaire, Depp, *thierry.rocher@education.gouv.fr*

MESURE DES COMPÉTENCES : QUAND L'INSTRUMENT FAIT NAÎTRE SON OBJET

Quel est le niveau des élèves ? Comment évolue-t-il ? Comment les élèves français se situent-ils par rapport à leurs camarades d'autres pays ? Au-delà de l'école, quel est le niveau d'adéquation entre compétence et emploi ? Ces questions font, et c'est bien naturel, l'objet de beaucoup d'attentions. Elles engendrent commentaires et débats, tant médiatiques qu'entre professionnels et spécialistes, pas toujours exempts de préjugés[1]. Elles appellent donc des réponses objectives, statistiquement fondées, et constituent ainsi un enjeu pour la statistique publique.

Mais qu'est-ce qu'un niveau de compétence, comment le mesurer ?

De multiples aspects du processus de mesure sont relativement classiques dans le champ des enquêtes statistiques. Cependant, la nature même de la variable mesurée distingue de façon très singulière ces évaluations, car les compétences ne s'observent pas directement. Seules les manifestations des compétences sont observables. Ce seront par exemple les résultats obtenus à un test standardisé : l'existence supposée de la compétence visée est alors matérialisée dans la réussite au test – plus précisément dans l'agrégation des résultats de chaque *item* du test.

> *« Les compétences ne s'observent pas directement. Seules les manifestations des compétences sont observables. »*

D'une certaine manière, on pourrait avancer que c'est l'opération de mesure elle-même qui définit concrètement l'objet de la mesure. D'ailleurs, dans le domaine de la psychométrie, qui s'intéresse à la mesure de dimensions psychologiques en général, c'est le terme de *« construit »* qui est employé pour désigner l'objet de la mesure : l'intelligence logique, la lecture, la mémoire de travail, etc. D'où le célèbre pied de nez attribué à Alfred Binet[2], l'inventeur de la discipline en France au début du XX^{e} siècle : à la question « *Qu'est-ce que l'intelligence ?* », il aurait répondu « *C'est ce que mesure mon test.* ».

Bien entendu, une majorité de statistiques peut être considérée comme une construction, basée sur des conventions[3]. Cependant, des distinctions avec celles ayant trait à l'évaluation de compétences peuvent être opérées, en lien avec le caractère tangible de la variable visée.

Par exemple, la réussite scolaire peut être appréhendée par la « réussite au baccalauréat » qui est mesurable directement, car elle est sanctionnée par un diplôme, donnant lieu à un acte administratif que l'on peut comptabiliser (Evain, 2020). Le « décrochage scolaire », quant à lui, est un concept qui doit reposer sur une définition précise, choisie parmi un ensemble de définitions possibles : ce choix conventionnel fait acte de construction. Une fois la définition établie, le calcul repose le plus souvent sur l'observation de variables administratives existantes, telles que la non ré-inscription dans un établissement scolaire.

1. Les débats suscités par la couverture médiatique des résultats des enquêtes PISA (*Programme international pour le suivi des acquis des élèves*) de l'OCDE en sont l'illustration la plus flagrante.
2. Alfred Binet (1857-1911) est un pédagogue et psychologue français. Il est connu pour sa contribution essentielle à la psychométrie.
3. *« On peut présenter la sociologie de la quantification comme perpétuellement tendue entre deux conceptions des opérations statistiques, l'une « réaliste métrologique » (l'objet existe antérieurement à sa mesure), et l'autre « conventionnaliste » (l'objet est créé par les conventions de la quantification : exemples du taux de pauvreté, du chômage, du quotient intellectuel ou de l'opinion publique). »* (Desrosières, 2008).

En comparaison, la mesure des compétences des élèves se présente comme une démarche de construction assez particulière. L'idée sous-jacente de la psychométrie consiste à postuler qu'un test mesure des performances qui sont la manifestation concrète d'un niveau de compétence, non observable directement. Ainsi, l'objet de la mesure est une **variable latente**. Cette approche n'est pas propre au domaine de la cognition. On retrouve ce type de variable en économie avec la notion de propension, en sciences politiques avec la notion d'opinion ou encore en médecine avec la notion de qualité de vie (Falissard, 2008).

Cette singularité n'est pas simplement d'ordre conceptuel, elle a des implications concrètes quand il s'agit de répondre objectivement, avec des indicateurs statistiques, à des questions du débat public, comme celle du niveau des élèves.

LE NIVEAU DES ÉLÈVES : QUESTION ANCIENNE, RÉPONSES RÉCENTES

Il y a plus de trente ans, le débat sur la baisse supposée du niveau scolaire faisait rage, attisé par des interrogations sur la nécessaire transformation du système éducatif (un système de masse peut-il être performant ?) ou sur le lien entre éducation et économie (les compétences des élèves comme levier dans la compétition économique internationale ? Voir (Goldberg et Harvey, 1983) aux États-Unis). En France, (Thélot, 1992), tout comme (Baudelot et Establet, 1989) soulignaient alors le manque criant de mesures directes et objectives des acquis des élèves.

Si d'aucuns étaient tentés de mobiliser les statistiques sur les examens, elles ne permettaient cependant pas de se prononcer sur l'évolution du niveau des élèves. En effet, chaque année les sujets d'examen changent, sans qu'il n'ait été établi de comparaison rigoureuse de leur difficulté. Si bien que, par exemple, la comparaison de deux taux de réussite au baccalauréat n'est pas pertinente pour mesurer une évolution dans le temps : si le taux de réussite augmente, est-ce parce que le niveau d'exigence est moins élevé ou bien parce que le niveau des élèves est meilleur ?

Jusque dans les années quatre-vingt-dix, les seules données disponibles permettant une comparaison temporelle rigoureuse étaient celles issues des tests « psychotechniques » passés pendant les « *trois jours* » organisés par le ministère de la Défense. Elles ne concernaient cependant pas tous les élèves.

Le recours à des tests standardisés est alors apparu comme la solution adaptée. Ce type de dispositif de mesure trouve ses origines en France, dans les travaux d'Alfred Binet et de ses collaborateurs au début du XX^{e} siècle : pourtant, la psychométrie y reste une discipline très méconnue aujourd'hui encore. Paradoxalement, les évaluations des élèves sont très présentes dans le système scolaire français, à travers les contrôles continus fréquents conduits par les enseignants. Des études docimologiques, menées depuis près d'un siècle, avec notamment les travaux de la commission Carnegie sur le baccalauréat en 1936, montrent pourtant que le jugement des élèves par les enseignants est en partie empreint de subjectivité et peut dépendre de facteurs étrangers au niveau de compétence des élèves (Piéron, 1963). La notation des élèves est ainsi susceptible de varier sensiblement selon les caractéristiques des enseignants, des contextes scolaires, ainsi que des élèves eux-mêmes.

En revanche, dans d'autres pays, la psychométrie s'est considérablement développée, notamment aux États-Unis, à travers des thématiques telle que la méritocratie scolaire (assurer un traitement équitable des élèves) ou bien l'intelligence, sujet ayant d'ailleurs conduit à certaines dérives idéologiques (Gould, 1997).

Ainsi, malgré une demande sociale forte et récurrente, la question du niveau des élèves et de son évolution a longtemps souffert d'un manque de cadrage conceptuel et méthodologique. Le recours à des dispositifs d'évaluations standardisées est relativement récent dans le paysage des enquêtes statistiques françaises.

AUJOURD'HUI, IL EXISTE UN VASTE SYSTÈME D'ÉVALUATIONS STANDARDISÉES…

Forte de ce constat, dans les années quatre-vingt-dix, la direction de l'Évaluation et de la Prospective du ministère de l'Éducation nationale[4] a conduit plusieurs études visant à mesurer l'évolution des acquis des élèves. Ces travaux avaient clairement pour objectif de répondre aux tenants de la faillite du système éducatif, souvent nostalgiques d'un modèle scolaire révolu. Cependant, ces premières enquêtes comparatives montraient quelques faiblesses méthodologiques.

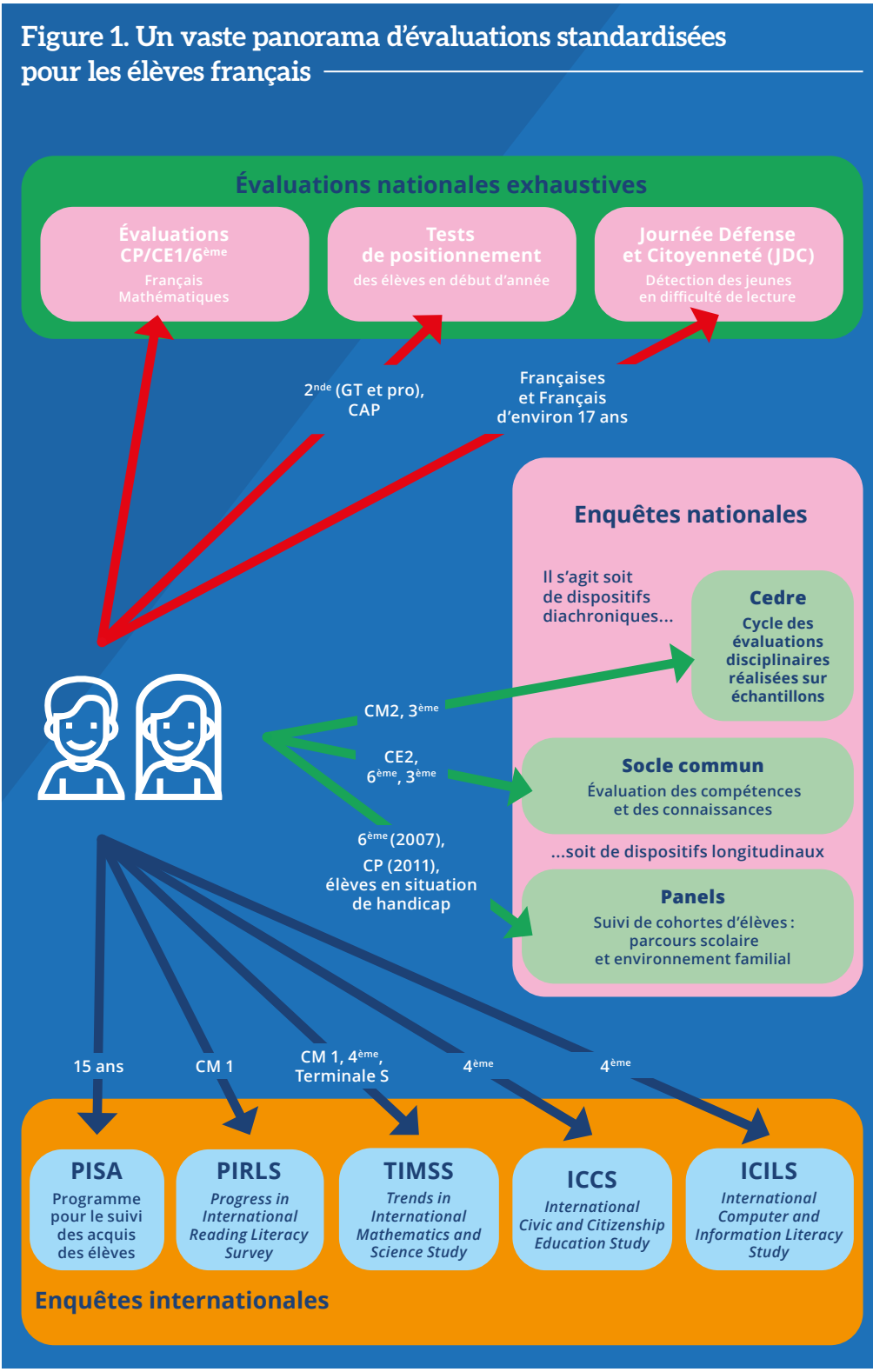
La France avait pourtant une longue expérience de campagnes d'évaluations des élèves, notamment à travers les évaluations nationales diagnostiques passées par tous les élèves de CE2 et de 6ème, à chaque rentrée scolaire, entre 1989 et 2007. Mais ces évaluations ne permettaient pas d'établir des comparaisons temporelles statistiquement robustes. D'une part, leur objectif premier, tout comme celui des examens, n'était pas de rendre compte de l'évolution du niveau d'acquisition des élèves dans le temps, mais de servir d'outils de repérage individuel des difficultés pour les enseignants. D'autre part, les connaissances dans le champ de la mesure en éducation et plus largement en psychométrie étaient très peu diffusées et vulgarisées ; le constat est encore actuel, bien que l'expérience de la Depp dans ce domaine se soit considérablement améliorée depuis une vingtaine d'années.

Progressivement, d'autres dispositifs d'évaluations construits pour permettre des comparaisons diachroniques se sont développés en France ***(figure 1)***. Plusieurs phénomènes relativement récents expliquent cet essor et cette multiplicité. Tout d'abord, la volonté de construire des indicateurs de suivi, pour le pilotage du système, est devenue de plus en plus prégnante, notamment sous l'impulsion de la LOLF[5], qui implique la construction d'indicateurs annuels de résultats, tels que le pourcentage d'élèves qui maîtrisent les compétences attendues, à différents niveaux scolaires.

Parallèlement, les dispositifs internationaux, tels que PISA[6] (OCDE, 2020), PIRLS[7] ou TIMSS[8] (Rocher et Hastedt, 2020) ont largement contribué à rendre incontournables les programmes d'évaluations standardisées à grande échelle (*Large-scale assessments*) dans le débat public sur l'École et dans les décisions en matière de politiques éducatives. Aujourd'hui, rares sont les papiers ou les discours sur le système éducatif qui ne renvoient pas aux évaluations internationales.

4. La DEP a été plus récemment transformée en direction de l'évaluation, de la prospective et de la performance (Depp), qui est le service statistique du ministère.
5. La loi organique relative aux lois de finances fixe le cadre des lois de finances en France. Promulguée en 2001, s'applique à toute l'administration depuis 2006 et vise à moderniser la gestion de l'État. Elle a favorisé le développement d'outils d'évaluation des résultats, sur l'ensemble de l'action administrative.
6. Programme international pour le suivi des acquis des élèves (*Programme for International Student Assessment*), mené par l'Organisation de coopération et de développement économiques (OCDE).
7. Programme international de recherche en lecture scolaire (*Progress in International Reading Literacy*) de l'IEA (*International Association for the Evaluation of Educational Achievement*) basée aux Pays-Bas.
8. *Trends in International Mathematics and Science Study* (TIMSS) est une enquête internationale de l'IEA sur les acquis scolaires en mathématiques et en sciences.

Figure 1. Un vaste panorama d'évaluations standardisées pour les élèves français



Enfin, depuis 2017, de nouvelles évaluations exhaustives ont été mises en place : elles concernent aujourd'hui tous les élèves de CP, de CE1, de 6ème et de 2nde, ainsi que de CAP[9] de lycée. Le déploiement de ces évaluations, qui concernent plus de trois millions d'élèves à chaque rentrée scolaire, a évidemment favorisé l'essor et la visibilité de ce type de dispositifs.

Ainsi, du point de vue des producteurs d'indicateurs dans le domaine des compétences des élèves, il est devenu primordial d'adopter un corpus méthodologique adapté permettant d'établir des mesures fiables dans le temps et dans l'espace. Or les indicateurs produits alimentent différents types d'usages, rendant plus complexe la configuration optimale d'un système d'évaluation de compétences.

… ADAPTÉ À DIFFÉRENTS USAGES

En effet, les utilisations des résultats d'évaluations standardisées des compétences des élèves ont connu différentes phases dans l'histoire, conduisant à une succession de dispositifs différents depuis quarante ans (Trosseille et Rocher, 2015). En particulier, les débats portent sur la configuration des évaluations nationales exhaustives, qui concernent tous les élèves d'un ou de plusieurs niveaux scolaires. La clarification précise des objectifs qui leur sont assignés est indispensable à l'efficacité de leur mise en œuvre. En combinant l'appréciation individuelle (diagnostic de difficulté) et le compte-rendu collectif (construction d'indicateurs statistiques), elles répondent à différents besoins. Ce *hiatus* était d'ailleurs déjà pointé par Alfred Binet (Rozencwajg, 2011) avec la distinction entre approche « clinique » et approche statistique.

Aujourd'hui, d'une façon générale, on identifie trois finalités différentes :

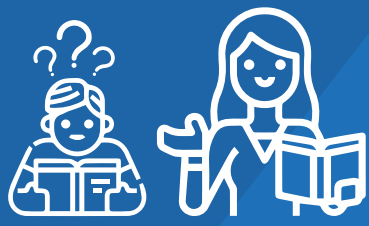
- fournir aux enseignants des repères sur les acquis de leurs élèves, compléter ainsi leurs constats et leur permettre d'enrichir leurs pratiques pédagogiques. Par exemple, en début de 6ème, les élèves passent un test de fluence (*i.e.* de rapidité de lecture), identique pour tous et préalablement calibré, permettant de repérer les élèves susceptibles d'être pénalisés lors de leur parcours au collège ;
- doter les « pilotes de proximité »[10] d'indicateurs leur permettant de mieux connaître les résultats des écoles et d'effectuer une vraie régulation. Par exemple, à partir des résultats obtenus aux évaluations nationales, un recteur peut situer son académie aux différents niveaux du parcours des élèves (CP, CE1, 6ème, 2nde), identifier des points de faiblesse et mettre en place des dispositifs d'actions pédagogiques ;
- disposer d'indicateurs permettant de mesurer, au niveau national, les performances du système éducatif, d'en apprécier les évolutions temporelles et d'en déduire des comparaisons internationales. Par exemple, le cycle des évaluations disciplinaires réalisées sur échantillon (Cedre) situe les élèves par rapport aux attendus des programmes scolaires, de manière fine, ce qui permet d'alimenter la réflexion sur d'éventuels ajustements de ces programmes.

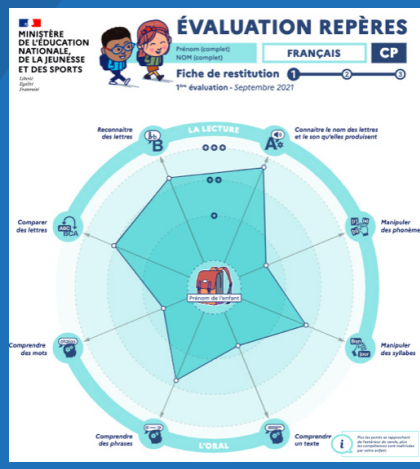
S'agissant des évaluations nationales exhaustives aujourd'hui, intentionnellement positionnées en début d'année scolaire, elles permettent à la fois d'aider à l'action pédagogique immédiate à partir des résultats individuels et à la fois d'alimenter des outils statistiques de suivi et de pilotage, notamment s'agissant des résultats du niveau scolaire précédent. À chaque niveau, les acteurs sont destinataires de résultats répondant à leurs besoins spécifiques ***(figure 2)***.

9. Le certificat d'aptitude professionnelle est un diplôme français d'études secondaires et d'enseignement professionnel.
10. Recteurs d'académie, directeurs académiques des services de l'Éducation nationale (DASEN, anciennement inspecteurs d'académie) ou inspecteurs de l'Éducation nationale (IEN).

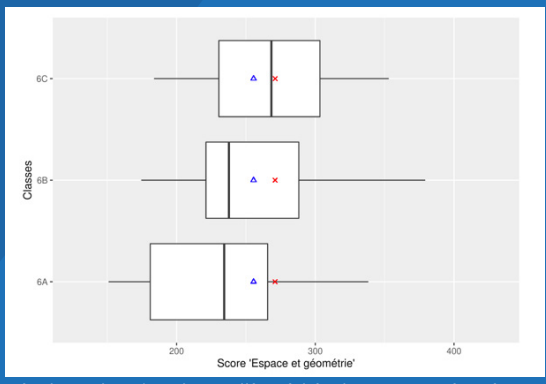
Figure 2. Les évaluations nationales exhaustives fournissent des indicateurs utiles aux différents acteurs éducatifs

Pour aider l'enseignant en CP à identifier les élèves en difficulté de lecture…





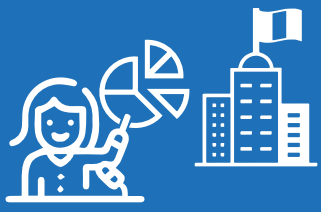
Résultat pour chaque élève de CP de l'évaluation en français de début d'année.

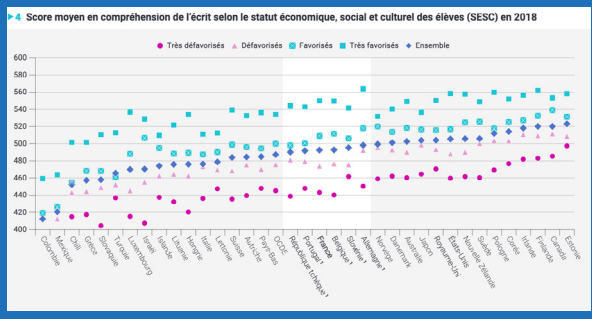


Résultats des 6èmes d'un collège à l'évaluation en géométrie.

… ou pour permettre au principal d'un collège et à son équipe pédagogique d'adapter leur action dans les classes de 6ème…

… ou pour que le Ministère adapte l'affectation des soutiens aux populations en difficulté scolaire…





Comparaisons internationales des scores en compréhension de l'écrit, selon le statut économique et social (PISA).

Depuis 2017, les évaluations nationales exhaustives concernent tous les élèves de CP, CE1, 6ème, 2nde, et de CAP.

… ET ANCRÉ DANS UN HÉRITAGE MÉTHODOLOGIQUE ROBUSTE

> *« Les programmes d'évaluations ont pour ambition d'établir une mesure objective, scientifique, des acquis des élèves, la plus indépendante possible des conditions d'observation, de passation, de correction. »*

Quel que soit le niveau d'usage, les programmes d'évaluations ont pour ambition d'établir une mesure objective, scientifique, des acquis des élèves, la plus indépendante possible des conditions d'observation, de passation, de correction. En ce sens, ces évaluations sont « standardisées ».

Ces opérations se situent au carrefour de deux traditions méthodologiques : celle de la psychométrie, pour ce qui relève de la mesure de dimensions psychologiques, en l'occurrence des acquis cognitifs ; et celle des enquêtes statistiques pour ce qui a trait aux procédures de recueil des données.

Si cette dernière tradition est largement partagée dans le champ de la statistique publique, celle relative à la psychométrie est nettement moins connue. Or, les modélisations statistiques proposées dans ce domaine sont très anciennes (cf. les analyses factorielles de Spearman en 1905) et donnent lieu encore aujourd'hui à une littérature très riche au niveau international. La diffusion de ce corpus méthodologique dans le cadre de dispositifs nationaux doit sans doute beaucoup à l'influence des grandes évaluations internationales (Rocher, 2015b). Les pays participants – et contributeurs actifs – ont bénéficié d'une acculturation à ces méthodes leur permettant d'opérer un transfert de technologie dans leurs systèmes nationaux d'évaluation.

Ces deux traditions – enquêtes statistiques et psychométrie – sont convoquées et combinées lors de la mise en œuvre d'une opération d'évaluation, qui suit un processus précis. Dans le cadre du Cedre (cf. *supra*), ce processus fait même l'objet d'une certification externe (voir sur le site de l'*Afnor*) basée sur un référentiel d'engagements de service pris à chaque étape, conférant ainsi au programme des qualités de reproductibilité et de transparence.

Le processus se déroule en six grandes étapes, depuis la définition du « construit » que l'on cherche à mesurer, jusqu'à la production de résultats contrôlés et redressés.

QUEL «CONSTRUIT» VISE-T-ON?

À la base du processus de mesure se situe le « construit », c'est-à-dire le concept visé par l'opération de mesure. Le construit est défini de manière précise dans un cadre conceptuel (*framework*). Ce cadre conceptuel peut être adossé à un plan de formation ou bien se référer à une théorie cognitive. Par exemple, le cycle Cedre mesure les acquis des élèves dans chaque discipline ; il repose ainsi sur ce qui est censé être enseigné à l'école, au regard de l'ensemble des programmes scolaires. Une illustration radicalement différente peut être donnée par l'identification de troubles d'apprentissage, comme la dyslexie. Les tests vont alors être élaborés pour détecter un déficit sur des dimensions très spécifiques.

Il est primordial que ce cadre soit le plus précis possible, au vu des définitions multiples que peut avoir un même objet. Par exemple, en quoi consistent les compétences en mathématiques ? Deux visions très différentes nous sont données par les évaluations internationales. D'un côté, l'évaluation TIMSS de l'IEA s'intéresse à la façon dont est structuré l'enseignement des mathématiques, pour élaborer l'évaluation selon des domaines assez partagés dans les différents *curricula* (nombres, géométrie, résolution de problèmes, etc.). L'évaluation va ainsi souvent porter sur des aspects « intra-mathématiques ». De son côté, l'enquête internationale PISA s'intéresse au concept de « *literacy* », c'est-à-dire la capacité des individus à appliquer leurs connaissances dans des situations de la vie réelle. La question est alors celle de pouvoir passer du monde réel au monde des mathématiques, et *vice et versa*. L'orientation est donc très structurante pour la suite, c'est-à-dire pour l'élaboration de l'instrument mais également pour l'interprétation des résultats.

Le cadre conceptuel décrit également la structure de l'objet, ou encore l' « univers des *items* », c'est-à-dire l'ensemble des *items* (plus petit élément de mesure) censés mesurer la dimension visée. C'est cette structuration qui définit au final l'objet mesuré.

LA CONCEPTION DES UNITÉS DE MESURE OU *ITEMS*

Une fois ce cadre posé, un groupe de concepteurs – composé d'enseignants, d'inspecteurs, possiblement de chercheurs – est donc chargé de construire un large ensemble d'*items*. Fruit d'un travail collectif, ils font l'objet de débats jusqu'à aboutir à un consensus.

Les *items* sont ensuite soumis à un « cobayage », c'est-à-dire à une passation auprès d'une ou plusieurs classes pour estimer leur difficulté, leur durée de passation minimale et recueillir les réactions éventuelles des élèves.

Les *items* passant avec succès l'étape du cobayage sont alors testés par un échantillon d'élèves représentatif du niveau visé (entre 500 et 2 000 élèves par *item*). Cette phase expérimentale a lieu un an avant la phase principale de l'évaluation, afin de respecter le positionnement de l'évaluation dans le calendrier de l'année scolaire.

L'ÉCHANTILLONNAGE DES ÉLÈVES

S'agissant des grandes enquêtes nationales ou internationales ***(figure 1)***, les échantillons sont composés de plusieurs milliers d'élèves (entre 4 000 et 30 000 selon les programmes et les cycles). Des problématiques classiques du domaine des sondages se posent alors, par exemple la définition du champ, les bases de sondage, les modalités de tirage, etc. Ces aspects sont documentés en détails dans les rapports techniques (Bret *et alii*, 2015 ; OCDE, 2020), en particulier concernant les grandes enquêtes internationales qui imposent le respect de nombreux standards dans ce domaine : il s'agit notamment de maximiser le taux de couverture de la population visée et de limiter les exclusions (par exemple, territoriales ou en raison de contraintes pratiques).

En général, les procédures de sondage procèdent par tirage à deux degrés, d'abord des établissements scolaires (ou directement des classes) et ensuite des élèves. Enfin, dans la mesure où plusieurs échantillons peuvent être tirés à partir des mêmes bases, la coordination de leur tirage est traitée avec précaution (Garcia *et alii*, 2015).

L'ADMINISTRATION DES ÉVALUATIONS

Dans le cadre de ces enquêtes sur échantillons, la passation des évaluations est le plus souvent assurée par des personnels extérieurs à l'établissement scolaire (par exemple : l'évaluation sur tablettes des compétences des élèves de l'école primaire ou les évaluations internationales TIMSS et PISA sur ordinateurs dans les collèges et lycées).

Pour les évaluations nationales exhaustives, en CP et CE1, ce sont les professeurs qui font passer les évaluations. Les consignes sont extrêmement précises, mais il est évident, dès lors que 45 000 professeurs sont concernés, qu'une certaine variabilité des conditions de passation, difficilement quantifiable, vient affecter l'erreur de mesure. Pour les évaluations 6ème et 2nde, les conditions sont en revanche plus favorables, grâce aux modalités de passation sur ordinateurs, qui garantissent une meilleure standardisation.

L'IMPLICATION DES ÉLÈVES

Comme dans toute enquête, il est important de s'interroger sur les dispositions des enquêtés. Dans le cas des évaluations standardisées, elles restent à faible enjeu (*low stakes)* pour les élèves y participant, même si elles renvoient à des enjeux politiques croissants.

Dans le système éducatif français, la notation tient une place prépondérante. Dès lors, face à une évaluation ne conduisant pas à une note, on peut s'interroger sur le degré de motivation des élèves. À partir des enquêtes du Cedre, une étude expérimentale a montré que les élèves s'investissaient davantage lorsqu'on leur annonce préalablement que le résultat obtenu conduira à une note.

LA «CORRECTION» DES RÉPONSES

Enfin, les réponses données par les élèves sont soit codées automatiquement (par exemple dans le cas de questions à choix multiples), soit codées par des correcteurs humains (par exemple dans le cas de productions écrites complexes). En cas de correction humaine, un processus de corrections multiples avec arbitrage est suivi. En effet, il s'agit de neutraliser les nombreux biais de correction qui peuvent apparaître et qui ont été documentés depuis plus d'un siècle par la docimologie, la science des examens (Piéron, 1963).

Enfin, l'analyse psychométrique permet d'identifier des *items* ayant un mauvais fonctionnement, par exemple les *items* n'étant pas corrélés à l'ensemble des *items* censés mesurer la même dimension. Ce processus suit une démarche très empirique, à façon, qui consiste à établir un ensemble d'*items* cohérent et le plus en phase avec le cadre conceptuel.

En guise d'illustration, l'opération Cedre Sciences a réalisé en 2017 l'expérimentation d'environ 400 *items*, pour en retenir 262 pour l'évaluation finale de 2018, dont 43 ont été repris à l'identique de l'enquête de 2007 et 31 de celle de 2013, afin d'assurer des comparaisons temporelles (Bret *et alii*, 2015).

QUELQUES CONCEPTS PSYCHOMÉTRIQUES AUTOUR DE LA NOTION DE VARIABLE LATENTE

Les éléments qui viennent d'être présentés sont potentiellement présents également dans d'autres domaines couverts par les enquêtes statistiques classiques. Comme nous l'avons indiqué en introduction, la spécificité des dispositifs d'évaluation de compétences tient plus particulièrement à l'objet de mesure, dont la matérialité se révèle uniquement à travers l'instrument de mesure. Il est ainsi convenu que l'instrument permet d'observer des performances, qui sont des manifestations concrètes de la compétence, variable qui ne nous est pas accessible directement : cette notion de **variable latente** est centrale en psychométrie.

Afin d'illustrer de façon pédagogique les grandes notions de psychométrie, un exemple classiquement utilisé porte sur la taille des individus ***(figure 3)***. La situation est la suivante : nous n'avons aucun moyen de mesurer directement la taille des individus d'un échantillon donné. Mais nous avons la possibilité de proposer un questionnaire, composé de questions appelant une réponse binaire (oui/non) et n'évoquant pas directement la taille. Nous nous plaçons ainsi artificiellement dans le cas de la mesure d'une variable latente que nous cherchons à approcher à l'aide d'un questionnaire, soit un dispositif de mesure apparemment comparable à celui d'une évaluation standardisée.

Ce questionnaire permet d'illustrer concrètement des concepts importants de psychométrie :

- la **validité** : le test mesure-t-il bien ce qu'il est censé mesurer ? En l'occurrence, il se trouve que la taille réelle des individus est fortement corrélée à un score calculé à partir des 24 *items*. Le score obtenu représente donc bien la variable latente visée ;
- la **dimensionnalité d'un ensemble d'*items*** : le calcul d'un score suppose que les *items* mesurent la même dimension, que le test est unidimensionnel. Cependant, il est clair que les *items* présentés ici ne mesurent pas purement la dimension taille, mais interrogent chacun une multiplicité de dimensions. L'idée est qu'un facteur commun prépondérant relie ces *items*, facteur lié à la taille. Différentes techniques existent pour déterminer si un test peut être considéré comme unidimensionnel.

PASSER DES UNITÉS À L'ÉCHELLE DE MESURE

Lorsqu'il s'agit de construire l'échelle de mesure, d'autres concepts sont mobilisés et peuvent également être illustrés à travers notre questionnaire sur la taille :

- les **fonctionnements différentiels** d'*items* : en guise d'illustration, à l'affirmation « À deux sous un parapluie, c'est souvent moi qui le tiens », 89 % des hommes répondent oui contre 52 % des femmes, soit un écart de 37 points, alors qu'en moyenne sur l'ensemble des *items*, la différence entre les hommes et les femmes est de 20 points seulement. La question est dite « biaisée » selon le sexe : la réponse donnée dépend d'une caractéristique de groupe et non pas seulement de la taille. L'étude des fonctionnements différentiels est fondamentale en matière de comparaison temporelle ou internationale, pour savoir si d'autres facteurs interviennent dans la réussite, au-delà du seul niveau de compétence ;
- le **pouvoir discriminant** des *items* ou la corrélation *item*-test permet de vérifier qu'un *item* mesure bien la dimension supposée. Par exemple, l'*item* « Dans un lit, j'ai souvent froid aux pieds. », repris d'un questionnaire similaire aux Pays-Bas, n'est pas corrélé avec les autres *items* sur l'échantillon français. Ainsi, cet *item* ne mesure pas la dimension taille en France mais plutôt une autre dimension décorrélée, telle que la frilosité… ;

Figure 3. Une illustration des grands concepts de psychométrie

Évaluer la taille des individus (la variable latente),
à l'aide d'un questionnaire de 24 *items*
auxquels il suffit de répondre par oui ou par non (extrait).

1 **Je dois souvent faire attention à ne pas me cogner la tête**

2 **Pour les photos de groupe, on me demande souvent d'être au premier rang**

3 **On me demande souvent si je fais du basket-ball**

4 **Dans la plupart des voitures, je suis mal assis(e)**

5 **Je dois souvent faire faire les ourlets quand j'achète un pantalon**

6 **Je dois souvent me baisser pour faire la bise**

Pour une description détaillée, consulter (Rocher, 2015a).

- l'**échelle de mesure** : le questionnaire ne permet pas de connaître la taille des individus, mais simplement de les classer selon une variable corrélée à la taille, en l'occurrence un score obtenu aux 24 *items*. Il est ainsi possible d'opérer des transformations linéaires sur ce score, ce qui ne modifie pas les rapports entre intervalles de scores entre individus. Typiquement les scores peuvent être standardisés de moyenne 0 et d'écart-type 1, mais le plus souvent ils sont transformés à des valeurs supérieures (moyenne 250 et écart-type 50 dans Cedre, ou moyenne 500 et écart-type 100 dans PISA) afin d'éviter des valeurs négatives.

UN BESOIN DE MODÉLISATION POUR RELIER LES OBSERVATIONS À LA VARIABLE LATENTE

Envisager les résultats à une évaluation comme résultant d'un processus de mesure d'une variable latente ne s'impose pas de lui-même. En effet, le calcul de scores à une évaluation peut sembler trivial : compter le nombre de bonnes réponses obtenues apparaît comme un indicateur adapté du niveau de compétences et il est tout à fait possible de considérer uniquement le nombre de points et de ne pas donner plus de significations à cette statistique qu'un score observé à un test.

> *« Distinguer ce qui relève de la difficulté du test de ce qui relève du niveau de compétence des élèves. »*

Mais cette démarche est très frustre d'un point de vue théorique et trouve vite des limites en pratique, car elle permet difficilement de distinguer ce qui relève de la difficulté du test de ce qui relève du niveau de compétence des élèves. En particulier, pour assurer la comparabilité entre différentes populations ou entre différentes épreuves, le recours à une modélisation plus adaptée, qui se situe au niveau des *items* eux-mêmes et non au niveau du score agrégé, est apparu nécessaire[11]. En particulier, les **modèles de réponse à l'*item*** (ou MRI), nés dans les années soixante, se sont imposés dans le champ des évaluations standardisées à grande échelle ***(encadré 1)***.

Ces modèles permettent de relier de manière probabiliste les réponses aux *items* et la variable latente visée. Ils sont très utiles dès lors qu'il s'agit de comparer les niveaux de compétence de différents groupes d'élèves. Cette problématique renvoie à la notion d'ajustement des métriques (*equating*). Il s'agit de positionner sur la même échelle de compétence les élèves de différentes cohortes, à partir de leurs résultats observés à des évaluations partiellement différentes. De nombreuses techniques existent et sont couramment employées dans les programmes d'évaluations standardisées (Kolen et Brennan, 2004). Typiquement, les comparaisons sont établies à partir d'*items* communs, repris à l'identique d'un moment de mesure à l'autre. Les modèles de réponse à l'*item* fournissent alors un cadre approprié, dans la mesure où ils distinguent les paramètres des *items*, qui sont considérés comme fixes, des paramètres des élèves, considérés comme variables.

11. Pour une lecture en français sur la théorie des tests, voir (Laveault et Grégoire, 2002).

Encadré 1. Des modèles probabilistes pour séparer niveau de compétence et difficulté de l'*item*

Pour un même objet de mesure, les questions (*items*) composant l'instrument de mesure peuvent être différentes. Dès lors, travailler sur un score agrégé trouve vite des limites et il est préférable de faire reposer l'analyse sur l'élément le plus élémentaire, c'est-à-dire l'*item*.

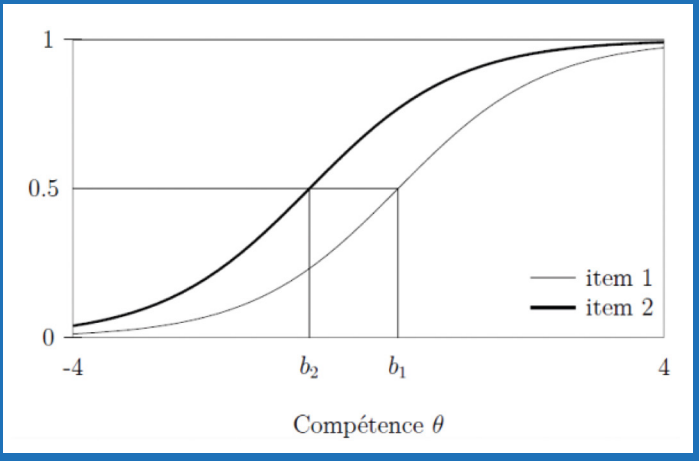
Les modèles de réponse à l'*item* (MRI), nés dans les années soixante, sont une classe de modèles probabilistes. Ils modélisent la probabilité qu'un élève donne une certaine réponse à un *item*, en fonction de paramètres concernant l'élève et l'*item*.

Dans le modèle le plus simple, proposé par le mathématicien danois George Rasch en 1960, la probabilité P_{ij} que l'élève i réussisse l'*item j* est une fonction sigmoïde du niveau de compétence θ_i de l'élève *i* et du niveau de difficulté b_j de l'*item j*. La fonction sigmoïde étant une fonction croissante (voir figure), il ressort que la probabilité de réussite augmente lorsque le niveau de compétence de l'élève augmente et diminue lorsque le niveau de difficulté de l'*item* augmente, ce qui traduit à l'évidence les relations attendues entre réussite, difficulté et niveau de compétence.

L'intérêt de ce type de modélisation, et ce qui explique son succès, c'est de séparer deux concepts-clés, à savoir la difficulté de l'*item* et le niveau de compétence de l'élève.

Ainsi, les **MRI ont un intérêt pratique pour la construction de tests :** si le modèle est bien spécifié sur un échantillon donné, les paramètres des *items* peuvent être considérés comme fixes et applicables à d'autres échantillons dont il sera alors possible de déduire les paramètres relatifs aux élèves, en l'occurrence leur niveau de compétence.

Autre avantage : le niveau de compétence des élèves et la difficulté des *items* sont placés sur la même échelle. Cette propriété permet d'interpréter le niveau de difficulté des *items* par rapprochement avec le continuum de compétence. Ainsi, les élèves situés à un niveau de compétence égal à b_j auront 50 % de chances de réussir l'*item*, ce que traduit visuellement la représentation des courbes caractéristiques des *items* selon ce modèle.



Note de lecture : la probabilité de réussir l'item (en ordonnées) dépend du niveau de compétence (en abscisse). Par définition, le paramètre de difficulté d'un item correspond au niveau de compétence ayant 50 % de chances de réussir l'item. Ainsi, l'item 1 en trait fin est plus difficile que l'item 2 en trait plein. La probabilité de le réussir est plus élevée quel que soit le niveau de compétence.

QUELLE UTILITÉ POUR LA STATISTIQUE PUBLIQUE?

Cet appareillage statistique et psychométrique permet d'élaborer des indicateurs robustes du niveau des élèves et surtout il permet d'établir des comparaisons temporelles et spatiales. Par exemple, la question de l'évolution du niveau scolaire dans le temps peut être abordée statistiquement grâce à la reprise à l'identique d'***items* dits d'ancrage**. En effet, la reprise identique de l'ensemble des *items* passés lors d'une précédente enquête n'est pas forcément pertinente, au regard de l'évolution des programmes scolaires, des pratiques, de l'environnement culturel, etc. Certains *items* doivent être retirés, d'autres ajoutés. Par conséquent, les élèves des deux cohortes passent une épreuve en partie différente. Dès lors, comment assurer la comparabilité des résultats ? Le simple calcul du nombre de bonnes réponses n'est plus pertinent et il faut recourir aux modélisations présentées ici. Ainsi, avec cette approche les enquêtes Cedre permettent d'établir des comparaisons du niveau en mathématiques des élèves depuis plus de dix ans et qui montrent d'ailleurs une dégradation préoccupante des résultats, à l'école primaire comme au collège (Ninnin et Pastor, 2020 ; Ninnin et Salles, 2020).

De leur côté, les évaluations internationales utilisent ces méthodologies pour assurer la comparabilité des difficultés des *items* d'un pays à l'autre. En effet, une hypothèse forte de ces enquêtes est que l'opération de traduction ne modifie pas la difficulté de l'*item*. Des procédures strictes de contrôle des traductions sont mises en œuvre. Cependant, des analyses montrent que la hiérarchie des paramètres de difficulté des questions posées est à peu près conservée pour des pays partageant la même langue, mais qu'elle peut être bouleversée entre deux pays ne parlant pas la même langue. Les modèles de réponse à l'*item* permettent de repérer ces sources de biais potentiels de comparabilité. Un exemple concret est donné par les enquêtes TIMSS auxquelles participent des dizaines de pays et qui ont récemment montré la place préoccupante de la France en mathématiques, de manière complémentaire et cohérente avec Cedre (Colmant et Le Cam, 2020 ; Le Cam et Salles, 2020).

EN PERSPECTIVE : ÉVALUER LES COMPÉTENCES TRANSVERSALES…

Les programmes d'évaluation font face à de nouvelles perspectives, liées à la demande d'évaluation de compétences plus complexes, ainsi qu'à l'essor du numérique.

Ainsi, aux difficultés méthodologiques évoquées précédemment viennent s'ajouter de nouveaux défis posés par une demande croissante (venant pour partie du monde économique) pour la mesure de dimensions beaucoup plus complexes que les compétences traditionnelles, académiques. On parle parfois de compétences transversales, de *soft skills*, compétences du XXIe siècle, de compétences socio-cognitives, etc. ***(encadré 2)***.

L'évaluation de ces dimensions constitue un vrai challenge. En effet, la définition de ces compétences (*framework*) n'est pas toujours très solide ou consensuelle. Ensuite, leur enseignement n'est pas toujours explicite, ce qui interroge sur la portée des résultats de l'évaluation. Enfin, leur structuration est complexe : elles impliquent le plus souvent des dimensions cognitives, mais également des attitudes, des dispositions, etc. Par exemple, évaluer l'esprit critique peut renvoyer à de multiples dimensions intriquées, telles que la compréhension, des composantes métacognitives, la curiosité, etc. Même si chaque composante est potentiellement évaluable, leur juxtaposition ne permet pas non plus de rendre compte d'un degré d'esprit critique. Des dispositifs plus « holistiques » doivent être imaginés.

Encadré 2. Les évaluations standardisées concernent des types de compétences très divers

	De quoi s'agit-il ?	Une illustration
Connaissances et compétences disciplinaires	La référence est celle des programmes scolaires qui définissent les attendus en matières d'apprentissage pour les différents cycles et niveaux scolaires.	**CEDRE Mathématiques 3ème** • Développer une expression algébrique • Appliquer le théorème de Thalès pour calculer des longueurs • etc.
Littératie et Numératie	Il s'agit de la capacité à mobiliser ses acquisitions scolaires pour agir dans la vie quotidienne.	**PISA Compréhension de l'écrit à 15 ans** • Localiser une information dans un texte • Relier des informations entre différentes sources • Évaluer la qualité et la crédibilité d'un texte • etc.
Compétences sociocognitives	Il s'agit d'un terme général pouvant regrouper de nombreuses dimensions spécifiques, parfois appelées aussi socio-comportementales, non cognitives ou conatives, provenant de la recherche en psychologie.	**Panels** Questionnaires dits « subjectifs » sur des thèmes tels que : • Motivation (ex : j'essaie de bien faire au collège parce que j'apprends des choses qui m'intéressent) • Sentiment de performance (ex : vous sentez-vous capable de réussir en mathématiques ?) • etc.
Compétences du XXIe siècle	Un ensemble de compétences censées être importantes face aux évolutions – notamment technologiques – de nos sociétés. Un des premiers cadres de référence est celui du « Partenariat pour les Compétences du 21e siècle » ou P21 qui définit les 4C : • Esprit Critique • Créativité • Collaboration • Communication. D'autres cadres se sont développés, intégrant par exemple deux autres C : • Citoyenneté • et *Character* (personnalité).	**Socle commun** Exemples de travaux en cours dans le champ des évaluations standardisées en France, *via* des conventions de recherche : • Esprit Critique : évaluation de la capacité des élèves à juger la véracité d'une information • Créativité : dans le langage, avec la création d'histoires ou en mathématiques, avec la recherche de solutions originales à des problèmes.

… ET INTÉGRER LA RÉVOLUTION NUMÉRIQUE

La révolution numérique entraîne des transformations profondes, y compris dans le domaine de l'évaluation des compétences des élèves. En 2015, les programmes d'évaluations standardisées ont entamé leur mue vers le format numérique. Aujourd'hui, dans le second degré, les évaluations sont toutes réalisées sur ordinateur et concernent chaque année près de deux millions d'élèves. Dans le premier degré, la situation est plus compliquée, en raison des équipements mal adaptés.

Le processus de mesure n'est pas bouleversé dans ses principes, mais l'évolution technologique apporte son lot de difficultés nouvelles. Ainsi :

- passer du papier/crayon au numérique pose des questions de comparabilité et d'éventuelles ruptures de séries ;
- l'aptitude des élèves à utiliser[12] ces nouveaux outils, voire la familiarité des élèves avec ces nouveaux environnements est mal connue ;
- utiliser le numérique amène des problématiques liées à la confidentialité ou à la sécurité.

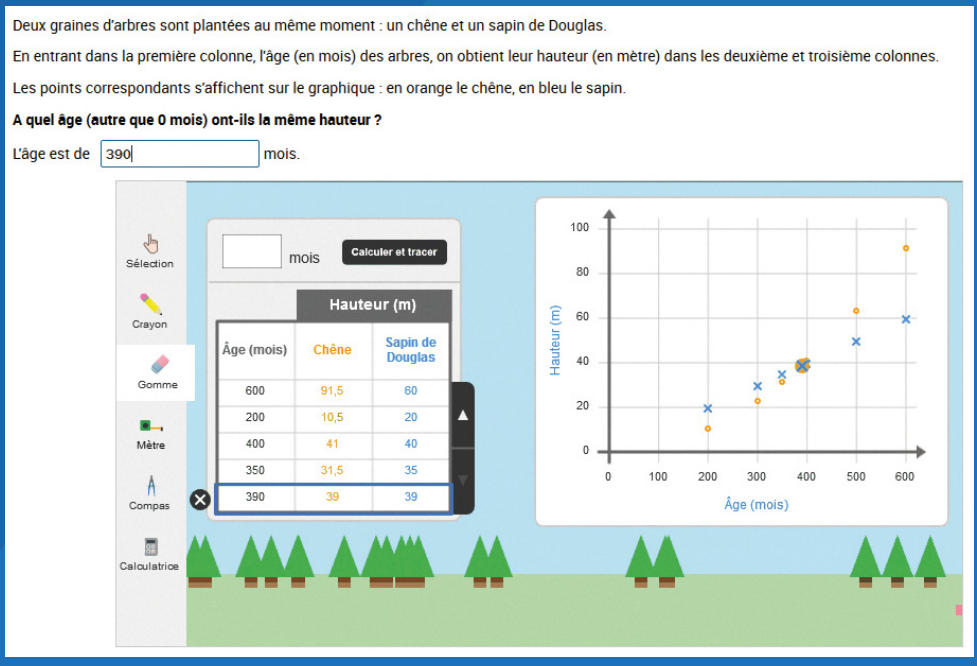
Mais *a contrario*, le numérique offre des fonctionnalités très intéressantes en matière d'évaluation (multimédia, accessibilité, etc.), des techniques plus sophistiquées (comme la possibilité d'introduire des processus adaptatifs), des situations interactives pour des expériences plus ludiques (*game-based*), etc.

Enfin, en matière d'analyse statistique, ces dispositifs permettent de recueillir beaucoup plus de données, à travers l'enregistrement des actions des élèves (« traces » des élèves) ***(encadré 3)***. Ces approches permettent déjà d'enrichir les analyses, et seront très utiles à la fois pour un retour individuel approfondi et pour des statistiques plus précises sur le niveau de compétences des élèves.

12. On parle alors d'« utilisabilité » en ergonomie informatique par exemple.

Encadré 3. L'analyse des traces numériques des élèves

Dans cet *item* interactif, l'élève doit réaliser une série d'essais pour déterminer le point de croisement entre deux fonctions : en entrant des valeurs dans un tableau, celles-ci sont positionnées automatiquement sur un graphique. L'élève peut utiliser différents outils numériques (crayon, gomme, etc.). À partir des traces laissées par ses différentes actions, une analyse basée sur les techniques de *data science* permet d'identifier des profils cognitifs pertinents (Salles *et alii*, 2020). Il est important de noter que cette étude n'est pas *data driven** – approche souvent vouée à l'échec dans ce domaine – mais s'est appuyée sur un cadre théorique didactique qui a guidé le processus d'analyse.



Source : Évaluation des compétences du socle commun en fin de 3ème.

** Le « pilotage par la donnée » supposerait de contextualiser ou de personnaliser l'outil à l'élève en fonction de ses caractéristiques.*

BIBLIOGRAPHIE

BAUDELOT, Christian et ESTABLET, Roger, 1989. *Le niveau monte*. Éditions du Seuil. ISBN 2-02-010385-0.

BRET, Anaïs, GARCIA, Émilie, ROCHER, Thierry, ROUSSEL, Léa et VOURC'H, Ronan, 2015. *Rapport technique de CEDRE, Cycle des Évaluations Disciplinaires Réalisées sur Échantillons. Sciences expérimentales 2013*, *Collège*. [en ligne]. Février 2015. MENESR-DEPP. [Consulté le 18 novembre 2021]. Disponible à l'adresse : *https://www.education.gouv.fr/media/10742/download*.

COLMANT, Marc et LE CAM, Marion, 2020. *TIMSS 2019 – Évaluation internationale des élèves de CM1 en mathématiques et en sciences : les résultats de la France toujours en retrait*. [en ligne]. Décembre 2020. MENESR-DEPP. Note d'information n°20.46. [Consulté le 18 novembre 2021]. Disponible à l'adresse : *https://www.education.gouv.fr/media/73349/download*.

DESROSIÈRES, Alain, 2008. *Gouverner par les nombres, L'argument statistique II*. [en ligne]. Presses des Mines. [Consulté le 18 novembre 2021]. Disponible à l'adresse : *https://books.openedition.org/pressesmines/341*.

EVAIN, Franck, 2020. Indicateurs de valeur ajoutée des lycées : du pilotage interne à la diffusion grand public, In : *Courrier des statistiques*. [en ligne]. 31 décembre 2020. Insee. N° N5, pp. 74-94. [Consulté le 18 novembre 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/5008710/courstat-5.pdf*.

FALISSARD, Bruno, 2008. *Mesurer la subjectivité en santé – Perspective méthodologique et statistique*. 2e édition. Éditions Elsevier-Masson, Issy-les-Moulineaux. ISBN 978-2-294-70317-1.

GARCIA, Émilie, LE CAM, Marion, ROCHER, Thierry et alii, 2015. Méthodes de sondage utilisées dans les programmes d'évaluation des élèves. In : *Éducation & Formations*. [en ligne]. Mai 2015. MENESR-DEPP. N°86-87, pp. 101-117. [Consulté le 18 novembre 2021]. Disponible à l'adresse : *https://cache.media.education.gouv.fr/file/revue_86-87/63/8/depp-2015-EF-86-87_424638.pdf*.

GOLDBERG, Milton et HARVEY, James, 1983. A Nation at Risk: The Report of the National Commission on Excellence in Education. In : *The Phi Delta Kappan*. [en ligne]. Vol. 65, n°1, pp. 14-18. [Consulté le 18 novembre 2021]. Disponible à l'adresse : *https://www.jstor.org/stable/20386898*.

GOULD, Stephen Jay, 1997. *La mal-mesure de l'homme*. Éditions Odile Jacob. ISBN 978-2-7381-0508-0.

KOLEN, Michael J. et BRENNAN, Robert L., 2004. *Test Equating, Linking, and Scaling: Methods and practices*. 3e édition. Éditions Springer-Verlag, New York. ISBN 978-1-4939-0317-7.

LAVEAULT, Dany et GRÉGOIRE, Jacques, 2002. *Introduction aux théories des tests en psychologie et en sciences de l'éducation*. 3e édition, janvier 2014. Éditions De Boeck, Bruxelles. ISBN 978-2-804170752.

LE CAM, Marion et SALLES, Franck, 2020. *TIMSS 2019 – Mathématiques au niveau de la classe de quatrième : des résultats inquiétants en France.* [en ligne]. Décembre 2020. MENESR-DEPP. Note d'information, n°20.47. [Consulté le 18 novembre 2021]. Disponible à l'adresse : *https://www.education.gouv.fr/timss-2019-mathematiques-au-niveau-de-la-classe-de-quatrieme-des-resultats-inquietants-en-france-307819*.

NINNIN, Louis-Marie et PASTOR, Jean-Marc, 2020. *Cedre 2008-2014-2019 Mathématiques en fin d'école : des résultats en baisse.* [en ligne]. Septembre 2020. MENESR-DEPP. Note d'information n°20.33. [Consulté le 18 novembre 2021]. Disponible à l'adresse : *https://www.education.gouv.fr/cedre-2008-2014-2019-mathematiques-en-fin-d-ecole-des-resultats-en-baisse-306336*.

NINNIN, Louis-Marie et SALLES, Franck, 2020. *Cedre 2008-2014-2019 Mathématiques en fin de collège : des résultats en baisse*. [en ligne]. Septembre 2020. MENESR-DEPP. Note d'information n°20.34. [Consulté le 18 novembre 2021]. Disponible à l'adresse : *https://www.education.gouv.fr/cedre-2008-2014-2019-mathematiques-en-fin-de-college-des-resultats-en-baisse-306338*.

OCDE, 2020. *PISA 2018 Technical Report.* [en ligne]. [Consulté le 18 novembre 2021]. Disponible à l'adresse : *https://www.oecd.org/pisa/data/pisa2018technicalreport/*.

PIÉRON, Henry, 1963. *Examens et docimologi*e. 1er janvier 1963. Presses universitaires de France.

ROCHER, Thierry et HASTEDT, Dirk, 2020. *International large-scale assessments in education: a brief guide*. [en ligne]. Septembre 2020. IEA Compass: Briefs in Education, Amsterdam, n°10. [Consulté le 18 novembre 2021]. Disponible à l'adresse : *http://files.eric.ed.gov/fulltext/ED608251.pdf*.

ROCHER, Thierry, 2015a. Mesure des compétences : méthodes psychométriques utilisées dans le cadre des évaluations des élèves. In : *Éducation et Formations*. [en ligne]. Mai 2015. MENESR-DEPP. N°86-87, pp. 37-60. [Consulté le 18 novembre 2021]. Disponible à l'adresse : *https://cache.media.education.gouv.fr/file/revue_86-87/63/8/depp-2015-EF-86-87_424638.pdf*.

ROCHER, Thierry, 2015b. PISA, une belle enquête : lire attentivement la notice. In : *Administration et Éducation*. [en ligne]. Association Française des Acteurs de l'Éducation. N°145, pp 25-30. [Consulté le 18 novembre 2021]. Disponible à l'adresse : *https://www.cairn.info/revue-administration-et-education-2015-1-page-25.htm*.

ROZENCWAJG, Paulette, 2011. La mesure du fonctionnement cognitif chez Binet. In : *Bulletin de psychologie*. [en ligne]. 2011/3, N°513, pp. 251-260. [Consulté le 18 novembre 2021]. Disponible à l'adresse : *https://www.cairn.info/revue-bulletin-de-psychologie-2011-3-page-251.htm*.

SALLES, Franck, DOS SANTOS, Reinaldo et KESKPAIK, Saskia, 2020. *When didactics meet data science: process data analysis in large-scale mathematics assessment in France*. [en ligne]. 29 mai 2020. IEA-ETS Research Institute Journal, Large-scale Assessments in Education, 8:7. [Consulté le 18 novembre 2021]. Disponible à l'adresse : *https://doi.org/10.1186/s40536-020-00085-y*.

THÉLOT, Claude, 1992. *Que sait-on des connaissances des élèves ?* Octobre 1992. Les Dossiers d'Éducation et formations. N°17.

TROSSEILLE, Bruno et ROCHER, Thierry, 2015. Les évaluations standardisées des élèves. Perspective historique. In : *Éducation et Formations*. [en ligne]. Mai 2015. MENESR-DEPP. N°86-87, pp. 15-35. [Consulté le 18 novembre 2021]. Disponible à l'adresse : *https://cache.media.education.gouv.fr/file/revue_86-87/63/8/depp-2015-EF-86-87_424638.pdf*.

LE DÉFI DE L'ÉLABORATION D'UNE NOMENCLATURE STATISTIQUE DES INFRACTIONS

*Benjamin Camus**

La France manquait d'une nomenclature statistique des infractions commune à tous les acteurs de la statistique pénale ; ministère de l'Intérieur et ministère de la Justice utilisaient des nomenclatures de diffusion différentes, ce qui empêchait de disposer de statistiques fines cohérentes tout au long de la filière pénale. La mise au point par l'ONU en 2015 d'une nomenclature internationale a fourni l'occasion de lancer ce chantier en France, lequel a abouti au printemps 2021.

Pour s'affranchir des différences de législations pénales, est retenue une approche fondée principalement sur le comportement de l'auteur de l'infraction. Partant d'une nomenclature juridique préexistante très détaillée, un groupe de travail interministériel a construit en cinq ans la Nomenclature française des infractions : la NFI est ainsi articulée avec la nomenclature internationale pour les grandes catégories, mais avec un détail plus pertinent dans le contexte français. Ce nouvel outil permettra de disposer enfin de statistiques fines comparables entre les deux ministères et de surcroît, susceptibles de comparaisons internationales.

France lacked a statistical nomenclature of offences common to all those involved in criminal statistics; the Ministry of the Interior and the Ministry of Justice used different nomenclatures for dissemination, which prevented the availability of fine-grained statistics that were consistent throughout the criminal justice system. The development of an international nomenclature by the UN in 2015 provided the opportunity to launch this project in France, which was completed in spring 2021.

In order to overcome the differences in criminal legislation, an approach based mainly on the behaviour of the offender has been adopted. Based on a very detailed pre-existing legal nomenclature, an inter ministerial working group has built the French Nomenclature of Offences over a period of five years: the NFI is thus linked to the international nomenclature for the main categories, but with a more relevant detail in the French context. This new tool will make it possible to finally have fine statistics that are comparable between the two ministries and, moreover, that can be compared internationally.

* Inspecteur général honoraire de l'Insee, président du groupe de travail interministériel sur la nomenclature statistique des infractions de 2019 à 2021. Ce qui est présenté dans cet article est le résultat du travail collectif de ce groupe. Que tous les contributeurs soient ici remerciés.

UNE STATISTIQUE PÉNALE INACHEVÉE

La statistique pénale est une des plus anciennes : les premières statistiques régulières en France sont diffusées à partir de 1827 sous la forme d'un « Compte général de l'administration de la justice criminelle » (Perrot, 1976). Ces statistiques ont suscité alors un large intérêt de la part des sociologues, car elles jetaient un éclairage sur la santé morale du pays ; c'est le début de la criminologie.

Mais ensuite cette « statistique morale » intéressera moins, elle ne progressera guère et restera fondée sur des sources administratives ; celle du ministère de la Justice sera complétée, puis supplantée largement dès les années soixante-dix, par celle de la Police et de la Gendarmerie qui appréhende la criminalité plus en amont.

Or cette approche par des sources administratives comporte des limites. Elles s'appuient sur des catégories juridiques et non analytiques, et reflètent aussi l'activité des services luttant contre la criminalité : à criminalité constante, renforcer les contrôles concernant l'usage des stupéfiants ou la circulation routière augmente mécaniquement le nombre d'infractions.

C'est seulement dans les années quatre-vingt-dix que furent mises en place des enquêtes statistiques dites de victimation, s'inspirant d'exemples anglo-saxons (Chambaz, 2018 ; Estival et Filatriau, 2019). On disposait alors d'une source qui mesure la délinquance subie sans le biais des statistiques administratives : les comparaisons internationales devenaient possibles, mais celles-ci restaient limitées aux grandes catégories des enquêtes de victimation et ne couvraient pas l'ensemble des crimes et délits. Or ceux-ci sont connus de façon très détaillée par les sources de la Justice et de la Police, mais avec des nomenclatures « métiers » différentes ***(encadrés 1 et 2)***.

> *« Alors que les principaux domaines économiques et socio-économiques disposent de nomenclatures régulièrement mises à jour, la statistique pénale manquait de ce qui constitue le fondement de tout travail statistique : une nomenclature statistique largement partagée entre tous les acteurs. »*

Alors que les principaux domaines économiques et socio-économiques disposent de nomenclatures régulièrement mises à jour (Amossé, 2020 ; Insee, 2008 ; Guibert, Laganier et Volle, 1971), la statistique pénale manquait de ce qui constitue le fondement de tout travail statistique : une nomenclature statistique largement partagée entre tous les acteurs.

En pratique, le ministère de l'Intérieur et le ministère de la Justice avaient développé des nomenclatures très détaillées mais incompatibles, ce qui entraînait des diffusions récurrentes de chiffres incohérents sur la délinquance, les statisticiens devant expliquer régulièrement les raisons des écarts de comptages. Sur le fond, il s'agit pourtant de suivre le même élément : le traitement pénal d'une infraction, de sa constatation à la réponse pénale apportée par la Justice ***(figure 1)***.

Avec l'informatisation du Casier judiciaire national dans les années 1980, le ministère de la Justice[1] a mis en place une codification fine des infractions dite NATINF (pour NATure d'INFraction) ; cette nomenclature de gestion a vocation à attribuer un code chiffré pour chaque infraction créée par la loi. La NATINF est vraiment très détaillée : dans l'état de la législation en vigueur, elle recense environ 900 crimes, 9 100 délits et 7 000 contraventions.

1. Plus précisément le pôle d'évaluation des politiques pénales de la direction des Affaires criminelles et des grâces (DACG).

Depuis plusieurs années, cette nomenclature est aussi utilisée dans les logiciels de la Police et de la Gendarmerie et figure ainsi dans tous les applicatifs de la filière pénale (Police, Gendarmerie et Justice). Cette référence de gestion commune portait en germe la possibilité d'élaborer des statistiques comparables entre les ministères de l'Intérieur et de la Justice.

La création du service statistique ministériel de la Sécurité intérieure en 2014 a contribué à la production de statistiques sur des périmètres infractionnels partagés entre les deux ministères. Des travaux méthodologiques ont été conduits en lien avec le service statistique ministériel de la Justice et la direction des Affaires criminelles et des grâces pour rapprocher les données des services sur le champ contentieux des stupéfiants (Clanché, Chambaz *et alii*, 2016) ainsi que sur celui des violences conjugales (Brunin, Guedj et Le Rhun, 2019).

Par ailleurs, pour répondre à certaines demandes institutionnelles, ces services ont échangé afin de disposer de champs infractionnels communs, par exemple pour les atteintes sexistes (Haut Conseil à l'égalité entre les femmes et les hommes), les atteintes à caractère raciste, xénophobe ou antireligieux (Commission nationale consultative des droits de l'homme, institution nationale de protection et de promotion des droits de l'homme) ou encore le blanchiment et le financement du terrorisme (Groupe d'action financière). Cependant, ces travaux ponctuels n'avaient pas pour ambition de constituer une réelle nomenclature visant à couvrir l'ensemble des champs contentieux.

Encadré 1. La filière pénale, de l'infraction à l'exécution de la peine –

La statistique pénale porte sur les infractions aux lois pénales et donc sur un domaine que l'on qualifie souvent de délinquance. Elle correspond à la notion de délinquance enregistrée principalement par les services de Police, de Gendarmerie ou de Justice et donc à des **statistiques administratives**.

De façon simplifiée, une infraction est d'abord constatée le plus souvent par un service de Police ou de Gendarmerie. Elle fait alors l'objet d'un procès-verbal décrivant ses caractéristiques : nature de l'infraction, lieu, auteur si celui-ci est connu, éventuelle victime, etc.

Puis ce procès-verbal est transmis à la Justice selon une filière pénale qui associe le ministère de l'Intérieur et de la Justice pour le traitement de la délinquance.

La justice pénale vise à sanctionner les auteurs d'infraction. Selon la gravité de l'infraction, les circuits et les peines sont différents :

- les infractions les plus graves qualifiées de crimes (homicides, viols, etc.) sont passibles d'une peine de prison supérieure ou égale à dix ans ;
- les infractions les moins graves (petits excès de vitesse, tapage nocturne, chasse sans permis, coups et blessures légers, etc.) sont susceptibles de contraventions ;
- les infractions de gravité moyenne sont qualifiées de délits (susceptibles de peines de prison de moins de dix ans).

Le premier niveau de la procédure pénale est celui du *ministère public*, dit aussi du « Parquet » qui reçoit les procès-verbaux des agents de Police judiciaire, soit principalement de la Police et de la Gendarmerie, mais aussi d'administrations diverses, et parfois des plaintes directes des particuliers.

Le Parquet apprécie la suite à donner aux affaires : classement sans suite (pas d'auteur connu, pas d'infraction juridiquement constituée, insuffisance des charges, etc.), alternative aux poursuites (rappels à la loi, indemnisations, etc.) ou poursuite devant une juridiction pénale, laquelle peut acquitter ou condamner ; ces condamnations sont ensuite inscrites au Casier judiciaire national.

L'exécution de la peine peut conduire à l'incarcération en cas de peine de prison ou à un suivi en milieu ouvert (bracelet électronique, etc.) dans le cadre de l'administration pénitentiaire ***(figure 1)***.

LE DÉFI D'UNE NOMENCLATURE INTERNATIONALE

En 2009, l'ONU mit en place une équipe dédiée pour élaborer une nomenclature des infractions. Son travail aboutit en 2015 avec la validation d'une « Classification internationale des infractions à des fins statistiques » (ICCS, pour *International Classification of Crimes for Statistical Purposes*) par l'ONUDC, Office des Nations Unies contre la Drogue et le Crime (ONUDC, 2015).

> *« Les infractions sont définies par le système juridique de chaque pays. »*

Une infraction est une atteinte aux valeurs de la société, or c'est la loi qui fixe les écarts inacceptables à ces valeurs, d'où la définition opérationnelle retenue par l'ICCS : les infractions sont « des comportements considérés comme illégaux, et qui à ce titre, sont punissables par la loi. Elles sont définies par le système juridique de chaque pays ».

Le problème est que les systèmes pénaux des pays sont très variés : droit romain des pays européens latins, *common law* des anglo-saxons, lois islamiques, droit chinois, etc.

Encadré 2. La statistique pénale est un reflet de l'activité délinquante dans la société

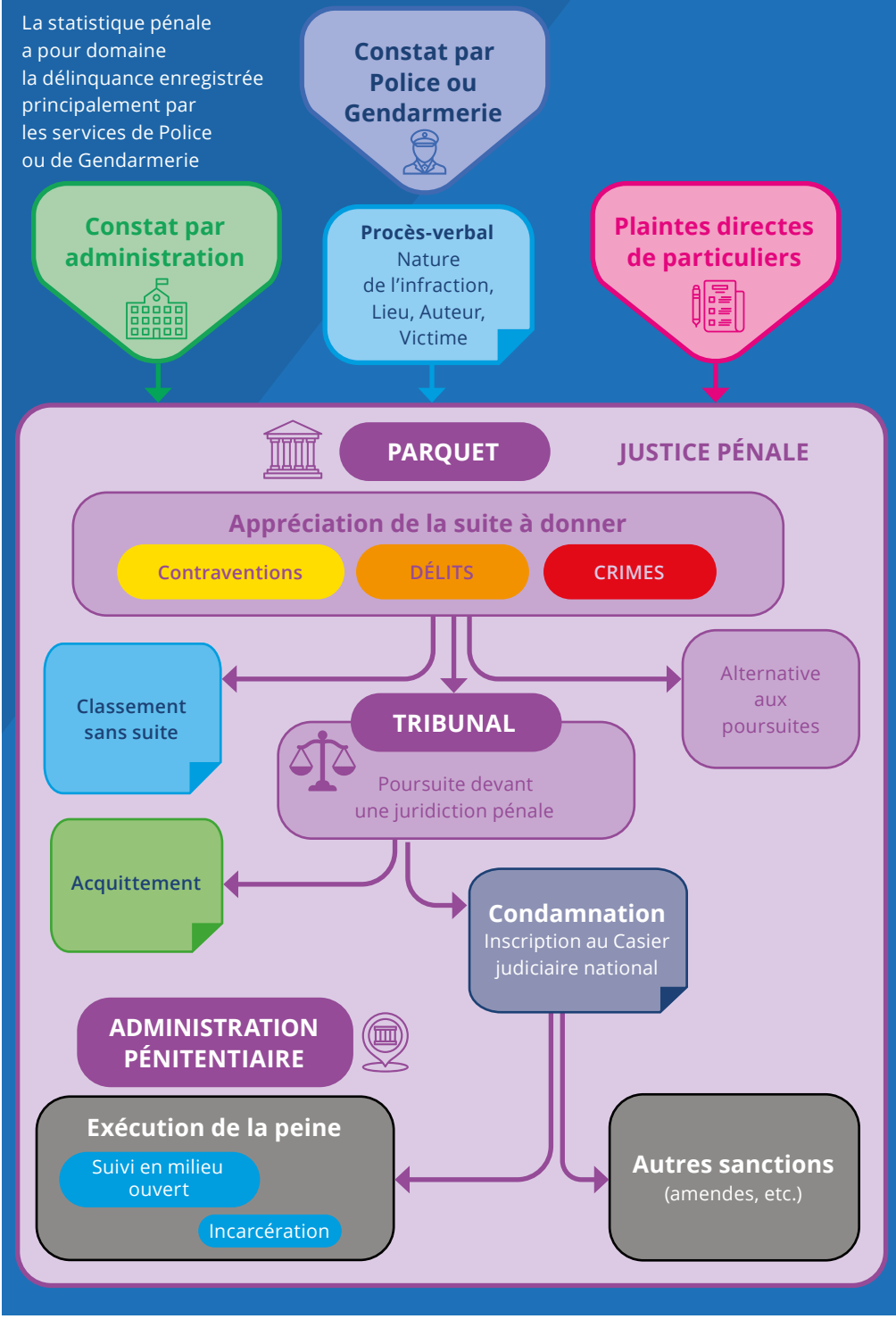
C'est tout d'abord une statistique de gestion. Elle mesure l'activité des services de sécurité et de justice, ce qui est essentiel pour ces fonctions régaliennes de l'État : cette statistique permet de justifier les moyens demandés par les ministères de l'Intérieur et de la Justice, elle sert aussi à répartir les ressources entre les nombreuses unités locales (commissariats, tribunaux, etc.), elle permet de mesurer l'efficacité des services ; elle éclaire la politique pénale en quantifiant l'impact, l'effectivité et l'efficacité des nombreuses lois pénales ; elle pourrait aussi appuyer des travaux prospectifs sur les flux d'affaires et d'auteurs qui transitent par la filière pénale pour adapter le dispositif sécurité/justice.

Ensuite, **la statistique pénale fournit une mesure de la délinquance enregistrée.** La statistique pénale est une réalité déformée et tronquée comme *via* un prisme, car on n'observe que la délinquance enregistrée donc d'une certaine gravité.

Depuis les années soixante-dix, les services de Police et de Gendarmerie diffusent **des indicateurs mensuels**, sortes de bulletins statistiques sur les crimes et délits constatés, ou sur l'évolution conjoncturelle de la délinquance le plus en amont possible de la filière pénale. Ces données désormais finement localisées permettent de dresser des cartes de délinquance selon les contentieux. Plus en aval, la Justice publie des statistiques selon les étapes du traitement judiciaire : affaires transmises, suites données, jugements prononcés, condamnations, incarcération ou suivi en milieu ouvert. Cette mesure de la délinquance comporte des biais connus : elle ne concerne que la délinquance enregistrée par les services de Police et de Gendarmerie, la délinquance subie mais non déclarée n'est connue que par les enquêtes de victimation (lesquelles ne couvrent qu'une partie de la délinquance, celle concernant des particuliers victimes directes) ; cette mesure dépend aussi de l'activité des forces de sécurité. Mais cette mesure est très précieuse : elle permet d'observer les tendances conjoncturelles de la délinquance et de connaître de façon très fine le type d'infraction et le profil des auteurs, notamment s'il s'agit de primo délinquants ou de récidivistes, d'où des analyses fécondes sur le comportement de récidive qui peut être vu comme un indicateur de l'efficacité de la filière pénale et qui est au cœur de l'analyse de la criminalité.

Figure 1. De l'infraction à l'exécution de la peine : la filière pénale simplifiée

La statistique pénale a pour domaine la délinquance enregistrée principalement par les services de Police ou de Gendarmerie



Pour s'affranchir des différences de législations pénales, l'ICCS a retenu une approche fondée principalement sur le comportement de l'auteur associé à une infraction pénale. Dans la terminologie, la nomenclature s'appuie sur la Déclaration universelle des droits de l'homme de 1948, ainsi que sur de nombreuses conventions internationales de l'ONU pour combattre le crime (trafic de drogue, traite des êtres humains, blanchiment d'argent, terrorisme, crime organisé, etc.) et parfois sur d'autres textes internationaux (par exemple, une directive européenne pour le délit d'initié). Ces éléments de droit internationaux permettent de dépasser le problème posé par l'existence de législations pénales très différentes.

Pour construire la classification ICCS, priorité a été donnée aux critères présentant un intérêt particulier pour les politiques en matière de prévention de la criminalité et de justice pénale. Interviennent ensuite les critères :

- de cible (personne, objet, milieu naturel, État, etc.), ce qui correspond à la notion française d'intérêts protégés ;
- de gravité (acte ayant entraîné la mort, etc.) ;
- ou de modes opératoires (avec violence, etc.).

La classification internationale comprend 11 sections avec une répartition par grands domaines et une potentielle hiérarchie dans l'ordre des sections ***(figure 2)*** :

- on isole d'abord tous les homicides et tentatives d'homicides (section 1) ;
- puis toutes les autres atteintes à la personne (section 2) ;
- en isolant les actes à caractère sexuel (section 3) ;
- puis les atteintes aux biens en distinguant les actes avec violence (section 4) ;
- ou sans violence (section 5) ;
- et enfin les atteintes à la société (sections 6 à 10) : la drogue (6), la fraude (7), les atteintes à l'ordre public (8), les atteintes à la sécurité publique (9) et enfin au milieu naturel (10) ;
- la section résiduelle 11 comprend essentiellement les actes relevant de la compétence universelle (comme les crimes contre l'humanité).

On comprend que les premières sections correspondent à des domaines bien cernés de la criminalité traditionnelle et que les dernières sections concernent des atteintes à la société dont la définition évolue plus dans le temps et varie dans l'espace ; il suffit de penser à l'exemple de la cybercriminalité, ignorée il y a quelques décennies, ou des atteintes aux bonnes mœurs (dépénalisation de l'homosexualité inégale selon les pays).

UNE POSSIBLE HIÉRARCHIE ENTRE LES SECTIONS

Tous les actes conduisant à la mort d'une personne sont regroupés dans la section 1 (sauf les crimes contre l'humanité) : par exemple, le viol suivi de mort est classé en section 1 et non en section 3 « actes préjudiciables à caractère sexuel » ; la mort suite à une action terroriste est également classée en section 1 et non en section 9 « atteintes à la sécurité publique et à la sûreté de l'État ».

Autre singularité : visant à couvrir l'ensemble du champ international des infractions, la classification comporte des infractions qui correspondent à des actes légaux dans certains pays et illégaux dans d'autres, voire en contradiction avec les Droits de l'homme (apostasie, prosélytisme, avortement, adultère, homosexualité par exemple). Toutefois, ces situations ne concernent qu'un nombre très limité de postes, d'où un impact marginal sur les comparaisons internationales.

LA DÉCLINAISON EN FRANCE DE LA NOMENCLATURE INTERNATIONALE

Il revenait à l'Insee de coordonner l'adaptation et la mise en œuvre de la nomenclature internationale ICCS dans le système statistique public français, afin que celle-ci devienne le cadre de référence de la production et de la diffusion de statistiques publiques dans les domaines de la sécurité et de la justice pénale.

À cet effet, un groupe de travail interministériel associant les principaux acteurs concernés a été mis en place en 2016[2]. Le groupe de travail devait mener un double chantier : renseigner au mieux les postes de l'ICCS, et définir une déclinaison nationale articulée avec l'ICCS et

Figure 2. Même premier niveau pour la nomenclature française (NFI) et la nomenclature internationale (ICCS)

2. Dans sa composition finale fixée en 2018, ce groupe comprenait le SSMSI (service statistique ministériel de la Sécurité intérieure) du ministère de l'Intérieur, la sous-direction de la Statistique des études (SDSE), service statistique ministériel de la Justice, le pôle d'évaluation des politiques pénales (PEPP) de la direction des Affaires criminelles et des grâces (DACG) du ministère de la Justice, des représentants des services opérationnels du ministère de l'Intérieur (direction générale de la Police nationale et direction générale de la Gendarmerie nationale).

pertinente en France. Ce travail devait permettre aussi de déterminer une nomenclature statistique agrégée commune aux ministères de l'Intérieur et de la Justice, qui n'existait pas encore. Au terme d'un cycle de 34 réunions, le groupe de travail a ainsi proposé en avril 2021 une première version de nomenclature française des infractions (NFI[3]).

La difficulté de l'exercice était de faire table rase des nomenclatures anciennes et de construire une nomenclature commune. De fait, les acteurs concernés ont bien joué le jeu, sans doute parce que l'ICCS proposait un cadre très structuré qui manquait aux nomenclatures métiers élaborées le plus souvent au fil de l'eau sans visée statistique. Le travail de réflexion mené pour définir l'ICCS était de grande qualité ; pour preuve, dès 2016, les spécialistes universitaires américains ont reconnu que l'ICCS présentait toutes les qualités attendues d'une nomenclature d'infractions et ont ainsi proposé de l'adopter comme cadre central d'une nomenclature pour les États-Unis, avec quelques aménagements pour tenir compte du contexte national (NASEM, 2016).

En termes de champ, un manuel d'implémentation de l'ICCS de 2019 (ONUDC, 2019) préconise de se limiter pour les pays de droit romain aux infractions les plus graves (crimes et délits). Le groupe de travail a cependant retenu une option plus large, en rajoutant le champ des contraventions, car la frontière entre les crimes et délits et le contraventionnel est variable dans le temps (exemple de la conduite sans permis) et que, de ce fait, ce champ large correspond le plus souvent au champ des statistiques actuellement diffusées par les ministères de l'Intérieur et de la Justice en France[4] ***(figure 3)***. Les comparaisons internationales ne seront possibles que sur le champ des crimes et des délits.

CINQ ANS POUR DÉFINIR UNE NOUVELLE NOMENCLATURE

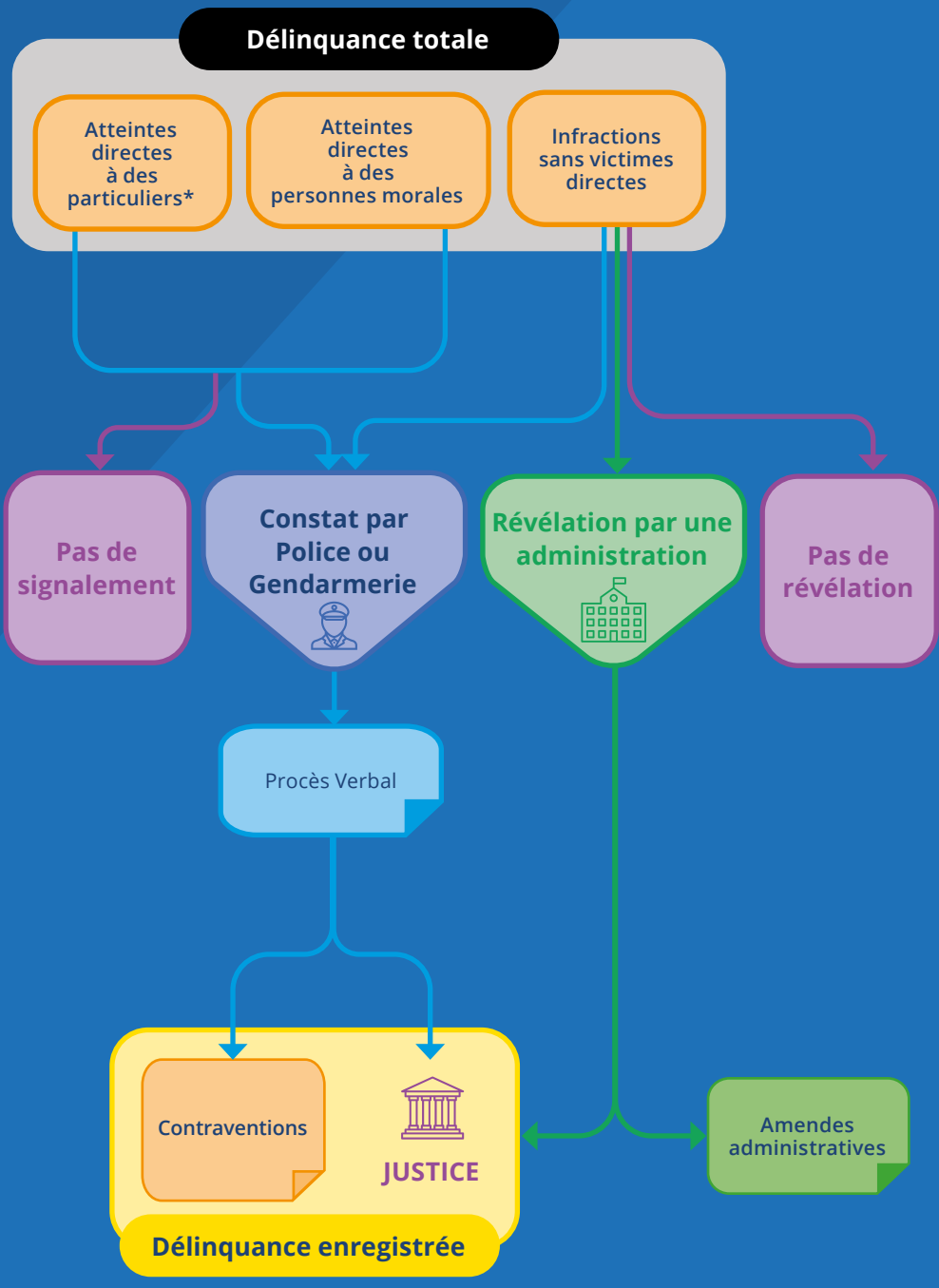
Une première étape a été d'établir une table de passage de la NATINF vers l'ICCS. Les experts de la DACG devaient donc analyser les quelque 17 000 postes élémentaires de la NATINF[5] utilisés dans les applicatifs de la filière pénale. Ce préalable est recommandé par l'ONU et Eurostat (Eurostat, 2017), il a été suivi notamment par les statisticiens allemands (Baumann, Kerner et Mischkowitz, 2016). C'est un travail complexe et minutieux qui nécessite de nombreux choix avec des affinements progressifs au fil de l'examen des différentes sections, ce qui explique la durée du groupe de travail.

> *« L'affectation d'une infraction précise à un poste de l'ICCS a souvent posé des problèmes de frontière. »*

L'affectation d'une infraction précise à un poste de l'ICCS a souvent posé des problèmes de frontière. Le groupe de travail a cherché à dégager **le domaine principal de l'infraction** pour constituer des catégories homogènes. Par exemple, le droit français de l'environnement comporte beaucoup d'infractions à caractère préventif : on les a affectées en section 10 « atteintes au milieu naturel » et non en section 8 « atteintes à l'ordre public, à l'autorité et aux dispositions juridiques de l'État », dans la mesure où la principale valeur sociale protégée est ici l'environnement ; cela permet ainsi de regrouper tous les contentieux de l'environnement.

3. La version définitive a été mise en ligne le 2 décembre de la même année (Ministère de l'Intérieur, 2021).
4. Les contraventions représentent une part importante des infractions dans certains domaines : consommation, environnement, circulation routière, atteinte à l'ordre public.
5. En vigueur ou abrogés.

Figure 3. De la délinquance totale à la délinquance enregistrée



** Champ des enquêtes de Victimation*

Autre critère d'affectation, **l'affectation au poste le plus précis**. Par exemple, la fraude fiscale est isolée dans la section 8 dans le poste 8041 « actes contraires aux dispositions fiscales » et non dans un poste plus large de la section 7 « fraude financière visant l'État » (70111), alors que les inclusions/exclusions de l'ICCS étaient contradictoires sur ce point.

La nomenclature juridique NATINF ne décrit pas complètement les comportements ; par exemple, les infractions de « vol avec circonstances aggravantes » peuvent ne pas faire apparaître certaines caractéristiques du comportement : le statut de la victime, le lieu d'infraction, la nature du bien volé, etc. C'est pourquoi, au-delà de la NATINF, on a dû recourir parfois à d'autres variables de gestion comme la codification en « index » (catégorie de diffusion) du ministère de l'Intérieur ou la nature des « circonstances aggravantes » pour le ministère de la Justice, par exemple pour bien distinguer entre les vols avec ou sans violence.

UNE NOMENCLATURE HYBRIDE À L'ARRIVÉE

Partant de la table de passage NATINF/ICCS, la nomenclature française des infractions a été articulée avec l'ICCS et adaptée au contexte français. On a construit ainsi une nomenclature hybride entre un découpage statistique international et une codification fine de la législation pénale.

> *« On a conservé de façon quasi systématique le découpage par section et division de l'ICCS. »*

La granularité de la table de passage NATINF/ICCS fixe pour l'essentiel les niveaux retenus pour la nomenclature française. La profondeur d'emboîtement des deux nomenclatures est différente selon les sections. On a conservé de façon quasi systématique le découpage par section et division de l'ICCS (11 niveaux 1 et 62 niveaux 2). Au sein des divisions, on a cherché à coller autant que possible aux subdivisions de l'ICCS ; quand ce n'était plus possible, on a introduit une ventilation simple avec des catégories d'actes bien isolés et d'importance en France, de façon à avoir une nomenclature statistique des infractions pertinente pour le pays. Par exemple, pour la section 10 (atteintes au milieu naturel), la nomenclature française des infractions reprend les subdivisions de l'ICCS mais en les détaillant davantage pour le poste résiduel « autres atteintes au milieu naturel », en rajoutant 8 subdivisions afin de distinguer le domaine concerné par les actes de prévention ***(encadré 3)***.

Cette démarche est analogue à celle des États-Unis qui ont défini en 2016 une nomenclature statistique des infractions reprenant les 11 divisions de l'ICCS, mais en restructurant les découpages dès le niveau des divisions (NASEM, 2016). Par exemple, dans la section 2, plutôt que la distinction entre agressions graves ou mineures, les Américains isolent les violences avec ou sans armes à feu. De façon générale, la NFI s'est beaucoup moins écartée des subdivisions de l'ICCS que la nomenclature américaine.

DES LIBELLÉS ADAPTÉS AU CONTEXTE NATIONAL

Les libellés de la nomenclature française reprennent ceux de la version française de l'ICCS, sauf pour tenir compte des pratiques de repérages des infractions et du droit pénal français pour aboutir à des libellés plus compréhensibles en France. De très nombreux libellés ont ainsi été modifiés par rapport à ceux de la version française de l'ICCS : « personne morale » plutôt qu'« entité morale » ; « vols avec violence ou menaces contre une personne » plutôt que « vols qualifiés » ; « harcèlement » plutôt qu' « actes visant à provoquer la peur ou la détresse émotionnelle ».

Encadré 3. Les deux premiers niveaux de la nomenclature française des infractions

Comme toute nomenclature dérivée de nomenclature internationale, la NFI a des niveaux hiérarchiques emboîtés : section, division, groupe, classe. Elle reprend les 11 sections de l'ICCS et très largement les divisions de l'ICCS, mais est plus détaillée pour les groupes et les classes. Le schéma suivant présente les deux premiers niveaux de la NFI, soit 11 sections et 59 divisions.

01 - Actes entraînant ou visant à entraîner la mort

- 01. A - Homicide intentionel
- 01. B - Tentative d'homicide intentionnel
- 01. C - Homicide non intentionel
- 01. Z - Autres actes entraînant ou visant à entraîner la mort

02 - Actes portant atteinte ou visant à porter atteinte à la personne

- 02. A - Atteintes volontaires à l'intégrité de la personne
- 02. B - Atteintes à la liberté
- 02. C - Esclavage ou exploitation
- 02. D - Traite des êtres humains
- 02. E - Extorsion ou chantage
- 02. F - Négligences ou comportements dangereux
- 02. G - Harcèlements
- 02. H - Diffamation ou injure
- 02. I - Discrimination
- 02. J - Atteintes à l'intimité de la personne
- 02. K - Abus de faiblesse

03 - Actes portant atteinte à la personne à caractère sexuel

- 03. A - Viol
- 03. B - Agression ou atteinte sexuelle
- 03. C - Violences sexuelles non physiques
- 03. D - Exploitation sexuelle

04 - Atteintes aux biens avec violence ou menace contre la personne

- 04. A - Vol avec violence ou menace
- 04. Z - Autres atteintes aux biens avec violence ou menace

05 - Atteintes au biens sans violence ni menace

- 05. A - Vol sans violence et abus de confiance
- 05. B - Atteinte au droit d'auteur (propriété littéraire et artistique)
- 05. C - Destructions ou dégradations volontaires
- 05. Z - Autres atteintes aux biens sans violence

06 - Actes faisant intervenir des stupéfiants ou d'autres substances psychoactives

- 06. A - Infractions à la législation sur les stupéfiants
- 06. B - Infractions à la législation sur l'alcool, le tabac ou les produits dopants
- 06. Z - Autres infractions liées aux substances vénéneuses

07 - Actes relevant de la fraude, de la tromperie et de la corruption

- 07. A - Fraude
- 07. B - Contrefaçon ou faux
- 07. C - Atteinte à la probité
- 07. D - Actes faisant intervenir le produit d'une infraction
- 07. E - Trafic de biens culturels

08 - Atteintes à l'ordre public et à l'autorité de l'État

- 08. A - Atteintes à l'ordre public
- 08. B - Atteintes aux mœurs
- 08. C - Atteintes à la liberté d'expression ou à ses limites
- 08. D - Infractions économiques ou financières
- 08. E - Infraction à la législation sur les jeux d'argent ou dopage animal
- 08. F - Atteintes au patrimoine historique et culturel
- 08. G - Infractions à la législation sur les étrangers
- 08. H - Atteintes à l'autorité
- 08. I - Infractions électorales ou au financement des partis politiques
- 08. J - Infractions à la législation du travail
- 08. Z - Autres atteintes à l'ordre public et à l'autorité de l'État

09 - Atteintes à la sécurité publique et à la sûreté de l'État

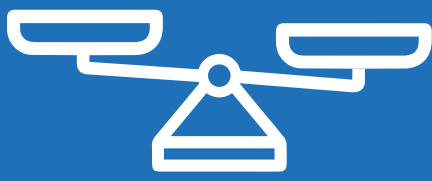
- 09. A - Infractions à la législation sur les armes et les explosifs
- 09. B - Atteintes à la santé et à la sécurité
- 09. C - Atteintes à un système informatique
- 09. D - Atteintes à la sûreté de l'État
- 09. E - Participation à une association de malfaiteurs
- 09. F - Terrorisme
- 09. G - Infractions à la réglementation routière sans dommage corporel ni matériel
- 09. Z - Autres atteintes à la sécurité

10 - Atteintes à l'environnement

- 10. A - Pollution de l'environnement
- 10. B - Déchets
- 10. C - Commerce ou détention d'espèces de faune ou de flore protégées ou interdites
- 10. D - Actes entraînant l'appauvrissement ou la dégradation des ressources naturelles
- 10. Z - Autres atteintes au milieu naturel

11 - Autres actes illégaux

- 11. A - Actes relevant de la compétence universelle
- 11. B - Infractions militaires



Le travail d'élaboration de la NFI s'est fait à partir de la NATINF, mais d'autres informations ont parfois été mobilisées. Comme on l'a vu précédemment, certains critères utilisés par l'ICCS ne figurent parfois pas dans le libellé de la NATINF, mais sont accessibles dans d'autres variables décrivant les actes incriminés et disponibles dans les fichiers de la filière pénale (par exemple, l'âge de la victime pour repérer des actes contre les mineurs, « les circonstances aggravantes », « l'index » des relevés de la Police ou de la Gendarmerie). La définition de la NFI tient compte de cette possibilité de codage plus fin sur certains segments de la filière pénale. On peut faire un parallèle avec la nomenclature des « Professions et Catégories Socioprofessionnelles » (PCS) qui peut être codée à un niveau différent selon les sources statistiques « employeur » ou « salarié ». Par exemple, seul le recours à la variable « index » dans les sources Police/Gendarmerie permet de distinguer les vols avec violence sur une personne, dans un lieu public, dans un lieu privé, dans une institution financière ou dans une institution non financière.

LES LIMITES HÉRITÉES DE LA NOMENCLATURE INTERNATIONALE

L'ICCS est une nomenclature internationale qui privilégie les domaines d'infraction à dimension internationale, qui ont fait l'objet de conventions, comme la drogue, la propriété intellectuelle, ou le crime organisé. À l'opposé, les domaines à dimension locale ou ceux où le droit pénal est moins développé sont moins bien couverts. Ainsi le manuel qui l'accompagne n'évoque ni les infractions au droit de l'urbanisme ou de la construction, ni les infractions liées aux moyens de transport autres que la circulation routière (ONUDC, 2015).

Lors de la définition de la NFI, on a donc cherché à dépasser cette limite en rajoutant des subdivisions qui complètent l'approche de la nomenclature de l'ONUDC. Ce faisant, elle nécessitera quelques précautions d'usage, notamment lors des comparaisons internationales.

Les domaines couverts par les cinq premières sections sont relativement faciles à cerner : atteintes aux personnes et aux biens. Il n'en est pas de même pour les atteintes à la société. Certes les domaines couverts par les sections 6 (actes faisant intervenir des drogues contrôlées ou d'autres substances psychoactives) et 10 (atteintes au milieu naturel) sont bien délimités, mais on observe de possibles chevauchements entre les sections 7 (actes relevant de la fraude, de la tromperie et de la corruption), 8 (atteintes à l'ordre public et à l'autorité de l'État) et 9 (atteintes à la sécurité publique et à la sûreté de l'État). Par exemple, l'exercice illégal d'une profession peut être rattaché au poste 07019 de l'ICCS (autres actes de fraude) ou au poste 08042 (actes contraires aux réglementations commerciales et financières) voire au poste 02071 (actes mettant en danger la santé) pour le cas de l'exercice illégal de la médecine.

Lors des comparaisons internationales, il faudra être vigilant sur le fait que les législations peuvent différer parfois sensiblement. Il y a des domaines où les comparaisons sont possibles sans difficulté particulière, comme les homicides ou tentatives d'homicide, et d'autres domaines, comme l'usage de stupéfiants, où les législations pénales peuvent diverger beaucoup. En pratique, une large part des infractions relèvent des premiers domaines.

UNE UTILISATION EN COMPLÉMENT D'AUTRES APPROCHES

Enfin, comme dans toute nomenclature, l'affectation d'une infraction à une et une seule catégorie est réductrice et ne permet pas de répondre à certains besoins nationaux d'analyse. Certaines infractions pourraient être rattachées à deux catégories, cette difficulté se retrouve lors de l'élaboration de la NFI. On pourrait isoler les catégories susceptibles d'une double approche. Une solution théorique optimale aurait été de retenir une NFI suffisamment détaillée pour permettre cette double approche. Par exemple, si les viols suivis de la mort étaient isolés dans la NFI, on pourrait alors calculer l'agrégat de tous les viols (avec ou sans mort). Mais cette solution appliquée de façon systématique serait très coûteuse (par multiplication de postes fins NFI de faible effectif), elle n'a donc été que très rarement retenue. Par exemple, les contrefaçons dangereuses pour la santé sont isolées dans les actes dangereux de la section 2 pour pouvoir être agrégées avec l'ensemble des contrefaçons de produits classées ailleurs dans la section 7 (actes relevant de la fraude, de la tromperie ou de la corruption).

> *« Pour aborder des problématiques transversales (comme le crime organisé) il faudra donc s'affranchir de la NFI et retenir des regroupements de NATINF. »*

Pour aborder des problématiques transversales (comme le crime organisé) il faudra donc s'affranchir de la NFI et retenir des regroupements de NATINF. La meilleure solution pour dépasser cette limite est de rajouter pour chaque infraction des « descripteurs supplémentaires » comme le préconise le manuel de l'ICCS ; une telle extension nécessite un important travail qui pourrait être organisé à l'avenir pour donner plus de souplesse à l'utilisation de la nomenclature, comme l'ont réalisé les Américains (NASEM, 2016).

UN OUTIL QUI RESTE À CALIBRER PAR L'USAGE

Il reste à pratiquer cette première version de la NFI pour en tester la pertinence et la robustesse. Seules des analyses par contentieux montreront les niveaux d'agrégation les plus pertinents. Il faudra aussi étudier les rétropolations possibles et documenter d'éventuelles ruptures de séries. L'articulation de la NFI avec l'ICCS facilitera la réponse aux questionnaires internationaux sur la criminalité et permettra notamment de renseigner certains indicateurs du développement durable de l'ONU, principalement ceux associés au 16e objectif « Paix, justice et institutions efficaces » (Clanché, 2019).

Cet exercice de définition d'une nomenclature nationale autorise aussi un retour critique vers l'ONUDC dans la perspective d'une future révision de l'ICCS. Par exemple, le partage entre les différentes rubriques des atteintes à la société (sections 7 à 10) reste à clarifier par endroit. Dans la section 7, la première division 701 (fraude financière concernant l'État) a un intitulé trompeur, car elle exclut la fraude fiscale. De même, il faut signaler de forts chevauchements entre les postes des sections 8 et 9. Parfois, le détail est excessif : dans la section 1, les postes d'intérêt sont les trois premières divisions et les cinq suivantes ne sont là que pour isoler certains types d'homicides peu comparables entre pays mais peu nombreux, on gagnerait à regrouper ces postes en une seule division (autres actes entraînant ou visant à entraîner la mort) comme on l'a fait en NFI. L'élaboration de la NFI a conduit à créer 30 % de postes en plus par rapport à l'ICCS ; l'ONU pourrait examiner si ces rajouts n'ont pas un intérêt dans un contexte plus large.

UN PAS VERS L'ÉLARGISSEMENT DU CHAMP DES ANALYSES QUANTITATIVES

Ainsi, cette nomenclature marque une étape décisive dans l'approche quantitative de la criminalité en élargissant et structurant le champ des études. En particulier, on pourra enrichir les analyses sur des domaines encore peu couverts de la criminalité : délinquance économique et financière, cybercriminalité, environnement, etc.

Pour conclure, souhaitons que cette nomenclature favorise le développement d'études quantitatives sur la délinquance au sens large (crimes, délits et contraventions). Le domaine reste peu étudié alors que désormais les sources statistiques des ministères de l'Intérieur et de la Justice sont devenues très riches et exploitables dans des catégories communes : ceci était attendu de façon légitime par les parlementaires, les médias et le grand public. Enfin, ces catégories autorisent des comparaisons internationales mettant en perspective les chiffres français.

BIBLIOGRAPHIE

AMOSSÉ, Thomas, 2020. La nomenclature socioprofessionnelle 2020 : Continuité et innovation, pour des usages renforcés. In : *Courrier des statistiques.* [en ligne]. 29 juin 2020. Insee. N° N4, pp. 62-80. [Consulté le 16 décembre 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/4497076/courstat-4-5.pdf*.

BAUMANN, Thomas, KERNER, Hans-Jürgen et MISCHKOWITZ, Robert, 2016. National Implementation of the new International Classification of Crimes for Statistical Purposes (ICCS). In : *WISTA – Scientific Journal*. [en ligne]. 1er mai 2016. Statistisches Bundesamt. N° 5-2016, pp. 102 et suivantes. [Consulté le 16 décembre 2021]. Disponible à l'adresse : *https://www.destatis.de/EN/Methods/WISTAScientificJournal/Downloads/national-implementation-052016.html*.

BRUNIN, Louise, GUEDJ, Hélène et LE RHUN, Béatrice, 2019. *Comparaison des statistiques Sécurité et Justice : Le contentieux des violences conjugales.* [en ligne]. Novembre 2019. SSMSI-SDSE-DACG. Interstats Méthode N°16. [Consulté le 16 décembre 2021]. Disponible à l'adresse : *https://www.interieur.gouv.fr/Interstats/Themes/Violences-physiques-ou-sexuelles/Interstats-Methode-N-16-Comparaison-des-statistiques-Securite-et-Justice*.

CHAMBAZ, Christine, 2018. De l'activité de la justice au suivi du justiciable, faire parler les données de gestion. In : *Courrier des statistiques*. [en ligne]. 6 décembre 2018. Insee. N° N1, pp. 45-57. [Consulté le 16 décembre 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/3646985/courstat-1-8.pdf*.

CLANCHÉ, François, 2019. La mesure de la sécurité et la satisfaction vis-à-vis des institutions en France, l'impulsion donnée par les objectifs de développement durable des Nations Unies. In : *Courrier des statistiques*. [en ligne]. 27 juin 2019. Insee. N° N2, pp. 33-45. [Consulté le 16 décembre 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/4168394/courstat-2-5.pdf*.

CLANCHÉ, François, CHAMBAZ, Christine, LETURCQ, Fabrice, LIXI, Clotilde, MAHUZIER, Ombeline, TURNER, Laure et VIARD-GUILLOT, Louise, 2016. *Pour une méthodologie d'analyse comparée des statistiques Sécurité et Justice : l'exemple des infractions liées aux stupéfiants.* [en ligne]. Décembre 2016. SSMSI-SDSE-DACG. Interstats Méthode N°8. [Consulté le 16 décembre 2021]. Disponible à l'adresse : *https://www.interieur.gouv.fr/content/download/123323/988596/file/IM8.pdf*.

ESTIVAL, Alexandre et FILATRIAU, Olivier, 2019. La mesure statistique de la délinquance. In : *Dalloz AJ pénal*. [en ligne]. Avril 2019. pp. 224-231. [Consulté le 16 décembre 2021]. Disponible à l'adresse : *https://www.interieur.gouv.fr/content/download/119499/958322/file/Article%20Dalloz%20AJ%20Penal%2004_2019.pdf*.

EUROSTAT, 2017. *EU guidelines for the International Classification of Crime for Statistical Purposes – ICCS.* [en ligne]. Octobre 2017. Manual and Guidelines. [Consulté le 16 décembre 2021]. Disponible à l'adresse : *https://ec.europa.eu/eurostat/documents/3859598/8305054/KS-GQ-17-010-EN-N.pdf/feefb266-becc-441c-8283-3f9f74b29156?t=1507884966000*.

GUIBERT, Bernard, LAGANIER, Jean et VOLLE, Michel, 1971. Essai sur les nomenclatures industrielles. In : *Économie et Statistique*. [en ligne]. Février 1971. N°20, pp. 23-36. [Consulté le 16 décembre 2021]. Disponible à l'adresse : *https://www.persee.fr/doc/estat_0336-1454_1971_num_20_1_6122*.

INSEE, 2008. Dossier spécial Nomenclatures. In : *Courrier des Statistiques*. [en ligne]. Novembre-décembre 2008. N° 125. [Consulté le 16 décembre 2021]. Disponible à l'adresse : *https://www.bnsp.insee.fr/ark:/12148/bc6p06xt47h*.

MINISTÈRE DE L'INTÉRIEUR, 2021. *La nomenclature française des infractions (NFI).* [en ligne]. 2 décembre 2021. [Consulté le 16 décembre 2021]. Disponible à l'adresse : *https://mobile.interieur.gouv.fr/Interstats/Actualites/La-nomenclature-francaise-des-infractions-NFI*.

NASEM, 2016. *Modernizing Crime Statistics: Report 1: Defining and Classifying Crime*. [en ligne]. Washington, DC, USA.The National Academies Press. [Consulté le 16 décembre 2021]. Disponible à l'adresse : *https://doi.org/10.17226/23492*.

ONUDC, 2015. *Classification internationale des infractions à des fins statistiques. Version 1.0*. [en ligne]. Mars 2015. [Consulté le 16 décembre 2021]. Disponible à l'adresse : *http://www.unodc.org/documents/data-and-analysis/statistics/crime/ICCS/ICCS_French_2016_web.pdf*.

ONUDC, 2019. *ICCS Implementation manual.* [en ligne]. Mars 2019. [Consulté le 16 décembre 2021]. Disponible à l'adresse : *https://www.unodc.org/unodc/en/data-and-analysis/statistics/iccs.html*.

PERROT, Michelle, 1976. Premières mesures des faits sociaux : les débuts de la statistique criminelle en France (1780-1830). In : *Pour une histoire de la statistique. Contributions Insee*. Economica/Insee, Tome 1, pp. 125-137. ISBN 2-7178-1260-1.

NOTES

Présentation du numéro N7

Septième numéro et troisième anniversaire pour la revue depuis sa renaissance. L'ambition est toujours d'y aborder un large panel des problématiques de la statistique publique. Sur une tonalité pédagogique, il s'adresse au statisticien, débutant ou expert, à l'étudiant et à l'enseignant, comme au citoyen que la « fabrique » des statistiques intéresse.
Les deux premiers articles traitent de l'intégration du multimode dans les enquêtes, abordant les questions de méthodes et d'outils pour tirer parti de cette nouvelle approche de la collecte de données. Une grande opération statistique se modernise : le recensement agricole est désormais en multimode. Les sources administratives exhaustives sont plus accessibles, mais sont-elles pour autant faciles à mobiliser ? Un exemple avec l'analyse fine du patrimoine immobilier des ménages.
Si la donnée forme la tonalité de ce numéro, une large place y est faite aux instruments qui la rendent exploitable et audible. La maîtrise du *Cloud computing* et des techniques de développement informatique sont mises en avant pour veiller à qualité de la production statistique. Le statisticien doit aussi être en capacité de jouer de concert avec d'autres disciplines académiques, comme la psychométrie dans l'évaluation des compétences des élèves. Enfin, la mise au point d'une nomenclature sur les infractions illustre l'utilité d'adopter un solfège commun pour ranger, classer et analyser les données.

ISSN 2107-0903
ISBN 978-2-11-1623453
9782111623453

Insee
Mesurer pour comprendre

STATISTIQUE PUBLIQUE