



N 4

COURRIER DES STATISTIQUES

Juin 2020

Rédaction en chef

Odile Rascol

Contribution

Insee : Christophe Alviset,
Didier Blanchet, Sébastien Faivre,
Pierre Lamarche, Nicolas Paliod,
Michael Sicsic, Patrick Sillard,
Ludovic Vincent

Cnam : Thomas Amossé

Drees : Pierre Cheloudko, Simon Fredon

SCB (Statistics Sweden) : Johan Erikson

Sdes : Jérôme Harnois, Henri Martin

Directeur de la publication

Jean-Luc Tavernier

Directeur de la collection

Pascal Rivière

Rédaction

Maryse Cadalanu, Pierre Glénat,
Odile Rascol, Pascal Rivière,
Nelly Verrière

Cartographie

Claire Godin, Thibaut Louza

Composition

Agence **LATITUDE** Nantes
5, rue Jacques Brel
« Les Reflets » Bâtiment A
44800 SAINT-HERBLAIN

0093/20

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

© Insee 2020 « Reproduction partielle
autorisée sous réserve de la mention
de la source et de l'auteur ».

Courrier des statistiques N4

SOMMAIRE

Présentation du numéro <i>Odile Rascol</i>	4
Des modèles de microsimulation dans un institut de statistique : pourquoi, comment, jusqu'où ? <i>Didier Blanchet</i>	6
Une décennie de modélisation du système de retraite : la genèse du modèle de microsimulation TRAJECTOIRE <i>Pierre Cheloudko et Henri Martin</i>	23
Ines, le modèle qui simule l'impact des politiques sociales et fiscales <i>Simon Fredon et Michaël Sicsic</i> ...	42
La nomenclature socioprofessionnelle 2020 : continuité et innovation, pour des usages renforcés <i>Thomas Amossé</i>	62
Pour les enquêtes auprès des ménages, l'Insee rénove ses échantillons <i>Patrick Sillard, Sébastien Faivre, Nicolas Paliod et Ludovic Vincent</i>	81
La troisième refonte du répertoire Sirene : trop ambitieuse ou pas assez ? <i>Christophe Alviset</i>	101
Le modèle de processus statistique en Suède : mise en œuvre, expériences et enseignements <i>Johan Erikson</i>	122
Le système statistique du logement : étendue et perspectives <i>Jérôme Harnois et Pierre Lamarche</i>	142

PRÉSENTATION DU NUMÉRO

Le Courrier des statistiques, dans sa nouvelle formule, publie son numéro « N4 » : N comme nouveau, N comme numérique. L'ambition de 2018 est maintenue, avec pour objectif d'explorer les différents aspects des métiers et méthodes de la statistique, qu'elle soit publique ou non, et avec plusieurs volontés. Une exigence de pédagogie tout d'abord : les sujets sont vite complexes, techniques, or les articles visent non seulement les statisticiens publics, mais aussi, potentiellement, les étudiants, universitaires, et plus généralement le « citoyen éclairé », qui doit être à même de comprendre les principes des méthodes utilisées. Une indispensable ouverture ensuite, en invitant des instituts statistiques étrangers, des producteurs de données intéressantes pour un usage statistique, des universitaires concernés par les questions de traitement et de qualité de la donnée. Dans la même veine, nous veillons à ce que les articles sortent de leur zone de confort, par exemple en insistant sur la nécessité de références, notamment au niveau international.

Ce quatrième numéro du Courrier des statistiques s'ouvre sur un sujet qui n'a pas été abordé jusque-là par la revue, avec trois articles consacrés à la microsimulation. On peut se demander si les méthodes de la microsimulation relèvent bien du domaine de la statistique publique, et c'est justement une des questions que soulève le premier article. **Didier Blanchet** y définit les notions de microsimulation, statique ou dynamique, la façon de procéder en pratique, et les principales composantes d'un modèle de microsimulation. Il y explicite les raisons pour lesquelles en France, cette méthode s'est imposée naturellement dans la statistique publique. Il s'agit donc d'un article général qui, même s'il cite souvent le modèle Destinie, pose en quelque sorte un cadre, indispensable pour la suite. Les deux autres papiers portent en effet à chaque fois sur un modèle de microsimulation spécifique.

Ainsi, **Pierre Cheloudko et Henri Martin** racontent l'histoire du modèle TRAJECTOIRE, qui modélise le système de retraite et inclut ainsi inévitablement une forte dimension temporelle. Le modèle est dynamique, et avec lui le mot simulation recouvre ainsi à la fois la simulation (déterministe) d'une législation et la simulation (aléatoire) d'une carrière future. L'article montre les étapes ayant permis d'aboutir au dispositif actuel, en commençant par un modèle agrégé portant sur des cas-types, puis un deuxième bloc de modélisation des départs à la retraite, apportant la dimension dynamique, et enfin TRAJECTOIRE qui modélise l'intégralité des carrières, avec des données au niveau individuel.

Simon Fredon et Michaël Sicsic présentent Ines, modèle statique qui permet de simuler les politiques sociales et fiscales, tout en intégrant des comportements. Ils mettent en évidence les différents axes de travail lorsqu'on élabore et maintient un modèle de microsimulation : suivre de près et prendre en considération l'évolution de la législation, se constituer des bases de données solides à partir de plusieurs sources, caractériser les dispositifs à simuler. On constate également la puissance de tels modèles, dont les usages sont variés : effet d'une hausse ou baisse des transferts sociaux et fiscaux, effets des réformes sur les inégalités, etc.

Les autres articles de ce numéro ont un point commun : d'une façon ou d'une autre, ils portent sur des « références » de la statistique publique.

La nomenclature des professions et des catégories socio-professionnelles tout d'abord : cet instrument majeur pour les statistiques sociales vient de faire l'objet d'une vaste refonte, ce qui constitue un événement rare. **Thomas Amossé** en narre la genèse, après avoir rappelé l'historique des classifications socio-professionnelles et posé le concept de nomenclature. Il ne s'agit pas juste d'une nouvelle version, mais aussi d'une nouvelle logique, qui tient

compte du fait que les usages sont multiples : elle doit donc être suffisamment plastique pour s'adapter à ces différents usages.

Le répertoire Sirene est aussi un référentiel, un pivot de la production statistique d'entreprise, mais aussi de la simplification administrative, avec son identifiant unique qui structure tout l'écosystème administratif autour des entreprises. Contrairement à l'article précédent où un résultat abouti était présenté, l'article de **Christophe Alviset** se situe en amont, en présentant le programme Sirene 4. Le projet ne vise pas à revoir Sirene de fond en comble, mais plutôt à l'améliorer sur un certain nombre de sujets essentiels en vue d'une industrialisation des traitements : identification automatique d'établissement, codification automatique d'activité, normalisation des échanges de données, contrôles. L'enjeu sera aussi d'accompagner la révolution numérique de l'administration au service des entreprises.

La nécessité de références vaut non seulement pour les données, mais aussi pour les processus de production statistique. Depuis une dizaine d'années, le GSBPM s'est imposé comme une norme au niveau européen, désormais largement partagée. Dans ce numéro N4, le choix a été fait de donner la parole à un institut statistique étranger, en l'occurrence le *Statistiska centralbyrån* (SCB) en Suède, car il a appliqué de manière très vaste et systématique une démarche très voisine de celle portée par la communauté statistique internationale avec le GSBPM. **Johan Erikson** y explique les riches enseignements tirés de cette pratique : les facteurs de réussite, les freins et les évolutions envisagées.

Les enquêtes auprès des ménages, lorsqu'elles sont effectuées en face-à-face, requièrent des techniques de tirages d'échantillon très sophistiquées liées à la nécessité de faciliter au plan pratique le déplacement des enquêteurs, tout en respectant les contraintes de maîtrise de probabilités d'inclusion et d'optimisation de la précision. Pour éviter de se poser de complexes questions méthodologiques lors de chaque tirage, on se constitue une référence, que l'on conserve dix ans, et au sein duquel on tire les échantillons pour toute nouvelle enquête en face-à-face : l'échantillon-maître. Comme l'expliquent **Patrick Sillard et alii**, ce dernier est désormais opérationnel depuis 2020. La nouveauté réside dans l'utilisation des sources fiscales et la coordination avec l'échantillon de l'enquête Emploi.

Le dernier article, contrairement aux précédents qui portent sur des référentiels transversaux, s'intéresse à un domaine spécifique : le logement. Et là aussi il s'agit de s'interroger sur la référence : quelles sont les données et statistiques de référence ? **Jérôme Harnois et Pierre Lamarche** abordent la question sous toutes ses facettes : la construction, l'équipement en logement, le prix, le coût d'occupation, l'état des logements et le mal-logement, etc. L'article ne se limite donc pas à pointer telle ou telle enquête ou source administrative : il adopte un point de vue original en visant une démarche systémique, structurée par le cycle de vie du logement, mais également contrainte par les limites des sources de données disponibles.

Odile Rascol
Rédactrice en chef, Insee

DES MODÈLES DE MICROSIMULATION DANS UN INSTITUT DE STATISTIQUE

POURQUOI, COMMENT, JUSQU'OU ?

*Didier Blanchet**

Les modèles de microsimulation sont incontournables pour l'analyse des politiques de transferts sociaux et fiscaux. Il peut s'agir de modèles statiques, centrés sur l'effet des transferts à date donnée, ou bien dynamiques, ce second type de modélisation s'imposant pour les politiques dont les effets se déploient dans la durée, comme le font les réformes des retraites.

Si l'usage de ces modèles est très répandu, il n'est pas systématique de les voir mis en œuvre dans des instituts de statistique. Le développement de ces outils au sein de l'Insee s'explique par le fait que l'institut est non seulement producteur de données brutes mais aussi producteur d'études et de projections. Analyser les politiques de redistribution et en prévoir l'impact à long terme est donc tout à fait dans ses missions.

Mais la microsimulation présente également des synergies intéressantes avec la production statistique proprement dite. On propose aussi quelques pistes de développement pour consolider la place de ces modèles au sein de l'ensemble des outils d'évaluation des politiques publiques.

 *Microsimulation models are key to the analysis of tax and social transfer policies. This may involve static models, focused on the effect of transfers as at a given date, or else dynamic models, with this second type of modelling being essential for policies with long-term effects, as is the case with pension reforms.*

Although these models are in widespread use, they are not automatically implemented in national statistical offices. The development of these tools within INSEE can be explained by the fact the institute produces not only raw data but also studies and projections. Analysing redistribution policies and forecasting their long-term impact is therefore firmly within its remit.

But microsimulation also offers interesting synergies with the actual production of statistics. So, we propose various avenues of development to strengthen the role of these models within the set of tools used to assess public policies.

* Directeur des études et synthèses économiques (Dese), Insee,
didier.blanchet@insee.fr

1 LA MICROSIMULATION : DEUX GRANDES CATÉGORIES DE MODÈLES

Il y a maintenant plus d'une vingtaine d'années que le recours à la microsimulation s'est développé à l'Insee, et son usage s'est aussi répandu dans une partie du système statistique public et d'autres administrations proches de l'Insee, principalement dans le domaine social. S'agissant de l'Insee, deux modèles coexistent et ont chacun su accéder à une forte notoriété dans leurs champs respectifs : le modèle Ines¹, appliqué à l'étude des transferts fiscaux et sociaux, et le modèle Destinie², spécialisé sur le domaine des retraites et autres sujets impactés par le vieillissement de la population (Bardaji *et alii*, 2003 ; Blanchet *et alii*, 2011). Ces modèles sont représentatifs de deux types de microsimulation, la microsimulation statique pour Ines, la microsimulation dynamique pour Destinie.

Que recouvrent ces deux termes ? La microsimulation statique appliquée à la fiscalité ou aux transferts sociaux, c'est s'appuyer sur un fichier de données individuelles à date donnée et appliquer aux revenus de ces individus les formules de calcul des prélèvements ou transferts fiscaux et sociaux. On chiffre ainsi à la fois qui sont les contributeurs ou bénéficiaires *in fine* de ces transferts et comment leurs profils évolueraient suite à des modifications de ces transferts. Dans sa forme la plus élémentaire, la démarche consiste donc uniquement en l'application comptable de barèmes réglementaires, ce qui n'empêche pas ces modèles d'être lourds à construire et à manipuler. Ils peuvent éventuellement être enrichis par la prise en compte de réponses comportementales à ces changements de règles, par exemple des modifications de l'offre de travail – on y reviendra en conclusion –, mais l'optique reste néanmoins une approche de statique comparative : on compare des photos de l'état du monde avec ou sans ces changements de politique (*figure 1*).

La microsimulation dynamique part du même genre de photo de groupe instantanée, mais la photo n'est que la première image d'un film au cours duquel les positions des individus vont évoluer, incluant la disparition de certains d'entre eux par décès et l'apparition de nouveaux individus assurant le renouvellement de la population simulée, comme dans la vraie vie (*figure 2*). C'est ainsi que Destinie et les autres modèles de même type simulent le vieillissement des personnes déjà retraitées, l'avancée dans la carrière d'actifs qui deviennent retraités à leur tour, et l'arrivée de nouveaux cotisants. Il le fait à coup d'hypothèses sur les comportements démographiques, les mobilités salariales et les changements de statut sur le marché du travail ainsi que sur les comportements de départ en retraite, en appliquant des cadres législatifs qui peuvent être ceux issus des réformes des retraites déjà actées, ou de nouveaux *scenarii* de réforme.

1 POURQUOI LE RECOURS À CES MÉTHODES ?

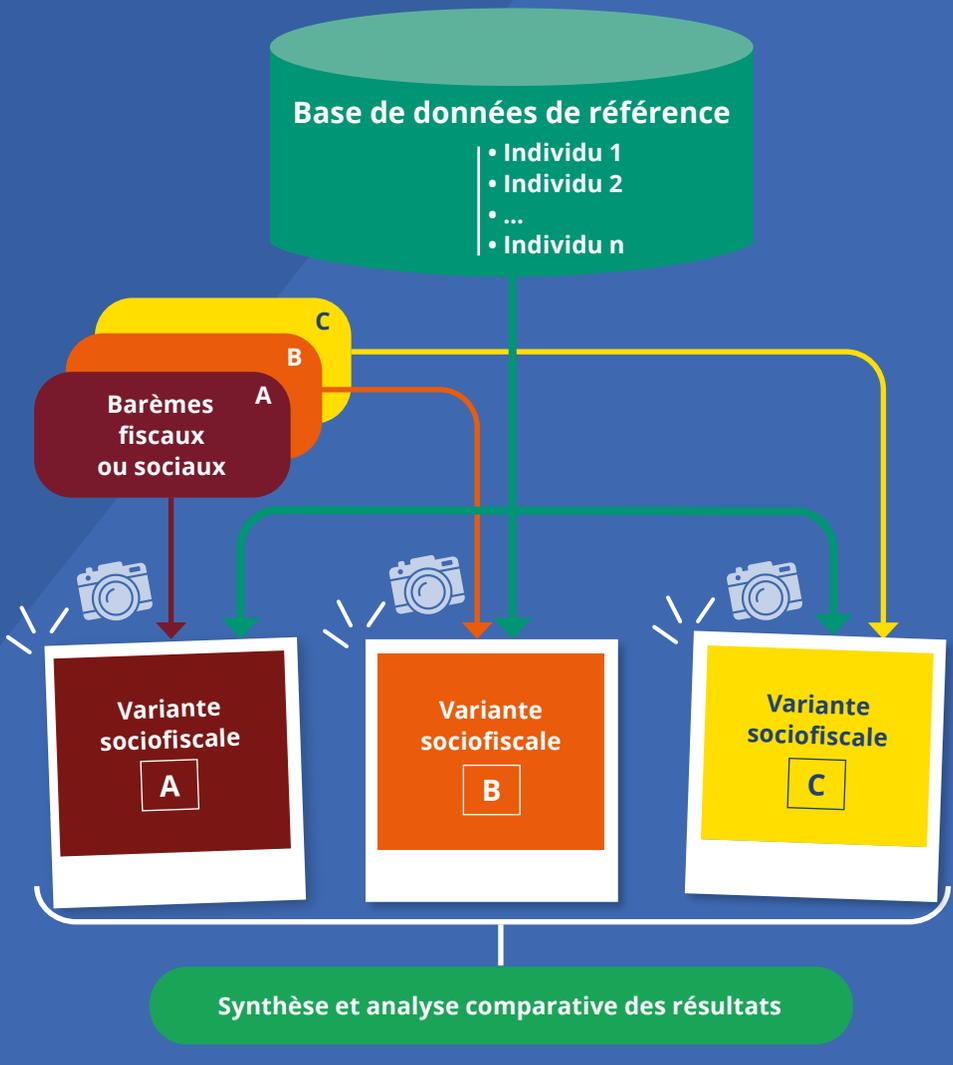
De telles techniques sont mobilisées dans de nombreuses disciplines : démographie, épidémiologie, modélisation des systèmes urbains ou de réseaux de transports, etc. Dans le domaine économique et social, la paternité de l'idée est en général attribuée à Guy H. Orcutt, plus connu des économètres pour la méthode de correction de l'auto-corrélation qui porte son nom, mais qui conçoit aussi, dès les années 1950, un projet de modélisation globale qui est d'entrée de jeu de type dynamique : une modélisation représentative de l'ensemble de

-
1. Ines est l'acronyme d'« Insee-Drees », les deux organismes qui développent conjointement le modèle (la Drees est le service statistique du ministère des Solidarités et de la Santé). Voir l'article de Simon Fredon et Michael Sicsic dans ce même numéro.
 2. Destinie : modèle Démographique Économique et Social de Trajectoires INDIVIDUELLES simulées.

l'évolution du système socio-économique, celles des individus et des ménages auxquels ils appartiennent, mais aussi celle des entreprises qui les emploient, et même celle des institutions publiques interférant avec les comportements de l'ensemble de ces acteurs (Orcutt, 1957). L'idée était de se démarquer des modèles macroéconomiques qui commençaient alors à se développer, en rappelant les limites de leurs raisonnements en termes de grandes catégories d'agents représentatifs.

Figure 1. La microsimulation statique appliquée à la fiscalité ou aux transferts sociaux

Dans une microsimulation statique, on compare des « photos » de l'état du monde avec ou sans changements de politique.



L'objectif d'Orcutt était très ambitieux pour l'époque. Il a essayé de lui donner corps tout au long de la suite de sa carrière (Watts, 1991) mais ce sont surtout les modèles statiques qui ont d'abord connu le plus fort développement. Leur utilité pour l'analyse des politiques de transferts fiscaux et sociaux va en effet de soi. On y est d'ailleurs souvent venu spontanément d'une manière totalement indépendante des propositions d'Orcutt. Dans ce domaine, l'approche agrégée des modèles macroéconomiques n'est certes pas sans intérêt, bien au contraire : ils permettent d'explorer l'impact de transferts de structure pas trop complexe, avec l'avantage de prendre en compte les effets d'équilibre général, tels que l'impact des transferts sur l'équilibre offre-demande ou sur le marché du travail. Mais ils ne seront jamais en mesure de simuler les effets détaillés de politiques dont les règles dépendent d'un grand nombre de caractéristiques individuelles, très souvent non linéaires, par exemple en raison d'effets de seuil ou de plafond ; le barème de l'impôt sur le revenu constitue l'exemple typique bien connu de tous. C'est cela que les modèles de microsimulation statique permettent de faire.

Ils auront pour limite symétrique de proposer des analyses en équilibre partiel, car ce qui y arrive à chaque agent est supposé sans effet sur les autres agents et donc sur le reste de l'économie. Mais au moins permettent-ils d'avoir des évaluations précises de ces effets dits « de premier tour », avant bouclages ou réactions comportementales. On ajoute que ce qui justifie ce recours à la microsimulation n'est pas uniquement l'intérêt pour les effets redistributifs : il est tout aussi nécessaire pour un chiffrage précis des coûts ou des bénéfices budgétaires globaux de ces politiques.

Les motivations ont été les mêmes pour la construction de modèles dynamiques appliqués aux retraites, Destinie n'ayant fait qu'embrayer sur une pratique déjà répandue dans d'autres pays. Projeter le financement d'un système de retraite et ce qu'il offre comme

« Projeter le financement d'un système de retraite et ce qu'il offre comme droits est possible sans descendre à un niveau de granularité très fin, tant qu'on est dans un travail de dégrossissage. »

droits est possible sans descendre à un niveau de granularité très fin, tant qu'on est dans un travail de dégrossissage consistant à envisager à grands traits les conséquences du vieillissement démographique. C'est ainsi qu'on avait procédé jusqu'au début des années quatre-vingt-dix, période de la rédaction du Livre blanc sur les retraites³, avec des modèles consistant à plaquer une représentation simplifiée du système de retraite sur des projections démographiques de type traditionnel qui

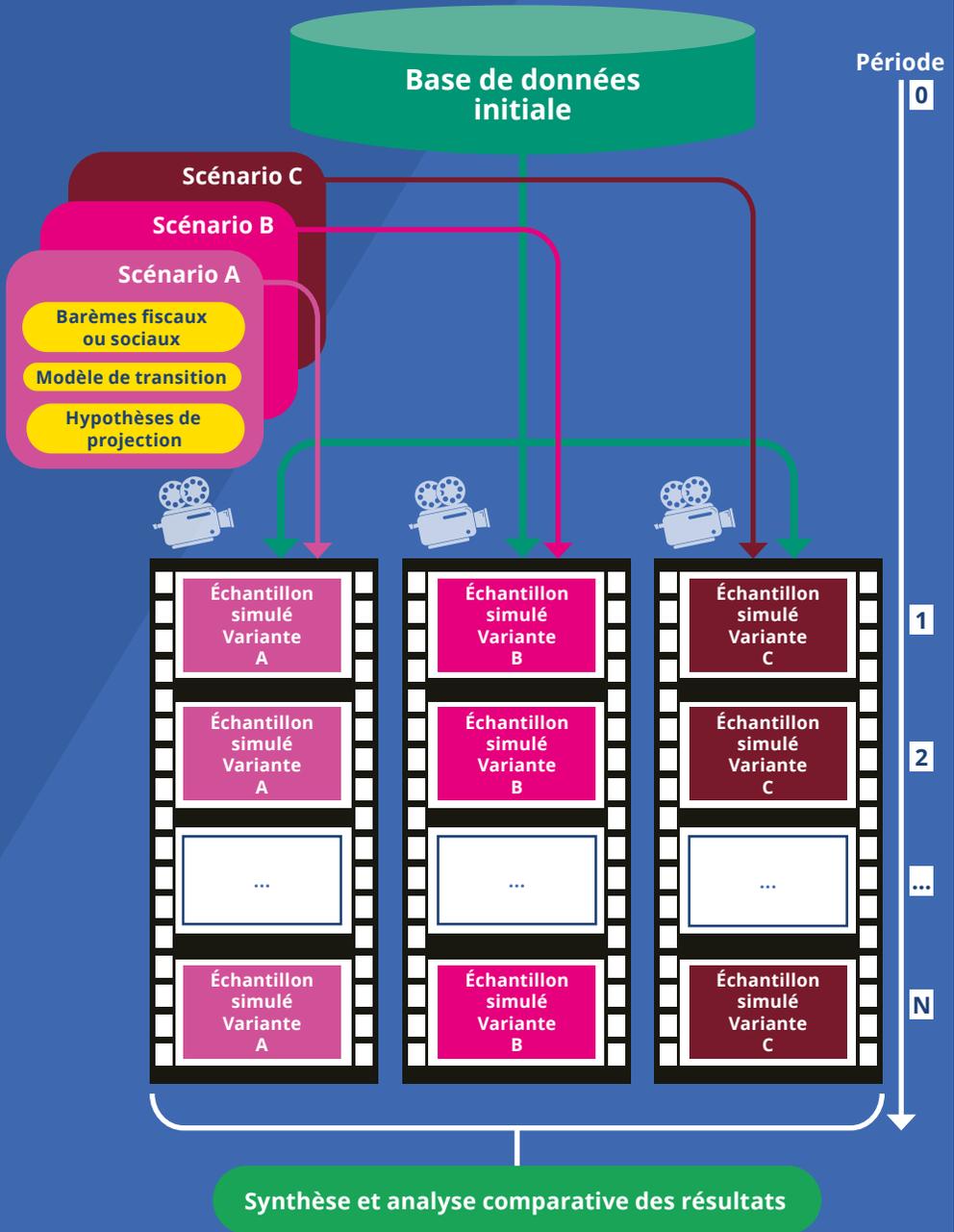
ne désagrègent la population que selon les deux critères de l'âge et du sexe. Procéder ainsi était suffisant pour simuler les effets globaux du vieillissement sur l'équilibre des retraites ou celui de décalages d'ensemble de l'âge du départ en retraite.

Mais la réforme de 1993 qui a commencé à répondre aux questions soulevées par ce Livre Blanc a obligé à changer d'échelle. Comment simuler en effet l'impact d'un calcul de la retraite sur les 25 plutôt que les 10 meilleures années sans raisonner individu par individu, compte tenu des effets très variables qu'a cette règle selon les types de carrière ?

3. Le Livre blanc sur les retraites d'avril 1991 (www.vie-publique.fr/sites/default/files/rapport/pdf/134000051.pdf) présentait l'ensemble des régimes de retraite, leurs perspectives d'évolution ainsi que des propositions pour gérer les effets du vieillissement démographique.

Figure 2. La microsimulation dynamique appliquée à la fiscalité ou aux transferts sociaux

Dans une microsimulation dynamique, on visualise l'évolution de la population échantillonnée sur une période. Les différents scénarii du modèle déterminent ainsi autant de versions des « films » qu'on pourra comparer ou synthétiser.



Et comment simuler l'impact du passage de 37,5 à 40 années de cotisation pour l'obtention du taux plein, sans connaissance de la distribution détaillée des durées cotisées à l'intérieur de chaque génération ? Multiplier l'examen de cas-types n'aurait fourni que des réponses partielles.

LE RECOURS À LA MICROSIMULATION DYNAMIQUE S'EST PETIT À PETIT IMPOSÉ

Répondre à ces questions par le recours à la microsimulation n'a pas été sans susciter certaines réticences. On a pu notamment s'inquiéter de la complexité et donc du manque de transparence de ce type d'outils. On a pu aussi s'inquiéter de voir la prospective s'appuyer sur des modèles aux résultats affectés d'un bruit stochastique⁴. La microsimulation dynamique c'est en effet faire évoluer les situations d'un échantillon d'individus à coups de tirages aléatoires selon les probabilités spécifiées par le modèle, puis faire confiance à la loi des grands nombres pour considérer que les résultats auxquels on arrive donnent une bonne image des évolutions à des niveaux plus agrégés, mais qui sera fatalement affectée d'un aléa de simulation qui n'existe pas dans les modèles déterministes agrégés.

Face à cette dernière réticence, on a pu faire valoir que cette forme d'imprécision était bien inférieure aux autres sources d'incertitude qui affectent toute projection de long terme. Il y avait du reste un peu de paradoxes à redouter cet élément d'aléa du modèle alors que nous sommes tout à fait habitués à ce que l'observation du présent soit elle aussi sujette à fluctuation d'échantillonnage : on admet que le taux de chômage courant ne soit mesuré qu'à quelques décimales près, pourquoi cette même marge d'imprécision aurait-elle été intolérable pour des projections des dépenses de retraite à 2040 ? Quant au défaut de transparence, les modèles ainsi construits sont certes difficilement appropriables, et le reproche d'opacité n'est donc pas totalement infondé. Mais c'est oublier que la transparence apparente des modèles plus simples se paye d'un grand nombre d'hypothèses simplificatrices

aux conséquences tout aussi peu transparentes : l'individu représentatif des modèles macroéconomiques est une abstraction qui, à l'usage, peut se révéler bien plus opaque que le fait d'agréger les simulations de milliers d'individus qu'on s'est efforcé de faire ressembler le plus possible aux individus du monde réel.

« À la question du « pourquoi » de la microsimulation, la réponse est tout simplement qu'il n'y a pas d'autre façon de procéder pour modéliser des systèmes complexes. »

Au final la pratique a donc fini par être acceptée et assumée. À la question du

« pourquoi » de la microsimulation, la réponse est tout simplement qu'il n'y a pas d'autre façon de procéder pour modéliser des systèmes complexes.

Mais ceci ne répond pas à la question du « pourquoi dans un institut de statistique ? ». Car d'autres lieux d'accueil naturel existent pour ce genre d'exercice, qu'il s'agisse du monde de la recherche, ou des administrations en charge des programmes sociaux concernés, pour leurs besoins de gestion de court-terme et de prévision à long terme.

4. i.e. le bruit dû au fait que la simulation utilise des répétitions de tirages au sort, qui est davantage que de l'aléa d'échantillonnage *stricto sensu*.

❶ POURQUOI DANS UN INSTITUT NATIONAL DE STATISTIQUE ? —

De fait, la microsimulation s'était aussi développée dans ces autres lieux avant d'arriver à l'Insee, et elle a continué de s'y développer après⁵ :

- ❶ en matière de microsimulation statique avait préexisté la double expérience du modèle Sysiff au Delta⁶ (Bourguignon *et alii*, 1988) et du modèle MIR à la direction de la Prévision. Sysiff peut être vu comme l'ancêtre du simulateur fiscal de Landais, Piketty et Saez devenu le modèle TaxIPP de l'Institut des Politiques Publiques (IPP) (Landais *et alii*, 2011). La tradition initiée par MIR s'est poursuivie au Trésor, qui dispose actuellement du modèle Saphir, et la Caisse Nationale d'Allocations Familiales a développé et longtemps exploité son propre modèle, Myriade, avant de finalement se joindre au développement et à l'utilisation d'Ines ;
- ❶ côté microsimulation dynamique, c'est plutôt l'Insee qui a été précurseur pour la France, mais la technique s'y est depuis propagée : dans un premier temps à la Caisse Nationale d'Assurance Vieillesse avec le modèle Prisme (Poubelle *et alii*, 2006), puis à la Drees avec le modèle TRAJECTOIRE⁷ et au Trésor avec le modèle Aphrodite, avant que d'autres caisses n'emboîtent le pas (le modèle Pablo du Service des retraites de l'État, le modèle Canopée développé par la Caisse des dépôts et consignations pour la CNRACL⁸, et également un modèle en cours de développement à l'Agirc-Arrco⁹), et enfin que le monde académique ne s'y lance également, avec le modèle PensIPP de l'IPP.

Dans la plupart des pays étrangers, c'est plutôt dans ces deux types de cadres, universitaire ou administratif, que se sont développés ces modèles.

Pourquoi une exception française ? Signalons pour être exact que l'exception n'est pas totale. La microsimulation s'était aussi développée à Statistique Canada, et à *Statistics*

“ Un usage assez dominant des modèles de microsimulation est la simulation *ex ante* de réformes, avec une mobilisation directe pour la préparation des politiques économiques. ”

Norway. Dans le cas de la France et de la Norvège, cette originalité s'explique évidemment par le fait que les instituts nationaux statistiques (INS) des deux pays ont pour caractéristique de combiner les fonctions de production statistique et d'étude. On peut donc considérer que la microsimulation s'est développée à l'Insee dans le cadre de sa mission d'études et de diagnostic.

Mais l'argument doit être précisé et complété, car on a bien compris de tout ce qui précède qu'un usage assez dominant des modèles de microsimulation est la simulation *ex ante* de réformes, avec une mobilisation directe pour la préparation des

politiques économiques. Or il s'agit d'un terrain sur lequel la statistique publique s'interdit en principe d'intervenir : son rôle est d'analyser et de décrire ce qui est ou a été, pas de dire ou de participer à la conception de ce qui devrait être. Ceci doit-il fragiliser la place qu'y ont acquise ces modèles de microsimulation ?

5. Sur tout cet historique voir (Legendre, 2019) : on y renvoie pour des références bibliographiques plus systématiques sur l'ensemble des modèles cités dans cet article.

6. Unité mixte de recherche ENS-CNRS-EHESS.

7. Voir l'article de Pierre Cheloudko et Henri Martin dans ce même numéro.

8. CNRACL : Caisse nationale de retraites des agents des collectivités locales.

9. L'Agirc (Association générale des institutions de retraite complémentaire des cadres) et l'Arrco (Association pour le régime de retraite complémentaire des salariés) gèrent des régimes de retraite complémentaires.

Plusieurs arguments plaident en sens inverse. Ils viennent très facilement à l'esprit pour la microsimulation statique. Ce n'est certes pas l'Insee qui est en charge de simuler les réformes fiscales à venir. L'institut évite même d'intervenir trop précocement sur les mesures les plus récentes. Mais, pour sa mission de diagnostic sur l'évolution de la société française, il a besoin d'évaluer ce qu'a été l'incidence des mesures passées.

Or la microsimulation est un outil qui sert aussi bien à l'évaluation *ex ante* qu'à tous les types d'évaluations *ex post*, à chaud comme à froid. C'est par la microsimulation qu'on est le plus facilement en mesure de retracer en détail ce qu'a été la contribution des mesures

« C'est le terme de *nowcasting* qui désigne cette pratique de combiner informations partielles et modèles pour donner de premières estimations de ce qui est en train de se produire. »

fiscales ou sociales passées au niveau de vie des différentes catégories de population. Elle peut le faire de manière précise lorsqu'on a des données microéconomiques complètes sur les revenus avant transferts, et elle permet aussi de le faire de façon avancée avant que ces données détaillées ne soient totalement disponibles. C'est de cette façon qu'Ines a permis de répondre à la demande d'indicateurs avancés de la répartition des revenus et du taux de pauvreté. Le recours à la microsimulation apparaît ici comme un instrument parmi d'autres de la production statistique. On peut la voir comme une extension de la pratique de l'imputation qui est

utilisée pour combler les lacunes temporaires ou définitives de l'observation directe. On peut tout aussi bien y voir l'application à la statistique sociale de l'approche par modèles, mobilisée de longue date pour l'estimation avancée des comptes nationaux : dans un cas comme dans l'autre, c'est le terme de *nowcasting* qui désigne cette pratique de combiner informations partielles et modèles pour donner de premières estimations de ce qui est en train de se produire.

L'argument peut sembler moins convaincant pour la microsimulation dynamique et la projection des retraites. Là, on est clairement dans l'exploration du futur et même d'un futur très lointain. Il se trouve néanmoins que l'exploration du futur fait elle aussi partie du portefeuille d'activités de l'Insee, ou tout du moins de certains aspects de ce futur.

Il y a des domaines dans lesquels cette exploration du futur ne se hasarde pas au-delà du très court terme, celui qui est couvert par la Note de conjoncture¹⁰, parce que l'institut considère que ses outils ne lui permettent pas mieux que de la prévision à six mois, pour des variables qu'on sait affectées d'une forte volatilité.

Mais les phénomènes démographiques sont caractérisés par une bien plus forte inertie et répondent à une logique très mécanique qui rend tout à fait possible la prévision de long terme, ou plus exactement la réalisation de « projections », c'est-à-dire la simulation de comment sont appelés à évoluer l'effectif total et la structure de la population selon que c'est tel ou tel scénario de mortalité, de fécondité et de flux migratoire qui serait appelé à se réaliser.

10. L'Insee produit une Note de Conjoncture (mars, juin, décembre) et un Point de Conjoncture (en octobre), voir <https://www.insee.fr/fr/information/1303532>.

Sur le fait de pratiquer ce style de prévision, l'Insee n'est pas du tout dans l'exception : c'est dans les INS que sont la plupart du temps réalisées les projections démographiques des États membres de l'Union européenne. Il y a par ailleurs longtemps que l'Insee a pris l'habitude de compléter ses projections de population totale par des projections de population active.

Ce que fait Destinie vient donc assez naturellement dans le prolongement de ces exercices très standard, il alimente même le second des deux pour ce qui concerne la projection des comportements d'activité aux âges élevés. Pendant longtemps, celle-ci a été faite comme

« L'impact des règles très complexes d'ouverture des droits ne peut être évalué que par des modèles qui prennent en compte le détail des situations individuelles. »

pour les autres tranches d'âge, par extrapolation logistique des chroniques de taux d'activité passée. Cette méthode est devenue clairement inadaptée une fois enclenché le processus de durcissement des conditions d'âge ou de durée pour la liquidation de la retraite. Il fallait trouver une façon de modéliser comment ces révisions avaient commencé et allaient continuer à affecter les comportements de départ en retraite. On retombe sur la raison citée en introduction pour expliquer comment le recours à la microsimulation s'est imposé pour l'ensemble des travaux sur les retraites. L'impact des règles très

complexes d'ouverture des droits ne peut être évalué que par des modèles qui prennent en compte le détail des situations individuelles. Ceci ne veut pas dire que ces modèles détiennent forcément la vérité sur les comportements futurs, mais au moins mobilisent-ils le maximum de l'information qui est disponible ou raisonnablement projetable.

Ce faisant, on dispose d'un outil qui permet de répondre à bien d'autres questions, et qui reste dans le champ de la mesure sans se hasarder du côté de la préconisation (**encadré 1**). Utiliser Destinie pour prévoir l'évolution des niveaux de vie relatifs des ménages de retraités et sa dépendance aux hypothèses de croissance économique, ce n'est pas contribuer à l'élaboration des réformes futures, c'est éclairer sur les effets de réformes déjà actées. Utiliser les trajectoires individuelles que le modèle est amené à reconstituer pour évaluer les propriétés redistributives intragénérationnelles du système actuel, ce n'est ni plus ni moins qu'une prolongation sur le cycle de vie des exercices d'évaluation de la redistribution instantanée menés en statique par le modèle Ines.

LE « COMMENT » : QUELLES SOURCES ?...

Il y a d'autres recoupements, similitudes et synergies avec l'activité de production statistique. On aborde ici la question du « comment ».

Selon qu'il est statique ou dynamique, un modèle de microsimulation a deux ou trois grandes composantes :

- ① un ou des modules reproduisant les législations qu'on cherche à simuler ;
- ① une base de données individuelles ;
- ① et, dans le cas de la microsimulation dynamique, le ou les modules de projection des situations individuelles, avec leurs paramétrages.

Reproduire les législations à simuler ne relève pas de la compétence statistique proprement dite. À la limite, lorsque c'est possible, un modèle de microsimulation peut directement s'appuyer sur les outils informatiques de calcul des droits utilisés par les administrations en charge d'appliquer ces législations.

Encadré 1. Destinie : trois apports à l'analyse du système de retraite

Les propriétés redistributives du système de retraite peuvent passer par deux canaux : des mécanismes « non contributifs » dont la redistribution est la raison d'être (*minima* de pension, validation des périodes de chômage, droits familiaux, etc.) ; ou les effets redistributifs du cœur « contributif » du système : ainsi, la règle des 25 meilleures années est jugée redistributive, car gommant l'impact des accidents de carrière sur le montant des retraites. Destinie fait apparaître une réalité plus complexe (Aubert et Bachelet, 2012). La prise en compte des 25 meilleures années profite surtout aux carrières longues ou ascendantes, qui ne conduisent pas aux retraites les plus petites. Le cœur du système serait plutôt anti-redistributif, *a fortiori* en tenant compte des écarts d'espérance de vie. Les mécanismes non contributifs amortissent mieux les inégalités accumulées durant la vie active. On a pu quantifier cet effet (Dubois et Marino, 2015) : pour les générations 1960 à 1970, de bas en haut de l'échelle des niveaux d'éducation, le « rendement » de la cotisation retraite serait dégressif de 1,5 % à un peu moins de 1 % pour les hommes, et de 2,5 % à 2 % pour les femmes (les avantages familiaux, des droits à réversion et une espérance de vie plus longue compensent leur désavantage en termes de salaires et de durée de carrière).

Avant la fin des années quatre-vingt, pour revaloriser les retraites ou les salaires entrant dans son calcul, on se référait à l'évolution du salaire moyen. **Le passage à l'indexation sur les prix** a eu deux effets : les salaires passés représentent un plus faible pourcentage du dernier salaire, faisant baisser le taux de remplacement initial, puis, après liquidation, le niveau de la retraite continue de décrocher par rapport au salaire moyen des actifs. Avec Destinie, on démontre à quel point cela a contribué à stabiliser la part des retraites dans le PIB (Marino, 2014) : d'environ 11 % au début des années quatre-vingt-dix, elle serait passée à 20 % en 2040 sans aucune réforme. L'indexation prix a rabattu le point d'arrivée de la courbe autour de 16 %. Mais elle génère une dépendance nouvelle aux hypothèses de croissance : si celle-ci redevenait très rapide, les retraites décrocheraient de beaucoup en proportion des salaires. À l'inverse, une productivité et des salaires réels stagnants rendraient neutre le passage de l'indexation sur les salaires à celle sur les prix. Peut-on **trouver des règles d'indexation permettant un pilotage du niveau relatif des retraites qui ne soit pas soumis aux aléas de la croissance économique** ? C'est un sujet que Destinie a également exploré (Dubois et Koubi, 2017), ainsi que, hors Insee, le modèle PensIPP (Blanchet, Bozio et Rabaté, 2016). Avec ou sans mise en place d'un système unifié, la sensibilité du système de retraite aux chocs macroéconomiques va nécessairement être un sujet tout à fait majeur de la période post Covid-19.

La stabilisation de la part des retraites dans le PIB, qui était anticipée avant le choc du Covid-19, devait résulter aussi du recul de l'âge de départ à la retraite (passage de 37,5 à 43 ans de la durée requise pour le taux plein avec les réformes de 1993, 2003 et 2014, et de 60 à 62 ans pour l'âge minimal d'ouverture des droits). Ce recul a déjà conduit à accroître le taux d'emploi des 60-64 ans, et à système inchangé, devait porter l'âge moyen de liquidation aux alentours de 64 ans en 2040. Une autre idée avait guidé l'un des aspects de la réforme de 2003 : la mise en place d'**une indexation sur l'espérance de vie**, chaque trimestre gagné se répartissant pour deux tiers/un tiers entre durée de la vie active et durée de la retraite. Destinie a servi à confirmer que la mesure ne suffisait pas à stabiliser ce ratio : elle est sans effet pour les individus validant la durée de cotisation requise dès l'âge d'ouverture des droits, ainsi que pour ceux à carrière courte et dans tous les cas contraints de liquider à l'âge maximal. La stabilisation tendancielle du ratio est liée à la remontée à 62 ans de l'âge d'ouverture des droits (Aubert et Rabaté, 2015). Sans autre réforme, le ratio resterait stationnaire dans un premier temps, puis reprendrait sa croissance après la stabilisation à 43 ans de la condition de durée, prévue pour la génération 1970.

Les deux autres volets sont en revanche très adhérents aux travaux de production statistique.

C'est clairement le cas pour ce qui est de **la constitution de la base de données de départ**. Le développement d'Ines a ainsi été étroitement lié à celui du dispositif statistique sur les revenus fiscaux et sociaux¹¹, qui a lui-même alimenté les modèles Myriade de la Drees et Saphir du Trésor. Pour les modèles de microsimulation dynamique, la prévision du futur suppose des fichiers comportant suffisamment d'information sur le passé.

La toute première version de Destinie s'appuyait sur l'enquête Patrimoine, seule source qui donnait à l'époque une information rétrospective sur la durée de carrière passée des individus, sous une forme néanmoins assez fruste. Les besoins de Destinie ont pesé en faveur de l'enrichissement de la source avec le recueil d'informations plus détaillées sur l'historique de carrière.

Destinie reste attaché à l'enquête Patrimoine, qui lui donne la possibilité de projeter non pas seulement des individus mais aussi les ménages auxquels ils appartiennent. Pour la projection des droits individuels, l'avantage est cependant du côté des panels administratifs, avec des fichiers bien plus complets et aux données bien plus précises. Le modèle TRAJECTOIRE est ainsi étroitement lié à l'échantillon inter-régimes de cotisants (EIC) mis au point par la Drees¹². Avoir un panel qui ne donne que des durées de cotisation passées n'est que d'une utilité très relative pour alimenter les constats sur la retraite, il ne trouve tout son intérêt que prolongé par des durées probables de cotisation futures, pour le calcul des droits afférents : c'est ce que TRAJECTOIRE apporte à l'EIC. En quelque sorte, le modèle est le prolongement logique et nécessaire du panel sur lequel il est basé.

S'agissant ensuite des **données de paramétrage de la projection**, les données rétrospectives contenues dans le panel de départ peuvent en faire partie : elles informent sur la façon dont les individus se sont comportés dans le passé, ce qui est une base pour la prévision des comportements futurs. Mais le modèle est souvent amené à puiser dans d'autres sources.

Dans Destinie, c'est par exemple de l'enquête Génération du Céreq¹³ et de l'EIC que sont tirées les matrices de probabilités servant à projeter les transitions sur le marché du travail, et c'est sur les données de l'échantillon démographique permanent (EDP) de l'Insee, combinées à celles des enquêtes Famille, Emploi et Formation et qualification professionnelle que sont basées celles qui gouvernent l'évolution des comportements démographiques. Les projections agrégées peuvent aussi alimenter le modèle, à leur façon, si on les utilise comme outils de calage d'une partie de ses résultats, offrant du même coup une solution partielle au problème de l'aléa stochastique (voir *supra*) associé à un fonctionnement « libre » du modèle (**encadré 2**).

11. Il s'agit de l'Enquête sur les revenus fiscaux (ERF), remplacée en 2006 par l'Enquête sur les revenus fiscaux et sociaux (ERFS).

12. L'EIC est un échantillon anonyme d'individus, reconstituant l'intégralité de la carrière année après année par rapprochement des données des différents régimes. Pour plus de détails, voir l'article de Simon Fredon et Michael Sicsic dans ce même numéro.

13. Le Centre d'études et de recherches sur les qualifications mène une enquête qui permet de reconstituer et d'analyser les parcours des jeunes au cours de leurs trois premières années de vie active.

Encadré 2. La question de la cohérence avec les projections agrégées

Quelles techniques de calage utiliser pour consolider les résultats d'un modèle de microsimulation et garantir leur cohérence avec les prévisions réalisées par des méthodes traditionnelles ? Prenons l'exemple des naissances et d'un modèle appliqué aux retraites :

Des projections démographiques de référence sont établies par la méthode dite « des composants », où le seul prédicteur de l'événement naissance est l'âge de la mère : ces projections déterminent le nombre annuel de naissances comme somme des effectifs féminins par âge, pondérés par les taux de fécondité ;

Dans le modèle de microsimulation, l'âge de la mère sera aussi un déterminant de la probabilité d'une naissance mais, il s'y ajoutera d'autres facteurs explicatifs : le statut matrimonial, le niveau d'éducation et la position sur le marché du travail, le nombre d'enfants déjà eus et la durée écoulée depuis la naissance du dernier d'entre eux. Or rien ne garantit *in fine* un nombre de naissances cohérent avec celui de la projection usuelle.

Il y a deux écoles face à une telle divergence : soit assumer l'écart en expliquant que, à modèle différent, il est normal d'avoir des résultats différents, soit chercher à éviter les incohérences apparentes. Ceci est possible en contraignant les probabilités conditionnées par les variables individuelles de sorte que, en moyenne, elles donnent le même nombre d'événements que la projection de référence. On parle alors d'« **alignement** ». On considère que le modèle n'est pas forcément plus performant que la projection standard pour donner ce nombre, qui relève d'ailleurs plus de la conjecture que de la prévision : dans le cas des naissances, la projection classique s'appuie sur des hypothèses de fécondité globale ; en revanche, le modèle de microsimulation ventile de manière plausible ces naissances dans la population féminine, en respectant les probabilités relatives d'avoir une naissance selon les caractéristiques des mères.

Le choix entre la simulation libre ou contrainte doit s'apprécier au cas par cas. Dans Destinie, on se cale sur les projections déterministes de population active pour les âges compris entre 25 et 55 ans. En revanche c'est Destinie qui a la main pour évaluer les proportions d'actifs au-delà de 55 ans, ou pour les moins de 25 ans, et c'est la projection traditionnelle qui se cale sur la microsimulation plutôt que l'inverse.

De tels calages apportent une réponse à l'une des réticences à la microsimulation : en contraignant les nombres d'évènements à respecter des cibles macro, on neutralise l'aléa d'échantillonnage et donc le caractère bruité des résultats. Néanmoins le remède doit être appliqué avec précaution. Si le modèle « libre » donne un nombre d'évènements spontanément très éloigné de la cible, forcer ce nombre peut distordre d'autres résultats du modèle non soumis à calage, voire accentuer leur instabilité.

Une façon moins risquée de réduire les biais d'échantillonnage, qualifiée de « **réduction de variance** », consiste à contraindre le nombre d'évènements simulé à être égal à sa valeur espérée sur la base des probabilités individuelles du modèle, sans l'aléa de tirage spontanément engendré par le modèle.

Enfin en répliquant un grand nombre de simulations indépendantes du modèle et en calculant la moyenne, il est encore plus simple de réduire le caractère bruité du modèle. L'inconvénient du procédé est qu'il démultiplie les temps de calcul, ce qui est plus ou moins bloquant selon la vitesse d'exécution du modèle.

📍 ...ET QUELS OUTILS INFORMATIQUES ?

La question du « comment » est aussi celle des outils informatiques. La capacité limitée des ordinateurs a longtemps empêché d'aller très loin dans la réalisation du programme originel d'Orcutt. C'est bien évidemment de moins en moins le cas, mais d'une façon qui reste très sensible aux solutions logicielles adoptées pour la construction du modèle.

Dans ce domaine, on n'a pas spontanément assisté au développement d'outils dédiés et optimisés comme il a pu s'en développer pour la modélisation macroéconomique ou l'analyse statistique. À ceci une raison très simple : la microsimulation est un exercice qui ne fait pas appel à une technicité très avancée. Une fois réunis et maîtrisés les fichiers de données individuelles, un bon langage généraliste et un générateur fiable de nombres pseudo-aléatoires suffisent pour se lancer dans la construction d'un modèle, le reste n'est qu'affaire de rigueur et de patience. Dans ce contexte, la tendance naturelle du modélisateur est d'opter pour le logiciel ou le langage de programmation dont il est le plus familier, quel qu'il soit.

Pour prendre à nouveau l'exemple de Destinie, il avait été initialement écrit en Turbo Pascal, puis il a tenté un virage avorté vers C++ au début des années deux mille, à la suite de quoi il s'est ensuite replié sur Perl et a tourné quelques années sous ce langage, avant de basculer vers un mélange de R et de C++. Le passage à R est cohérent avec la place que ce langage est en train de prendre dans l'ensemble de l'Insee. Si R est le langage de référence pour le reste des travaux statistiques et d'études conduits dans l'institut, il est assez naturel d'y recourir aussi pour la microsimulation, évitant aux utilisateurs de devoir maîtriser deux langages. Mais son inconvénient est la plus grande lenteur d'exécution des programmes ; c'est ici qu'intervient la bascule à C++ pour une partie de son code.

L'argument de la rapidité d'exécution avait aussi conduit les concepteurs de Myriade à opter directement pour le même langage C++, néanmoins abandonné depuis avec le ralliement de la Cnaf à Ines : ce modèle est programmé en SAS, tout comme TRAJECTOIRE à la Drees et Prisme à la Cnav. Mais pour Ines se pose maintenant la question d'un basculement à R, comme l'a fait Destinie.

Il est souvent regretté que cette grande variété d'approches conduise beaucoup de modélisateurs à réinventer la roue, voire à la réinventer plusieurs fois lorsqu'ils passent d'un langage à un autre. Même si chaque modèle est spécifique, on se dit qu'il doit en effet exister des briques mutualisables. Les procédures de calage ou d'alignement présentées dans **l'encadré 2** en sont un exemple évident, mais il y a eu des tentatives de construire des cadres de programmation plus systématiques, devant permettre au modélisateur de se soucier avant tout du contenu de son modèle en perdant le minimum de temps sur les questions purement informatiques. Statistique Canada avait ainsi développé l'outil ModGen, écrit en C++. De développement plus récent et qui rencontre un certain succès est le progiciel LIAM 2, porté par les microsimulateurs du Bureau belge du plan (de Menten *et alii*, 2014). Cette problématique du choix de l'outil de programmation est donc en train d'évoluer.

Comme pour le reste de la production statistique, il y a un équilibre à trouver entre mutualisation et préservation d'un principe de subsidiarité autorisant le modélisateur à choisir ce qui lui paraît le plus adapté pour le modèle qu'il souhaite construire.

1 LA MICROSIMULATION À L'INSEE : JUSQU'OU ?

Ayant abordé les questions du « pourquoi » et du « comment », reste celle des perspectives. Y a-t-il de nouvelles directions vers lesquelles faire prospérer ces outils ? Au risque d'un peu d'utopie, on peut revenir au programme originel d'Orcutt, mais aussi développer la question des liens avec d'autres pratiques de l'évaluation des politiques publiques, un sujet auquel l'Insee contribue par d'autres voies que la microsimulation et avec lesquelles des complémentarités pourraient être développées.

Le programme d'Orcutt tout d'abord. La microsimulation dynamique appliquée aux seules retraites n'est qu'un aspect parmi d'autres de ce que pouvait englober ce programme, d'ailleurs probablement pas celle à laquelle Orcutt avait pensé à l'époque. Or il se trouve justement que, sur ce créneau spécifique des retraites, Destinie a perdu une partie de l'avantage que lui avait temporairement procuré son statut de précurseur : il est moins précis que les modèles assis sur les données administratives de la Drees, de la Cnav ou des autres caisses qui se sont lancées dans le même genre de démarche. Il lui reste en revanche l'avantage de simuler des ménages et pas uniquement des individus, ce qui lui permet des calculs de niveau de vie et l'étude de dispositifs conditionnés aux ressources totales de ces ménages, pas seulement de chacun de leurs membres pris séparément. Cette spécificité le rend incontournable pour l'étude de la réversion, et elle est aussi utilisée pour un sujet très lié aux retraites mais néanmoins distinct, celui de la dépendance et de sa prise en charge, pour lesquels la dimension ménage est très importante.

Sans aller jusqu'à la modélisation sociale « totale » qui était envisagée par Orcutt, une voie d'évolution peut donc être de chercher à faire de Destinie un outil de référence pour nourrir la prospective à long terme des niveaux et des conditions de vie, sous un grand nombre

« faire de Destinie un outil de référence pour nourrir la prospective à long terme des niveaux et des conditions de vie. »

d'aspects. Ceci peut se faire par construction progressive de modules additionnels, car un des avantages de la méthode, c'est aussi la possibilité d'y greffer des modules additionnels tirant parti de ce que simule le cœur du modèle et le complétant par la simulation de variables annexes. Un module épargne et patrimoine est ainsi en cours de construction, centré sur le patrimoine financier, auquel un complément naturel

serait la simulation du patrimoine immobilier. D'autres thèmes sont envisageables à la commande, par exemple la simulation de l'exposition à la précarité énergétique, déjà traitée au CGDD¹⁴ dans un cadre statique, celui du modèle Prometheus (Thao-Khamsing et alii, 2016).

La place de la microsimulation dans l'évaluation des politiques publiques pour finir. Cette place est un peu paradoxale. La microsimulation est par nature une pièce centrale de tout dispositif d'évaluation, *ex ante* comme *ex post*, et pourtant peu reconnue par les économistes se réclamant de l'évaluation. Pour une majorité d'économistes du domaine, évaluer une politique publique, c'est mesurer empiriquement son impact sur une ou quelques variables

14. Le commissariat général au développement durable (CGDD) relève du ministère de la Transition écologique et solidaire. Il abrite le service des données et études statistiques (Sdes) i.e. le service statistique des ministères chargés de l'environnement, de l'énergie, de la construction, du logement et des transports.

d'intérêt – les *outcomes* – en s'appuyant dans le cas idéal sur de l'expérimentation contrôlée, ou à défaut sur des quasi-expériences ayant affecté de façon différenciée des groupes de population suffisamment comparables pour pouvoir jouer les rôles de groupe test et de groupe témoin.

Les approches expérimentales ou pseudo expérimentales sont évidemment des aspects importants de l'évaluation, mais les y réduire n'a pas de sens. Toutes les politiques ne sont pas évaluable en ces termes. Privilégier une seule méthode d'évaluation c'est se condamner à ignorer un grand nombre de sujets. Même pour des sujets pour lesquels cette méthode est pertinente, elle ne répond qu'à une partie de la question : savoir que telle incitation financière a modifié l'offre de travail de telle ou telle proportion de personnes n'informe pas à soi tout seul sur le rapport coût-bénéfice de la mesure ou ses effets redistributifs.

Ceci plaide donc pour reconnaître la complémentarité des approches (Blanchet *et alii*, 2016). Le reproche régulièrement adressé aux modèles de microsimulation statiques est qu'ils sont le plus souvent purement comptables, sans hypothèses de réactions comportementales. Les modèles de microsimulation dynamique introduisent des comportements, mais souvent très mécaniques ou exogènes, comme lorsqu'ils se bornent à reproduire les probabilités de transitions entre états constatées dans le passé. Il y a donc place pour y mettre des hypothèses de comportement dont le paramétrage se nourrirait des résultats d'évaluations empiriques *ex post*, mais pas que, car on peut aussi être amené à projeter des réponses comportementales pour lesquelles l'évaluation *ex post* rigoureuse ne sera pas disponible avant longtemps, voire jamais. L'exemple de l'âge de départ en retraite en est le meilleur exemple : on ne peut pas attendre que les réformes aient produit tous leurs effets sur les âges de liquidations pour intégrer ces effets dans les modèles de projection.

Reste l'ambition d'avoir des modèles qui soient non seulement à comportements endogènes mais qui soient également bouclés. Dans le projet d'Orcutt, la prise en compte du bouclage passait par la récursivité. À la date t , les agents réagissent indépendamment les uns des autres à l'état courant du monde qu'ils constatent autour d'eux, il en résulte à l'instant suivant un nouvel état du monde auquel ils réagissent à nouveau et ainsi de suite. Pour que cela rende compte du monde réel, il suffit de construire des modèles à pas temporel suffisamment fin, quand les moyens de calcul le permettent. Des tentatives ont été faites en France pour appliquer cette logique au fonctionnement du marché du travail (Barlet *et alii*, 2010 ; Ballot *et alii*, 2016), tirant la microsimulation vers une autre catégorie de modèles, ceux que la littérature économique préfère qualifier de modèles d'agents, pour bien marquer la différence avec une microsimulation jugée trop comptable.

Il est trop tôt pour dire jusqu'où ce type d'approche serait en mesure de concurrencer la modélisation macro-économétrique traditionnelle, mais les macro-économistes eux-mêmes sont de plus en plus sensibles à la nécessité de sortir de leurs représentations excessivement agrégées : il n'y a pas que pour les constats statistiques qu'il est nécessaire d'aller au-delà de la moyenne, c'est pour l'ensemble des outils de modélisation que la même question peut se poser.

AUBERT, Patrick et BACHELET, Marion, 2012. Disparités de montant de pension et redistribution dans le système de retraite français. In : *L'Économie française – Comptes et dossiers, édition 2012*. [en ligne]. 4 juillet 2012. Collection Insee références, pp. 40-62. [Consulté le 24 avril 2020]. Disponible à l'adresse :

https://www.insee.fr/fr/statistiques/fichier/1373977/ECOFRA12c_D1_retrait.pdf.

AUBERT, Patrick et RABATÉ, Simon, 2015. Durée passée en carrière et durée de vie en retraite : quel partage des gains d'espérance de vie ?. In : *PIB par habitant en longue période – Assurance dépendance – Durée de la retraite et espérance de vie – Mesure de la concentration spatiale*. [en ligne]. 19 février 2015. Économie et Statistique / Economics and Statistics, N°474-2014, pp. 69-95. [Consulté le 24 avril 2020]. Disponible à l'adresse :

<https://www.insee.fr/fr/statistiques/fichier/1377630/ES474C.pdf>.

BALLOT, Gérard, KANT, Jean-Daniel et GOUDET, Olivier, 2016. Un modèle multi-agents du marché du travail français, outil d'évaluation des politiques de l'emploi : l'exemple du contrat de génération. In : *Revue économique*. [en ligne]. Juillet 2016. Vol. 67, n°4, pp. 733-771. [Consulté le 24 avril 2020]. Disponible à l'adresse :

<https://www.cairn.info/revue-economique-2016-4-page-733.htm>.

BARDAJI, José, SÉDILLOT, Béatrice et WALRAET, Emmanuelle, 2003. Un outil de prospective des retraites : le modèle de microsimulation Destinie. In : *Micro-simulation : l'expérience française*. [en ligne]. La Documentation française – Économie & prévision, N°160-161, pp. 193-213. [Consulté le 24 avril 2020]. Disponible à l'adresse :

<https://www.cairn.info/revue-economie-et-prevision-2003-4-p-193.htm>.

BARLET, Muriel, BLANCHET, Didier et LE BARBANCHON, Thomas, 2010. Microsimulation et modèles d'agents : une approche alternative pour l'évaluation des politiques d'emploi. In : *Bas salaire et marché du travail*. [en ligne]. 20 août 2010. Économie et Statistique, N°429-430 – 2009, pp. 51-76. [Consulté le 24 avril 2020]. Disponible à l'adresse :

<https://www.insee.fr/fr/statistiques/fichier/1377394/ES429C%20.pdf>.

BLANCHET, Didier, BOZIO, Antoine et RABATÉ, Simon, 2016. Quelles options pour réduire la dépendance à la croissance du système de retraite français ?. In : *Revue économique*. [en ligne]. Vol. 67, n°2016/4, pp. 879-911. [Consulté le 24 avril 2020]. Disponible à l'adresse :

<https://www.cairn.info/revue-economique-2016-4-page-879.htm>.

BLANCHET, Didier, BUFFETEAU, Sophie, CRENNER, Emmanuelle et LE MINEZ, Sylvie, 2011. Le modèle de microsimulation Destinie 2 : principales caractéristiques et premiers résultats. In : *Les systèmes de retraite et leurs réformes : évaluations et projections*. [en ligne]. 20 octobre 2011. Économie et Statistique, n°441-442 – 2011, pp. 101-121. [Consulté le 24 avril 2020]. Disponible à l'adresse : https://www.persee.fr/doc/estat_0336-1454_2011_num_441_1_9615.

BLANCHET, Didier, HAGNERÉ, Cyrille, LEGENDRE, François et THIBAUT, Florence, 2016. Évaluation des politiques publiques, ex post et ex ante : l'apport de la microsimulation. In : *Revue économique*. [en ligne]. Juillet 2016. Vol. 67, n°4, pp. 685-696. [Consulté le 24 avril 2020]. Disponible à l'adresse :

<https://www.cairn.info/revue-economique-2016-4-page-685.htm>.

BOURGUIGNON, François, CHIAPPORI, Pierre-André et SASTRE-DESCALS, José, 1988. SYSSIFF : a simulation of the French tax benefit system. In : ATKINSON, A. B. et SUTHERLAND, Holly, 1988. *Tax benefit models*. Décembre 1988. STICERD Occasional Paper, n°10, p. 395. ISBN : 8-85328-103-3.

DEMENTEN, Gaëtan, DEKKERS, Gijs, BRYON, Geert, LIÉGEOIS, Philippe et O'DONOGHUE, Cathal, 2014. *LIAM2 : a New Open Source Development Tool for Discrete-Time Dynamic Microsimulation Models*. [en ligne]. 30 juin 2014. Journal of Artificial Societies and Social Simulation (JASSS), n°17. [Consulté le 24 avril 2020]. Disponible à l'adresse : <http://jasss.soc.surrey.ac.uk/17/3/9.html>.

DUBOIS, Yves et KOUBI, Malik, 2017. *Règles d'indexation des pensions et sensibilité des dépenses de retraites à la croissance économique et aux chocs démographiques*. [en ligne]. 30 mars 2017. Insee, Documents de travail, n°G2017-02. [Consulté le 24 avril 2020]. Disponible à l'adresse : <https://www.insee.fr/fr/statistiques/fichier/2672106/G2017-02.pdf>.

DUBOIS, Yves et MARINO, Anthony, 2015. Le taux de rendement interne du système de retraite français : quelle redistribution au sein d'une génération et quelle évolution entre générations ? In : *Microsimulation appliquée aux politiques fiscales et sociales*. [en ligne]. 17 décembre 2015. Économie et Statistique / Economics and Statistics, n°481-482, pp. 77-95. [Consulté le 24 avril 2020]. Disponible à l'adresse : <https://www.insee.fr/fr/statistiques/fichier/1305193/ES481D.pdf>.

LANDAIS, Camille, PIKETTY, Thomas et SAEZ, Emmanuel, 2011. *Le modèle de micro-simulation TAXIPP – version 0.0*. [en ligne]. Janvier 2011. Édition Institut des Politiques Publiques. [Consulté le 24 avril 2020]. Disponible à l'adresse : <https://www.ipp.eu/wp-content/uploads/2013/01/guide-methodIPP-janv2011-taxipp00.pdf>.

LEGENDRE, François, 2019. L'émergence et la consolidation des méthodes de microsimulation en France. In : *Numéro spécial – 50^e anniversaire*. [en ligne]. Économie et Statistique / Economics and Statistics, n°510-511-512 – 2019, pp. 201-217. [Consulté le 24 avril 2020]. Disponible à l'adresse : https://www.insee.fr/fr/statistiques/fichier/4253061/510_511_512_Legendre_FR.pdf.

MARINO, Anthony, 2014. *Vingt ans de réformes des retraites : quelle contribution des règles d'indexation ?*. [en ligne]. 15 avril 2014. Insee Analyses n°17. [Consulté le 24 avril 2020]. Disponible à l'adresse : <https://www.insee.fr/fr/statistiques/1521315>.

ORCUTT, Guy H., 1957. A new type of socio-economic system. In : *The Review of Economics and Statistics*. [en ligne]. N°39(2), pp. 116-123. [Consulté le 24 avril 2020]. Disponible à l'adresse : https://www.microsimulation.org/IJM/V1_1/IJM_1_1_2.pdf.

POUBELLE, Vincent, ALBERT, Christophe, BEURNIER, Paul, COUHIN, Julie et GRAVE, Nathanaël, 2006. Prisme, le modèle de la Cnav. In : *Retraite et société*. [en ligne]. Juin 2006. La Documentation française, n°48, pp. 202-215. [Consulté le 24 avril 2020]. Disponible à l'adresse : <https://www.cairn.info/revue-retraite-et-societe1-2006-2-page-202.htm>.

THAO KHAMSING, Willy, CECI-RENAUD, Nila et GUILLOT, Lola, 2016. Simuler l'impact social de la fiscalité énergétique : le modèle Prometheus (PROGRAMME de Microsimulation des Énergies du Transport et de l'Habitat pour ÉVALUations Sociales) – Usages et méthodologie. In : *Études et documents*. [en ligne]. Février 2016. Service de l'Économie, de l'Évaluation et de l'Intégration du développement Durable (SEEIDD), n°138. [Consulté le 24 avril 2020]. Disponible à l'adresse : <http://temis.documentation.developpement-durable.gouv.fr/docs/Temis/0083/Temis-0083851/22397.pdf>.

WATTS, Harold W., 1991. Distinguished Fellow : An Appreciation of Guy Orcutt. In : *The Journal of Economic Perspectives*. [en ligne]. Winter 1991. Vol. 5, n°1, pp. 171-179. [Consulté le 24 avril 2020]. Disponible à l'adresse : <https://www.aeaweb.org/articles?id=10.1257/jep.5.1.171>.

UNE DÉCENNIE DE MODÉLISATION DU SYSTÈME DE RETRAITE

LA GENÈSE DU MODÈLE DE MICROSIMULATION TRAJECTOIRE

Pierre Cheloudko* et Henri Martin**

Le système de retraite français se caractérise par sa très grande complexité réglementaire et par une très grande diversité des profils de carrière de ses assurés. Dans ce contexte, et pour assurer au mieux ses missions de production d'informations statistiques et d'éclairage des politiques publiques en matière de retraite, la Drees, le service statistique du Ministère des Solidarités et de la Santé, collecte des données auprès des différents régimes sur un échantillon de cotisants, et sur un échantillon de retraités. Depuis la fin des années deux-mille, la Drees a développé un modèle de microsimulation dynamique basé sur ces deux échantillons inter-régimes. TRAJECTOIRE est né de la juxtaposition de trois modules : le premier modélise les carrières professionnelles, le deuxième simule les âges de départ à la retraite et le troisième calcule les pensions de retraite. Le développement de chaque nouveau module a été rendu nécessaire par les limites que présentaient les outils pré-existants. Modèle particulièrement riche, TRAJECTOIRE a depuis fait l'objet de nombreuses exploitations. Beaucoup d'autres institutions se sont dotées de modèles similaires : l'Insee, les principaux régimes de retraite, certaines administrations et instituts de recherche. Si cette diversité entraîne une fragmentation des ressources, elle est aussi une richesse pour la connaissance et le pilotage du système de retraite français.

 *The French pension system is characterised by its highly complex regulatory framework and by the very wide diversity of career profiles of its members. In this context, and in order to best carry out its missions of producing statistical information and shedding light on public policies in the area of pensions, DREES, the Health and Solidarity Ministerial Statistical Department, collects data from the various schemes on a sample of contributors and on a sample of pensioners. Since 2010, DREES has been developing a dynamic microsimulation model based on these two inter-regime samples. TRAJECTOIRE was created through the juxtaposition of three modules: the first models professional careers, the second simulates retirement ages and the third calculates retirement pensions. The development of each new module was made necessary by the limitations of the pre-existing tools. As a particularly rich model, TRAJECTOIRE has since been used on numerous occasions. Many other institutions, such as INSEE, the main pension schemes and certain government departments and research institutes, have adopted similar models. Although this diversity leads to a fragmentation of resources, it is also an asset for understanding and managing the French pension system.*

* Chargé d'études, bureau Retraites, direction de la Recherche, des études, de l'évaluation et des statistiques (Drees), pierre.cheloudko@sante.gouv.fr

** Au moment de la rédaction de cet article, chef du pôle Simulations microéconomiques, bureau Retraites, direction de la Recherche, des études, de l'évaluation et des statistiques (Drees), henri.martin@developpement-durable.gouv.fr

UN SYSTÈME DE RETRAITE FRANÇAIS D'UNE GRANDE COMPLEXITÉ

La réglementation française juxtapose plus d'une quarantaine de régimes de retraite différents entre lesquels les règles d'accumulation des droits, les règles de calcul des pensions, voire les conditions pour partir à la retraite ne sont pas intégralement harmonisées¹ (**encadré 1**). Au sein même de chaque régime la réglementation peut varier avec la date de naissance de l'assuré et éventuellement avec la date à laquelle il fait valoir ses droits à retraite. Les projections financières des régimes sont en outre très dépendantes des paramètres macroéconomiques et démographiques : le taux de chômage et le taux de croissance, l'immigration, la fécondité et l'espérance de vie.

À cette complexité réglementaire s'ajoute la grande diversité des situations individuelles. Certains assurés ont une durée de carrière supérieure à 45 ans alors que pour d'autres elle est de moins d'un an, notamment pour les personnes qui n'ont été que de passage sur le territoire français. Certaines carrières peuvent être plus ou moins « fractionnées », au sens où l'assuré connaîtra, ou non, des périodes de non-emploi ; se caractériser par des niveaux de revenus d'activité très différents ; se dérouler dans des régimes de retraite différents, par exemple quand un assuré commence sa carrière dans le secteur privé avant d'être recruté dans la fonction publique ; ou encore comprendre ou non certains événements biographiques : la naissance d'un enfant, un mariage, etc. Enfin, les carrières évoluent au fil des générations : les jeunes entrent de plus en plus tardivement sur le marché du travail ; les périodes de non-emploi sont devenues plus fréquentes avec l'augmentation du chômage et des formes particulières d'emploi ; les taux de fécondité et de nuptialité diminuent ; l'âge à la maternité augmente, etc.

Pour tenir compte de ces différents éléments de complexité, plusieurs organismes ont développé des outils qui modélisent plus finement à la fois la pluralité des parcours des assurés sur un horizon temporel relativement long et la réglementation en matière de retraite.

« Pour tenir compte de ces différents éléments de complexité, plusieurs organismes ont développé des outils qui modélisent plus finement à la fois la pluralité des parcours des assurés sur un horizon temporel relativement long et la réglementation en matière de retraite. »

Au sein du service statistique public, l'Insee et la Direction de la recherche, des études, de l'évaluation et des statistiques (Drees), ont apporté leur pierre à cet édifice. Leur objectif est d'assurer au mieux leurs missions de production et de diffusion d'informations statistiques et d'analyses économiques, afin d'éclairer notamment les choix politiques auprès du législateur. Cette contribution retrace, à travers une approche historique,

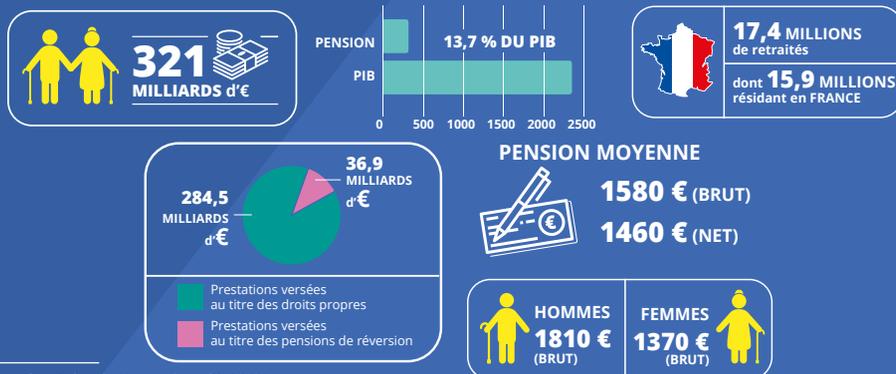
les étapes successives par lesquelles la Drees s'est progressivement dotée d'un modèle de microsimulation dynamique sur les retraites : le modèle TRAJECTOIRE (**encadré 2**). Il met en même temps l'accent sur les outils développés en parallèle par d'autres institutions.

1. Historiquement, l'affiliation à ces régimes dépendait de l'emploi occupé par l'assuré et de son secteur d'activité.

Encadré 1. Le système de retraite français : plus de 40 régimes « obligatoires »

Les assurés sont affiliés à un ou plusieurs régimes de retraite selon leur statut professionnel. Ils acquièrent des droits dans ces régimes qui leur versent ensuite, une fois ces droits liquidés, une ou plusieurs pensions de retraite. Depuis juillet 2017, les retraités poly-pensionnés des régimes dits « alignés » (MSA salarié, Cnav et RSI, aujourd'hui devenu le SSI), qui percevaient auparavant une pension de chacun de ces régimes perçoivent désormais une pension unique au titre de l'ensemble de leurs droits cumulés dans ces régimes.

Les prestations versées en 2018



Données tirées de (Arnaud et alii, 2019).

Populations concernées	Régimes de base ou intégrés (gérés par l'État)	Régimes complémentaires ou additionnels (gestion déléguée aux partenaires sociaux)
Salariés du secteur privé	Caisse nationale d'assurance vieillesse (Cnav), régime général	Arrco (salariés d'employeurs privés) Agirc (cadres)
Salariés agricoles	Mutualité sociale agricole (MSA) aligné sur le régime général	Ircantec (agents non titulaires de l'État et des collectivités publiques)
Artisans, industriels et commerçants	Sécurité sociale des indépendants (SSI) alignée sur le régime général	SSI complémentaire. Avant 2013, il y avait un régime complémentaire pour les artisans et un autre pour les commerçants
Professions libérales	Régime de base (10 sections professionnelles regroupées au sein de la CNAVPL) Caisse nationale des barreaux français (CNBF)	Régimes complémentaires obligatoires selon la section professionnelle Régimes supplémentaires pour les praticiens et auxiliaires médicaux conventionnés
Non-salariés agricoles	Mutualité sociale agricole (MSA)	Régime complémentaire obligatoire
Fonctionnaires civils et militaires de l'État	Régime spécial géré par le Service des retraites de l'État (SRE)	Retraite additionnelle de la fonction publique (RAFP)
Fonctionnaire des hôpitaux et des collectivités locales	Régime spécial géré par la Caisse nationale de retraites des agents des collectivités locales (CNRACL)	Retraite additionnelle de la fonction publique (RAFP)
Agents des entreprises publiques et autres régimes spéciaux	Régimes spéciaux de la SNCF, de la RATP, des industries électriques et gazières (CNIEG), de la Banque de France, de l'Opéra de Paris, de la CRPCEN (clercs et employés de notaires), des mines, etc.	

À L'ORIGINE : MODÈLES AGRÉGÉS ET MICROSIMULATION AVEC DESTINIE

Chronologiquement, le besoin de modélisation a d'abord concerné les services gestionnaires des régimes de retraite eux-mêmes, préoccupés par la soutenabilité de leur trajectoire financière à court, moyen et long terme et par le pilotage de leur trésorerie. Ceux-ci ont donc développé les premiers outils de modélisation sur les retraites, sous la forme de modèles agrégés, chacun se focalisant sur ses problématiques propres à son régime. Ces premiers modèles dits « régime » se limitent au champ de la sous-population d'affiliés correspondante. La qualité et la finesse de l'information produite par ces outils est logiquement très variable d'un régime à l'autre, reflétant ainsi la très grande disparité des ressources humaines et financières dont disposent les organismes gestionnaires. Mais la principale limite de ces modèles est inhérente aux objectifs qui ont sous-tendu leur développement : en se concentrant sur le périmètre d'un seul régime, ils ne modélisent que partiellement le système de retraite.

C'est en partie pour dépasser cette limite que l'Insee a engagé, dès le début des années quatre-vingt-dix, le développement du premier modèle de microsimulation « tous régimes » : le modèle Destinie² (Blanchet et Chanut, 1998). L'objectif est de simuler les carrières puis les pensions de retraite d'un échantillon d'individus résidant sur le territoire métropolitain. Le champ du modèle qui se limitait initialement aux assurés du régime général (les salariés du secteur privé) s'est élargi progressivement aux autres grands régimes de retraite (régimes de la fonction publique et régime des artisans et commerçants) si bien qu'il couvre aujourd'hui environ 90 % des masses financières de prestations versées.

L'enquête Patrimoine³ de l'Insee a semblé aux acteurs de l'époque être la source la plus adaptée sur laquelle adosser le modèle Destinie, car elle contient un « calendrier rétrospectif » dans lequel les répondants détaillent l'historique de leurs périodes d'activité depuis le début de leur carrière professionnelle. Au-delà de la couverture de l'ensemble

« Le principal apport du modèle Destinie résidait dans la modélisation du système de retraite à un niveau individuel, par le recours à la technique de microsimulation, »

du système de retraite, dans un contexte où l'ensemble des modèles existant fonctionnaient sur une base agrégée, le principal apport du modèle Destinie résidait dans la modélisation du système de retraite à un niveau individuel, par le recours à la technique de microsimulation, seule à même de tenir compte de manière satisfaisante de la complexité des paramètres du système et de permettre des analyses fines en termes de redistribution⁴. Là où un modèle agrégé se limite à la production de résultats en population générale, Destinie permet de décliner

ces résultats sur de nombreuses sous-populations de retraités : les hommes, les femmes, les retraités les plus aisés ou au contraire les moins aisés, les retraités aux carrières longues ou aux carrières courtes, les retraités aux carrières linéaires ou ceux dont les carrières sont fragmentées, etc.

2. Destinie : modèle Démographique Économique et Social de Trajectoires INDIVIDUELLES simulÉes.

3. L'enquête Patrimoine a pour objectif de décrire les actifs financiers, immobiliers et professionnels des ménages. Elle comporte un volet biographique familial et professionnel.

4. Voir l'article de Didier Blanchet sur les modèles de microsimulation, dans ce même numéro.

Pendant plusieurs années, Destinie a été l'unique outil permettant des projections et des simulations « tous régimes confondus » au niveau individuel. C'est pourquoi, l'Insee a été à plusieurs reprises sollicité par le législateur au titre de ses missions d'études économiques, notamment lors de la préparation de la réforme des retraites de 2003. À partir de la fin des années deux-mille, afin de répondre à une demande sans cesse croissante de projections, et d'évaluations de projets de réforme dont les modalités étaient elles-mêmes sans cesse plus complexes, la Drees a engagé le développement de nouveaux outils de modélisation. En parallèle, des investissements comparables ont eu lieu à la direction générale du Trésor (DG Trésor) ainsi que dans plusieurs régimes de retraite (*encadré 1*).

DEUX ÉCHANTILLONS « TOUS RÉGIMES » DE LA DREES : COTISANTS ET RETRAITÉS

En tant que service statistique du ministère des Solidarités et de la Santé, la Drees joue un rôle central dans la production d'informations statistiques (bases de données et études) sur le système de retraite dans son ensemble. Entre autres, elle a en charge la conception, la collecte et la production de deux panels inter-régimes issus des fichiers de gestion des différents régimes de retraite : l'**échantillon de cotisants (EIC)** et l'**échantillon inter-régime de retraités (EIR)**. Chacune de ces sources rassemble, pour un échantillon de la population des affiliés au système de retraite français, l'ensemble des informations connues par la quasi-totalité des régimes de retraite français : montants de pensions, trimestres validés, points accumulés, etc.

« L'échantillon de cotisants, comporte des informations sur les droits à retraite cumulés par les affiliés, qu'ils soient, ou non, déjà partis à la retraite. »

L'échantillon de cotisants, comporte des informations sur les droits à retraite cumulés par les affiliés, qu'ils soient, ou non, déjà partis à la retraite. Tous les éléments utiles au calcul de la pension de retraite (revenus d'activité et points portés au compte, trimestres validés, etc.) y sont renseignés, ainsi que le parcours d'acquisition de ces droits sur un pas annuel (nombre de trimestres ou de points acquis pour chaque année de carrière). Cette source permet une connaissance globale des droits à la retraite

acquis par les actifs dans les différents régimes de retraite. L'échantillon est produit à une périodicité quadriennale depuis 2001 et la dernière vague disponible, la quatrième du panel, contient des informations sur la situation des assurés du système de retraite au 31 décembre 2013 (EIC 2013)⁵. L'échantillon, extrait du répertoire national d'identification des personnes physiques (RNIPP), est constitué de personnes nées les premiers jours de chaque trimestre des années paires. Le taux de sondage est compris entre 2,2 % et 4,4 % selon la génération.

L'échantillon de retraités, de son côté contient des informations détaillées concernant les bénéficiaires d'une pension de retraite (de droit direct et/ou de réversion) ou d'invalidité : nature et montant des pensions versées, trimestres validés, âge de départ à la retraite, etc. C'est l'unique source qui permet de reconstituer la pension globale des retraités en rapprochant les informations fournies par les différents régimes de retraite. L'échantillon est produit à une périodicité quadriennale depuis 1988 et la dernière vague

5. L'édition 2017 de l'EIC était en cours de collecte à la date de rédaction de l'article.

disponible (la huitième du panel) contient des informations sur la situation des retraités au 31 décembre 2016 (EIR 2016). L'échantillon, également extrait du RNIPP, est constitué de personnes nées les premiers jours de chaque trimestre de certaines années. Il est construit de manière à être représentatif de l'ensemble des retraités. Le taux de sondage est compris entre 0,8 % et 9,9 % selon la génération.

Ainsi, l'échantillon de cotisants est un « film » qui décrit la carrière professionnelle des assurés jusqu'à leur départ à la retraite alors que l'échantillon de retraités est une « photographie » qui décrit les pensions versées aux retraités à une date donnée.

DES SOURCES PRIVILÉGIÉES POUR MODÉLISER LE SYSTÈME DE RETRAITE

Avec ses deux panels, la Drees a constitué des bases de données sur le système de retraite uniques en leur genre. L'une des finalités initiales de l'échantillon inter-régime de cotisants était de fournir des données pour produire des projections concernant les retraites futures. L'EIC retrace finement les trajectoires professionnelles et les droits acquis par un échantillon d'assurés. Certains individus étant communs aux deux échantillons, il est également possible de mettre en perspective les carrières et parcours professionnels des assurés avec leur situation au moment de la retraite (âge de départ à la retraite, niveau de la pension, etc.) et donc de connaître les choix des assurés en termes d'âge de départ à la retraite.

Enfin, les travaux de collecte, de production, de structuration et d'enrichissement ont amené les équipes de la Drees à développer une expertise méthodologique, à la fois sur la construction de ces sources mais aussi sur la législation en matière de retraites. Cette expertise s'est avérée cruciale pour mener à bien le développement du modèle de microsimulation TRAJECTOIRE.

Au début des années deux-mille, la Drees disposait déjà d'une maquette agrégée pour procéder à des projections à long terme d'effectifs de retraités et de masses de prestations versées (Bac *et alii*, 2003). Cependant, les résultats n'étaient pas désagrégés par régimes et ne tenaient donc pas compte de la réglementation propre à chacun d'eux. Cet outil s'avérait

« Au début des années deux-mille, la Drees disposait déjà d'une maquette agrégée pour procéder à des projections à long terme d'effectifs de retraités et de masses de prestations versées »

alors insuffisant pour modéliser la complexité des paramètres du système de retraite et pour disposer de résultats fins et détaillés par sous-populations. Aussi, dotée de nouvelles sources particulièrement riches (le premier EIC date de 2001), et guidée par son ambition d'assurer au mieux ses missions d'éclairage des politiques publiques et de production d'études sur le système de retraite, la Drees a entrepris, dès la fin des années

deux-mille, le développement d'un modèle de microsimulation dynamique du système de retraite français. La charge de travail que représentait une telle initiative a impliqué de l'aborder par étapes successives.

❶ PREMIÈRE PIERRE À L'ÉDIFICE : UN MODULE DE CALCUL DES PENSIONS...

Régulièrement sollicitées pour évaluer des évolutions des règles de calcul des pensions de retraite, les équipes de la Drees ont commencé par développer un programme SAS permettant de calculer le montant des pensions de retraite dans les principaux régimes (Lermechin, Duc et Burricand, 2011). En entrée de ce programme, les données de carrière des assurés sont connues : trimestres validés, points acquis, revenus d'activité portés au compte, âge de départ à la retraite, etc. À l'origine, l'objectif était simplement de calculer les pensions en appliquant la réglementation en vigueur, mais au fil du temps des scénarios réglementaires alternatifs, paramétrables par l'utilisateur ont été développés. Ce module, baptisé « CALIPER » (CALcul Inter-régimes des PEnsions de Retraite), a été mobilisé pour conduire des travaux extrêmement variés (*figure 1*).

❷ ... POUR TESTER LA COHÉRENCE ENTRE LES DEUX ÉCHANTILLONS INTER-RÉGIMES...

Son application la plus immédiate consistait à fournir en entrée du module les données individuelles de carrière contenues dans l'échantillon de cotisants et à calculer les pensions de retraite qui en découlent. En théorie, pour les individus communs aux deux échantillons inter-régimes, les montants de pension calculés par CALIPER devaient être similaires à ceux renseignés dans l'échantillon de retraités. Des travaux méthodologiques ont ainsi pu être entrepris pour vérifier la cohérence entre ces deux sources et expertiser l'origine des écarts. Ces travaux ont permis d'identifier plusieurs anomalies, et d'améliorer sensiblement la qualité de l'échantillon de cotisants *via* les redressements appliqués aux données brutes fournies par les régimes.

❸ ... ÉVALUER L'IMPACT DES ÉVOLUTIONS DANS LES RÈGLES DE CALCUL...

Le module CALIPER rendait également possible la production de scénarios de changements des règles de calcul des pensions de retraite à partir des données de l'échantillon de cotisants. Par exemple, à l'occasion de la réforme des retraites de 1993, la période prise en compte pour le calcul du revenu servant de référence au calcul des pensions servies par le régime général

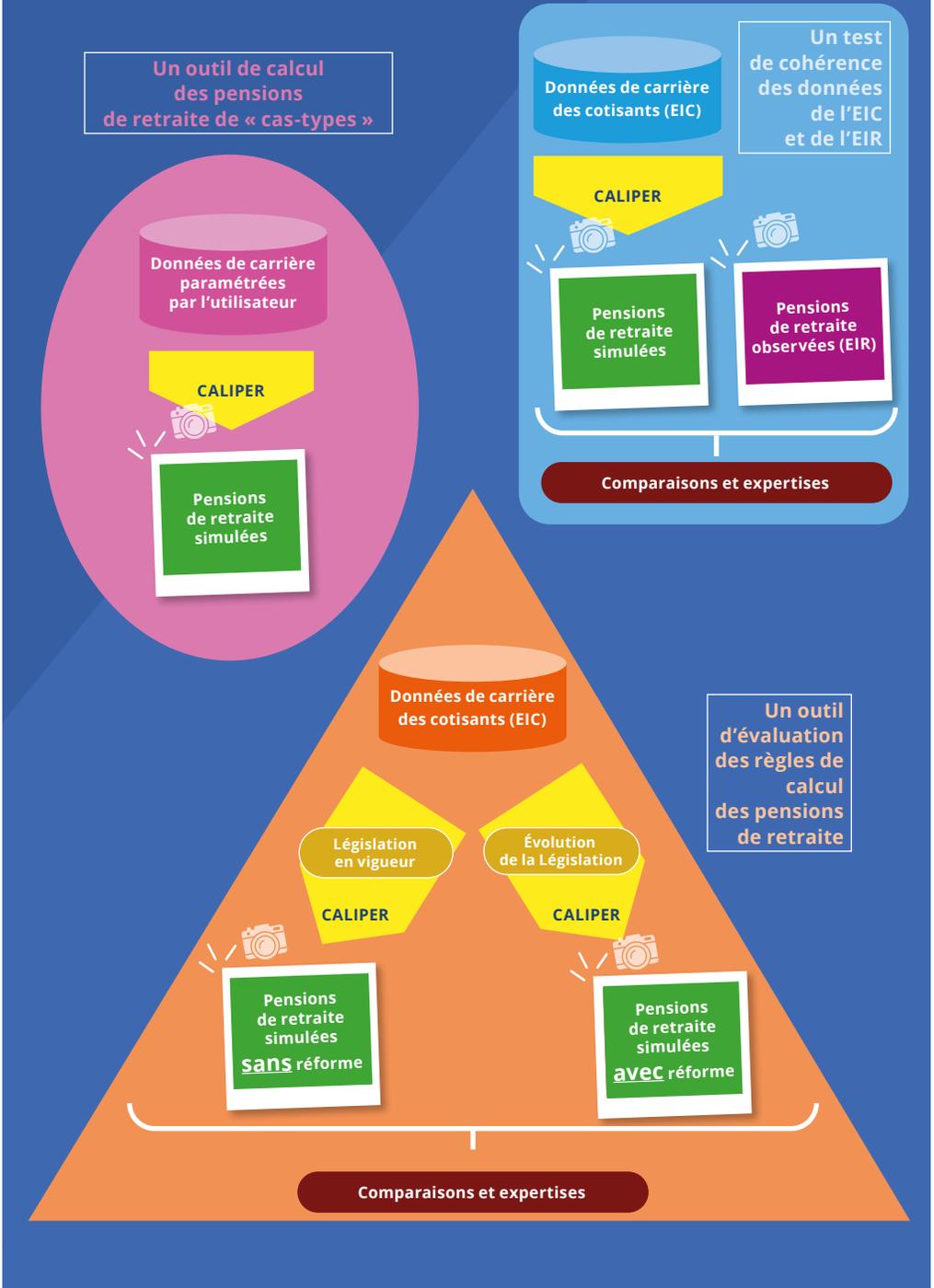
a été augmentée progressivement au fil des générations, passant des 10 meilleures années de revenus aux 25 meilleures années. De plus, ce calcul a été annualisé alors qu'il était auparavant réalisé sur une base trimestrielle. L'outil CALIPER a été mobilisé par la Drees pour fournir une estimation de ces mesures (Lermechin, Duc et Burricand, 2011). Un autre exemple

« Un autre exemple est la production de simulations de mise en place d'un régime unique harmonisant les règles de calcul des droits et des pensions entre les régimes alignés »

est la production de simulations de mise en place d'un régime unique harmonisant les règles de calcul des droits et des pensions entre les régimes alignés⁶ (Aubert *et alii*, 2012).

6. On appelle régimes alignés la Cnav, le RSI, aujourd'hui devenu le SSI et la MSA salarié, voir [encadré 1](#).

Figure 1. Les trois principales applications du module CALIPER



L'objectif était de mettre en lumière, en raisonnant à rebours, les spécificités qui caractérisaient alors les règles de calcul des pensions des retraités poly-pensionnés de ces régimes (c'est-à-dire percevant des pensions versées par au moins deux de ces régimes). Enfin, plus récemment, une analyse précise des effets de la mesure d'évolution des règles d'éligibilité et de calcul du mécanisme de minimum de pension a été conduite à partir de CALIPER (Chantel et Plouhinec, 2014).

🔗 ... ET CALCULER LES PENSIONS DE RETRAITE DE « CAS-TYPES » —

Outre les données contenues dans l'échantillon de cotisants, il est possible de fournir en entrée du module CALIPER des carrières fictives, dont le paramétrage est laissé à la discrétion de l'utilisateur. Cette possibilité a ouvert la voie aux approches par « cas-types »

qui consistent, pour des carrières perçues comme illustratives de celles de certaines sous-populations, à appliquer différentes règles de calcul des pensions de retraite.

« L'approche par cas-types a été mobilisée par la Drees pour simuler l'application des règles de calcul des pensions du régime général à des carrières de fonctionnaires »

Ces analyses constituent un complément utile aux analyses descriptives sur données réelles, car elles permettent de dissocier clairement les effets qui résultent de l'application mécanique de la réglementation, de ceux qui sont la conséquence d'autres paramètres :

évolution des carrières, paramètres macroéconomiques, etc. L'approche par cas-types a été mobilisée par la Drees pour simuler l'application des règles de calcul des pensions du régime général à des carrières de fonctionnaires (Aubert et Plouhinec, 2017), pour étudier l'effet des changements de régimes de retraite d'affiliation en cours de carrière sur le montant de la pension (Lermeschin, Duc et Burricand, 2011) ou encore pour quantifier l'impact, dans différents régimes, des périodes de temps partiel sur le montant de retraite (Aubert, Plouhinec et Prouet, 2015).

Cette approche a également été reprise par le Conseil d'orientation des retraites (COR) qui, depuis 2014, publie dans son rapport annuel plusieurs indicateurs issus de simulations sur cas-types et dont la production est assurée annuellement par la Drees à partir de l'outil CALIPER : par exemple, le taux de remplacement entre la première pension de retraite et le dernier salaire de carrière. Ce suivi permet d'isoler les effets mécaniques des changements de réglementation et des paramètres économiques sur divers indicateurs. Le COR a initialement construit sept « cas-types » : un salarié cadre du secteur privé, un salarié non-cadre du secteur privé, un salarié qui connaît des périodes de chômage, une salariée qui connaît des interruptions de carrière pour élever ses enfants, un fonctionnaire de catégorie B, un enseignant fonctionnaire et un cadre A+ de la fonction publique. Ces cas-types du COR se sont institutionnalisés dans le débat public sur les retraites. Le décret de 2014 relatif au Comité de suivi des retraites (CSR)⁷ a ainsi consacré comme référence le cas-type n°2 (le salarié non-cadre du secteur privé). Le décret précise ainsi que pour ce dernier, « les recommandations [...] ne peuvent tendre à diminuer en deçà des deux tiers le rapport [...] entre la moyenne des avantages de vieillesse perçus l'année de la liquidation et le revenu moyen d'activité perçu pendant la dernière année d'activité ».

7. Voir la référence dans les fondements juridiques.

Néanmoins, CALIPER, considéré isolément, se révélait un outil insuffisant pour modéliser pleinement le système de retraite. En effet, les carrières des assurés (trimestres validés, revenus d'activité, âge de départ à la retraite, etc.) constituant une donnée d'entrée du module, CALIPER ne permettait pas de simuler les effets d'évolutions réglementaires susceptibles de modifier ces carrières. Cela concerne par exemple les mesures de décalage de l'âge légal d'ouverture de droits qui ont pour conséquence un prolongement des carrières des assurés (augmentation du nombre de trimestres validés, nouveaux revenus d'activité, etc.)

① UN DEUXIÈME MODULE POUR MODÉLISER LES FINS DE CARRIÈRE

Dès 2010, la Drees a été sollicitée par le législateur dans le cadre de la préparation de réformes, par exemple la mesure de relèvement de 60 à 62 ans de l'âge légal d'ouverture des droits à retraite. C'est pourquoi ses équipes se sont dotées d'un nouvel outil conçu pour modéliser la fin de carrière des assurés (Aubert, Duc et Ducoudré, 2010). En pratique, il s'agit des années de carrière postérieures aux 54 ans des assurés. Ce choix est fondé sur les données issues de l'enquête Emploi de l'Insee qui montrent une diminution des taux d'emploi à partir de cet âge alors qu'ils demeurent à un niveau élevé auparavant.

① PROMESS + CALIPER : PREMIERS PAS VERS UN MODÈLE «DYNAMIQUE»

Par rapport au seul module CALIPER, l'apport principal de PROMESS fut de rendre possibles des simulations de scénarios modifiant les paramètres d'âge⁸ et de durée⁹ du système de retraite et donc les âges de départ à la retraite de certains assurés. Le modèle a par exemple été utilisé pour estimer *ex post* les effets des réformes de 1993 et 2003 qui comportaient chacune une augmentation de la durée d'assurance requise pour le taux plein (Aubert, Duc et Ducoudré, 2012).

“ PROMESS a également été mobilisé pour étudier les questions d'équité liées aux diverses réformes, en détaillant les impacts différenciés de ces réformes sur les âges de départ à la retraite des différentes catégories d'assurés. ”

PROMESS a également été mobilisé pour étudier les questions d'équité liées aux diverses réformes, en détaillant les impacts différenciés de ces réformes sur les âges de départ à la retraite des différentes catégories d'assurés (Aubert, Duc et Ducoudré, 2013). En outre, en prolongeant les carrières observées dans les données de l'échantillon de cotisants, il a également permis de conduire des simulations sur des générations qui

n'étaient pas encore parties à la retraite et qui étaient donc susceptibles d'être concernées par des projets de réformes.

-
8. L'âge légal d'ouverture des droits à partir duquel il est possible de partir à la retraite et l'âge d'annulation de la décote à partir duquel la pension de retraite est calculée à taux plein quel que soit le nombre de trimestres validés par l'assuré.
9. Pour bénéficier d'une pension complète un assuré doit avoir validé un certain nombre de trimestres.

Enfin, avec PROMESS et CALIPER, la Drees était désormais dotée d'un modèle de simulation « dynamique », au sens où ce n'était plus seulement un ensemble d'indicateurs à la date de départ à la retraite qui était calculé, comme dans le seul module « statique » CALIPER,

mais une série temporelle de ces indicateurs¹⁰. PROMESS simule en effet, sur un pas trimestriel, la chronique des situations occupées par chaque assuré sur le marché du travail : emploi affilié aux différents régimes de retraite et revenus d'activité correspondant, chômage, maladie, invalidité pouvant ouvrir droit à l'acquisition de trimestres gratuits, inactivité, etc. C'est ici une particularité, et en même temps une des difficultés de l'exercice de modélisation du système de retraite : la dimension temporelle est incontournable.

« C'est ici une particularité, et en même temps une des difficultés de l'exercice de modélisation du système de retraite : la dimension temporelle est incontournable. »

MODÉLISER L'INTÉGRALITÉ DES CARRIÈRES : NAISSANCE DE TRAJECTOIRE

La Drees a donc entrepris, à partir de 2013, de développer **un module de modélisation des carrières professionnelles** pour la partie des carrières antérieure aux 54 ans des assurés. L'objectif de ce module est de simuler, pour chaque assuré de l'échantillon de cotisants, la chronique d'affiliation aux différents régimes de retraite et la chronique des droits à retraite acquis (trimestres, revenus d'activité et points portés au compte) pour la partie non-observée de sa carrière¹¹.

Le module a été paramétré à partir des données de l'échantillon de cotisants sur les années les plus récentes d'observation de manière à répliquer dans le futur les transitions les plus récentes observées sur le marché du travail. En pratique des probabilités de transition sont estimées et répliquées entre de nombreux états possibles (emploi affilié aux différents régimes de retraite, différentes situations de non-emploi comme la maladie ou le chômage, etc.) pour des sous-populations définies par leur âge, leur sexe, leur pays de naissance (selon que l'assuré soit né en France ou à l'étranger) et leur âge d'entrée sur le marché du travail. Par exemple, le module fait l'hypothèse qu'une femme âgée de 30 ans, née en France, entrée relativement tôt sur le marché de l'emploi et qui est affiliée à la MSA salarié au cours de l'année T aura 60 % de chances d'être de nouveau affiliée à la MSA salarié au cours de l'année T+1, aura 15 % de chance de connaître une période de non-emploi (chômage, maladie, invalidité, maternité) en T+1 et aura 25 % de chance de changer de régime d'affiliation en T+1.

À ce module de simulation des carrières devait s'ajouter **un module démographique**, car l'échantillon de cotisants ne couvre pas toutes les générations¹² : il fallait donc « créer » les individus de ces générations manquantes de manière à être représentatif de la population des assurés. Ces chantiers ont été menés entre 2012 et 2014 (Duc *et alii*, 2013).

Le modèle de microsimulation « TRAJECTOIRE » (TRAJectoire de Carrières Tous REGimes) est donc né de la juxtaposition du nouveau module de modélisation des carrières (qui modélise la chronique des droits à retraite pour la partie des carrières non-observée

10. Voir l'article de Didier Blanchet sur les modèles de microsimulation, dans ce même numéro.

11. C'est-à-dire postérieure à la date du dernier millésime disponible de l'EIC.

12. À l'origine il ne contenait que les individus appartenant à une génération paire sur deux. Depuis le millésime 2013, toutes les générations paires sont échantillonnées.

dans l'EIC), du module PROMESS (qui modélise les âges de fin d'emploi et de départ à la retraite) et du module CALIPER (qui calcule les pensions de retraite) (*figure 2*).

Depuis, le modèle fonctionne et évolue comme un ensemble unifié si bien que les frontières entre les modules historiques qui le composent tendent à s'effacer. Le module PROMESS a par ailleurs fait l'objet d'une importante refonte pour en faire un véritable modèle « micro ».

DES SOLLICITATIONS MULTIPLES ET DES INITIATIVES INTERNES

La richesse du modèle TRAJECTOIRE (indicateurs de carrière et de retraite) explique les nombreuses sollicitations dont la Drees a fait l'objet, de la part de diverses administrations,

« Le modèle est aussi régulièrement sollicité pour répondre à de nombreuses questions soulevées lors des séances du Conseil d'orientation des retraites (COR) ou du Comité de suivi des retraites (CSR). »

notamment afin d'éclairer les politiques publiques. TRAJECTOIRE a ainsi été fortement mobilisé dans le cadre de la préparation de la réforme des retraites de 2014, notamment concernant la mise en œuvre de la liquidation unique des régimes alignés (*encadré 1*). Le modèle est aussi régulièrement sollicité pour répondre à de nombreuses questions soulevées lors des séances du Conseil d'orientation des retraites (COR) ou du

Comité de suivi des retraites (CSR). Il a été mis à contribution pour produire une estimation des effets des réformes du système de retraite¹³ :

- réforme de 2010 ayant décalé de 60 à 62 ans l'âge légal d'ouverture de droits et de 65 à 67 ans l'âge du taux plein automatique ;
- réforme de 2014 ayant augmenté la durée d'assurance requise pour liquider une pension au taux plein ;
- et une évaluation des droits familiaux de retraite¹⁴ (trimestres acquis au titre de la naissance et de l'éducation des enfants, majoration de 10 % pour les parents d'au moins trois enfants, droits à retraite pour les parents qui réduisent leur activité professionnelle suite à la naissance d'un enfant, etc.).

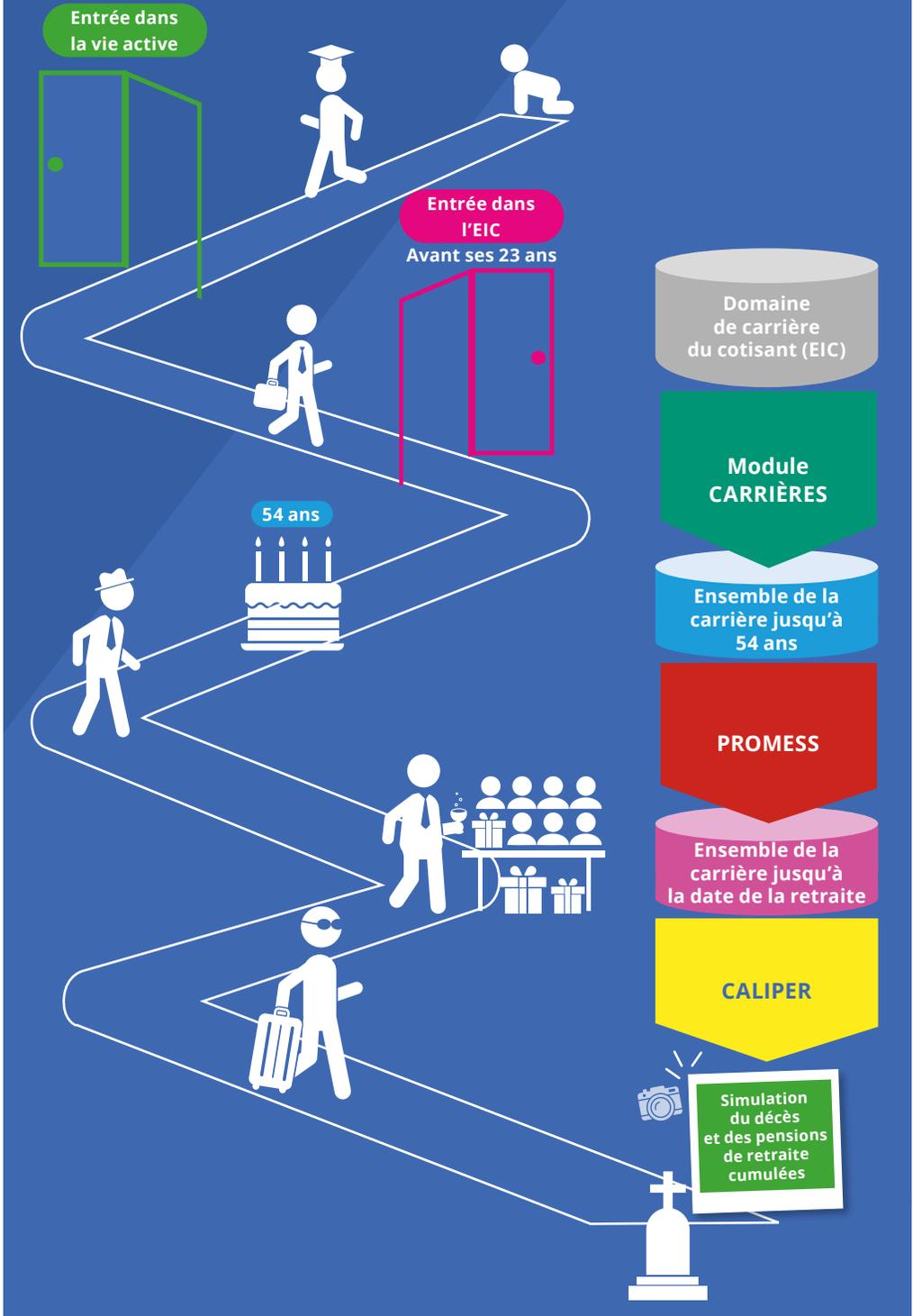
Le modèle TRAJECTOIRE a également fait l'objet d'exploitations à l'initiative interne de la Drees. Ces dernières consistent essentiellement en des analyses fines des effets des réformes des retraites passées, déclinées sous plusieurs angles : effets sur les pensions des assurés, sur les durées de retraite, sur les masses financières des régimes, sur des indicateurs d'équité, etc (Duc, 2015 ; Duc, Martin et Tréguier, 2016).

Des études ont aussi été conduites sur l'effet des règles du système de retraite et des redistributions qu'elles génèrent. Ainsi, les différences de réglementations entre fonction publique et secteur privé ont été étudiées à partir de simulations issues du modèle TRAJECTOIRE (Duc, Solard et Tréguier, 2017). Des travaux sont en cours concernant les effets redistributifs et l'équité du système de retraite.

13. Ces travaux ont été publiés en 2016 dans le rapport annuel de la Cour des comptes consacré à l'application des lois de financement de la Sécurité sociale.

14. Rapport commandé à Bertrand Fragonard, Président du Haut Conseil à la famille.

Figure 2. Fonctionnement simplifié de TRAJECTOIRE et des modules CALIPER et PROMESS



UN CHANTIER PERMANENT

Depuis sa naissance en 2015, TRAJECTOIRE évolue et s'enrichit de manière continue. Afin d'être en permanence en mesure de simuler la réglementation en vigueur, il doit faire l'objet de nouveaux développements à l'occasion de chaque modification réglementaire dans le domaine des retraites. Or ces modifications sont devenues très fréquentes sur la dernière décennie, qu'elles concernent les régimes de base ou les régimes complémentaires : augmentation de l'âge légal d'ouverture des droits, ou de la durée d'assurance requise pour le taux plein, entrée en vigueur de la liquidation unique dans les régimes alignés, introduction de coefficients de solidarité dans les régimes complémentaires Agirc et Arrco, etc.

De plus, l'entretien d'un modèle de microsimulation comme TRAJECTOIRE implique de conduire régulièrement des investissements méthodologiques : comparaison des résultats issus des simulations aux données observées dans différentes sources (notamment les deux échantillons inter-régimes) et aux simulations issues d'autres modèles, expertise des écarts, méthodologie afin d'améliorer la modélisation, etc.

« Les modules déjà existants font donc l'objet d'investissements et d'enrichissements réguliers. »

Les modules déjà existants font donc l'objet d'investissements et d'enrichissements réguliers. Par exemple, de nouveaux régimes de retraite sont régulièrement ajoutés au module CALIPER (le dernier en date est le régime spécial de la SNCF), le module de modélisation des départs à la retraite

a été repensé, le non-recours aux droits à retraite a été modélisé, le jeu de pondérations a été refondu, etc.

Par ailleurs, TRAJECTOIRE doit faire l'objet d'un rebasage complet lorsque de nouveaux millésimes des sources sur lesquelles il est adossé (les deux échantillons inter-régimes) sont disponibles¹⁵. Il doit également s'adapter aux évolutions de ces sources (par exemple, à l'occasion du millésime 2013, la couverture de l'échantillon des cotisants a été enrichie).

Enfin, malgré l'importance des travaux déjà menés, de nouveaux investissements seraient nécessaires afin d'affiner les résultats du modèle : modéliser les différences de mortalité selon les caractéristiques des retraités par exemple, ou encore modéliser le cumul emploi-retraite. Des besoins se font jour également pour simuler de manière plus exhaustive le système de retraite : modéliser les pensions de réversion par exemple, ce qui suppose de modéliser des mariages au sein de la population d'assurés, ou modéliser le compte professionnel de prévention.

MICROSIMULATION DES RETRAITES : UN PAYSAGE DIVERS ET FRAGMENTÉ...

Au cours des années deux-mille-dix, en même temps que la Drees développait TRAJECTOIRE, de nombreux autres modèles de microsimulation dynamiques sur les retraites voient

15. L'EIR 2016 est disponible depuis mi-2019. L'EIC 2017 est prévu pour fin 2020.

le jour. Les principaux organismes gestionnaires de régimes se sont dotés ainsi d'un tel outil. C'est le cas notamment :

- ❶ de la Caisse nationale d'assurance vieillesse (Cnav), dès la fin des années deux-mille, sur le champ des salariés du secteur privé (modèle Prisme) ;
- ❶ du Service des retraites de l'État (SRE) en 2015, pour les fonctionnaires d'État (modèle Pablo) ;
- ❶ de la Caisse nationale de retraite des agents des collectivités locales (CNRACL), à partir de 2019, pour les fonctionnaires relevant des fonctions publiques territoriale et hospitalière (modèle Canopée).

Parmi ces modèles, celui de la Cnav se singularise car, au fil du temps, il a progressivement intégré à son champ, outre le régime général, les régimes alignés ainsi que les régimes complémentaires Agirc et Arrco. Des développements sont en cours depuis 2019 pour faire de Prisme un modèle « tous régimes » intégrant notamment les régimes de la fonction publique.

L'Agirc-Arrco a également entrepris la construction d'un modèle de microsimulation au début de l'année 2018. En parallèle, la DG Trésor abandonne en 2016 sa maquette de cas-types pondérés pour développer le modèle de microsimulation Aphrodite, adossé à la base biographique du modèle Destinie de l'Insee mais utilisant sa propre modélisation pour simuler les âges de liquidation et les droits à pension des assurés. L'Institut des politiques publiques (IPP) engage dès 2014 le projet PENSIPP dont l'objectif est d'aboutir à un modèle de microsimulation également adossé à la base biographique de Destinie.

Le paysage actuel des modèles de microsimulation dynamiques « tous régimes » du système de retraite français se caractérise donc par une grande diversité des acteurs : l'Insee (modèle Destinie), la Drees (modèle TRAJECTOIRE), la DG Trésor (modèle Aphrodite), l'IPP (modèle PENSIPP), et la Cnav (modèle Prisme). Cette diversité, a pour inconvénient majeur d'entraîner un éclatement des ressources (humaines notamment), et parfois une certaine confusion quand les résultats des différents modèles divergent. Mais elle est cependant précieuse pour la connaissance et le pilotage du système de retraite.

❶ ... MAIS SOURCE DE RICHESSES

Paradoxalement, l'intérêt de l'accumulation des modèles découle d'abord des atouts et des limites de chacun d'entre eux. À titre d'exemple, les modèles adossés à l'enquête Patrimoine de l'Insee (Destinie, Aphrodite et PENSIPP) reposent sur des informations déclaratives, alors que les sources sur lesquels s'appuient TRAJECTOIRE et Prisme sont directement issues des fichiers de gestion des régimes de retraite.

L'avantage des données administratives, par rapport aux données d'enquête, réside dans leur fiabilité mais aussi dans la taille des échantillons qui permet des analyses plus fines par sous-populations. En outre, elles sont exhaustives concernant les droits à retraite alors que le champ de l'enquête Patrimoine se réduit à la population des résidents en France métropolitaine¹⁶.

16. À titre indicatif, au 31 décembre 2017, sur 17,2 millions de retraités, 1,6 million ne résidaient pas en France, soit environ 9 % (Arnaud [et alii](#), 2019).

À l'inverse, les fichiers administratifs sont moins riches en données socio-démographiques et renseignent généralement peu sur le ménage au sein duquel l'assuré réside. L'enquête Patrimoine collecte des informations relatives au ménage (ressources, nombre d'enfants, etc.) : ceci permet aux modèles qui y sont adossés d'aborder des problématiques comme l'évolution des niveaux de vie par âge au fil des générations, ou encore la prise en charge de la dépendance¹⁷.

Enfin cette richesse vient aussi de la grande diversité des acteurs, qui se distinguent par leur positionnement institutionnel dans le paysage français et donc par les missions

qu'ils ont vocation à remplir.

Les besoins diffèrent en effet, entre un gestionnaire de régime de retraite, une administration centrale ou un institut de recherche. Ces différences de finalité peuvent conduire à des développements propres à chaque modèle, qui se complètent et permettent d'éclairer une grande variété de problématiques en matière de retraite.

« Ces différences de finalité peuvent conduire à des développements propres à chaque modèle, qui se complètent et permettent d'éclairer une grande variété de problématiques en matière de retraite. »

🌐 ... FACE À DES DÉFIS IMPORTANTS

Aujourd'hui, les acteurs qui contribuent à la richesse de l'analyse doivent faire face à de nouveaux investissements importants.

D'une part, les sollicitations du législateur sont toujours plus nombreuses et toujours plus complexes à traiter : demandes de projections sur une temporalité de plus en plus fine (notamment concernant la montée en charge des évolutions réglementaires), simulations de scénarios réglementaires très éloignés de la réglementation actuelle dans le cadre des travaux visant à construire un système de retraite universel par points, etc.

D'autre part, les sources disponibles pour enrichir ces modèles se diversifient et se complexifient en même temps que leur volumétrie augmente. En plus des deux échantillons inter-régimes de la Drees, de nouveaux fichiers administratifs sont depuis quelques années disponibles et exploitables à des fins statistiques. C'est le cas notamment de l'Échange inter-régime de retraite, géré par la Cnav, qui comporte des données quasiment exhaustives sur les retraités du système français, mais aussi des données du groupement d'intérêt public « Union Retraite » qui détaillent les carrières d'un large échantillon d'assurés¹⁸. Enfin, à moyen terme, la constitution du répertoire de gestion des carrières unique, pilotée par l'Agirc-Arrco et qui rassemblera des informations exhaustives concernant les carrières des assurés pourrait fournir une source de choix pour les modèles de microsimulation¹⁹.

17. Voir l'article de Didier Blanchet sur les modèles de microsimulation, dans ce même numéro.

18. Il s'agit des données produites dans le cadre des campagnes de droit à l'information en matière de retraite.

19. Le RGCU a intégré les données de carrière du régime général en 2019. L'intégration des données de la MSA et du SSI doit intervenir en 2021. Les autres régimes de retraite doivent suivre d'ici au 31 décembre 2022.

Ces bouleversements, qui impliquent des investissements méthodologiques massifs, posent la question du dimensionnement des équipes qui ont en charge la maintenance, le développement et l'exploitation des modèles. Le projet de réforme actuel, qui prévoit, en cas d'adoption, la mise en place progressive d'un régime unique par points, offrira l'opportunité d'envisager une plus grande mutualisation des ressources allouées à la modélisation du système de retraite. Il impliquera également une réflexion concernant le rôle futur des différents organismes aujourd'hui en charge du suivi du système de retraite français (Conseil d'orientation des retraites, Comité de suivi des retraites, Drees, Insee, régimes de retraite), avec des répercussions possibles sur les différents modèles.

Encadré 2. Fiche technique du modèle TRAJECTOIRE

- **Champ** : générations nées entre 1942 et 2013 ;
- **Taux d'échantillonnage** : 2,2 %, soit environ 20 000 personnes par génération et 1,6 million de personnes pour l'ensemble du modèle ;
- **Jeu de pondération** calculé pour être représentatif des données du RNIPP sur un croisement année de naissance, pays de naissance et sexe. Le champ constitué des personnes enregistrées dans le RNIPP est en effet très proche de celui des assurés du système de retraite ;
- **Langage de programmation** : SAS. Réflexions en cours en vue du passage à un autre langage parmi plusieurs options possibles (R, Python, etc.) ;
- **Volumétrie du modèle** : 150 Gigaoctets ;
- **Temps d'exécution** : 24 heures (en pratique souvent moins car chaque simulation n'implique pas nécessairement de mobiliser l'ensemble des modules du modèle).

BIBLIOGRAPHIE

ARNAUD, Franck *et alii*, 2019. *Les retraités et les retraites*. [en ligne]. 6 juin 2019. Panoramas de la Drees Social, édition 2019. [Consulté le 29 avril 2020]. Disponible à l'adresse : <https://drees.solidarites-sante.gouv.fr/IMG/pdf/panorama-retraite-2019.pdf>.

AUBERT, Patrick, BARATON, Manuella, DUC, Cindy et CROGUENNEC, Yannick, 2012. *Les polypensionnés*. [en ligne]. 30 août 2012. Drees, Dossier Solidarité et Santé n°32. [Consulté le 29 avril 2020]. Disponible à l'adresse : <https://drees.solidarites-sante.gouv.fr/IMG/pdf/dossier32.pdf>.

AUBERT, Patrick, DUC, Cindy, et DUCOUDRÉ, Bruno, 2010. *Le modèle PROMESS : Projection « méso » des âges de cessation d'emploi et de départ à la retraite*. [en ligne]. Décembre 2010. Drees, Documents de travail, Série Études et Recherches, n°102. [Consulté le 29 avril 2020]. Disponible à l'adresse : <https://drees.solidarites-sante.gouv.fr/etudes-et-statistiques/publications/documents-de-travail/serie-etudes-et-recherche/article/le-modele-promess-projection-meso-des-ages-de-cessation-d-emploi-et-de-depart-a>.

AUBERT, Patrick, DUC, Cindy et DUCOUDRÉ, Bruno, 2012. Projeter l'impact des réformes des retraites sur les sorties d'activité : Une illustration par le modèle PROMESS. In : *Revue française des affaires sociales*. [en ligne]. La Documentation française. N° 4, pp. 84-105. [Consulté le 29 avril 2020]. Disponible à l'adresse : <https://www.cairn.info/revue-francaise-des-affaires-sociales-2012-4-page-84.htm>.

AUBERT, Patrick, DUC, Cindy et DUCOUDRÉ, Bruno, 2013. French Retirement Reforms and Intragenerational Equity in Retirement Duration. In : *Economist*. [en ligne]. Mai 2013. Springer Verlag, n°3, vol. 161, pp. 277-305. [Consulté le 29 avril 2020]. Disponible à l'adresse : <https://halshs.archives-ouvertes.fr/halshs-01435711/document>.

AUBERT, Patrick et PLOUHINEC, Corentin, 2017. Les différences de retraite entre secteur public et secteur privé : une analyse par simulations sur carrières types. In : *Âges et générations*. [en ligne]. 1^{er} mars 2017. Économie et Statistique / Economics and Statistics, n°491-492, pp. 25-42. [Consulté le 25 mars 2020]. Disponible à l'adresse : https://www.insee.fr/fr/statistiques/fichier/2647347/491-492_Aubert-Plouhinec_FR.pdf.

AUBERT, Patrick, PLOUHINEC, Corentin et PROUET, Emmanuelle, 2015. Les effets du temps partiel sur le taux de remplacement dans les secteurs privé et public. In : *Retraite et Société*. [en ligne]. La Documentation française, 2015/2, n°71, pp. 113-126. [Consulté le 29 avril 2020]. Disponible à l'adresse : <https://www.cairn.info/revue-retraite-et-societe1-2015-2-page-113.htm>.

BAC, Catherine, BONNET, Carole, BONTOUT, Olivier et CORNILLEAU, Gérard, 2003. *Exercices de simulation de l'évolution à long terme du système de retraite – Ensemble des variantes réalisées pour le Conseil d'Orientation des Retraites*. [en ligne]. Octobre 2003. Drees, Document de travail, Collection Études, n°38. [Consulté le 29 avril 2020]. Disponible à l'adresse : <https://drees.solidarites-sante.gouv.fr/IMG/pdf/serieetud38.pdf>.

BLANCHET, Didier et CHANUT, Jean-Marie, 1998. Les retraites individuelles à long terme : une projection par microsimulation. In : *Économie et statistique*. [en ligne]. Septembre 1998. N°315, pp. 95-106. [Consulté le 29 avril 2020]. Disponible à l'adresse : https://www.persee.fr/doc/estat_0336-1454_1998_num_315_1_2643.

CHANTEL, Cécile et PLOUHINEC, Corentin, 2014. *La réforme du minimum contributif applicable en 2012. Une simulation de son impact à partir de l'échantillon interrégimes de retraités de 2008*. [en ligne]. Avril 2014. Drees, Dossiers Solidarité et Santé, n°54. [Consulté le 29 avril 2020]. Disponible à l'adresse : <https://drees.solidarites-sante.gouv.fr/IMG/pdf/dss54.pdf>.

DUC, Cindy, 2015. *Les réformes des retraites depuis 1993 augmentent à terme l'âge moyen de départ de deux ans et demi*. [en ligne]. Avril 2015. Drees, Études et Résultats, n°915. [Consulté le 29 avril 2020]. Disponible à l'adresse : https://drees.solidarites-sante.gouv.fr/IMG/pdf/er915_toile.pdf.

DUC, Cindy, LEQUIEN, Laurent, HOUSSET, Félix et PLOUHINEC, Corentin, 2013. *Le modèle de microsimulation TRAJECTOIRE, Trajectoire de carrières tous régimes*. [en ligne]. Mai 2013. Drees, Document de travail, Série Sources et Méthodes, n°40. [Consulté le 29 avril 2020]. Disponible à l'adresse : https://drees.solidarites-sante.gouv.fr/IMG/pdf/seriesource_method40.pdf.

DUC, Cindy, MARTIN, Henri, et TRÉGUIER, Julie, 2016. *Les réformes des retraites de 2010 à 2015. Une analyse détaillée de l'impact pour les affiliés et pour les régimes*. [en ligne]. Décembre 2016. Les Dossiers de la Drees, n°9. [Consulté le 29 avril 2020]. Disponible à l'adresse : <https://drees.solidarites-sante.gouv.fr/IMG/pdf/dd9.pdf>.

DUC, Cindy, SOLARD, Gwennaël et TRÉGUIER, Julie, 2017. *Les différences de retraite entre secteurs public et privé : résultats de simulations*. [en ligne]. Mai 2017. Les Dossiers de la Drees, n°16. [Consulté le 29 avril 2020]. Disponible à l'adresse : <https://drees.solidarites-sante.gouv.fr/IMG/pdf/dd16.pdf>.

LERMECHIN, Hugues, DUC, Cindy et BURRICAND, Carine, 2011. *Présentation et applications de l'outil CALIPER (CALcul Interrégimes des PEnsions de Retraite)*. [en ligne]. Novembre 2011. Drees, Document de travail, Série Études et Recherches, n°111. [Consulté le 29 avril 2020]. Disponible à l'adresse : <https://drees.solidarites-sante.gouv.fr/IMG/pdf/serieetud111-2.pdf>.

① FONDEMENTS JURIDIQUES

Décret n° 2014-654 du 20 juin 2014 relatif au comité de suivi des retraites. *Légifrance*. [en ligne]. [Consulté le 29 avril 2020]. Disponible à l'adresse : <https://www.legifrance.gouv.fr/eli/decret/2014/6/20/2014-654/jo/texte>.

INES, LE MODÈLE QUI SIMULE L'IMPACT DES POLITIQUES SOCIALES ET FISCALES

Simon Fredon* et Michaël Sicsic**

Le modèle Ines, coproduction de l'Insee et de la Drees et plus récemment de la Cnaf, simule les principaux prélèvements et prestations sociales en France. Fondé sur un échantillon d'individus et de ménages issu de l'enquête Revenus fiscaux et sociaux, le modèle de microsimulation mobilise un ensemble de données riches en informations sur les caractéristiques socio-démographiques, le logement, l'emploi et les revenus des personnes. L'article relate l'histoire d'Ines, décrit les données utilisées, les dispositifs sociaux et fiscaux simulés et le fonctionnement du modèle ainsi que ses principales caractéristiques.

Les études réalisées avec ce modèle de microsimulation permettent d'éclairer le débat public dans les domaines de la redistribution monétaire, de la fiscalité ou de la protection sociale. Les trois premières applications décrites sont des études annuelles, qui fournissent une analyse ex post des effets du système socio-fiscal et de ses évolutions : analyse de la redistribution, bilan des réformes mises en œuvre chaque année, publication d'indicateurs avancés de pauvreté et d'inégalités. Les deux suivantes concernent des analyses ex ante : un cahier de variantes mettant en œuvre un ensemble de réformes socio-fiscales fictives, et des éclairages apportés sur des projets de réformes. Deux exemples d'études originales et plus ponctuelles concluent ce panorama.

 *The INES model, co-produced by INSEE and DREES and more recently the National Family Allowance Fund, simulates the main taxes and benefits in France. Based on a sample of individuals and households from the Tax and Social Revenues Survey, the microsimulation model uses a data set rich in information on people's socio-demographic characteristics, housing, employment and income. This article tells the story of INES, describes the data used, the tax and benefit transfers simulated and how the model functions, as well as its main characteristics.*

The studies carried out using this microsimulation model help to shed light on the public debate in the fields of monetary redistribution, taxation or social security protection. The first three applications described are annual studies, which provide an ex-post analysis of the effects of the tax and benefit system and its evolutions: analysis of redistribution, assessment of the reforms implemented each year and nowcasting of poverty and inequality indicators. The following two concern ex-ante analyses: simulations of fictitious tax and benefit reforms and regular political orders on reform projects. Two examples of original and more selective studies conclude this overview.

* Chargé d'études au bureau redistribution et évaluation, Direction de la Recherche, des Études, de l'Évaluation et des Statistiques (Drees),
simon.fredon@sante.gouv.fr

** Chargé d'études à la division Études Sociales, Insee,
michael.sicsic@insee.fr

L'analyse de l'impact de transferts sociaux et fiscaux est un champ important de l'évaluation des politiques publiques. Si de nombreuses études évaluent l'effet d'un transfert particulier, il est difficile d'avoir une vision globale du système socio-fiscal et de ses répercussions sur les niveaux de vie. En effet, la législation des prélèvements et des prestations sociales en France est complexe (fondée sur de nombreux codes : le code général des impôts, de l'action sociale et des familles, ou de la sécurité sociale), et les interactions entre transferts socio-fiscaux, nombreux. Les modèles de microsimulation permettent d'identifier les effets de ces différents dispositifs sur les revenus des individus et de leur famille, en tenant compte de leurs caractéristiques spécifiques. Ces modèles restituent une analyse globale des transferts en fonction des configurations familiales, du niveau de vie des ménages ou encore du statut d'activité des individus. Ils permettent également de prendre en compte toutes les interactions qui existent entre les dispositifs. Les modèles de microsimulation ont ainsi acquis « *une place centrale, mais un peu méconnue, dans le domaine de l'analyse des politiques sociales et fiscales* » (Legendre, 2019).

Cet article présente le modèle de microsimulation Ines¹. De la famille des modèles de microsimulation statiques², Ines simule les principaux prélèvements et prestations en France. Il permet de réaliser des études approfondies afin d'éclairer le débat dans les domaines de la redistribution monétaire, la fiscalité ou la protection sociale.

📌 UN MODÈLE POUR SIMULER LA LÉGISLATION SOCIO-FISCALE —

Les modèles de microsimulation « statiques » simulent pour chaque personne les prélèvements qu'elle paye et les prestations qu'elle reçoit à partir des barèmes de la législation, ce qui permet d'en déduire son revenu disponible. Pour ce faire, ils s'appuient sur des données réelles (âge, composition familiale, revenus, statut d'activité des individus, etc.). Il est possible de faire varier les barèmes et d'étudier l'effet de ces variations sur le revenu disponible des ménages ou sur les finances publiques, et donc *in fine* d'apporter des éclairages utiles sur les effets de ces politiques publiques.

Ces modèles permettent ainsi d'évaluer l'effet d'une variation des barèmes de certains prélèvements ou prestations sociales sur le niveau de vie des personnes concernées, de repérer les gagnants et les perdants des réformes, ou encore de mesurer l'effet des réformes sur le taux de pauvreté, les inégalités de niveaux de vie, et les finances publiques. Pour cela, on compare deux situations : celle dans laquelle les barèmes ont été modifiés, par rapport à celle « contrefactuelle », généralement la législation actuellement appliquée. Ces modèles sont dits statiques, car ils tiennent compte de l'effet mécanique de la modification du barème, mais pas des effets de comportement pouvant découler des réformes (modification de l'offre de travail, par exemple) et des effets de bouclages macroéconomiques (*via* une variation de la consommation, puis du PIB et de l'emploi par exemple).

C'est dans ce but que le modèle Ines a vu le jour en 1995. Développé par l'Insee (David *et alii*, 1999), il est ensuite rejoint par la Drees au début des années deux-mille et la Cnaf en 2016 (**encadré 1**). Dès le début de son histoire, le modèle Ines a été utilisé comme outil

1. Ines est l'acronyme d'« Insee-Drees », les deux organismes qui développent conjointement le modèle : la Cnaf cogère également le modèle avec l'Insee et la Drees depuis 2016 (**encadré 1**).

2. Voir l'article de Didier Blanchet dans ce même numéro pour une discussion générale sur les modèles de microsimulation.

d'aide à la décision pour calibrer les grandes réformes sur la fiscalité et la redistribution. Ainsi, le modèle a été utilisé pour chiffrer la mise en place de la prime pour l'emploi, nouveau crédit d'impôt, en 2001, la construction et la mise en place du revenu de solidarité active (RSA) en 2009³, de la Prestation d'accueil du jeune enfant en 2014 (Paje), et de la prime d'activité en 2016. En 2019, il a servi à calibrer les mesures de réponse à l'urgence sociale suite au mouvement des gilets jaunes et permet d'alimenter les travaux autour du revenu universel d'activité (RUA).

Encadré 1. Ines, une déjà longue histoire

Le modèle Ines est né en 1995, dans la division des Études sociales de l'Insee, suite aux recommandations du rapport Guibert(i). Deux autres modèles avaient été développés en France, mais avec chacun des lacunes(ii) :

- le modèle d'impôt sur le revenu (MIR) ne simule que... l'impôt sur le revenu ;
- la base de données du modèle Sysiff repose sur l'enquête Budget de famille et présente un défaut de robustesse, car les revenus sont renseignés sur une base déclarative.

Dès sa conception, Ines pallie ces insuffisances en simulant les principaux prélèvements et prestations à partir de l'enquête Revenus fiscaux (ERF) de 1990 : cette source regroupe des données fiscales et les informations socio-démographiques du recensement. Le modèle est présenté dès 1996(iii) et documenté en 1999 (David *et alii*, 1999). Il est rapidement utilisé dans le débat politique pour faire un bilan de la politique familiale et évaluer le projet de mise sous condition de ressources des allocations familiales, dont les résultats sont publiés (Thélot et Villac, 1998)(iv).

Au début des années deux-mille, la Drees (Direction de la recherche, des études, de l'évaluation et des statistiques, service statistique du ministère en charge de la Santé et des Solidarités) rejoint l'Insee pour développer le modèle Ines(v). La nouvelle version se base sur l'ERF de 1996, appariement entre l'enquête Emploi et les déclarations fiscales ; elle est présentée en 2003 (Albouy *et alii*, 2003). La collaboration permet de mutualiser les coûts de développement et de maintenance du modèle, qui en sort « *plus fiable que le précédent du fait de procédures de validations croisées et des échanges de savoirs entre les deux équipes* » (Bessis et Cotton, 2019).

En 2006, l'ERF fait place à l'enquête Revenus fiscaux et sociaux (ERFS), fondée sur l'enquête Emploi en continu, les données fiscales et sociales.

Une nouvelle étape importante se déroule en 2016 avec la mise à disposition publique du code source du modèle sur le site de l'Adullact (Association des développeurs et utilisateurs de logiciels libres pour les administrations et les collectivités territoriales), et l'arrivée de la Cnaf (Caisse nationale des allocations familiales) qui rejoint les équipes Insee-Drees pour cogérer le modèle.

(i) Ce rapport a été commandé en 1994 par Michel Glaude, directeur de la direction des statistiques démographiques et sociales de l'Insee de l'époque, et par ailleurs auteur des premières expériences de microsimulation à l'Insee.

(ii) Voir l'article de Didier Blanchet dans ce même numéro ainsi que (Legendre, 2019) et (Bessis et Cotton, 2019).

(iii) Lors d'une journée d'étude à l'Insee, qui a donné lieu au premier dossier de revue consacré en France aux méthodes de microsimulation (Insee, 1998).

(iv) Ce rapport, commandé par le Premier ministre, contient les premiers résultats publiés à partir d'Ines et une première présentation sommaire de ce modèle. Les conséquences du rapport seront l'annulation de la mise sous condition de ressources des allocations familiales et le plafonnement de l'avantage dû au quotient familial.

(v) Au passage, la signification de l'acronyme devient désormais « Insee-Drees », elle était auparavant une contraction de « Insee-étude sociales ».

3. (Bessis et Cotton, 2019) détaillent le rôle des modèles de microsimulation Ines, Saphir (Direction Générale du Trésor) et Myriade (Cnaf) dans la création du RSA.

LES DONNÉES EN ENTRÉE DU MODÈLE

Contrairement aux maquettes de cas types qui simulent la législation pour un individu fictif, les modèles de microsimulation se basent sur l'observation de situations réelles, et sont donc très liés aux données utilisées. Depuis 2006, le modèle Ines utilise principalement les données de l'enquête Revenus fiscaux et sociaux (ERFS) (**encadré 2**). Cette source contient des centaines d'informations détaillées sur les revenus, l'emploi ou la situation familiale pour un échantillon de plus de 50 000 ménages et 130 000 individus. L'ERFS est la source privilégiée par l'Insee pour l'analyse des niveaux de vie. Cette richesse d'information permet de pouvoir simuler finement les prestations, cotisations et contributions sociales qui dépendent de nombreuses variables, pas toujours présentes dans les seules sources fiscales : contour de la famille, historique de la situation sur le marché du travail, nombre d'heures travaillées, type d'emploi et d'entreprise, loyers, lieu d'habitation, situation de handicap, etc.

« L'échantillon de l'ERFS est représentatif de la population française vivant en logement ordinaire et en France métropolitaine. »

L'échantillon de l'ERFS est représentatif de la population française vivant en logement ordinaire et en France métropolitaine (**encadré 2**). Cela permet de produire des analyses en répartissant la population en dix groupes de taille égale selon leur niveau de vie (des 10 % les plus modestes aux 10 % les plus aisés), ou selon le statut d'activité ou

la configuration familiale. Certains résultats peuvent être donnés à des niveaux plus fins, en divisant la population en 20 par exemple, afin d'estimer l'effet de certaines mesures spécifiques qui touchent surtout les plus aisés, comme la transformation de l'impôt de solidarité sur la fortune (ISF) en impôt sur la fortune immobilière (IFI) (Paquier *et alii*, 2019). Mais l'effet estimé aux extrémités est moins robuste que ceux estimés par dixième et la taille de l'échantillon ne permet pas de détailler davantage, par centième par exemple, contrairement aux sources exhaustives.

Notons enfin que les données portant sur les revenus de l'année N ne sont disponibles qu'en N+2 : pour les besoins du modèle (simulation de la législation de l'année N+2), cela nécessite de « vieillir » les données de l'ERFS (voir *infra*).

LES DISPOSITIFS SOCIAUX ET FISCAUX SIMULÉS PAR LE MODÈLE INES

Pour chaque personne, foyer ou ménage, en fonction du niveau d'analyse requis, le modèle Ines mobilise les données de l'ERFS et leur applique l'ensemble des règles de la législation. Il est par exemple ainsi possible d'identifier les ménages⁴ éligibles à la prime d'activité, en utilisant la composition familiale (nombre d'enfants à charge, couple ou famille monoparentale), la situation mensuelle des personnes par rapport à l'emploi reconstitué à partir d'un « calendrier d'activité » et les revenus du ménage. En mobilisant toutes ces informations pour chaque ménage potentiel, le modèle permet de simuler un droit à la prestation qui est intégré ensuite au revenu disponible de ce ménage, en prenant en compte la possibilité d'un non-recours à la prestation, c'est-à-dire que les personnes ne demandent pas celle-ci bien qu'ils en aient le droit.

4. Pour les prestations sociales, le contour de la famille est souvent différent de celui du ménage (on parle de « foyer RSA » ou « foyer prime d'activité »). Par simplicité, dans cet article, nous ne parlons que de ménages.

Les revenus sont cruciaux pour simuler les différents transferts. Les revenus utilisés en entrée du modèle et qui demeurent constants en cas de modification du montant des différents transferts reçus ou payés sont les revenus nets déclarés dans la déclaration fiscale (on parle de point fixe du modèle). À partir de ces revenus déclarés, le modèle Ines permet de simuler la plupart des transferts fiscaux et sociaux⁵ comme l'illustre la **figure 1**. Sont simulées les cotisations et contributions sociales, qu'elles soient à la charge de l'employeur, du salarié ou de l'indépendant. De même l'ensemble des autres prélèvements et subventions

Encadré2. Les données utilisées par le modèle Ines

● L'enquête Revenus fiscaux et sociaux (ERFS)

Pour chaque année N, l'ERFS est constituée par un appariement entre les répondants à l'enquête Emploi du 4^e trimestre et les sources fiscales de l'année, à savoir les déclarations de revenus de l'année N (remplies en mars N+1), la taxe d'habitation au 1^{er} janvier de l'année N, ainsi que les fichiers de la caisse nationale des allocations familiales (Cnaf), la caisse nationale de l'assurance vieillesse (Cnav) et la caisse centrale de la mutualité sociale agricole (CCMSA) qui donnent les prestations sociales versées. Notons que le modèle n'utilise pas directement les données de l'ERFS relatives à l'impôt sur le revenu (IR) ou aux prestations mais les simule*.

La dernière version du modèle (2018) utilise le millésime 2016 de l'enquête. L'échantillon sur lequel se base l'ERFS 2016, tiré dans les fichiers de la taxe d'habitation, est constitué de 120 000 individus pour 54 000 ménages répondants, ménages dits « ordinaires » de France métropolitaine : sont ainsi exclus les ménages vivant en collectivité (foyers, prisons, hôpitaux, etc.), de même que les personnes vivant dans des habitations mobiles (mariniers, etc.) et les sans-domicile.

● Les autres bases de données utilisées par le modèle

L'enquête Budget de famille de l'Insee est utilisée pour imputer les données de consommation à partir desquelles la TVA acquittée par les ménages est simulée dans Ines. Il s'agit d'une enquête réalisée depuis 1979 sur la consommation des ménages, l'objectif étant de mesurer le plus précisément possible les dépenses mais aussi les ressources des ménages résidant en France (Métropole et DROM). Elle couvre l'ensemble des ménages dits « ordinaires ».

L'enquête Logement est utilisée dans Ines pour imputer des loyers, qui sont absents de l'ERFS. L'enquête a pour objet de décrire les conditions de logement des ménages et leurs dépenses en logement et contient donc les loyers et charges des locataires ainsi que de nombreuses autres informations.

Afin de simuler l'impôt sur la fortune (ISF) et l'impôt sur la fortune immobilière (IFI), il est nécessaire de disposer des patrimoines des individus. Pour cela, nous mobilisons un appariement statistique à partir des enquêtes Patrimoine de l'Insee. Ces enquêtes décrivent les biens immobiliers, financiers et professionnels des ménages ainsi que leur endettement, à partir d'un échantillon tiré dans les fichiers de la taxe d'habitation et d'autres sources fiscales. Pour mieux appréhender les hauts patrimoines, on mobilise également les fichiers spécifiques à l'ISF et à l'IFI, récemment mis à disposition par la DGFIP (Paquier *et alii*, 2019).

* Les données de l'ERFS sur certaines prestations sont toutefois utilisées dans le modèle, par exemple l'Allocation adulte handicapé (AAH) pour mieux prendre en compte les personnes en situation de handicap.

5. On simule les cotisations et les contributions sociales en appliquant la législation à un revenu brut, lui-même reconstitué à partir du revenu net et des taux de cotisation et contribution.

assis sur la masse salariale sont simulés, comme le crédit d'impôt compétitivité-emploi (CICE) ou la taxe sur les salaires ou les transports. Pour l'impôt direct le modèle simule l'impôt sur le revenu, l'ensemble des crédits d'impôt et les prélèvements forfaitaires⁶. Récemment le développement d'un module sur le patrimoine des ménages (Paquier

« Récemment le développement d'un module sur le patrimoine des ménages a permis la simulation de l'impôt de solidarité sur la fortune (ISF) et de l'impôt sur la fortune immobilière (IFI). »

et alii, 2019) a permis la simulation de l'impôt de solidarité sur la fortune (ISF) et de l'impôt sur la fortune immobilière (IFI). Enfin le module de taxation indirecte (André et alii, 2016) permet d'ajouter à l'analyse des prélèvements tels que la taxe intérieure de consommation sur les produits énergétiques (TICPE) ou les taxes sur le tabac.

Du côté des prestations versées aux ménages, le modèle permet de simuler les prestations familiales avec ou sans condition de ressources⁷, les aides personnelles au logement, les minima sociaux⁸, la prime d'activité et la garantie jeune. Pour certaines de ces prestations comme la prime d'activité, du non-recours est pris en compte. Enfin certains transferts en nature peuvent aussi être simulés⁹, pour une analyse d'un revenu disponible « élargi » (figure 1).

Le plus souvent, ces transferts sont simulés selon la législation de l'année antérieure, mais ils peuvent également être simulés selon une autre législation (voir *infra*).

LE FONCTIONNEMENT DU MODÈLE AUJOURD'HUI

Le modèle Ines simule ces transferts socio-fiscaux grâce à un enchaînement de plus de 100 programmes informatiques¹⁰. Pour ce faire, il prend en entrée une version de l'ERFS qui est d'abord retravaillée dans la partie « amont » du modèle, puis utilisée dans la partie « aval » pour simuler les différents transferts du système socio-fiscal d'une année de législation donnée, à partir des barèmes législatifs. L'« amont » permet de garantir que le format des données sera exploitable par le reste du modèle. L'« aval » contient l'ensemble des programmes simulant la législation. Dans cette partie, on applique les paramètres de la législation (et éventuel non-recours) aux données sorties de la partie amont. C'est à cette opération que nous faisons référence quand nous parlons de « simulation ». Elle se distingue des « imputations » présentes dans l'amont, qui ont vocation à ajouter ou à se caler sur des données externes.

En sortie, le modèle permet d'obtenir les masses financières et les effectifs des bénéficiaires ou personnes assujetties aux différents transferts, le revenu disponible et le niveau de vie des ménages¹¹, ainsi que différents indicateurs usuels d'inégalité et de pauvreté (figure 2).

6. La taxe d'habitation est directement présente dans les données de l'ERFS ; la taxe foncière n'est pas simulée.

7. i.e. allocations familiales, allocation de soutien familial, allocation d'éducation de l'enfant handicapé, prime de naissance et allocation de base de la prestation d'accueil du jeune enfant, complément familial, allocation de rentrée scolaire et complément libre choix d'activité.

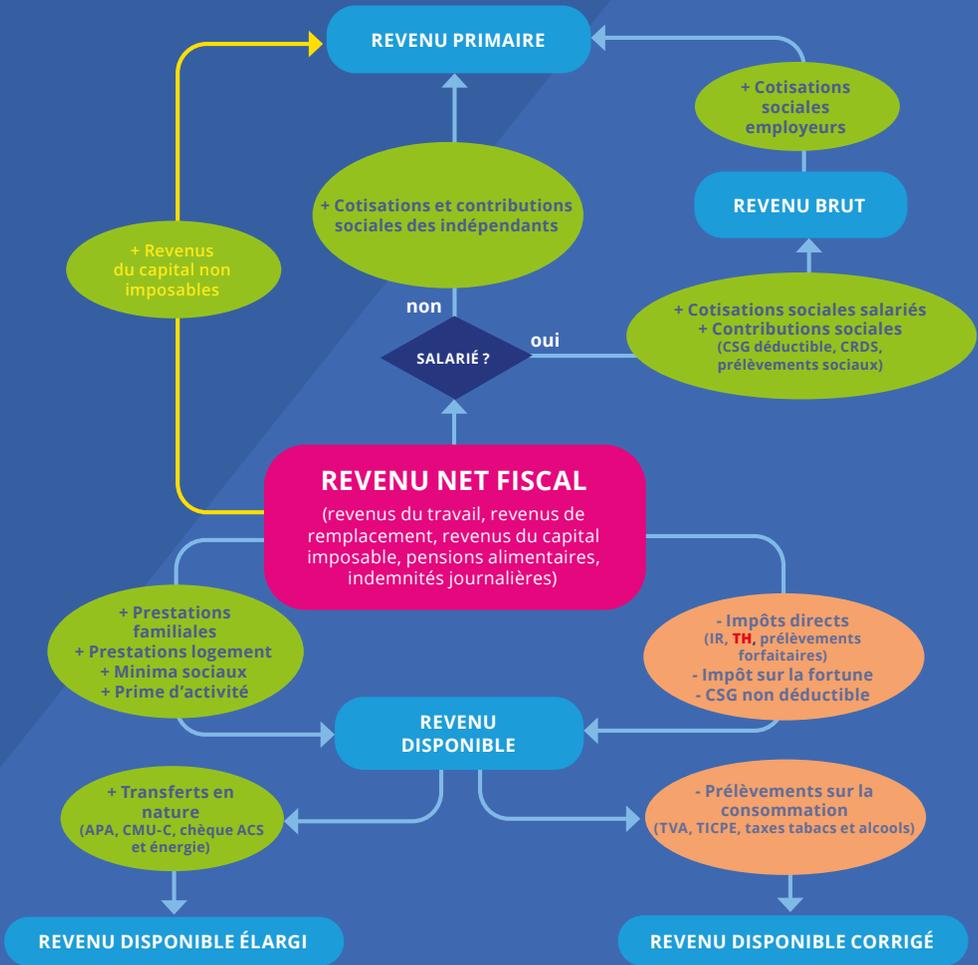
8. i.e. revenu de solidarité active, allocation de l'adulte handicapé, allocation supplémentaire d'invalidité, allocation de solidarité aux personnes âgées, allocation de solidarité spécifique et garantie jeune (depuis 2017).

9. i.e. allocation personnalisée d'autonomie, couverture maladie universelle complémentaire et chèques aide au paiement d'une complémentaire santé (Sireyjol, 2016), chèque énergie, ou bourses du secondaire.

10. Ces programmes sont codés en SAS. Une migration du modèle vers le langage libre R est en cours.

11. Le niveau de vie d'un ménage correspond à son revenu disponible rapporté à un nombre d'unité de consommation, permettant de prendre en compte la composition du ménage et les économies d'échelle.

Figure 1. Les différents concepts de revenu et les dispositifs socio-fiscaux simulés avec Ines



Note : en caractères bleus figurent les dispositifs fiscaux ou sociaux qu'Ines arrive à simuler.

Signification des sigles et acronymes :

- ACS : aide au paiement d'une complémentaire santé
- APA : allocation personnalisée d'autonomie, mesure sociale en faveur des personnes âgées et dépendantes
- CMU-C : couverture maladie universelle complémentaire
- CRDS : contribution pour le remboursement de la dette sociale, impôt créé dans le but de résorber l'endettement de la Sécurité sociale
- CSG : contribution sociale généralisée, impôt qui participe au financement de la sécurité sociale et de l'assurance chômage
- IR : impôt sur le revenu
- TH : taxe d'habitation
- TICPE : taxe intérieure de consommation sur les produits énergétiques
- TVA : taxe sur la valeur ajoutée

Le fait que les transferts soient simulés permet de comparer l'ensemble des résultats du modèle dans différentes situations en fonction de changements législatifs et d'analyser les effets redistributifs.

Pour tourner, le modèle a besoin de deux paramètres importants : l'année de référence (l'année de l'ERFS en entrée) et l'année de législation que l'on souhaite simuler. Dans son utilisation courante ce couple est généralement composé de la dernière version de l'ERFS disponible (année N) et de l'année de législation correspondant à deux ans après (année N+2).

« Concrètement, ce vieillissement se fait en deux étapes, une étape de calage en modifiant les poids de chaque ménage dans l'ERFS et une étape de dérive des revenus. »

Quelle que soit l'année de législation choisie, les données de l'ERFS, sont « vieilles » de deux ans dans la partie amont du modèle, afin de représenter au mieux la population des années N+1 et N+2. Pour cela, le modèle prend en compte une multitude d'évolutions socio-démographiques (pyramide des âges, les catégories socio-professionnelles, etc.) ou économiques (inflation, évolution des salaires, etc.) entre l'année de référence et deux ans plus tard.

Concrètement, ce vieillissement se fait en deux étapes, une étape de calage en modifiant les poids de chaque ménage dans l'ERFS¹² et une étape de dérive des revenus¹³ :

- 1. on s'assure d'abord par exemple qu'on obtient le même nombre d'enfants, de chômeurs et de couples que dans les sources Insee de l'année N+2, puis on fait évoluer les revenus entre l'année N et l'année N+2. Par exemple, si le nombre de chômeurs a augmenté entre l'année N et l'année N+2, on augmente le poids des chômeurs dans l'échantillon dans la première étape¹⁴ ;
- 2. pour la deuxième étape, on applique une évolution différente selon le type de revenu¹⁵ (salaires, pensions de retraites ou de chômage, revenus financiers, revenus agricoles, etc.) et selon l'année. Entre l'année N et N+1, on utilise en général les données fiscales récemment disponibles, et entre N+1 et N+2 on utilise les données Insee de la comptabilité nationale, de la Banque de France (pour les revenus financiers), ou d'enquêtes de la statistique publique¹⁶.

Le modèle est mis à jour chaque année entre février et octobre afin de simuler la législation de l'année précédente (**figure 3**). En 2020, le modèle évolue ainsi pour intégrer la législation 2019. Cette mise à jour consiste d'une part à actualiser l'ensemble des paramètres des différents barèmes législatifs. Par exemple, on met à jour en 2020 le montant forfaitaire de l'allocation adulte handicapé, qui est passé de 860 euros à 900 euros en novembre 2019. Elle implique d'autre part un travail de codage des nouveaux dispositifs sociaux et fiscaux, qui peut s'avérer conséquent quand des réformes d'ampleur voient le jour.

12. Pour cela on utilise une méthode classique de calage sur marge.

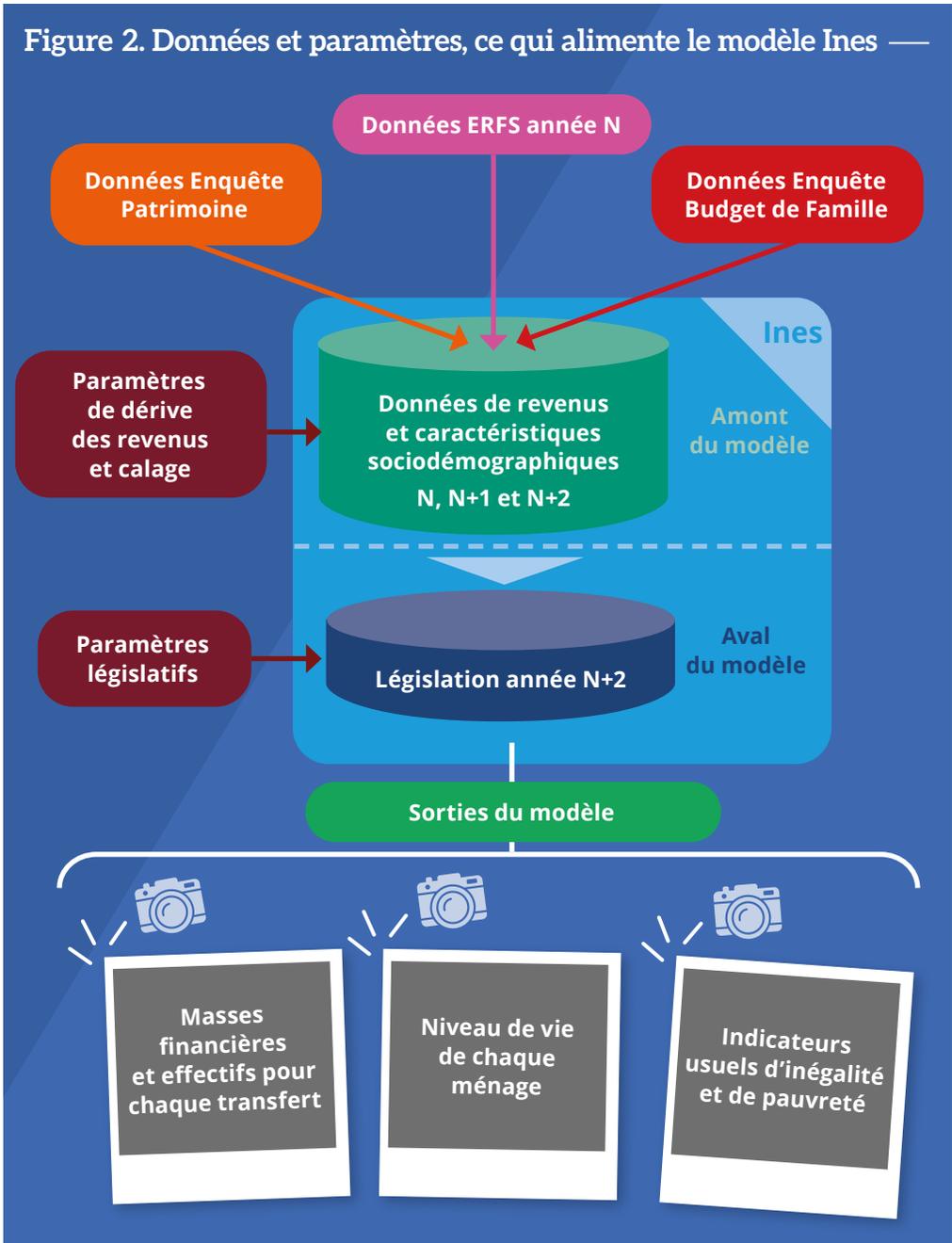
13. Ces deux étapes sont elles-mêmes liées : on prend en compte le calage dans l'étape de dérive des revenus.

14. Le modèle ne simule pas de transitions entre statuts d'activité, (par exemple entre le travail et le chômage), comme le font d'autres modèles comme Euromod.

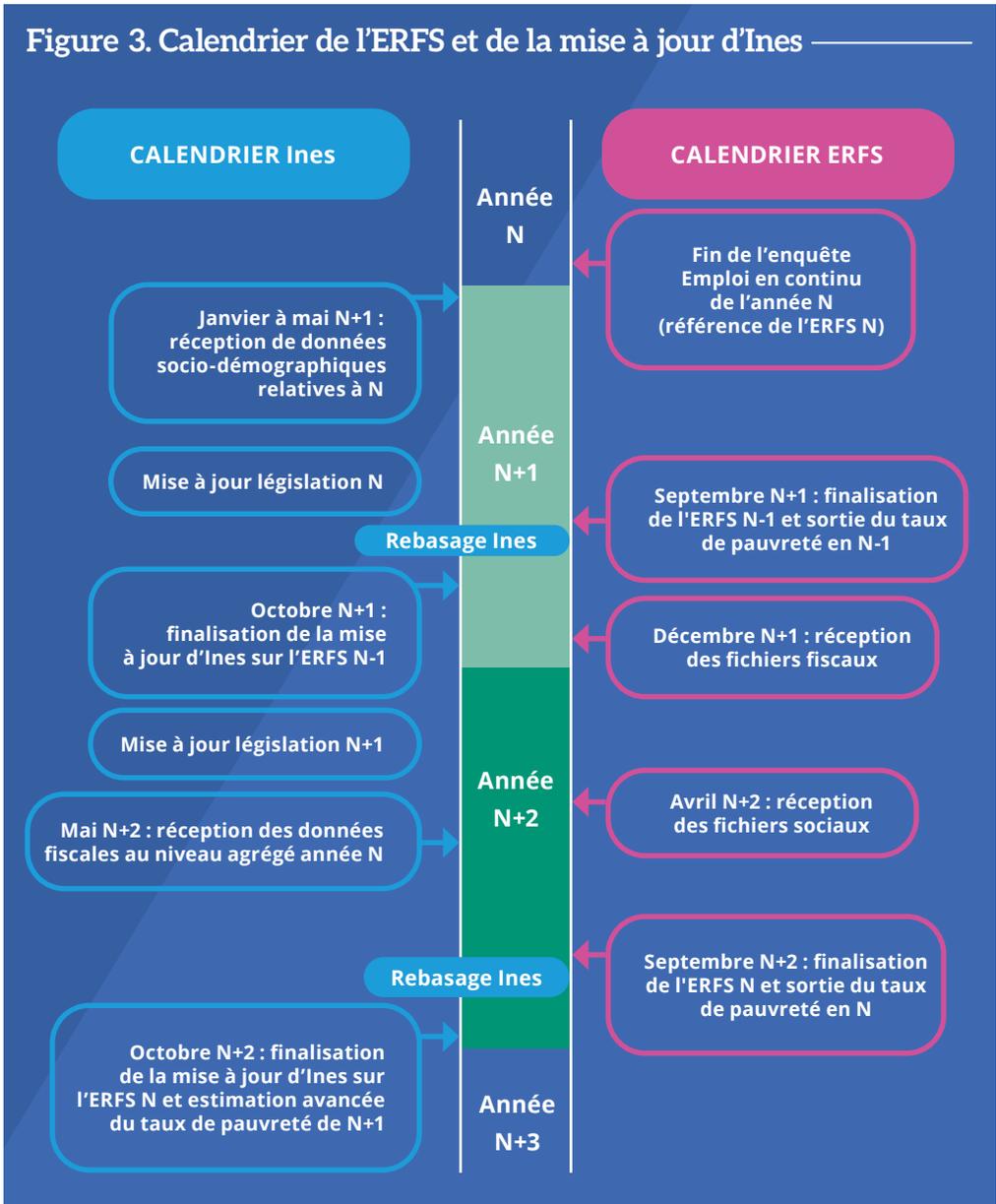
15. Une trentaine de types de revenus sont différenciés.

16. Par exemple, l'enquête ACEMO pour les salaires qui permet de tenir compte de la catégorie socio-professionnelle et du secteur d'activité de chacun des salariés. On utilise également l'inflation pour certains revenus.

Figure 2. Données et paramètres, ce qui alimente le modèle Ines —



En plus de ces mises à jour annuelles, le modèle est modifié en continu pour être amélioré ou corrigé. Des modules complémentaires sont en outre régulièrement créés. Facultatifs dans l'exécution du modèle, ils permettent d'étendre le champ des mesures couvert et d'étoffer les sorties, avec notamment l'analyse des taxes indirectes (principalement TVA et les droits d'accise sur le tabac et l'alcool) et des prélèvements sur le patrimoine (impôt sur la fortune, ISF, et impôt sur la fortune immobilière, IFI).



📌 UN MODÈLE STATIQUE MAIS POUVANT INTRODUIRE DES EFFETS DE COMPORTEMENTS

Dans son utilisation la plus courante, le modèle ne prend en compte aucun effet de comportement dans l'évaluation des réformes socio-fiscales, il est purement statique.

Or, des réactions comportementales peuvent résulter des réformes mises en œuvre sur les transferts simulés : par exemple, une baisse de la consommation de tabac suite à la hausse de la fiscalité, une hausse des dividendes versés suite à la mise en place du prélèvement forfaitaire unique (PFU) en 2018 ou encore une modification de l'offre de travail suite aux réformes de la prime d'activité.

Les effets de bouclages macro-économiques ne sont pas non plus traités, comme l'impact positif d'une hausse des revenus sur la consommation, le PIB et en retour sur l'emploi et les revenus. Introduire ces incidences nécessite en effet de faire de nombreuses hypothèses, en s'appuyant sur des travaux souvent divergents, si sérieux soient-ils. L'analyse statique est moins contestable et ainsi l'absence de prise en compte d'effets de comportement est « *quasi systématique par la plupart des gouvernements et des institutions indépendantes chargées de réaliser des chiffrages de mesures économiques* » (Bach et alii, 2020).

Il faut ainsi bien avoir en tête que les effets mesurés sont en général des effets de court terme, ne prenant pas en compte les effets dynamiques. Certaines études ont toutefois estimé des élasticités comportementales à partir du modèle Ines (Fugazza et alii, 2003 ; Lehmann et alii, 2013 ; Sicsic, 2019). L'évaluation de la mise en place de la prime pour l'emploi en 2001 a ainsi été réalisée en prenant en compte des effets sur l'offre de travail (Fugazza et alii, 2003). Des travaux actuels mesurent l'impact du PFU et de la réforme de l'ISF en 2018 en intégrant des effets induits par ces dispositifs sur les montants du patrimoine et les dividendes.

📌 UNE DOCUMENTATION DÉTAILLÉE ET UN MODÈLE EN OPEN SOURCE

Ines est le premier modèle de microsimulation sur données représentatives à être entré dans l'ère du libre : depuis 2016 le code source du modèle et sa documentation sont publiques, et disponibles en ligne sur le site de l'Adullact (Association des développeurs et utilisateurs de logiciels libres pour les administrations et les collectivités territoriales) (voir <https://adullact.net/projects/ines-libre>).

L'accès aux données de l'ERFS demeure fortement contrôlé du fait des données fiscales présentes. Mais toute personne ayant ce droit d'accès peut désormais utiliser le modèle pour réaliser des études ou évaluer des réformes. Cette ouverture donne au modèle « *une position centrale illustrée par exemple par son utilisation par l'OFCE* » (Legendre, 2019). L'OFCE¹⁷ fait en effet partie des tout premiers utilisateurs du modèle et saluait en 2017 cette mise à disposition comme une « *révolution silencieuse dans la statistique publique* » (Madec et Timbeau, 2017).

17. L'Observatoire français des conjonctures économiques est un organisme indépendant de recherche, de prévision et d'évaluation des politiques publiques créé par l'État en février 1981, au sein de la Fondation nationale des sciences politiques.

La richesse de la documentation est fondamentale : tant l'équipe de développement que les utilisateurs « libres » doivent pouvoir comprendre le modèle et connaître la qualité de ses résultats en fonction des dispositifs étudiés.

À la fin de la période de mise à jour, l'équipe de développement produit une note de validation de la version stabilisée du modèle. Cette version est aussi appelée modèle « figé » de l'année de législation qui vient d'être intégrée. Grâce à cette version « figée », ou *tag* dans le vocabulaire des outils de gestion des versions, l'évolution constante du modèle ne perturbe pas la reproductibilité des résultats. Ceci est particulièrement utile quand le modèle est utilisé pour répondre à la demande, car les chiffrages doivent pouvoir être reproduits avec une certaine constance d'une semaine sur l'autre, voire d'une année sur l'autre.

La note annuelle de validation permet de faire un état des lieux des forces et faiblesses du modèle : en particulier, elle compare les sorties du modèle avec des effectifs de bénéficiaires de prestations et des masses versées ou payées, issues de sources externes et appelées « cibles externe » ; ces dernières sont corrigées pour se ramener à un champ équivalent à celui de l'ERFS.

« Dans la plupart des cas, lorsque les dispositifs sont simulés avec une moindre précision, c'est parce qu'ils s'adressent à un public très spécifique que les données ne permettent pas d'identifier parfaitement bien. C'est notamment le cas de l'allocation aux adultes handicapés. »

Les résultats des simulations ne sont en effet pas « calés » sur des sources externes, ce qui permet de juger leur qualité. On montre ainsi que les simulations de l'impôt sur le revenu, des allocations familiales, des allocations logement ou de la prime d'activité sont très proches des cibles, tandis que les

simulations de l'allocation aux adultes handicapés (AAH) ou des prestations pour les jeunes enfants (CMG, PrePare, AEEH) sont de moins bonne qualité. Dans la plupart des cas, lorsque les dispositifs sont simulés avec une moindre précision, c'est parce qu'ils s'adressent à un public très spécifique que les données ne permettent pas d'identifier parfaitement bien. C'est notamment le cas de l'allocation aux adultes handicapés.

Une documentation plus générale du modèle peut être trouvée par les utilisateurs inscrits sur le site de l'Adullact dans un « wiki » spécifique au modèle Ines. De plus, chaque développement de nouveau module ou de nouveaux transferts simulés dans Ines donne lieu à des documents de travail documentant la méthode et les résultats de ces ajouts¹⁸.

LES UTILISATIONS RÉCURRENTES DU MODÈLE INES

Depuis le milieu des années quatre-vingt-dix, l'agrandissement de l'équipe Ines et l'enrichissement progressif du modèle par l'ajout de nouveaux modules pouvant étoffer les simulations, ont permis une diversification des utilisations et des publications liées à celui-ci. On détaille par la suite cinq utilisations importantes du modèle Ines. Les trois premières

18. Cela a notamment été le cas pour la simulation de la CMU-C et des chèques ACS (Sireyjol, 2016), des prélèvements indirects (André *et alii*, 2016), de l'ISF et de l'IFI (Paquier *et alii*, 2019).

se concentrent sur une analyse du système socio-fiscal tel qu'il existe ou de l'effet de réformes qui ont eu lieu (analyse *ex post*), alors que les deux suivantes servent à étudier les effets de pistes de réformes ou de réformes fictives (analyse *ex ante*). On présente également deux exemples d'études plus ponctuelles dans l'**encadré 3**.

❶ Comment le système socio-fiscal redistribue et réduit les inégalités ?

Un des principaux objectifs du modèle Ines est de décrire la redistribution opérée par le système socio-fiscal. Depuis les années deux-mille, une photographie annuelle de la redistribution est publiée dans l'ouvrage « France, portrait social » de l'Insee¹⁹. Les montants des différents transferts par décile de niveau de vie y sont indiqués, ainsi que le revenu disponible avant et après redistribution, ce qui permet de chiffrer la baisse des inégalités due au système socio-fiscal. Ainsi, en 2018, les 10 % de personnes les plus pauvres disposent d'un niveau de vie moyen avant redistribution de 3 290 euros par an, contre 73 130 euros pour les 10 % les plus aisées, soit 22,2 fois plus (Insee, 2019). Après redistribution, ce rapport est réduit à 5,6.

Encadré 3. Deux exemples d'études originales à partir d'Ines

Nous présentons ici deux études ponctuelles et originales faites à partir d'Ines, qui répondent à des questions économiques complémentaires aux analyses redistributives statiques habituelles qui ont été développées précédemment.

Ines a été utilisé pour évaluer les taux marginaux effectifs de prélèvement (TMEP) en 2014, qui sont une mesure des incitations au travail (Fourcot et Sicsic, 2017). Pour calculer des TMEP, il faut pouvoir mesurer l'effet d'une hausse des revenus sur les transferts socio-fiscaux, ce qui nécessite le recours à un modèle de microsimulation. On simule à partir d'Ines les prestations sociales et les prélèvements de chaque ménage, d'abord dans une situation contrefactuelle, puis dans une situation fictive dans laquelle les revenus sont augmentés. (Sicsic, 2018) a étendu l'analyse en calculant également les gains – ou incitations – monétaires au retour à l'emploi (à partir des taux de prélèvement du retour à l'emploi), et ce sur longue période en remontant à la fin des années quatre-vingt-dix. L'étude montre que les taux marginaux ont diminué pour les très bas revenus mais augmenté pour les plus hauts revenus et que les taux de retour à l'emploi ont diminué dans tout le premier tiers de la distribution. Le profil des taux marginaux en fonction du niveau de revenu a évolué d'un U vers un tilde. Cette étude s'inscrit dans une lignée d'analyses des TMEP par microsimulation, initiés en France par (Bourguignon, 1998), et réalisés à l'international (Jara et Tumino, 2013).

Ines a également été mobilisé pour évaluer les effets redistributifs de la TVA, à court et moyen terme (André et Biotteau, 2019). Il est bien établi que la TVA est un prélèvement anti redistributif en simulation statique (André et alii, 2016). L'innovation de cette étude est de prendre en compte les effets de second tour qui ne sont en général pas analysés. En effet, une hausse initiale des taux de TVA se traduit l'année même par une augmentation des prix des produits soumis à la TVA, puis dans un second temps, les salaires et les autres revenus s'ajustent partiellement, de même que les barèmes des prestations sociales, du fait de l'indexation - formelle ou non - sur l'inflation des revenus, prestations et loyers. À partir d'Ines les auteurs simulent la TVA et les effets directs et indirects. Ce travail nécessite notamment de simuler l'effet du changement des barèmes liés aux revalorisations. Les auteurs montrent que la TVA reste un prélèvement légèrement inégalitaire à moyen terme, malgré les effets de compensation liés aux revalorisations des revenus.

19. Dans la vue d'ensemble de l'ouvrage jusqu'à 2014, puis dans la fiche « Redistribution monétaire » ensuite.

Ce bilan redistributif permet de chiffrer également la contribution de chaque transfert à la réduction des inégalités²⁰. On montre ainsi que, si les prestations sociales représentent des transferts monétaires deux fois moins importants que les prélèvements, elles contribuent en 2018 pour 63 % à la réduction des inégalités du fait de leur ciblage vers les plus modestes, contre 37 % pour les prélèvements.

L'avantage d'un modèle de microsimulation est triple pour l'analyse de la redistribution, il permet de :

(i) faire des bilans redistributifs sur longue période comme cela a été le cas par le passé avec le modèle Ines : voir (Murat *et alii*, 2001) sur la période 1990-1998, (Amar *et alii*, 2007) sur la période 1996-2006, et (Eidelman *et alii*, 2013) sur la période 1990-2010 ;

(ii) prendre en compte les transferts non présents dans les données de l'ERFS : par exemple les cotisations sociales non contributives dans l'analyse habituelle²¹, l'impôt sur la fortune, voir un champ élargi de la redistribution à l'assurance maladie ou l'éducation dans certaines publications (Amar *et alii*, 2008) ;

(iii) donner des résultats plus récents que les données ERFS reposant sur des années antérieures.

❶ **Quels effets des réformes socio-fiscales mises en œuvre chaque année sur les inégalités ?**

L'application précédente permet d'avoir un panorama centré sur le champ des transferts sociaux-fiscaux pertinents pour l'analyse de la redistribution. On peut aussi souhaiter analyser l'effet de l'ensemble des changements législatifs et réglementaires sur le niveau de vie des ménages et sur divers indicateurs d'inégalité.

Pour cela, chaque année depuis 2014, l'Insee et la Drees publient un bilan redistributif et budgétaire des nouvelles mesures sociales et fiscales entrées en vigueur l'année précédente dans l'ouvrage « France, portrait social ». Le bilan intègre dans la même analyse des mesures mises en œuvre l'année précédant la publication (par exemple sur l'année 2018 pour l'édition 2019 de l'ouvrage ; (Biotteau *et alii*, 2019)) relevant des lois de finances et de financement de la Sécurité sociale, initiales et rectificatives, ainsi que de décrets et des accords nationaux interprofessionnels.

Le principe de cette évaluation est de comparer le revenu disponible obtenu avec la nouvelle législation par rapport à une législation dite « contrefactuelle » si aucune modification législative n'était intervenue dans l'année. La différence entre la situation effective et la situation contrefactuelle correspond uniquement à l'effet des réformes, indépendamment des évolutions conjoncturelles qui ont eu lieu dans le même temps et des effets des réformes décidées précédemment. La comparaison permet d'identifier les ménages dont le niveau de vie augmente ou diminue du fait des réformes et de les décrire en fonction de leur position dans l'échelle des niveaux de vie ou du statut d'activité. Elle permet également de mesurer l'effet de chaque mesure sur les finances publiques.

Les mesures prises en compte sont les modifications directes du mode de calcul des transferts permettant de passer d'un revenu « brut » au revenu disponible. Dans certaines analyses les taxes indirectes sont prises en compte.

20. Voir (André *et alii*, 2015) pour la méthode de décomposition.

21. Contrairement à d'autres publications (Guillaud *et alii*, 2019), seuls les transferts socio-fiscaux non contributifs sont pris en compte dans l'analyse de la redistribution.

❶ Comment mesurer les inégalités et la pauvreté plus rapidement ?

L'Insee publie à partir de l'ERFS le taux de pauvreté et les principaux indicateurs d'inégalités de niveau de vie relatifs à l'année N en septembre de l'année N+2. Pour évaluer plus rapidement l'efficacité des politiques publiques de lutte contre la pauvreté et les inégalités, l'Insee met en œuvre grâce à une méthode fondée sur Ines, appelée *nowcasting*, afin de produire à l'automne N+1 des indicateurs avancés sur l'année N (figure 3).

« L'exercice est plus global que le précédent, il prend non seulement en compte les réformes mises en œuvre mais aussi l'évolution de la conjoncture, les revalorisations habituelles ou encore l'effet des anciennes réformes. »

Ainsi, le taux de pauvreté 2018 a été estimé en octobre 2019 (Cornuet et Sicsic, 2019). L'exercice est plus global que le précédent, il prend non seulement en compte les réformes mises en œuvre mais aussi l'évolution de la conjoncture, les revalorisations habituelles ou encore l'effet des anciennes réformes. Cette méthode est également utilisée par Eurostat pour réaliser des publications avancées des indicateurs d'inégalités ou par des universitaires (Gasior et Rastrigina, 2017).

Pour faire cet exercice, l'ERFS d'une année N est mobilisée une première fois pour simuler l'année N, et une seconde fois pour simuler l'année N+1 (législation et évolution des différents types de revenus entre N et N+1)²². L'évolution des indicateurs entre N et N+1 qui est publiée s'appuie donc sur l'utilisation d'un seul millésime de l'ERFS. Ce faisant, la méthode gagne en robustesse, car elle permet de s'affranchir des aléas d'échantillonnage d'une ERFS à l'autre. C'est la cinquième année que l'Insee publie cet exercice d'estimation avancée et les résultats ont toujours été très proches des résultats définitifs publiés un an après (Fontaine et Sicsic, 2015).

❷ Quel est l'effet d'une hausse ou baisse des transferts socio-fiscaux ?

Comme indiqué par (Legendre, 2019), un des *outputs* importants des modèles de microsimulation est de pouvoir simuler des réformes socio-fiscales fictives.

Depuis 2018, l'Insee met à disposition, pour la première fois en France, un cahier de variantes de réformes de prestations et prélèvements élaboré à partir du modèle Ines. Il simule l'effet d'une hausse et d'une baisse fictives de 1 %, 3 % et 5 % des principaux transferts sur les niveaux de vie des ménages par décile, les inégalités, les finances publiques, les bénéficiaires des transferts (voir Fontaine et Sicsic, 2018) pour la méthodologie). Cela montre par exemple qu'une hausse de 5 % du montant forfaitaire du RSA diminuerait le taux de pauvreté de 0,1 point et l'intensité de la pauvreté de 0,7 point et coûterait 850 millions d'euros aux finances publiques. Ce cahier prend la forme d'une publication et d'un fichier présentant toutes les variantes (Cornuet et Sicsic, 2019).

La nature de l'exercice est différente du bilan annuel des effets des réformes socio-fiscales : les effets sont calculés à partir de mesures fictives, définies de façon théorique, et non sur des mesures réellement entrées en vigueur, mais les résultats des deux exercices sont cohérents.

22. L'évaluation décalée est de bonne qualité car à l'automne N+2, la grande majorité des données de cadrage sont disponibles pour N+1. Voir (Fontaine et Fourcot, 2015) pour plus de détails sur la méthode.

📍 L'aide à la décision et l'évaluation des politiques publiques

Le modèle Ines est régulièrement mobilisé à la demande des administrations et des organismes en charge du suivi et de l'évaluation des politiques publiques (Parlement, Haut Conseil de la famille, de l'enfance et de l'âge, Cour des comptes, comité d'évaluation de la fiscalité du capital, groupe d'experts sur le SMIC, etc.) : la simulation apporte des éclairages *ex ante* sur les effets des réformes

des prélèvements et des prestations, c'est-à-dire avant de décider de la mise en œuvre. Cette utilisation répond à l'un des objectifs principaux du modèle Ines depuis sa création, et continue chaque année de représenter une part importante du travail produit par les équipes de développement, notamment à la Drees.

« En décembre 2018, la Drees a ainsi été amenée à simuler plusieurs pistes de réforme permettant de cibler l'aide annoncée par le Président de la République (100 € mensuels) sur le public visé (les salariés au niveau du Smic). »

Récemment le modèle a été fortement mobilisé pour répondre à l'urgence sociale du mouvement des gilets jaunes.

En décembre 2018, la Drees a ainsi été amenée à simuler plusieurs pistes de réforme permettant de cibler l'aide annoncée par le Président de la République (100 € mensuels) sur le public visé (les salariés au niveau du Smic). Ces travaux ont contribué à l'arbitrage en faveur d'une augmentation du bonus individuel de la prime d'activité²³. Les équipes ont à nouveau été sollicitées au moment de l'évaluation de cette réforme par l'Assemblée Nationale à l'été 2019.

Ines est aussi utilisé depuis près de trois ans pour apporter des éclairages techniques sur les conséquences d'une fusion de certaines prestations et minima sociaux. Ce projet de revenu universel d'activité (RUA) est encore à l'étude au moment de l'écriture de cet article.

Ces travaux nécessitent de simuler très précisément les transferts faisant l'objet de la réforme et participent donc à l'amélioration du modèle. Ils sont généralement publiés sous forme de notes annexées à des rapports publics.

📍 VERS UNE EXTENSION DES SOURCES MOBILISÉES PAR INES ? —

Une grande force du modèle Ines est l'expérience accumulée depuis sa création il y a plus de 20 ans. Le modèle a servi à l'élaboration ou la modification de nombreux transferts socio-fiscaux affectant le niveau de vie des ménages et a pour cela fait l'objet d'améliorations constantes pour simuler le plus précisément ces transferts. Les utilisations du modèle se sont diversifiées au fur et à mesure des années et le modèle est utilisé aujourd'hui, au-delà l'analyse traditionnelle de la redistribution, pour fournir des estimations avancées des indicateurs d'inégalités, évaluer les réformes passées, ou encore apporter des éclairages sur celles qui sont en projet ou qui pourraient avoir lieu. Ces applications sont également l'occasion d'étoffer et d'améliorer le modèle.

Fort de cette expérience, le modèle Ines s'appuie sur une documentation importante, gage de sa fiabilité. Un des atouts d'Ines, qui assure la précision des simulations, est lié aux données utilisées, en particulier les informations très riches de l'enquête Emploi.

23. Adresse à la Nation du Président de la République du 10 décembre 2018, annonçant des mesures de réponse à l'urgence économique et sociale, notamment une hausse de salaire de 100 € au niveau du Smic.

C'est toutefois aussi une de ses limites, car la taille des échantillons d'enquête sur lesquels se fonde le modèle est réduite et ne permet pas de donner des résultats pour des petites sous-populations. Les nouvelles bases de données administratives exhaustives qui sont en train d'être mises à disposition de la Statistique publique²⁴ seraient une source d'amélioration à terme sur ce point. Ces données, notamment celles de la DSN qui sont mensuelles, pourraient également améliorer la connaissance des trajectoires et revenus infra-annuels. Cela pourrait permettre de gagner en précision pour la simulation de certains transferts pour lesquels une connaissance des trajectoires de revenus est essentielle, dans un contexte où le système socio-fiscal évolue pour prendre en compte les ressources de façon plus « contemporaine »²⁵. Une modification de la méthode de vieillissement, peu adaptée lors de crises économiques de grande ampleur comme celle provoquée par le Covid-19, pourrait également être une piste d'évolution du modèle. De quoi prédire encore un long avenir au modèle Ines !

24. Comme les bases Fidéli (Fichiers démographiques sur les logements et les individus) produites par l'Insee, ou la Déclaration Sociale Nominative (DSN).

25. C'est le cas de la réforme du prélèvement à la source de l'impôt sur le revenu, ou de celle à venir de la « contemporanéisation » des ressources pour le calcul des allocations logements, *i.e.* l'assise sur les revenus de l'année n-1 voire du trimestre passé, au lieu de n-2.

BIBLIOGRAPHIE

ALBOUY, Valérie, BOUTON, François, LE MINEZ, Sylvie et PUCCI, Muriel, 2003. Le modèle de microsimulation INES : un outil d'analyse des politiques socio-fiscales. In : *Dossiers solidarité et santé. La microsimulation des politiques de transferts sociaux et fiscaux à la Drees : objectifs, outils et principales études et évaluations*. [en ligne]. Juillet-septembre 2003. N°3, pp. 23-43. [Consulté le 20 avril 2020]. Disponible à l'adresse :

<https://drees.solidarites-sante.gouv.fr/IMG/pdf/dossier200303.pdf>.

AMAR, Élise, MARICAL, François, LAIB, Nadine et MIROUSE, Benoît, 2007. 1996-2006 : 10 ans de réforme du système de redistribution. In : *France Portrait social, édition 2007*. [en ligne]. 1^{er} novembre 2007. Collection Insee Références, pp. 81-97. [Consulté le 20 avril 2020]. Disponible à l'adresse : <https://www.insee.fr/fr/statistiques/fichier/1373172/fporsoc07c.pdf>.

AMAR, Élise, BEFFY, Magali, MARICAL, François et RAYNAUD, Émilie, 2008. Les services publics de santé, éducation et logement contribuent deux fois plus que les transferts monétaires à la réduction des inégalités de niveau de vie. In : *France, portrait social, édition 2008*. [en ligne]. 1^{er} novembre 2008. Collection Insee Références, pp.85-101. [Consulté le 20 avril 2020]. Disponible à l'adresse :

<https://www.insee.fr/fr/statistiques/fichier/1372948/FPORSOC08i.PDF>.

ANDRÉ, Mathias et BIOTTEAU, Anne-Lise, 2019. *Effets de moyen terme d'une hausse de TVA sur le niveau de vie et les inégalités : une approche par microsimulation*. [en ligne]. 11 février 2019. Insee, Documents de travail, n°F1901-G2019/01. [Consulté le 20 avril 2020].

Disponible à l'adresse :

<https://www.insee.fr/fr/statistiques/fichier/version-html/3714024/F1901-G2019-01.pdf>.

ANDRÉ, Mathias, BIOTTEAU, Anne-Lise et DUVAL, Jonathan, 2016. *Module de taxation indirecte du modèle Ines - Hypothèses, principes et aspects pratiques*. [en ligne]. Novembre 2016. Drees, Documents de travail, Série Sources et méthodes, n°60. [Consulté le 20 avril 2020]. Disponible à l'adresse : <https://drees.solidarites-sante.gouv.fr/IMG/pdf/dt60.pdf>.

ANDRÉ, Mathias, CAZENAVE, Marie-Cécile, FONTAINE, Maëlle, FOURCOT, Juliette et SIREYJOL, Antoine, 2015. *Effet des nouvelles mesures sociales et fiscales sur le niveau de vie des ménages : méthodologie de chiffrage avec le modèle de microsimulation Ines*. [en ligne]. 8 décembre 2015. Insee, Documents de travail, n° F1507. [Consulté le 20 avril 2020]. Disponible à l'adresse : <https://www.insee.fr/fr/statistiques/fichier/1303384/F1507.pdf>.

BACH, Laurent, BOZIO, Antoine, FABRE, Brice, GUILLOUZOUIC, Arthur, LEROY, Claire et MALGOUYRES, Clément, 2020. *Des analyses fiscales simplistes ?* [en ligne]. 28 février 2020. Le blog des économistes de l'IPP. [Consulté le 20 avril 2020]. Disponible à l'adresse :

<https://blog.ipp.eu/2020/02/28/des-analyses-fiscales-simplistes/>.

BESSIS, Franck et COTTON, Paul, 2019). Modèles de microsimulation et monopole de l'expertise économique : de nouveaux rapports entre gouvernants et gouvernés ? In : *L'État des économistes. La science économique face à la puissance publique (XX^e-XXI^e siècles)*. [en ligne]. Novembre 2019, Amiens, France. [Consulté le 20 avril 2020]. Disponible à l'adresse :

<https://hal.archives-ouvertes.fr/hal-02553413>.

BIOTTEAU, Anne-Lise, FREDON, Simon, PAQUIER, Félix, SCHMITT, Kevin, SICSIC, Mickaël et VERGIER, Noémie, 2019. Les personnes les plus aisées sont celles qui bénéficient le plus des mesures socio-fiscales mises en œuvre en 2018, principalement du fait des réformes qui concernent les détenteurs de capital. In : *France, portrait social, édition 2019*. [en ligne]. Collection Insee Références, pp. 133-155. [Consulté le 20 avril 2020]. Disponible à l'adresse : https://www.insee.fr/fr/statistiques/fichier/4238451/FPS2019_D3.pdf.

BOURGUIGNON, François, 1998. *Fiscalité et redistribution*. [en ligne]. Rapport du Conseil d'Analyse Économique n°11, Éditions La Documentation française. [Consulté le 20 avril 2020]. Disponible à l'adresse : <http://www.cae-eco.fr/IMG/pdf/011.pdf>.

CORNUET, Flore et SICSIC, Michaël, 2019. *Estimation avancée du taux de pauvreté et des indicateurs d'inégalités. En 2018, les inégalités et le taux de pauvreté augmenteraient*. [en ligne]. Octobre 2019. Collection Insee Analyses, n°49. [Consulté le 20 avril 2020]. Disponible à l'adresse : <https://www.insee.fr/fr/statistiques/fichier/version-html/4231288/ia49.pdf>.

DAVID, Marie-Gabrielle, LHOMMEAU, Bertrand et STARZEC, Christophe, 1999. *Le Modèle de Microsimulation INES*. [en ligne]. Août 1999. Insee, Documents de travail, n°F9902, tome 2. [Consulté le 20 avril 2020]. Disponible à l'adresse : <https://www.epsilon.insee.fr/jspui/bitstream/1/5703/1/f9902.pdf>.

EIDELMAN, Alexis, LANGUMIER, Fabrice et VICARD, Augustin, 2013. Prélèvements et transferts aux ménages : des canaux redistributifs différents en 1990 et 2010. In : *Économie et Statistique*. [en ligne]. 29 août 2013. Insee, n°459, pp. 5-26. [Consulté le 20 avril 2020]. Disponible à l'adresse : <https://www.insee.fr/fr/statistiques/fichier/1377610/ES459A.pdf>.

FONTAINE, Maëlle et FOURCOT, Juliette, 2015. *Nowcasting du taux de pauvreté par la micro-simulation*. [en ligne]. Décembre 2015. Insee, Documents de travail, n° F1506. [Consulté le 20 avril 2020]. Disponible à l'adresse : <https://www.insee.fr/fr/statistiques/fichier/1304142/F1506.pdf>.

FONTAINE, Maëlle et SICSIC, Michaël, 2015. *Des indicateurs précoces de pauvreté et d'inégalités - Résultats expérimentaux pour 2014*. [en ligne]. 23 décembre 2015. Collection Insee Analyses, n°23. [Consulté le 20 avril 2020]. Disponible à l'adresse : <https://www.insee.fr/fr/statistiques/fichier/version-html/1304063/ia23.pdf>.

FONTAINE, Maëlle et SICSIC, Michaël, 2018. *L'effet d'une variation du montant de certains transferts du système socio-fiscal sur le niveau de vie : résultats sur 2016 à partir du modèle de microsimulation Ines. (Cahier de variantes)*. [en ligne]. 30 août 2018. Insee, Documents de travail, n° F1806. [Consulté le 20 avril 2020]. Disponible à l'adresse : <https://www.insee.fr/fr/statistiques/fichier/version-html/3604001/F1806.pdf>.

FOURCOT, Juliette et SICSIC, Michaël, 2017. Les taux marginaux effectifs de prélèvement pour les personnes en emploi en France en 2014. [en ligne]. Février 2017. Insee, Documents de travail, n° F1701. [Consulté le 20 avril 2020]. Disponible à l'adresse : <https://www.insee.fr/fr/statistiques/fichier/2581807/F1701.pdf>.

FUGAZZA, Marco, LE MINEZ, Sylvie et PUCCI, Muriel, 2003. L'influence de la PPE sur l'activité des femmes en France : une estimation à partir du modèle Ines. In : *Économie et prévision*. [en ligne]. n°160-161, pp. 79-102. [Consulté le 20 avril 2020]. Disponible à l'adresse : https://www.persee.fr/doc/ecop_0249-4744_2003_num_160_4_6922.

GASIOR, Katrin et RASTRIGINA, Olga, 2017. *Nowcasting : timely indicators for monitoring risk of poverty in 2014-2016*. [en ligne]. 10 mai 2017. EUROMOD Working Paper Series EM7/17. [Consulté le 20 avril 2020]. Disponible à l'adresse : <https://www.euromod.ac.uk/publications/nowcasting-timely-indicators-monitoring-risk-poverty-2014-2016>.

GUILLAUD, Elvire, OLCKERS, Matthew et ZEMMOUR, Michaël, 2019. Four levers of redistribution: The impact of tax and transfer systems on inequality reduction. [en ligne]. 13 janvier 2019. The Review of Income and Wealth. DOI : 10.1111/roiw.12408. [Consulté le 20 avril 2020]. Disponible à l'adresse : <http://groupelavigne.free.fr/guillaud2019.pdf>.

INSEE, 1998. Les modèles de microsimulation. In : *Économie et Statistique / Economics and Statistics*. [en ligne]. Septembre 1998. N°315, pp. 29-116. [Consulté le 20 avril 2020]. Disponible à l'adresse : https://www.persee.fr/issue/estat_0336-1454_1998_num_315_1.

INSEE, 2019. Fiche 4.4 Redistribution monétaire. In : *France, portrait social, édition 2019*. [en ligne]. 19 novembre 2019. Collection Insee Références, pp. 200-201. [Consulté le 20 avril 2020]. Disponible à l'adresse : <https://www.insee.fr/fr/statistiques/fichier/4238781/FPORSOC19.pdf>.

JARA, H. Xavier et TUMINO, Alberto, 2013. Tax-benefit systems, income distribution and work incentives in the European Union. In : *International journal of microsimulation*. [en ligne]. N° 6(1), pp. 27-62. [Consulté le 20 avril 2020]. Disponible à l'adresse : https://www.microsimulation.org/IJM/V6_1/3_IJM_6_1_Jara_Tumino.pdf.

LEHMANN, Étienne, MARICAL, François et RIOUX, Laurence, 2013. Labor Income Responds Differently to Income-Tax and Payroll-Tax Reforms. In : *Journal of Public Economics*. [en ligne]. 16 octobre 2012. Vol. 99, pp. 66-84. [Consulté le 20 avril 2020]. Disponible à l'adresse : <https://halshs.archives-ouvertes.fr/halshs-00856246/document>.

LEGENDRE, François, 2019. L'émergence et la consolidation des méthodes de microsimulation en France. In : *Économie et Statistique / Economics and Statistics*. [en ligne]. 18 décembre 2019. N°510-511-512, pp. 207-220. [Consulté le 20 avril 2020]. Disponible à l'adresse : https://www.insee.fr/fr/statistiques/fichier/4253159/ES_510_511_512_FR.pdf.

MADEC, Pierre et TIMBEAU, Xavier, 2017. *Statistique publique : une révolution silencieuse*. [en ligne]. Le blog de l'OFCE. [Consulté le 20 avril 2020]. Disponible à l'adresse : <https://www.ofce.sciences-po.fr/blog/statistique-publique-une-revolution-silencieuse/>.

MURAT, Fabrice, ROTH, Nicole et STARZEC, Christophe, 2001. L'évolution de la redistributivité du système socio-fiscal entre 1990 et 1998 : une analyse à structure constante. In : *France, portrait social, édition 2000-2001*. V. 2000/01. pp. 133-150. ISSN : 1279-3671.

PAQUIER, François, SCHMITT, Kévin, SICSIC, Michaël, 2019. *Simulation des effets redistributifs de la transformation de l'ISF en IFI à l'aide du modèle Ines*. [en ligne]. 18 décembre 2019. Insee, Documents de travail, n°F1908. [Consulté le 20 avril 2020]. Disponible à l'adresse : <https://www.insee.fr/fr/statistiques/fichier/version-html/4267455/F1908.pdf>.

SICSIC, Michaël, 2018. Les incitations au travail en France entre 1998 et 2014. In : *Économie et Statistique / Economics and Statistics*. [en ligne]. 10 janvier 2019. N°503-504, pp. 13-35. [Consulté le 20 avril 2020]. Disponible à l'adresse : https://www.insee.fr/fr/statistiques/fichier/3652229/503-504_Sicsic-FR.pdf.

SICSIC, Michaël, 2019. *Les incitations fiscales au travail et à la R&D et leurs effets sur le marché du travail*. Paris : École doctorale des sciences économiques et gestion, sciences de l'information et de la communication. Thèse de doctorat de sciences économiques.

SIREYJOL, Antoine, 2016. *La CMU-C et l'ACS réduisent les inégalités en soutenant le pouvoir d'achat des plus modestes – Impact redistributif de deux dispositifs d'aide à la couverture complémentaire santé*. [en ligne]. Octobre 2016. Les Dossiers de la Drees, n°7. [Consulté le 20 avril 2020]. Disponible à l'adresse : http://www.epsilon.insee.fr/jspui/bitstream/1/63233/5/dossiers_7.pdf.

THÉLOT, Claude, VILLAC, Michel, 1998. *Politique familiale – Bilan et perspectives. Rapport à la Ministre de l'emploi et de la solidarité et au Ministre de l'économie, des finances et de l'industrie*. 22 juillet 1998. Paris, Éditions La Documentation française. ISBN : 978-2-11-090989-7.

LA NOMENCLATURE SOCIOPROFESSIONNELLE 2020

CONTINUITÉ ET INNOVATION, POUR DES USAGES RENFORCÉS

Thomas Amossé*

À l'occasion de la dernière révision de la nomenclature des professions et catégories socioprofessionnelles (PCS), l'article revient sur ces objets centraux mais méconnus de la statistique publique que sont les nomenclatures. Indispensables aux statisticiens pour nommer et organiser la réalité qu'ils ont à décrire, elles contribuent à stabiliser l'espace cognitif et pratique. Mais, au fil du temps, elles doivent rester en phase avec l'état du monde et faire sens pour les acteurs qui les manipulent.

La rénovation de la PCS illustre les deux enjeux que pose la révision périodique des nomenclatures : la tension entre nécessité d'actualisation et maintien de comparaison dans le temps ; leur appropriation par les usagers, qu'ils soient producteurs ou utilisateurs de statistiques. La solution retenue fait œuvre de souplesse, mêlant continuité et innovation : elle réaffirme les principes de la nomenclature et conserve inchangées les catégories socioprofessionnelles historiques ; mais elle propose en parallèle des catégorisations complémentaires pour l'analyse des positions sociales et une actualisation du niveau détaillé des professions. S'appuyant sur un processus de codage simplifié et une présentation pédagogique par le biais d'un site internet dédié, la nomenclature PCS rénovée constitue un dispositif complet, à même de favoriser une gamme plus large d'analyses, actuelles et historiques, du monde du travail et de la structure sociale.

 *On the release of the latest revision to the French Classification of Professions and Socio-Professional Categories (PCS), this article reconsiders those central but underrated subjects of official statistics: classifications, also sometimes known as nomenclatures. They are essential to statisticians in naming and organising the reality they have to describe and they help stabilize the cognitive and practical space. But over time, they must remain in step with the state of the world and make sense to the parties handling them.*

The updating of the PCS illustrates the two key issues at stake in the periodic revision of classifications: the conflict between the need for an update, set against the wish to maintain a comparison over time; and their appropriation by users, whether they produce or make use of statistics. The chosen solution provides flexibility, combining continuity and innovation: it reaffirms the classification principles and keeps the historic socio-professional categories unchanged; but at the same time, it proposes additional categorisations for analysing social position, and an update to the detailed level of classification of professions. Based on a simplified coding process and educational presentation via a dedicated website, the updated PCS classification constitutes a complete system allowing a wider range of current and historic analyses of the world of work and the social structure.

* Chercheur au Conservatoire national des arts et métiers (Cnam), Laboratoire interdisciplinaire pour la sociologie économique (Lise), Centre d'études de l'emploi et du travail (CEET), thomas.amosse@lecnam.net

Les nomenclatures occupent une place singulière dans la statistique économique et sociale. Sans elles, impossible de rendre compte d'une réalité sensible trop diverse pour être directement observée, mesurée et comprise. Au même titre que les indicateurs (croissance, prix, chômage, etc.), elles définissent les catégories élémentaires du langage statistique : ce sont les cases des tableaux, les barres des graphiques. Les nomenclatures sont donc indispensables, comme le sont les mots du dictionnaire. Elles demeurent pourtant méconnues.

« Les nomenclatures sont donc indispensables, comme le sont les mots du dictionnaire. Elles demeurent pourtant méconnues. »

Entre les formalisations mathématiques qui les accompagnent et les institutions qui les portent, elles sont « *la face obscure du travail scientifique comme du travail politique* » (Desrosières, 2010). De la même manière que la maîtrise pratique d'une langue ne suppose pas la connaissance approfondie

de sa grammaire et de son étymologie, l'efficacité des nomenclatures ne nécessite pas la maîtrise de leurs principes et de leur origine. C'est même en partie l'inverse : le succès des catégories statistiques tient à leur évidence apparente. Il faut que leur nom et la réalité qu'elles désignent paraissent naturels, que l'on puisse oublier, au moins temporairement, la construction dont elles résultent.

LES NOMENCLATURES STATISTIQUES NE SONT PAS DES ÉDIFICES ABSTRAITS

Elles ne sont pas sans lien avec les catégories institutionnelles (État, droit, science) ou ordinaires, qui renvoient à la parfois longue histoire des pratiques et élaborations humaines. Comme l'écrit François Héran (Héran, 1984), « [Le statisticien ou le démographe] *opèrent [...] sur des catégories prédécoupées, sur du prêt-à-compter, ils peuvent donc se permettre d'en oublier la genèse* ». En établissant des conventions d'équivalence, *i.e.* des règles de classement délimitant des espaces où les situations individuelles peuvent être considérées comme équivalentes, les nomenclatures adoptent le temps des institutions et des normes. Ce sont des « *investissements de forme* » (Thévenot, 1986), dont l'horizon temporel se mesure en décennies plutôt qu'en années. Elles permettent de mesurer les évolutions de court ou moyen terme, mais ne peuvent refléter les transformations structurelles, qui imposent une révision des catégories¹.

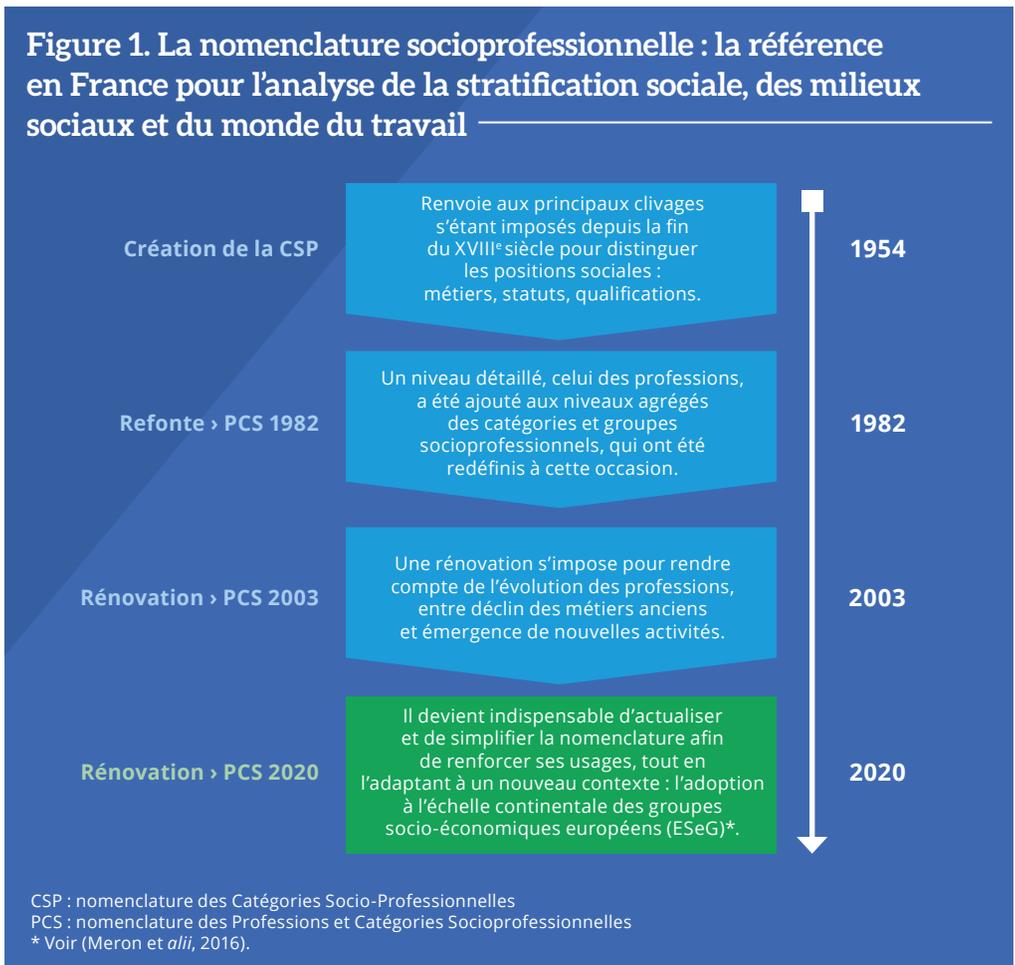
Les professions et catégories socioprofessionnelles (PCS) sont emblématiques des réflexions auxquelles ont donné lieu les nomenclatures statistiques en France (**figure 1**). L'examen de leur histoire, de leur origine à leur dernière rénovation² (Desrosières, 1977 ; Amossé, 2013 ; Amossé, 2019), fournit l'occasion de revenir sur l'objet complexe que représentent plus largement les nomenclatures, scientifiques puis statistiques, par un resserrement progressif de la focale. Dans le cas des PCS, elles posent deux principaux enjeux : leur évolution, tendue entre la nécessité d'analyses actualisées et le souhait de comparaisons temporelles ; leur appropriation par les usagers, qu'ils soient producteurs ou utilisateurs de statistiques.

1. Ces révisions se révèlent nécessaires par la déconnexion des temporalités statistiques et réelles, qui font écho aux trois temporalités (du droit, des bases de données et du monde réel) qu'analyse Isabelle Boydens (Boydens, 1999).
2. La rénovation récente de la PCS a pris la forme d'un groupe de travail du Conseil national de l'information statistique (Cnis), dont les travaux se sont déroulés en 2018 et 2019.

UNE NOMENCLATURE, C'EST NOMMER ET ORGANISER...

Littéralement, une nomenclature désigne l'ensemble des termes employés dans un domaine (une science, une technique, un art, etc.). Le mot dérive des racines latines *nomen* (« nom ») et *calare* (« appeler »), qui ont donné *nomenclator* dans la Rome antique (« esclave indiquant à son maître le nom des personnes qu'il devait saluer ») et *nomenclateur* à l'époque moderne (« personne qui donne le nom aux choses et aux êtres vivants »)³. Établir une nomenclature, c'est ainsi en premier lieu nommer la réalité dans sa diversité, les termes devant renvoyer à l'ensemble des entités jugées pertinentes pour décrire un domaine. Ce n'est qu'ensuite, avec les naturalistes du XVIII^e siècle (Linné, Buffon, etc.) ayant établi les premières nomenclatures scientifiques modernes que s'impose l'idée qu'une méthode est nécessaire pour organiser noms et entités.

Figure 1. La nomenclature socioprofessionnelle : la référence en France pour l'analyse de la stratification sociale, des milieux sociaux et du monde du travail



3. Au XVII^e siècle, dans sa fable *Le cas de conscience*, La Fontaine écrit ainsi : « Dieu, par sa bonté profonde, un beau jour mit dans le monde Apollon son serviteur ; et l'y mit justement comme Adam le nomenclateur, lui disant : Te voilà, nomme. ».

Synonymes de nomenclature, les termes de classification (qui est préféré par les anglophones) et de taxinomie (du grec *taxi* signifiant « ordre ») témoignent davantage de cette dimension : les éléments d'une nomenclature doivent être classés, ordonnés. Il s'agit de regrouper les entités désignées par des noms en adoptant une méthode ou des règles, qui reflètent tout à la fois la manière dont la réalité paraît d'elle-même organisée et la vision que l'on porte sur elle.

L'étymologie révèle quelques-unes des questions que posent les nomenclatures. Ces questions, Alain Desrosières les a analysées dans le chapitre *Classer et coder* de son ouvrage *La politique des grands nombres* (Desrosières, 2010). Il y décrit l'opposition entre les deux taxinomies des êtres vivants proposées par Linné et par Buffon. La première est liée à un ordre divin et érigée en système : toute la nature s'y trouve mise en tableau selon les caractères généraux retenus par Linné, qui sont présentés comme s'imposant de la réalité elle-même et sont systématiquement utilisés pour distinguer les espèces, genres, familles, etc. Dans la seconde, Buffon adopte une méthode où il construit une représentation de la nature en sélectionnant pas à pas les traits pertinents permettant de distinguer localement une espèce d'une autre. En accordant de l'importance à la typicité des familles d'espèce communes, et à leurs noms, la méthode de Buffon produit une nomenclature que l'on peut qualifier de naturelle, typique et nominaliste, alors que le système de Linné est une classification logique, critérielle et réaliste.

Cette opposition, qui renvoie respectivement à des raisonnements par l'exemple et par la généralité, traverse l'histoire des nomenclatures et des classifications. Comme le note Desrosières, « le taxinomiste théoricien est spontanément attiré par la démarche de Linné, et méfiant à l'égard de la méthode de Buffon ». De fait, la science moderne a contribué au développement d'approches systématiques et théoriques dans l'élaboration des

classifications ; par exemple, à la fin du XVIII^e siècle, le *Mémoire sur la nécessité de réformer et de perfectionner la nomenclature de la chimie* de Lavoisier visait à remplacer les anciennes dénominations issues de l'alchimie.

« Pour autant, les avancées scientifiques ne se font pas dans une seule direction, ni sans heurts. Des théories émergent, d'autres déclinent. »

Pour autant, les avancées scientifiques ne se font pas dans

une seule direction, ni sans heurts. Des théories émergent, d'autres déclinent. Dans la science, mais aussi dans l'administration, et plus largement dans l'ensemble des activités économiques, sociales ou humaines, des visions concurrentes se développent au cours de l'histoire, qui fournissent plusieurs manières de penser et d'organiser la réalité. Leur nombre est limité, compte tenu des coûts cognitifs et pratiques qu'accompagnent l'élaboration et l'appropriation de telles représentations du monde. Mais la perspective de disposer dans un domaine d'une classification unique, fondée sur des critères systématiques et liée à une théorie unifiée de la réalité, apparaît souvent illusoire. Fruits de la sédimentation et de l'hybridation de clivages à la pertinence souvent locale, s'appuyant sur des exemples typiques et des noms issus du langage ordinaire, les nomenclatures élaborées au fil du temps s'apparentent de fait bien plus à la méthode de Buffon qu'au système de Linné.

📍 ... REFLÉTER ET INSTITUER, DE DIFFÉRENTES MANIÈRES, LA RÉALITÉ

La pluralité et l'hybridation de règles de dénomination et principes de classement ressortent tout particulièrement de l'histoire des taxinomies statistiques : ces dernières désignent dans cet article les nomenclatures dont la mise en œuvre statistique apparaît centrale, étant acquis que toute nomenclature peut être utilisée à des fins statistiques, c'est-à-dire pour établir un comptage des entités qui y sont classées⁴.

Ainsi, dès son adoption à la fin du XIX^e siècle, la classification internationale des maladies renvoie de façon imbriquée à deux principes concurrents, qui témoignent à la fois des objectifs attendus en matière de connaissance des causes de décès et des contraintes

« À l'heure actuelle, pas moins de cinq nomenclatures d'information médicale visent à répondre à la diversité des besoins d'organiser les connaissances et les pratiques en matière de santé. »

pratiques de codage : selon le premier, topographique, les médecins légistes classent les décès d'après les symptômes observés, et leur localisation corporelle ; le second, étiologique, qui a la préférence des épidémiologistes, suppose d'identifier la maladie cause initiale des décès (Fagot-Largeault, 1990). À l'heure actuelle, pas moins de cinq nomenclatures d'information médicale visent à répondre à la diversité des besoins d'organiser les connaissances et les pratiques en matière de santé : ce

sont des nomenclatures décrivant des diagnostics (CIM⁵), des techniques utilisées par les professionnels de santé (CCAM, CSARR⁶) ou des dispositifs médicaux et des médicaments (LPP, ATC⁷).

Dans le domaine économique, l'existence de différentes nomenclatures statistiques est également fréquente : par exemple, les règles budgétaires prévoient que la comptabilité de l'État peut être présentée selon le service émetteur ou le service bénéficiaire des dépenses, mais aussi selon la nature de la dépense (personnel, fonctionnement, etc.) ou encore le domaine fonctionnel (ou l'activité) auquel elle correspond. Toujours en matière économique, l'histoire des branches de l'industrie a révélé la multiplicité des manières de classer les activités productives depuis le XVIII^e siècle (Guibert, Laganier et Volle, 1971) : au fil du temps, ont ainsi été successivement privilégiées des règles s'appuyant sur les matières premières utilisées, les techniques de production et l'usage des produits⁸, principes de structuration qui se trouvent encore aujourd'hui articulés dans les nomenclatures d'activité et de produit.

-
4. En règle générale, celles-ci sont adossées à des systèmes d'information, comme l'est par exemple la catégorie juridique, enregistrée dans le répertoire Sirene, ou encore la nomenclature des usages de l'eau, qui est liée au système d'information sur l'eau (Sandre).
 5. Classification dans laquelle les principes topographiques et étiologiques demeurent toujours imbriqués aujourd'hui (des catégories relatives aux causes externes de mortalité et aux facteurs influant sur l'état de santé ayant été ajoutées).
 6. CCAM : classification commune des actes médicaux ; CSARR : Catalogue spécifique des actes de rééducation et réadaptation.
 7. LPP : liste des produits et prestations remboursables par l'Assurance Maladie ; ATC : système de classification anatomique, thérapeutique et chimique (en anglais : *Anatomical Therapeutic Chemical Classification System*).
 8. Un rétroviseur peut par exemple être classé en fonction du principal matériau dont il est composé (le secteur de la fabrication du verre) ou de son usage (le secteur de l'équipement automobile).

La diversité des nomenclatures statistiques, ou des principes sur lesquels elles s'appuient, s'explique par le lien complexe qu'elles entretiennent avec la réalité qu'elles représentent. Elles ne peuvent être considérées comme des instruments neutres, de simples miroirs d'une réalité existant en dehors de la mise en forme qu'elles en fournissent. Elles la reflètent mais aussi l'instituent : en permettant de passer du singulier au général, elles servent de référence pour comprendre le monde qui nous entoure, de points d'appui pour formuler des jugements, de guides pour prendre des décisions. Ainsi, les nomenclatures contribuent à stabiliser, pour un temps du moins, l'espace cognitif et pratique. Et elles peuvent le faire de multiples manières. Résultant d'une négociation entre acteurs⁹, ce sont des constructions politiques qui peuvent suivre plusieurs principes et prendre des formalisations diverses : elles peuvent être autonomes ou articulées à d'autres, continues ou discrètes, plates ou emboîtées, hiérarchisées ou non (Boeda, 2008).

« Ces nomenclatures montrent qu'il n'existe pas une unique manière de rendre compte de la structure des emplois et des positions sociales. »

Ces caractéristiques ont tout particulièrement été mises en évidence par l'analyse comparée des nomenclatures socioprofessionnelles dans différents pays et à différentes périodes. Des échelles de statut socio-économique ou de prestige discutées outre Atlantique aux catégories pluridimensionnelles non ordonnées adoptées en France en passant par la classification hiérarchique britannique, ces nomenclatures montrent qu'il n'existe pas une unique manière de rendre compte de la structure des emplois et des positions sociales.

En France, la récente rénovation des PCS témoigne en outre de deux enjeux centraux qui s'y sont posés et que la suite de l'article va illustrer : la tension entre stabilité et actualisation dans l'évolution des nomenclatures ; leur appropriation pour une diversité d'usages et par une diversité d'utilisateurs.

🕒 L'EMPREINTE DU TEMPS : POURQUOI ET COMMENT RÉNOVER LES PCS ?

L'histoire récente de la nomenclature socioprofessionnelle française témoigne de « la situation singulière d'un outil de représentation du monde censé demeurer inchangé quand rien ne l'est, ni l'institution qui en assure la gestion, ni les acteurs qui en font usage, ni les catégories sociales, ordinaires ou juridiques, auxquelles il renvoie, ni, enfin, les théories sociologiques qui lui donnent une assise conceptuelle, qui permettent son interprétation ou équipent sa critique. » (Amossé, 2013). C'est d'ailleurs pour cette raison que chacune de ses refontes puis rénovations successives a été précédée d'une enquête réalisée auprès de ses utilisateurs¹⁰. Leurs conclusions ont été similaires : le monde change, mais les catégories n'en perdent pas pour autant toute pertinence. Ainsi, la mission préalable d'état des lieux conduite à l'occasion de la dernière rénovation indique que la nomenclature continue de constituer « un « langage commun » connu et reconnu dans des univers professionnels variés et un outil répondant à une large diversité d'objectifs et d'attentes des utilisateurs. » (Penissat, Perdoncin et Bodier, 2018).

9. Comme l'indique Desrosières, « il faut se réunir pour convenir ce qui équivaut, car l'équivalence n'est jamais donnée d'avance » (Desrosières, 2014).

10. Ces enquêtes ont été conduites par Alain Desrosières en 1975, Hedda Faucheux et Guy Neyret en 1998 et Étienne Penissat, Anton Perdoncin et Marceline Bodier en 2018.

Certes le modèle de représentation des hiérarchies professionnelles des entreprises s'est quelque peu transformé avec le développement des grilles à critères classant des conventions collectives dans les années quatre-vingt-dix. De même, les nouvelles règles de gestion de l'emploi public, et les reclassements catégoriels qui les ont accompagnées (professeures ou professeurs des écoles, officières ou officiers de police, infirmières ou infirmiers, etc.), ont modifié la structure des emplois dans la fonction publique. Pour autant, ces évolutions ont été graduelles et partielles : les branches traditionnelles où les petites entreprises sont plus nombreuses, comme le bâtiment, se sont par exemple montrées attachées aux catégories historiques des conventions collectives, dites Parodi¹¹, qui délimitent les groupes de « cadre », d'« ouvrier », etc.¹² Les consultations opérées en amont des deux dernières rénovations ont permis de mesurer l'ampleur, réelle mais limitée, des changements ayant affecté les catégories instituées de classement des emplois servant de fondement à la nomenclature socioprofessionnelle. La perspective de remanier la nomenclature en profondeur, dans sa structure ou ses principes, a ainsi été repoussée.

À côté du souhait, souligné par nombre d'utilisateurs, de maintenir inchangées les catégories historiques pour permettre des comparaisons temporelles, des demandes d'évolution ont toutefois été exprimées. S'est ainsi fait jour la tension classique entre le maintien

« S'est ainsi faite jour la tension classique entre le maintien des catégories à des fins de séries longues et leur nécessaire actualisation pour rendre compte des évolutions socio-économiques. »

des catégories à des fins de séries longues et leur nécessaire actualisation pour rendre compte des évolutions socio-économiques. Pour justifier d'une modification de la nomenclature ont notamment été évoquées des transformations concernant le niveau détaillé des professions : la diffusion des technologies numériques ; le

développement des activités liées à la transition écologique ; la convergence de métiers exercés par les agents publics et les salariés du privé. Des évolutions plus larges ont également été mentionnées, qui affectent par exemple la recomposition des groupes « ouvrier » et « employé » (et la nécessité de rendre compte de leur segmentation en fonction du niveau de qualification), la hiérarchisation du groupe « cadre » (en distinguant les cadres dirigeants et en précisant le positionnement des enseignants), une nouvelle délimitation et différenciation des travailleurs indépendants (où seraient isolées les « zones grises » de l'emploi).

Pour répondre à ces attentes partiellement contradictoires, la rénovation de 2020 a retenu une solution originale. D'une part, il a été décidé d'actualiser les rubriques de profession, afin de fournir une grille de lecture actualisée du monde du travail au niveau détaillé de la nomenclature, tout en maintenant inchangées¹³ des catégories socioprofessionnelles historiques au niveau agrégé pour préserver leur comparabilité temporelle. D'autre part, une catégorisation sociale complémentaire à la nomenclature historique a été élaborée – **le schéma de classes d'emploi** –, qui fournit une grille d'analyse des positions sociales rendant compte des clivages émergents de la structure des emplois. Comme nous allons le voir, ces évolutions respectent l'histoire et les principes de la nomenclature tout en assurant une certaine souplesse dans ses usages possibles.

11. Elles portent le nom d'Alexandre Parodi qui, étant alors ministre du travail, leur donna naissance en 1945.

12. Modifier les édifices normatifs que sont les grilles de classification suppose de renégocier les positions relatives des salariés les uns par rapport aux autres dans la hiérarchie des salaires, ce qui peut avoir un coût important en termes d'équilibre et d'organisation des entreprises.

13. En dehors de la reformulation de certains intitulés pour rendre compte de l'évolution de leur composition et de la perspective de leur version en écriture inclusive.

LES PROFESSIONS : PRINCIPES RÉAFFIRMÉS, RUBRIQUES ACTUALISÉES

Dans le prolongement du principe d'enregistrement statistique des compromis sociaux de classement qu'a retenu la nomenclature depuis 1982 (Desrosières et Thévenot, 2002), la notion de profession au sens des PCS correspond à la diversité des manières dont les activités de travail sont organisées, instituées, qualifiées et décrites dans leurs univers professionnels. Elle rend ainsi non seulement compte du contenu du travail, mais aussi de son environnement économique ou institutionnel.

Cette notion synthétique renvoie à la sociologie pragmatique et à l'économie des conventions, qui ont identifié une pluralité de manières de déclarer sa profession et d'organiser les activités de travail (Boltanski et Thévenot, 1991 ; Kramarz, 1991). Mais elle fait également écho à la sociologie nord-américaine des professions et aux travaux plus récents portant sur les micro-classes (Grusky et Sørensen, 1998).

Ainsi précisée dans ses résonances théoriques, la notion de profession dans la PCS apparaît toujours adaptée à la situation française : les différences institutionnelles (par exemple selon le statut indépendant / salarié ou la nature publique / privée de l'employeur)

« En pratique, les rubriques de profession de la nomenclature socioprofessionnelle correspondent à la fois à une activité et à un contexte professionnel. »

continuent spécifiquement d'influencer les situations objectives de travail, comme les représentations ordinaires (Hugrée et De Verdalle, 2015 ; Hugrée, Penissat et Spire, 2015). La profession de la PCS constitue une approche complémentaire, du moins au plan des principes, à celle suivie par la nomenclature de professions de l'Organisation internationale du travail¹⁴, qui renvoie au contenu de l'activité (tâches effectuées, fonctions exercées ou

compétences mises en œuvre). En pratique, les rubriques de profession de la nomenclature socioprofessionnelle correspondent à la fois à une activité et à un contexte professionnel, que permettent d'identifier un ensemble de libellés spécifiques en plus d'une combinaison de variables définissant l'environnement de travail (le statut, le nombre de salariés, la position professionnelle et la nature de l'employeur).

C'est en réaffirmant ces principes et en s'appuyant sur un examen empirique des effectifs et libellés des professions existantes que le niveau fin de la nomenclature (appelés les P 2020) a été actualisé afin de rendre compte de la dynamique des métiers et environnements de travail : il comporte désormais un nombre réduit de rubriques (à priori 310, contre 486 précédemment).

De tailles et constructions plus homogènes, les rubriques rénovées garantissent une représentation plus équilibrée des professions majoritairement occupées par les femmes et par les hommes¹⁵ et une plus grande comparabilité des professions du secteur privé et

14. La CIPD ou classification internationale type des professions (ISCO en anglais).

15. Cette homogénéisation des effectifs de profession selon leur genre correspond au double mouvement de regroupement des professions en déclin numérique et à la scission de professions en expansion, qui sont respectivement plus souvent masculines et féminines (Amossé, 2004).

de la fonction publique où la réalité des tâches, et non seulement la position administrative, est désormais prise en compte. L'explicitation d'un niveau intermédiaire (126 professions regroupées, définies par les trois premières positions des P 2020) vise par ailleurs à rendre plus lisible l'organisation des professions et leur articulation avec les catégories agrégées de la nomenclature.

La version actualisée et réorganisée du niveau détaillé de la nomenclature s'accompagne de plus d'un nouveau niveau d'analyse, au plus près des situations de travail. Ce niveau correspond aux libellés de profession déclarés dans les enquêtes : standardisés parce qu'issus d'une liste de plusieurs milliers d'items utilisés dans le processus de codage rénové (cf. *infra*), ils permettront d'étudier des domaines professionnels transversaux aux rubriques de professions de la nomenclature, sous réserve d'un accès garantissant le respect du secret statistique et la sécurisation des traitements. Dans le cadre de la rénovation, quatre regroupements *ad hoc* de libellés ont ainsi été constitués : les enseignants, les professions du numérique, les métiers verts et les cadres dirigeants, professionnels et experts de haut niveau (Amossé, 2019).

UN SCHEMA DE CLASSE D'EMPLOI COMPLÉMENTAIRE AUX CATÉGORIES HISTORIQUES

Au niveau agrégé de la nomenclature aussi, la rénovation s'est attachée à répondre à la tension entre continuité et rupture. Les catégories et groupes socioprofessionnels ne proposant qu'une manière, certes pertinente mais parfois datée, de représenter la société, un schéma de classes et sous-classes d'emploi a été élaboré.

Conçu pour se distinguer le plus clairement possible de la nomenclature historique¹⁶, ce schéma répond à la nécessité de mieux représenter des clivages à l'importance croissante dans la structuration des emplois et la détermination des positions sociales : l'opposition

entre indépendants et salariés, entre salariés très ou peu qualifiés¹⁷, précaires ou stables, du public ou du privé, dans un métier technique ou du tertiaire. Il entend ainsi permettre d'analyser la transformation de la société salariale à l'œuvre depuis plusieurs décennies (**encadré 1**).

« Au regard de leur qualification, peu de professions méritent d'être déplacées dans les classes d'emploi par rapport aux groupes de la nomenclature historique. »

Les analyses réalisées montrent que, au regard de leur qualification, peu de professions méritent d'être déplacées dans

les classes d'emploi par rapport aux groupes de la nomenclature historique (« cadre et profession intellectuelle supérieure », « profession intermédiaire », « employée ou employé » et « ouvrière ou ouvrier »). En nombre limité, ces professions se distinguent toutefois clairement de leurs groupes historiques, à l'image des professeuses ou professeurs des écoles qui apparaissent nettement plus qualifiés que l'ensemble des professions intermédiaires (dans le schéma de classes, elles font de ce fait partie des emplois salariés de

16. À la différence d'elle, il suit une construction logique et critique et ne renvoie pas à des catégories existantes dans le monde du travail.

17. Quatre niveaux de qualification sont distingués pour les professions exercées en tant que salarié, selon une notion composite mêlant diplôme requis, position occupée et niveau de rémunération observés en moyenne dans la profession.

niveau supérieur). Ces analyses indiquent ainsi à la fois que les groupes socioprofessionnels sont loin d'avoir perdu toute pertinence pour rendre compte de la stratification sociale mais aussi que les classes d'emploi sont plus adaptées à des travaux suivant cette perspective.

D'autres exploitations statistiques, conduites pour valider le schéma de classes d'emploi, confirment la pertinence des catégories et groupes socioprofessionnels, qui demeurent plus fortement associés aux origines sociales, positions sociales du conjoint et situations géographiques que les classes d'emploi (et que la nomenclature socio-

« Le schéma de classes d'emploi rend quant à lui mieux compte de la diversité des situations en matière d'âge, de vie en couple, de logement et de revenu. »

économique européenne, ESeG). Cette situation s'explique notamment par les catégories symboliques, historiquement et sociologiquement typées, en haut comme en bas du spectre social, que sont les agriculteurs et les ouvriers, les professions libérales et les chefs d'entreprise. Le schéma de classes d'emploi rend quant à lui mieux compte de la diversité des

situations en matière d'âge, de vie en couple, de logement et de revenu. Il apparaît ainsi bien complémentaire à la nomenclature historique, en étant d'effectifs plus équilibrés sur l'ensemble du spectre social (en raison de la gradation plus régulière selon le niveau de qualification) et en donnant à voir des pans qui y étaient jusqu'à présent pas ou peu représentés, comme le sont par exemple les fractions publiques et précaires des emplois de chaque niveau de qualification.

DES BOUQUETS DE CATÉGORISATIONS, OU LA RUPTURE DANS LA CONTINUITÉ

Avec le schéma de classes d'emploi, la rénovation propose une solution originale pour permettre à la fois des comparaisons temporelles et une actualisation des grilles d'analyse de la société. C'est ainsi un bouquet de catégorisations sociales liées à la nomenclature socioprofessionnelle qui pourra répondre aux différentes problématiques d'analyse des utilisateurs : les groupes et catégories historiques, dont la pertinence sociologique a été confirmée et qui continueront d'être particulièrement utiles pour les analyses historiques et géographiques ; l'ensemble des groupes et sous-groupes de l'ESeG, dont la vocation première est de nourrir des comparaisons entre les différentes situations nationales à l'échelle européenne ; et enfin les classes et sous-classes d'emploi, qui pourront être utilisées pour examiner les formes et conséquences de la segmentation des emplois. À ce bouquet, on peut en outre ajouter la **PCS Ménage**, autre innovation de la rénovation, qui fournit une alternative à l'utilisation de la personne de référence pour étudier la position sociale des ménages (**encadré 2**).

Mais, les utilisateurs n'ont pas seulement besoin d'un bouquet de catégorisations sociales. De fait, tout en garantissant leur statut de référentiel, la rénovation de la nomenclature a repensé les différents niveaux détaillés de la nomenclature rénovée : les libellés standardisés et leurs regroupements *ad hoc*, les professions et les professions regroupées composent une palette élargie d'outils pour analyser la structure des emplois et situations de travail, que complètent les principales nomenclatures étrangères (ISCO) et françaises (les Fap de

Encadré 1. Le schéma des classes et sous-classes d'emploi

I. Emplois indépendants

- I1. Emplois indépendants de niveau supérieur
- I2. Emplois indépendants de niveau intermédiaire
- I3. Emplois de petits indépendants, avec salarié ou aide familiale
- I4. Emplois de petits indépendants, sans salarié ou aide familiale

A. Emplois salariés de niveau supérieur

- A1. Emplois salariés de niveau supérieur d'orientation technique, en CDI
- A2. Emplois salariés de niveau supérieur d'orientation tertiaire, en CDI
- A3. Emplois salariés de niveau supérieur, fonctionnaires*
- A4. Emplois salariés de niveau supérieur, en contrat à durée limitée

B. Emplois salariés de niveau intermédiaire

- B1. Emplois salariés de niveau intermédiaire d'orientation technique, en CDI
- B2. Emplois salariés de niveau intermédiaire d'orientation tertiaire, en CDI
- B3. Emplois salariés de niveau intermédiaire, fonctionnaires*
- B4. Emplois salariés de niveau intermédiaire, en contrat à durée limitée

C. Emplois salariés qualifiés**

- C1. Emplois salariés qualifiés d'orientation ouvrière, en CDI
- C2. Emplois salariés qualifiés d'orientation employée, en CDI
- C3. Emplois salariés qualifiés, fonctionnaires*
- C4. Emplois salariés qualifiés, en contrat à durée limitée

D. Emplois salariés peu qualifiés**

- D1. Emplois salariés peu qualifiés d'orientation ouvrière, en CDI
- D2. Emplois salariés peu qualifiés d'orientation employée, en CDI
- D3. Emplois salariés peu qualifiés, fonctionnaires*
- D4. Emplois salariés peu qualifiés, en contrat à durée limitée (ou auprès de particuliers)

* Le terme fonctionnaire renvoie aux fonctionnaires et contractuels en CDI de la fonction publique.

** Afin d'éviter toute ambiguïté avec les emplois de niveau intermédiaire et supérieur, la mention « d'exécution » peut être ajoutée dans ces deux intitulés : « emplois salariés d'exécution qualifiés » et « emplois salariés d'exécution peu qualifiés ».

Nota : la classe I a été conçue de façon symétrique aux quatre classes suivantes (A, B, C, D), qui ne comportent que des salariés. Aussi, pour mener des analyses prenant comme premier principe de classement la position sur l'échelle sociale, il est possible de mettre I1 avec la classe A (alors appelée A*), I2 avec la classe B (B*), I3 avec la classe C (C*) et I4 avec la classe D (D*).

	Emplois salariés	Emplois indépendants (Classe I)
Classe A*	Classe A	I1
Classe B*	Classe B	I2
Classe C*	Classe C	I3
Classe D*	Classe D	I4

la Dares¹⁸ et le Rome¹⁹ de Pôle Emploi). C'est ainsi un second bouquet de catégorisations qui résulte de la rénovation, chacune ayant un usage préférentiel :

- ❶ les libellés de professions collectés sur liste pour des investigations situées au plus près des activités de travail ;
- ❶ le niveau des professions et professions regroupées pour les analyses structurelles du monde du travail ;
- ❶ les Fap et le Rome pour l'étude de la dynamique du marché de l'emploi ;
- ❶ l'ISCO pour des comparaisons internationales.

Se dégage ainsi une nouvelle architecture d'ensemble (**figure 2**) : toujours structurée autour de son épine dorsale, qui articule de façon emboîtée les professions, catégories et groupes socioprofessionnels, la nomenclature se trouve complétée à la fois aux niveaux agrégés et détaillés²⁰. L'articulation du professionnel et du social que prévoit de façon rigide la nomenclature depuis 1982 se trouve ainsi assouplie, et de nouveaux niveaux d'analyse sont proposés : plus détaillés avec les libellés collectés, intermédiaires avec les professions regroupées (qui peuvent s'apparenter à des micro-classes) et agrégés avec le schéma de classe d'emplois.

« La nomenclature se trouve complétée à la fois aux niveaux agrégés et détaillés. »

En parallèle de ces évolutions, d'autres innovations, techniques cette fois, ont été adoptées, qui visent à encourager l'utilisation de la nomenclature.

❶ RENFORCER LES USAGES PAR UN PROCESSUS SIMPLIFIÉ DE CODAGE...

Si l'histoire de la nomenclature socioprofessionnelle a porté une attention particulière à sa dimension pratique (Desrosières et Thévenot, 2002 ; Desrosières, 2010), ces travaux se sont principalement intéressés aux pratiques de codage et aux contraintes qu'elles imposent pour les producteurs en termes de lisibilité et simplicité des catégories statistiques. La question plus large des conditions d'appropriation de la nomenclature par ses utilisateurs a bien moins été investie, en dehors des enquêtes ponctuelles réalisées avant la refonte de 1982 et les rénovations de 2003 et 2020.

C'est pour cette raison, et en écho d'ailleurs aux souhaits exprimés lors de l'état des lieux, que le groupe en charge de la dernière rénovation a mis l'utilisateur au centre de ses réflexions et intégré de façon centrale l'objectif de renforcer les usages de la nomenclature. Il s'est attaché à y répondre non seulement en simplifiant sa production à l'aide d'un processus allégé de codage mais aussi en facilitant son appropriation grâce à un site internet dédié.

18. La direction de l'animation de la recherche, des études et des statistiques, (Dares) a élaboré la nomenclature des Familles professionnelles (Fap).

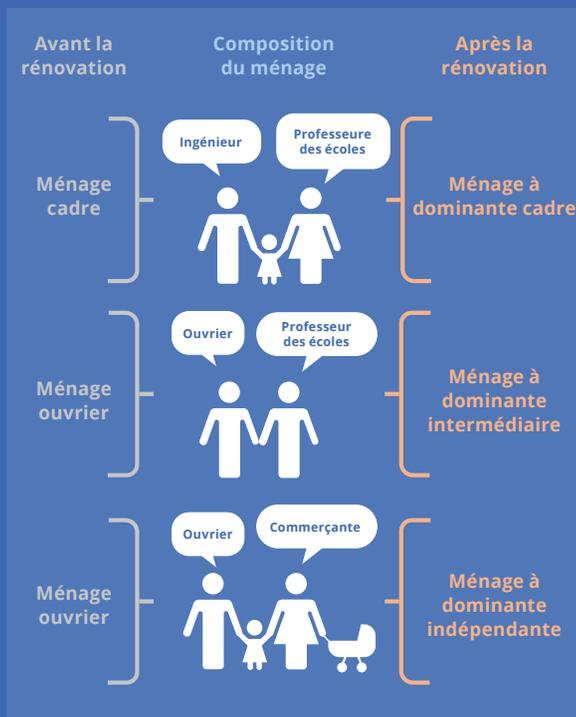
19. Rome : Répertoire Opérationnel des Métiers et des Emplois utilisé par le Pôle Emploi pour coder les emplois recherchés par les demandeurs ainsi que les offres déposées par les entreprises.

20. La dernière révision de la classification internationale des maladies a suivi la même logique, en considérant que la CIM-10 serait le noyau d'une « famille » de classifications permettant de couvrir d'autres besoins d'analyse.

Encadré 2. La PCS Ménage, un nouveau concept plus adapté aux évolutions récentes

Comment rendre compte de la position socioprofessionnelle d'un ménage à partir de celle de ses membres, occupants d'un logement, membres d'un foyer budgétaire ou d'une cellule familiale ? Dans un contexte où l'emploi des femmes continue de se développer et où les situations de bi-activité au sein des couples sont devenues la norme (au moins au sens statistique), on touche aux limites de l'utilisation de la personne de référence pour résumer la position sociale d'un ménage. Car donner la préférence à seulement l'un des individus du ménage ne rend pas compte des différences de position sociale liées aux configurations familiales.

En outre, la représentation est biaisée selon que la personne de référence est définie en fonction de son sexe ou de son âge (vision androcentrée) ou de sa contribution aux ressources du ménage (vision économiste, alors que la nomenclature s'appuie sur d'autres dimensions : sociales, culturelles, etc.) Une nouvelle nomenclature devenait nécessaire.



La nouvelle PCS Ménage comporte sept groupes et 16 sous-groupes, qui sont codés à partir des groupes socioprofessionnels des deux principaux adultes du ménage (la personne de référence et son éventuel conjoint, ou éventuelle conjointe).

Bien que non stratifiée par construction, la PCS Ménage fournit de fait une grille de lecture largement hiérarchisée de la société.

La distinction des ménages homo / hétérogames permet de préciser de façon transversale les positions sociales : c'est particulièrement le cas en haut de la structure sociale, avec les couples de cadres et chefs d'entreprise.

En bas, c'est la distinction des ménages composés d'un seul adulte (de niveau employé ou

ouvrier) qui délimite la fraction disposant des plus faibles niveaux de ressource avec les ménages uniquement composés de personnes inactives (non retraitées).

Au milieu, les ménages à dominante indépendante occupent une place pivot (en raison notamment de leur patrimoine).

La méthode suivie jusqu'à présent pour coder la nomenclature a initialement été établie pour des enquêtes passées sous format papier. Elle reposait ainsi sur des libellés de professions spontanément déclarés ou saisis en clair et sur un grand nombre d'informations dites « annexes »²¹ utilisées pour compléter les libellés, souvent lacunaires, et permettre *in fine* l'obtention d'un code (de façon automatique, à l'aide de l'environnement Sicore-PCS²², ou en reprise manuelle). Cette méthode avait ses limites : coût de collecte lié au nombre des variables annexes, coût de reprise dû aux difficultés de codage de libellés imprécis. Un nouveau protocole a donc été élaboré, qui s'appuie sur les possibilités offertes par le développement des outils numériques, et notamment des moteurs de recherche « intelligents » permettant la recherche en auto-complétion²³ d'un libellé dans une liste en comprenant plusieurs milliers.

Grâce à une application de collecte efficace (passation fluide, y compris en auto-administration) et à l'établissement d'une liste complète de libellés de profession enrichis et

« Le volume attendu de libellés non retrouvés dans la liste est faible. »

standardisés de façon à pouvoir être tous codés, le volume attendu de libellés non retrouvés dans la liste (qui continueront d'être codés comme actuellement) est faible²⁴.

Cette évolution constitue une amélioration sensible dans la mesure où l'information collectée est décisive pour la qualité du codage, quelle que soit la procédure de codification envisagée²⁵. De plus, le nombre de variables « annexes » peut être fortement réduit, puisqu'elles se limitent désormais au statut, à la taille de l'entreprise pour les indépendants et à la position professionnelle en fonction de la nature de l'employeur pour les salariés. La réduction du nombre de variables nécessaires pour coder s'accompagne par ailleurs de la simplification et de l'harmonisation de leur formulation, afin de permettre une meilleure comparabilité des codes obtenus dans l'ensemble des enquêtes, selon une logique de passation en « multi-mode »²⁶.

Ce dispositif rénové de collecte s'accompagne d'une simplification du processus de codification, grâce à un programme suivant une écriture matricielle qui permet d'affecter de façon déterministe un code à tout couple (libellé, variable annexe). De ce fait, les améliorations attendues ne concernent pas uniquement les professionnels des services producteurs de la statistique publique, qui verront normalement diminuer les coûts de collecte et s'améliorer la qualité de la codification. Le processus rénové de codage se traduira aussi en pratique par des guides de documentation plus accessibles, notamment un guide alphabétique permettant de visualiser l'ensemble des règles d'affectation pour les libellés de la liste.

21. Le statut, le nombre de salariés de l'entreprise, la position professionnelle, la nature de l'établissement, mais aussi le secteur d'activité et la fonction exercée, ou encore de façon plus exceptionnelle le sexe pour les aides familiaux, l'orientation agricole, la surface agricole utilisée et la région d'exploitation pour les agriculteurs.

22. Sicore est un outil de codification automatique utilisé depuis 25 ans dans la statistique publique française pour coder toutes sortes de libellés dans une nomenclature donnée : PCS, NAF pour les activités économiques, codes communes, etc. (Rivière, 1995). L'environnement Sicore-PCS est la version utilisée par l'Insee pour coder la PCS dans les enquêtes auprès des ménages.

23. L'auto-complétion ou complétion automatique est une fonctionnalité informatique permettant à l'utilisateur de limiter la quantité d'informations qu'il saisit, en se voyant proposer un complément qui pourrait convenir à la chaîne de caractères qu'il a commencé à taper.

24. En première estimation, il est de l'ordre de quelques points de pourcentage dans les enquêtes en face-à-face.

25. En la matière, à l'issue d'une comparaison internationale menée par l'Insee en 2018, aucun algorithme ou outil ne paraissait à cette date plus efficace que l'environnement Sicore.

26. Le développement des enquêtes en « multi-mode » (collecte par internet, par téléphone, en face-à-face) et des outils permettant de concevoir les questionnaires ont par exemple amené l'Insee à simplifier et harmoniser les questionnements (Cotton et Dubois, 2019 ; Koumarios et Sigaud, 2019).

La mise à disposition de l'ensemble de ces outils (application de collecte de libellé, formulation des variables annexes, programme de codification, guides de documentation²⁷) devrait ainsi permettre à une plus large communauté d'utilisateurs (statisticiens d'instituts de sondage, équipes de recherche) de pouvoir obtenir simplement une variable de PCS de qualité dans les enquêtes qu'ils produisent. C'est précisément un des objectifs du futur site internet dédié à la nomenclature qui de favoriser de tels usages.

📍 ... ET L'APPROPRIATION DE LA NOMENCLATURE GRÂCE À UN SITE INTERNET DÉDIÉ

L'état des lieux ainsi que les consultations réalisées dans le cadre de la rénovation ont souligné l'importance de pouvoir disposer d'un espace numérique d'information et d'échange qui permette, en lien avec insee.fr, de répondre à l'ensemble des attentes des utilisateurs.

« L'importance de pouvoir disposer d'un espace numérique d'information et d'échange qui permette, en lien avec insee.fr, de répondre à l'ensemble des attentes des utilisateurs. »

Il est ainsi attendu qu'un tel site fournisse :

- ❶ les éléments permettant de comprendre l'histoire et les grands principes de la nomenclature, et la manière dont elle s'articule avec ses pendants en France comme à l'étranger ;
- ❷ un outil de navigation afin d'entrer plus précisément dans sa logique et de parcourir par le menu les rubriques qui la composent et les modes d'agrégation qu'elle propose ;
- ❸ des statistiques pour analyser et comprendre la société à partir de la nomenclature socioprofessionnelle, d'une sélection de tableaux et graphiques à un moteur de recherche de micro-données en passant par la réalisation de tris en ligne ;
- ❹ les instruments nécessaires pour collecter et coder la nomenclature, ses différents agrégats ainsi que la nomenclature internationale des professions (ISCO) et la catégorisation socio-économique européenne (ESeG).

Ce sont précisément ces objectifs qui ont guidé l'élaboration de l'arborescence prévisionnelle de ce site, qui doit être progressivement mis en place au cours des années 2020 et 2021 : il sera ainsi *a priori* organisé en quatre rubriques principales (Découvrir ; Naviguer ; Décrire ; Coder). Fruit d'un partenariat entre différentes institutions (l'Insee, Progedo²⁸, le laboratoire Printemps de l'Université Versailles Saint-Quentin et le Cnam-CEET²⁹), il a vocation à répondre

27. À la date de rédaction de cet article, l'ensemble des éléments constitutifs de la documentation autour des usages de la PCS 2020 étaient en cours d'élaboration.

28. Progedo (*PROduction et GEstion des DONnées*) est une Très Grande Infrastructure de Recherche chargée d'impulser et structurer une politique publique des données pour la recherche en sciences sociales.

29. Le Centre d'études de l'emploi et du travail (CEET) est un programme transversal du Conservatoire national des arts et métiers (Cnam) visant à développer la recherche pluridisciplinaire sur le travail et l'emploi, dans une perspective académique et de réponse à la demande sociale.

à l'ensemble des attentes des producteurs et utilisateurs de la PCS, qu'il s'agisse du grand public, de journalistes, de représentants d'associations ou de syndicats, d'enseignants, de chercheurs, ou de statisticiens de la sphère publique et privée.

Au final, à l'issue de sa dernière rénovation, la nomenclature socioprofessionnelle, plus lisible, offre plus de souplesse d'usage. Au niveau détaillé, ses principes ont été réaffirmés en même temps que ses rubriques ont été actualisées et complétées par de nouvelles possibilités d'analyse. Il en va de même au niveau agrégé, où de nouvelles grilles d'analyse ont été ajoutées aux catégories historiques. La simplification du processus de codage et le site internet dédié parachèvent l'ensemble, de façon à ce que cette nomenclature emblématique de la statistique publique française voie sa position centrale renforcée.

BIBLIOGRAPHIE

AMOSSÉ, Thomas, 2004. Professions au féminin : représentation statistique, construction sociale. In : *Travail, genre et sociétés*. [en ligne]. Éditions La Découverte, 2004/1, n°11, pp. 31-46. [Consulté le 18 mai 2020]. Disponible à l'adresse : <https://www.cairn.info/revue-travail-genre-et-societes-2004-1-page-31.htm>.

AMOSSÉ, Thomas, 2013. La nomenclature socio-professionnelle : une histoire revisitée. In : *Annales. Histoire, Sciences Sociales*. [en ligne]. Éditions de l'EHESS, 2013/4, 68^e année, pp. 1039-1075. [Consulté le 18 mai 2020]. Disponible à l'adresse : <https://www.cairn.info/revue-Annales-2013-4-page-1039.htm>.

AMOSSÉ, Thomas, 2019. *La rénovation de la nomenclature socioprofessionnelle (2018-2019)*. [en ligne]. Décembre 2019. Rapport de groupe de travail du Cnis n°156. [Consulté le 18 mai 2020]. Disponible à l'adresse : <https://www.cnis.fr/wp-content/uploads/2020/01/Rapport-n%C2%B0-156.pdf>.

BOEDA, Michel, 2008. Les nomenclatures statistiques : pourquoi et comment. In : *Courrier des statistiques*. [en ligne]. Novembre-décembre 2008. N°125, pp. 5-11. [Consulté le 18 mai 2020]. Disponible à l'adresse : <http://www.epsilon.insee.fr/jspui/bitstream/1/122565/1/cs125.pdf>.

BOLTANSKI, Luc et THÉVENOT, Laurent, 1991. *De la justification. Les économies de la grandeur*. 12 avril 1991. Gallimard, Collection NRF Essais. ISBN : 978-2-07-072254-9.

BOYDENS, Isabelle, 1999. *Informatique, normes et temps. Évaluer et améliorer la qualité de l'information : les enseignements d'une approche herméneutique appliquée à la base de données LATG de l'ONSS*. 31 décembre 1999. Éditions Bruylant. ISBN : 978-2-8027-1268-8.

COTTON, Franck et DUBOIS, Thomas, 2019. Pogues, un outil de conception de questionnaires. In : *Courrier des statistiques*. [en ligne]. 19 décembre 2019. N°N2, pp. 17-28. [Consulté le 18 mai 2020]. Disponible à l'adresse : <https://insee.fr/fr/statistiques/fichier/4254216/courstat-3-3.pdf>.

DESROSIÈRES, Alain, 1977. Éléments pour l'histoire des nomenclatures socio-professionnelles. In : *Pour une histoire de la statistique*. Insee, Journées d'études sur l'histoire de la statistique, 23-25 juin 1976, Vauresson, Contributions, tome 1, pp. 155-231.

DESROSIÈRES, Alain, 2010. *La politique des grands nombres. Histoire de la raison statistique*. 19 août 2010. Éditions La Découverte, collection Poche / Sciences humaines et sociales, n°99. ISBN : 978-2-7071-6504-6.

DESROSIÈRES, Alain, 2014. *Prouver et gouverner. Une analyse politique des statistiques publiques*. 3 avril 2014. Éditions La Découverte, collection Sciences humaines. ISBN : 978-2-7071-8249-4.

DESROSIÈRES, Alain et THÉVENOT, Laurent, 2002. *Les catégories socioprofessionnelles*. 24 octobre 2002. Éditions La Découverte, collection Repères, 5^e édition. ISBN : 978-2-7071-3856-9.

FAGOT-LARGEAULT Anne, 1990. *Les causes de la mort. Histoire naturelle et facteur de risque*. Éditions Vrin, Lyon, Collection Science, histoire, philosophie. ISBN : 978-2-7116-9611-1.

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 18 mai 2020]. Disponible à l'adresse : https://www.persee.fr/doc/estat_0336-1454_1971_num_20_1_6122.

- GRUSKY, David B. et SØRENSEN, Jesper B., 1998. Can Class Analysis Be Salvaged ? In : *American Journal of Sociology*. [en ligne]. Mars 1998. N°5, vol.103, pp. 1187-1234.
- HÉRAN, François, 1984. L'assise statistique de la sociologie. In : *Économie et statistique*. [en ligne]. Juillet-Août 1984. N°168, pp. 23-35. [Consulté le 18 mai 2020]. Disponible à l'adresse : https://www.persee.fr/doc/estat_0336-1454_1984_num_168_1_4881.
- HUGRÉE, Cédric et DE VERDALLE, Laure, 2015. Incontournables statuts. « Fonctionnaires » et « indépendants » à l'épreuve des catégorisations ordinaires du monde social. In : *Sociologie du travail*. [en ligne]. Avril-juin 2015. Vol. 57, n°2, pp. 200-229. [Consulté le 18 mai 2020]. Disponible à l'adresse : <https://journals.openedition.org/sdt/1640>.
- HUGRÉE, Cédric, PENISSAT, Étienne et SPIRE, Alexis, 2015. Les différences entre salariés du public et du privé après le tournant managérial des États en Europe. In : *Revue française de sociologie*. [en ligne]. 2015/1, vol.56, pp. 47-73. [Consulté le 18 mai 2020]. Disponible à l'adresse : <https://www.cairn.info/revue-francaise-de-sociologie-2015-1-page-47.htm>.
- MERON, Monique, AMAR, Michel, BABET, Charline, BOUCHET-VALAT, Milan, BUGEJA-BLOCH, Fanny, GLEIZES, François, LEBARON, Frédéric, HUGRÉE, Cédric, PENISSAT, Étienne et SPIRE, Alexis, 2016. *ESeG = European Socio economic Groups. Nomenclature socio-économique européenne*. [en ligne]. Février-mars 2016. Insee, DSDS, Document de travail, n°F1604. [Consulté le 18 mai 2020]. Disponible à l'adresse : <https://www.insee.fr/fr/statistiques/fichier/2022135/F1604.pdf>.
- KRAMARZ, Francis, 1991. Déclarer sa profession. In : *Revue française de sociologie*. [en ligne]. N°32-1, pp. 3-27. [Consulté le 18 mai 2020]. Disponible à l'adresse : https://www.persee.fr/doc/rfsoc_0035-2969_1991_num_32_1_4018.
- KOUMARIANOS, Heïdi et SIGAUD, Éric, 2019. Eno, un générateur d'instruments de collecte. In : *Courrier des statistiques*. [en ligne]. 19 décembre 2019. N°N2, pp. 29-44. [Consulté le 18 mai 2020]. Disponible à l'adresse : <https://insee.fr/fr/statistiques/fichier/4254218/courstat-3-4.pdf>.
- PENISSAT, Étienne, PERDONCIN, Anton et BODIER, Marceline, 2018. *La PCS et ses usages, état des lieux et défis*. [en ligne]. Octobre 2018. Rapport de groupe de travail du Cnis n°151. [Consulté le 18 mai 2020]. Disponible à l'adresse : <https://www.cnis.fr/wp-content/uploads/2019/05/rapportcomplet151.pdf>.
- RIVIÈRE, Pascal, 1995. SICORE, un outil et une méthode pour le chiffrage automatique à l'Insee. In : *Courrier des statistiques*. [en ligne]. Août 1995. N°74, pp. 65-69. [Consulté le 18 mai 2020]. Disponible à l'adresse : <https://www.epsilon.insee.fr/jspui/bitstream/1/14345/1/cs74.pdf>.
- THÉVENOT, Laurent, 1986. Les investissements de forme. In : *Les conventions économiques*. [en ligne]. Cahiers de Centre d'étude de l'emploi, n°29, pp. 21-71, Presses Universitaires de France, Paris. [Consulté le 18 mai 2020]. Disponible à l'adresse : <http://gspm.ehess.fr/docannexe.php?id=556>.

POUR LES ENQUÊTES AUPRÈS DES MÉNAGES, L'INSEE RÉNOVE SES ÉCHANTILLONS

Patrick Sillard, Sébastien Favier**, Nicolas Paliod***, Ludovic Vincent*****

Les enquêtes auprès des ménages de l'Insee sont menées, majoritairement, en face-à-face, via un réseau d'enquêteurs implanté sur tout le territoire. Leur échantillonnage s'appuie sur deux zonages structurants : les secteurs de l'enquête Emploi et, pour toutes les autres enquêtes, les unités primaires de l'échantillon-maître. Depuis les années soixante, l'Insee utilisait la même base de sondage pour ses deux échantillons : le recensement de la population, qui a ainsi imprimé son rythme quasi-décennal aux refontes de l'échantillon-maître et de l'échantillon Emploi. Au début des années deux-mille, l'institut a rénové la méthode de recensement, devenue collecte annuelle sur des échantillons rotatifs de communes et de logements. Une nécessaire adaptation pour l'échantillon-maître a été menée en 2009 ; à la même époque, l'enquête Emploi choisissait d'utiliser des sources fiscales comme base de sondage. Fort de ces deux expériences, l'institut a engagé son projet de renouvellement de l'échantillon-maître et de l'échantillon Emploi en 2017 : utilisant comme base de sondage les fichiers fiscaux, adoptant des méthodes d'équilibrage spatial, le chantier a abouti en 2019 à la coordination des deux grands échantillons de l'Insee, une première dans l'histoire des enquêtes ménages de la statistique publique. L'article décrit les choix méthodologiques opérés : ils seront mis en œuvre pour les enquêtes en face-à-face pendant les dix prochaines années.

 *INSEE conducts face-to-face most of its household surveys, via a network of interviewers located throughout the country. Their sampling is based on two structuring zoning: the sectors of the Labour force survey and, for the other surveys, the primary units of the master sample. Since the 1960s, INSEE has been using the same sampling frame for the both samples: the population census, which has thus given its almost decennial rhythm to the redesign of the master sample and the Labour force survey's sample. In the early two-thousand's, the institute renovated the census method, which became annual collection on rotating samples of municipalities and housings. A necessary adaptation for the master sample was carried out in 2009; at the same time, the Labour force survey chose to base its sampling frame on Tax sources. On the strength of these two experiences, the institute began to renew the master sample and the Labour force survey's sample in 2017: using Tax files as a sampling frame, and adopting spatial balancing methods, the project led in 2019*

* Chef du département des Méthodes statistiques, direction de la Méthodologie et de la coordination statistique et internationale (DMCSI), Insee, patrick.sillard@insee.fr

** Chef de la division Sondages, DMCSI, Insee, sebastien.favier@insee.fr

*** Expert méthodologue, division Sondages, DMCSI, Insee, nicolas.paliod@insee.fr

**** Chef du projet Nautile, division Sondages, DMCSI, Insee, ludovic.vincent@insee.fr

to the coordination of the two large INSEE samples, a first in the history of household surveys in official statistics. The article describes the methodological choices made: they will be implemented for face-to-face surveys over the next ten years.

Pour construire les échantillons des enquêtes auprès des ménages, l'Insee s'appuie sur un échantillon-maître : sa constitution est structurante pour un institut statistique, car les choix méthodologiques qui sont opérés à cette occasion conditionnent la qualité des enquêtes menées pendant les dix années qui suivront et l'organisation de leur collecte. Longtemps associé aux recensements de la population, l'échantillon-maître a dû d'abord s'adapter aux évolutions méthodologiques du recensement rénové. La disponibilité des sources fiscales est ensuite apparue comme une alternative sérieuse. Leur utilisation pour l'enquête Emploi a confirmé dès 2009 l'intérêt pour l'Insee d'investir dans un nouvel échantillon-maître basé sur les fichiers fiscaux¹ : gains en précision des enquêtes, mais aussi efficacité de la collecte en face-à-face, facilitation du multi-mode, coordination accrue entre les échantillons pour limiter la charge statistique sur les ménages, etc. Cet article résume les étapes qui ont conduit l'Insee à rénover l'échantillon-maître qui sera mis en œuvre à compter de 2020.

ÉCHANTILLON ET QUALITÉ DES ENQUÊTES AUPRÈS DES MÉNAGES

La statistique publique documente les faits économiques et sociaux à l'aide de nombreuses sources de données. Parmi celles-ci, les enquêtes auprès des ménages sont pratiquées de longue date². Elles permettent, en une seule opération, la production d'un vaste panorama des caractéristiques des individus et de leurs interactions avec la thématique d'enquête : l'emploi, la sécurité et les conditions de vie, la santé, les revenus et le patrimoine, le logement, les transports sont autant de domaines qui sont l'objet de grandes enquêtes

« En s'appuyant sur un échantillon de quelques milliers de ménages, on peut prétendre inférer une connaissance en population générale de quelques dizaines de millions. »

de la statistique publique. Leur apport sur la connaissance des faits sociaux est irremplaçable : en s'appuyant sur un échantillon de quelques milliers de ménages, on peut prétendre inférer une connaissance en population générale de quelques dizaines de millions. Mais ceci sous réserve de s'inscrire dans un cadre de référence scientifique rigoureux : celui de la théorie probabiliste des sondages. Ce cadre associe des principes de sélection

aléatoire des individus dans l'enquête, à une méthodologie d'estimation de statistiques sans biais et de précision connue, fondée sur cette enquête. Par exemple, l'enquête Emploi en continu³ de l'Insee, avec un échantillon de 90 000 logements⁴, permet de connaître le taux de chômage trimestriel à $\pm 0,3$ point de pourcentage, dans un intervalle de confiance à 95 %.

1. L'Insee produit annuellement des Fichiers démographiques sur les logements et les individus (Fidéli) à partir des données fiscales sur les personnes physiques (essentiellement impôt sur le revenu, taxe d'habitation, taxe foncière).
2. L'Insee mène des enquêtes auprès des ménages depuis plus de 60 ans.
3. EEC, ou enquête Emploi dans la suite de l'article.
4. L'unité d'observation dans la base de sondage est le logement tandis que l'unité d'intérêt est le ménage, résidant dans ce logement.

Ce principe de sélection aléatoire est le seul qui garantit l'inscription des analyses de l'enquête dans le cadre théorique précédent. Cependant il exige de disposer *a priori* d'une base de sondage, c'est-à-dire d'une liste des ménages ou individus de l'ensemble de la population, ou à défaut, d'un sous-ensemble, lui-même issu d'une sélection aléatoire dans la population générale. Quand aucune base de sondage n'est disponible, des méthodes alternatives comme la méthode des quotas, ont été développées. Elles sont très peu utilisées dans le champ de la statistique publique, car les estimateurs qu'elles permettent de construire ne vérifient pas spontanément les « bonnes » propriétés évoquées *supra*. En effet, ceux-ci ne sont non biaisés et de précision connue que sous l'hypothèse, généralement invérifiable, que conditionnellement aux variables sur lesquelles les quotas sont construits, les réponses apportées aux questionnaires sont aléatoires (Deville, 1991). Le principe de sélectionner des ménages ou des individus aléatoirement dans une base de sondage est général aux enquêtes de la statistique publique, quel que soit le mode

« Le principe de sélectionner des ménages ou des individus aléatoirement dans une base de sondage est général aux enquêtes de la statistique publique, quel que soit le mode de collecte retenu. »

de collecte retenu : intermédié en face-à-face ou par téléphone, auto-administré par questionnaire internet ou papier.

L'interrogation intermédiée en face-à-face est, à ce jour, le mode dominant⁵ pour les enquêtes auprès des ménages⁶ gérées par l'Insee. Ce mode comporte en effet des avantages,

même s'il est plus onéreux. Il permet par exemple de mieux s'assurer de l'identité de la personne interrogée. Il permet surtout d'apprécier si l'enquêté comprend bien le contexte de l'enquête, s'il interprète convenablement les questions, et de prévenir, par l'échange direct avec l'enquêteur, ses éventuelles réserves. Il facilite aussi l'administration d'un questionnement parfois plus long que les autres modes de collecte. Le face-à-face est aussi, naturellement, plus adapté pour atteindre des personnes qui n'auraient pas accès à internet ou qui répondent peu au téléphone. Les autres modes de collecte, téléphone et internet, se développent, mais le questionnement en face-à-face restera vraisemblablement, pour les années prochaines, un mode important dans les enquêtes de la statistique publique.

1 APPORTS, FONCTIONNEMENT ET LIMITES D'UN ÉCHANTILLON-MAÎTRE

La collecte en face-à-face soulève des questions logistiques d'accès aux enquêtés et de limitation des déplacements des enquêteurs. En effet, les ménages susceptibles d'être enquêtés sont répartis sur tout le territoire national : s'ils étaient sélectionnés en population générale à chaque enquête, les coûts de déplacement des enquêteurs seraient prohibitifs⁷.

5. L'article a été rédigé avant les événements liés à la crise de la Covid-19. Il est possible que la part du face-à-face dans les modes de collecte s'en trouve durablement réduite.
6. La question ne se pose pas dans les mêmes termes en statistique d'entreprises, où la collecte des enquêtes est en général en mode auto-administré. Voir (Haag et Husseini-Skalitz, 2019).
7. En application du cadre d'emploi des enquêteurs de l'Insee, instauré en 2013, les temps de déplacement font partie intégrante du temps de travail, ce qui justifie l'importance accordée à leur optimisation.

Pour pallier ces inconvénients, on prédéfinit des zones géographiques sur lesquelles les ménages enquêtés ont vocation à être sélectionnés, quelle que soit l'enquête. Pour répondre au principe de sélection aléatoire, il convient que les zones géographiques soient elles-mêmes sélectionnées au hasard : on va donc couvrir le territoire national avec une partition géographique de zones, appelées **unités primaires**. Les unités primaires sélectionnées définissent alors les **zones de l'échantillon-maître** de l'Insee. Lorsqu'une enquête est organisée, les ménages à enquêter sont sélectionnés au hasard parmi ceux qui résident dans les zones de l'échantillon-maître (*figure 1*).

Afin de limiter la charge de réponse, lorsqu'un ménage est sélectionné pour une enquête, on cherche à faire en sorte qu'il ne puisse plus l'être par la suite (sauf évidemment dans le cadre d'une enquête panel). Mais de ce fait, les zones de l'échantillon-maître s'épuisent progressivement et il convient, au bout d'une dizaine d'années, de sélectionner un nouvel échantillon-maître. On verra par la suite que ce renouvellement devient également nécessaire pour des raisons d'efficacité de l'échantillonnage initial : en dix ans, la population évolue, se déplace, se concentre différemment sur le territoire ; les critères de représentativité initialement retenus se dégradent progressivement.

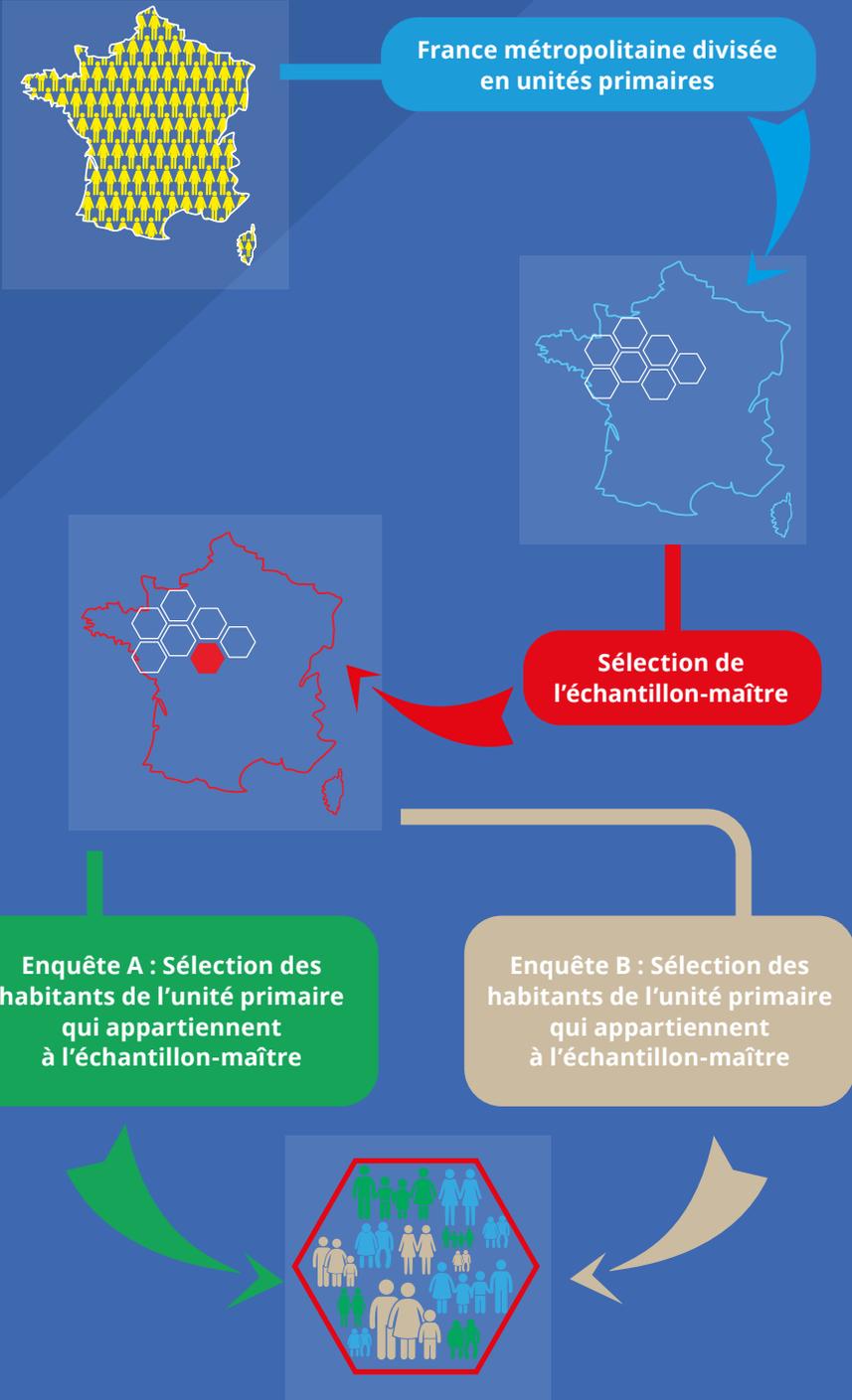
Depuis les années soixante, la **base de sondage** qui permettait de constituer les unités primaires des enquêtes de l'Insee était issue du recensement de la population. Ce recensement, appelé « général » jusqu'en 1999 parce qu'il était exhaustif, a vu sa méthodologie changer à partir de 2004 avec le recensement rénové ou en continu : le recensement quasi-décennal a laissé la place à une collecte tournante, sur cinq ans, et distinguant les communes selon leur taille ; en dessous du seuil de 10 000 habitants, les communes sont recensées exhaustivement tous les cinq ans ; au-dessus de ce seuil, la collecte a lieu chaque année sur un échantillon de 8 % de logements. Cette évolution a conduit à faire évoluer fortement la méthode de constitution de l'échantillon-maître dès 2010 (Christine et Favier, 2009).

UN RENOUVELLEMENT NÉCESSAIRE, DANS UN CONTEXTE NOUVEAU

En 2017, l'Insee disposait d'un échantillon-maître dont les zones, sélectionnées en 2010, s'épuisait peu à peu, et étaient susceptibles, à terme, de moins bien représenter le territoire. Un projet a donc été lancé, visant non seulement à renouveler cet échantillon-maître, mais aussi à en revisiter la méthodologie pour l'adapter à de nouvelles contraintes et à de nouvelles exigences. Car entre-temps, l'institut avait engrangé une expérience nouvelle : l'utilisation des sources fiscales depuis 2009, pour le tirage de l'échantillon de l'enquête Emploi, en lieu et place du recensement (voir *infra*).

En recherche de gains de productivité et de minimisation de ses coûts, la statistique publique se devait aussi de trouver des solutions méthodologiques permettant d'optimiser la précision de ses enquêtes. Car les réflexions s'inscrivaient dans un contexte européen qui évoluait. Le nouveau règlement-cadre européen sur les statistiques sociales IESS (*Integrated European Social Statistics*) fixe désormais un cadre général aux productions statistiques des États membres (Cases, 2019). Il impose, entre autres, des niveaux de précision à atteindre (ou des effectifs minimum pour les échantillons d'enquête) pour certains indicateurs-clés, obtenus par estimation sur données d'enquête, comme le taux de chômage ou le taux de pauvreté, au niveau national ou au niveau régional selon le cas.

Figure 1. Le procédé d'échantillon-maître



Dans les parties qui suivent, la méthodologie de ce nouvel échantillon-maître métropolitain⁸ est présentée, en particulier s'agissant du choix de la base de sondage, de la définition des unités primaires de l'échantillon-maître et de leur tirage. Un dernier paragraphe expose les travaux réalisés sur le renouvellement de l'échantillon de l'enquête Emploi que l'Insee a décidé, pour la première fois en 2019, de tirer de manière coordonnée à l'échantillon-maître.

📍 LE CHOIX DE LA BASE DE SONDAGE : OPPORTUNITÉS ET ARBITRAGES

L'échantillon d'enquête élémentaire consiste en une liste de ménages ou d'individus que l'enquêteur doit être en mesure d'interroger. Il faut pour cela qu'il puisse disposer d'éléments lui permettant d'entrer en contact avec le ménage sélectionné : adresse de résidence, caractérisation du logement, numéro de téléphone, courriel, etc. La base de sondage doit donc comprendre, outre une liste des individus et ménages, des variables permettant de les contacter.

Elle peut aussi utilement contenir des variables de caractéristiques socio-démographiques des ménages. En effet, si des variables de revenu, de profession, de situation par rapport au marché du travail, etc. sont disponibles dans la base de sondage, alors, dans la mesure où elles

« Le choix de la base de sondage résulte donc d'un arbitrage entre fraîcheur, couverture par rapport à l'exhaustif et qualité de la description des ménages et individus qui y figurent. »

sont corrélées aux variables d'intérêt de l'enquête, elles permettent d'élaborer un plan de sondage qui optimise la précision des estimations issues d'enquêtes, à ressource de collecte donnée.

Le choix de la base de sondage résulte donc d'un arbitrage entre fraîcheur, couverture par rapport à l'exhaustif et qualité de la description des ménages et individus qui y figurent. Nous allons voir comment la bascule a été progressivement opérée entre le choix du recensement et celui de la source fiscale.

📍 JUSQU'EN 2009, LE RECENSEMENT, GÉNÉRAL PUIS RÉNOVÉ...

Lorsque l'Insee réalisait des recensements généraux, le dernier datant de 1999, la base de sondage des enquêtes ménages de l'Insee était constituée simultanément à ces opérations⁹. Les unités primaires étaient sélectionnées à l'issue du recensement exhaustif et les coordonnées (nom, adresse, etc.) des habitants de ces zones étaient conservées pour que ceux-ci puissent, le cas échéant, être interrogés. Un nouveau recensement venait rafraîchir une base de sondage qui, vieillissant au rythme des déménagements, était, au bout d'une décennie, menacée d'obsolescence. Le dernier échantillon-maître constitué sur ce format a perduré, pour le tirage d'enquêtes, jusqu'en 2009. Cette année-là, après dix années d'usage de la base issue du dernier recensement exhaustif de 1999, la nouvelle base de sondage a été ancrée sur le dispositif du recensement en continu.

8. Les DOM font l'objet d'un traitement complémentaire spécifique qui n'est pas abordé dans cet article.

9. La périodicité quasi-décennale des échantillons-maîtres est en grande partie liée à celle des recensements généraux de population.

Comme auparavant, une liste d'unités primaires avait été constituée : mais cette fois, les personnes susceptibles d'être enquêtées une année donnée étaient celles qui avaient été *effectivement recensées* dans ces unités, lors de la *précédente* campagne annuelle du recensement. Au cours des années deux-mille-dix, la plupart des enquêtes auprès des ménages ont été tirées dans cette base. Ce dispositif bénéficiait de la fraîcheur de l'information acquise peu avant le tirage dans le cadre de la dernière campagne annuelle de recensement.

Mais un inconvénient est apparu au fil du temps, lié au mode de tirage des communes dans le recensement. En dessous de 10 000 habitants, l'enquête annuelle de recensement est fondée sur le schéma d'un échantillon de communes rotatif, destiné à les couvrir exhaustivement en 5 ans¹⁰. Cet échantillon annuel de communes résulte d'un tirage équilibré sur des données du recensement de 1999. Après une vingtaine d'années d'exercice, les évolutions démographiques et sociales ont éloigné cet échantillon de la situation équilibrée originelle. Ce défaut, dont l'ampleur dépend de la variable d'intérêt, se traduit en général par une hausse de l'erreur moyenne quadratique d'un rapport 2 à 20, par rapport à un tirage dans une base exhaustive (Favre-Martinoz, 2015).

La fin des années 2010 se rapprochant, le renouvellement de l'échantillon-maître de 2009 s'imposait donc et les premiers travaux méthodologiques ont été lancés en 2017.

... EN 2017, LE CHOIX DES SOURCES FISCALES

L'Insee disposait, depuis une décennie déjà, d'une solide expérience dans l'usage des sources fiscales pour le tirage d'échantillon d'enquête. En effet, l'échantillon de l'enquête Emploi de 2009 avait été tiré grâce aux fichiers de la taxe d'habitation (Loonis, 2009).

« L'échantillon de l'enquête Emploi de 2009 avait été tiré grâce aux fichiers de la taxe d'habitation. »

Ces données offrent une vision quasi-exhaustive des ménages résidant en France et de leurs habitations. Ces logements sont également l'objet de la taxe foncière et du suivi cadastral¹¹. Ils sont donc localisés très finement, à la parcelle cadastrale, soit une précision de quelques mètres des coordonnées géographiques associées. Ceci permet ainsi de constituer des grappes pour l'enquête Emploi, par proximité géographique. C'est

cette particularité qui a justifié, en 2009, de fonder l'échantillon de l'enquête Emploi sur les sources fiscales.

D'autres enquêtes ménages ont été tirées depuis lors dans la source fiscale, notamment celles ayant trait au patrimoine des ménages. L'usage de ces sources se généralisant, l'Insee a mis en place, en 2016, une production annuelle issue de la mise en cohérence statistique des sources fiscales. Les fichiers ainsi créés — appelés Fidéli ou fichiers démographiques sur les logements et les individus — constituent une base d'information de premier choix. Par rapport aux données brutes issues de la gestion fiscale, la source Fidéli porte une attention particulière aux questions de double-compte et de couverture de champ, notamment

10. En dessous du seuil de 10 000 habitants, pour une commune donnée, l'opération de recensement est quinquennale et exhaustive : tous les logements et les habitants de la commune sont recensés la même année.

11. Il s'agit du suivi par les services fiscaux des biens soumis à propriété privée, donc à la taxe foncière.

en rapprochant au maximum les concepts qui y sont utilisés de ceux du recensement. Par exemple, une identification des communautés¹² est réalisée dans le cadre du processus Fidéli, tandis que celles-ci ne sont pas à proprement parler repérées dans les données fiscales brutes (Merly-Alpa et Sillard, 2019).

Par ailleurs, Fidéli comprend un grand nombre de variables intéressantes pour procéder au sondage puis à la collecte des enquêtes : variables de contact (adresse des logements, courriels et coordonnées téléphoniques), variables de description des logements et de composition des ménages, variables de revenu. Celles-ci peuvent être utilisées *ex ante*, pour l'équilibrage ou la stratification, et *ex post*, pour les redressements. Les principales différences entre Fidéli et le recensement portent sur les variables de profession et de catégorie sociale, de diplôme et de revenus : les premières ne figurent que dans le recensement, tandis que la dernière ne figure que dans Fidéli. Les revenus sont connus pour être corrélés avec de nombreuses variables socio-démographiques, donc leur présence dans la base de sondage constitue une avancée importante. Au final, il apparaît donc que Fidéli constitue un candidat solide pour servir de base de sondage principale, en substitution du recensement. C'est le choix qui a été fait, pour la construction de l'échantillon-maître et de l'échantillon de l'enquête Emploi en 2020.

CONSTRUIRE LES UNITÉS PRIMAIRES : CONTRAINTE DE TAILLE ET DE GÉOGRAPHIE

L'Insee emploie un réseau d'environ 850 enquêteurs qui peuvent raisonnablement assurer la collecte d'une vague d'enquête auprès des ménages, répartie sur environ 500 unités primaires. Les enquêteurs se voient attribuer une zone de collecte, constituée généralement autour d'une unité primaire de l'échantillon-maître. Ces zones doivent donc être d'étendue adaptée pour pouvoir être couvertes par un enquêteur. Si différentes configurations ont été étudiées, il est apparu plus simple, y compris pour l'organisation sur le terrain, de constituer les unités primaires sous forme de réunions de communes contiguës.

Environ 500 000 ménages sont interrogés par l'Insee chaque année. Afin de limiter la charge de collecte pesant sur les ménages, les unités primaires sont aussi conçues pour éviter de ré-interroger un ménage qui resterait dans le même logement sur une période de cinq ans.

Compte tenu de ces différentes contraintes, les unités primaires de l'échantillon des enquêtes ménages comprennent au minimum 2 500 logements, assurant ainsi une réserve en cas d'utilisation plus large.

Enfin, pour faciliter la gestion des enquêteurs et de leur zone de collecte, chaque unité primaire est intégralement contenue dans un département. Ceci permet en outre de déconcentrer à ce niveau le problème d'optimisation sous-jacent.

12. Pour le recensement, la communauté est « un ensemble de locaux d'habitation relevant d'une même autorité gestionnaire, dont les habitants partagent à titre habituel un mode de vie commun » : maisons de retraite, cités universitaires, communautés religieuses, casernes, prisons, etc.

① SUIVRE DES CHEMINS POUR TROUVER LE PLUS COURT

Le problème a été formulé ainsi : pour chaque département, on parcourt les barycentres des communes du département, en partant de l'un d'entre eux sélectionné au hasard, en adoptant le plus court chemin revenant au point de départ et ne passant qu'une fois par commune. Ce chemin fermé, qui ne dépend que de la commune de départ, est calculé par un algorithme solution au problème du voyageur de commerce¹³ (Applegate *et alii*, 2006). On peut répéter le calcul autant de fois qu'il y a de communes dans le département.

Puis, pour chaque chemin calculé, on détermine les unités primaires associées, en parcourant les communes le long du chemin et en saturant progressivement la contrainte sur le nombre de logements de l'unité primaire, qui doit être supérieur à 2 500. Lorsqu'une unité de plus de 2 500 logements est constituée, on ouvre la constitution d'une nouvelle unité en passant à la commune suivante sur le chemin (*figure 2*). À chaque chemin, correspond donc un découpage du département en unités primaires respectant la contrainte sur le nombre de logements.

On définit ensuite **l'extension géographique d'une unité primaire**, par la moyenne des distances le long de la route entre le centre de la commune principale¹⁴ et celui des autres communes de l'unité, pondérées par le nombre de logements des communes non principales. Au niveau d'un département, la moyenne de l'extension géographique des unités primaires qui le recouvrent, détermine **une valeur-objectif**, qui caractérise le découpage obtenu et le chemin qui a été utilisé.

L'extension géographique associée à un découpage s'entend aussi comme une distance moyenne d'accès aux logements des unités primaires ainsi constituées, depuis leur centre. La question intéresse directement les enquêteurs sur le terrain : l'extension géographique de la zone qu'ils devront parcourir pour leur collecte aura un impact direct sur le coût et la durée de leurs déplacements ; il convient donc de la minimiser.

Finalement, parmi l'ensemble des chemins calculés, on retient celui dont la valeur-objectif est la plus petite : ce chemin correspond à une extension géographique moyenne des unités primaires, la plus petite parmi les différents découpages construits pour le département considéré. En moyenne, du fait de l'application du schéma d'optimisation précédent, l'extension géographique des unités primaires métropolitaines a diminué de 25 % par rapport à l'échantillon-maître de 2010.

La méthode ne permet cependant pas de prendre en compte certaines spécificités locales liées à la présence d'une frontière nationale, à la saturation de certains réseaux ou la fermeture de certaines voies, lesquelles peuvent moduler l'appréciation de l'extension géographique. La carte des unités primaires a donc été amendée ponctuellement après examen par les directions régionales de l'Insee. *In fine*, les travaux ont conduit à la partition de la France métropolitaine en 5 128 unités primaires (*figure 3*).

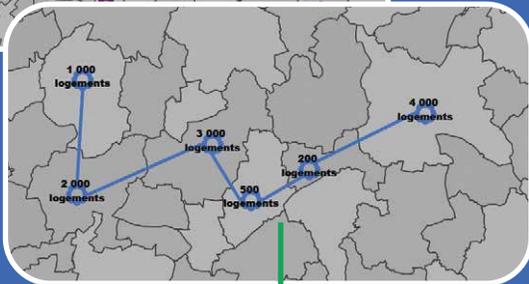
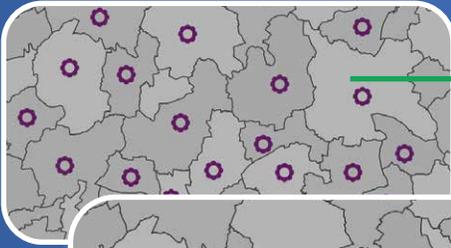
13. Problème d'optimisation qu'on pourrait résumer ainsi : étant donné une liste de villes, et des distances entre toutes les paires de villes, quel est le plus court chemin qui visite chaque ville une et une seule fois et qui se termine dans la ville de départ ?

14. Déterminée comme étant la commune de l'unité primaire comprenant le plus grand nombre de logements.

Figure 2. Comment constituer des unités primaires le long d'un chemin donné

1

Les points figurent les centres géographiques des communes.
Des distances le long de la route et centre à centre ont été calculées pour tous les couples de communes du département.

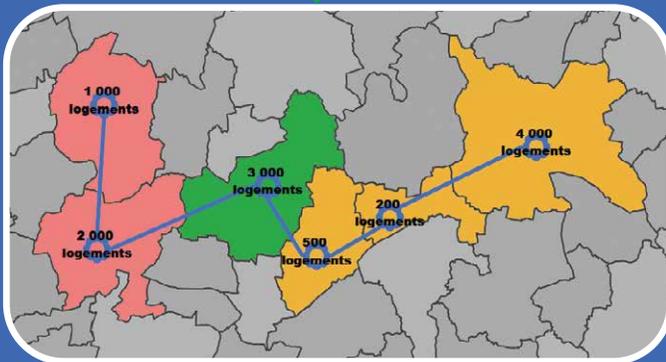


2

On part d'une première commune, sélectionnée au hasard : dans l'exemple, cette commune comprend 1 000 logements.

On cherche les chemins qui partent de cette commune et passent par toutes les autres communes du département, mais une fois seulement.

Parmi ces chemins, l'algorithme permet de calculer celui qui est le plus court.



3

Puis, le long de ce chemin, on constitue des regroupements – les unités primaires – en saturant successivement la contrainte de 2 500 logements par unité.

Dans l'exemple, le long de ce chemin, trois unités sont constituées : la première en rouge, la deuxième en vert et la troisième en jaune.

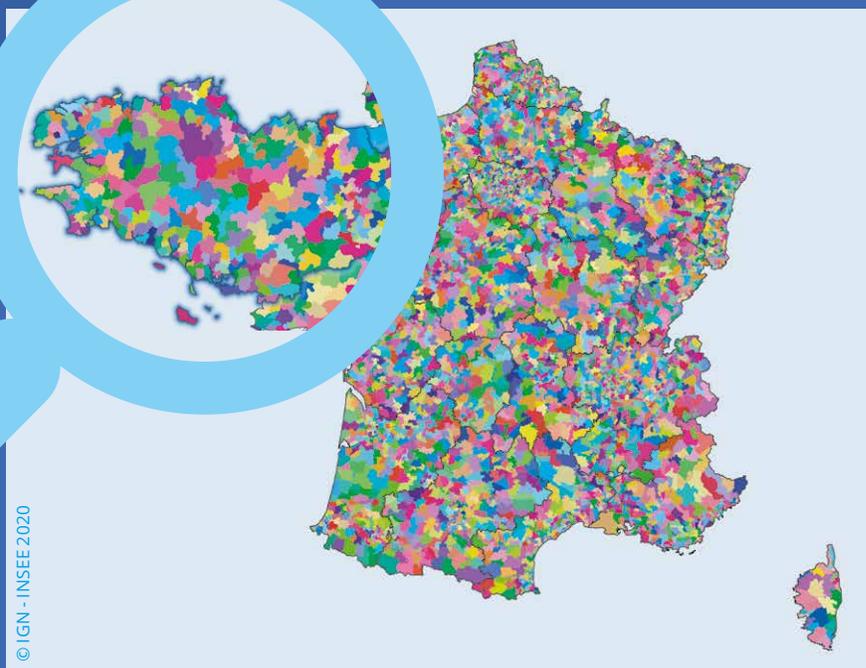
OPTIMISER LA PRÉCISION AVEC LE TIRAGE DE L'ÉCHANTILLON-MAÎTRE...

Un des enjeux principaux, pour l'Insee, est de maîtriser la précision de ses enquêtes. Comme on l'a vu, celles-ci sont tirées, en pratique, en deux étapes non simultanées. La première étape, dite de « tirage de premier degré » consiste à sélectionner les unités primaires de l'échantillon-maître. Cette première étape a lieu une fois tous les dix ans. Puis, pour chaque enquête, on tire, dans le cadre du « tirage de deuxième degré », parmi les ménages résidents de chaque unité primaire de cet échantillon-maître, les quelques dizaines qui seront effectivement interrogés au titre de cette enquête. La précision de l'enquête résulte de ces deux étapes de tirage : la sélection des unités primaires et celle des ménages parmi les résidents de ces unités primaires.

Les deux étapes de tirage sont indépendantes, donc la variance d'un estimateur, par exemple de moyenne ou de total fondé sur l'échantillon d'enquête recueilli, se décompose en deux termes distincts qui s'additionnent.

Plus précisément, on montre (Tillé, 2019) que la variance d'un estimateur de moyenne ou de total se décompose en deux termes : un terme de variance, dit de « premier degré » lié à la variance des totaux de la variable, observés au sein de chaque unité primaire ; à ce premier terme s'ajoute un second, dit de « deuxième degré », lié à la somme des variances

Figure 3. Avec le nouvel échantillon-maître, la métropole compte 5 128 unités primaires



intra-unités primaires. La variance de premier degré est indépendante de celle de second degré ; la variance de second degré est, elle aussi, essentiellement indépendante de celle de premier degré, hormis par la présence, dans son expression, d'un facteur d'échelle inversement proportionnel au taux de sondage de premier degré : si on réduit le nombre d'unités primaires sélectionnées, toutes choses égales par ailleurs, la variance de second degré augmente en proportion. Mais cette interaction est modeste et maîtrisable, de sorte qu'on peut donc découpler l'optimisation de variance. Ce qui s'avère très pratique puisque les tirages de premier et second degré ne sont pas réalisés simultanément.

L'optimisation du tirage de premier degré se résume donc à construire un plan de sondage optimal, c'est-à-dire qui minimise la variance de premier degré, associée aux quelque 500 unités primaires à sélectionner parmi les 5128 du territoire métropolitain.

📍 ...EN UTILISANT LES STATISTIQUES CONNUES SUR LES UNITÉS-PRIMAIRES...

Pour construire ce plan, on dispose des variables de la base de sondage (Fidéli), mais aussi de toutes les variables issues des productions de statistiques sociales de l'Insee au niveau de chaque commune, sommables sur des regroupements de communes. C'est d'ailleurs un des

« D'autres solutions auraient pu être envisagées, mais en complexifiant la production d'un ensemble de statistiques s'y rapportant. »

intérêts majeurs de constituer les unités primaires sous la forme de regroupements de communes, tandis que d'autres solutions auraient pu être envisagées, mais en complexifiant la production d'un ensemble de statistiques s'y rapportant.

De la sorte et en pratique, on dispose d'un grand nombre d'informations connues *a priori* sur les unités primaires : effectifs par sexe et âge issus du recensement, catégorie socio-professionnelle, diplôme, scolarisation, activité, composition des ménages, salaires, revenus, bénéfices industriels, agricoles et commerciaux, nombres d'assujettis à l'impôt sur la fortune, caractéristiques du logement, régions d'appartenance, etc.

Or la connaissance de ces statistiques est déterminante pour la variance de premier degré. En effet, le premier degré de sondage est, par nature, non affecté par d'éventuelles non-réponses, puisque toute unité primaire de l'échantillon-maître sera couverte par la collecte en toute circonstance. De ce fait, les plans de sondages les plus intéressants sont les plans équilibrés sur certaines variables : dans ces plans, les échantillons vérifient systématiquement l'égalité des estimateurs de moyennes avec leurs vraies valeurs sur l'ensemble de la population, ces vraies valeurs étant connues par ailleurs, par exemple issues du recensement.

📍 ... POUR PRIVILÉGIER L'ÉQUILIBRAGE (OU S'EN APPROCHER)...

Des algorithmes d'échantillonnage permettent de ne sélectionner que des plans équilibrés, ou éventuellement approchés sur le plan du strict équilibrage si celui-ci n'est pas rigoureusement possible. C'est le cas de la méthode du Cube (Deville et Tillé, 2004). L'équilibrage est intéressant dès qu'il existe une corrélation statistique entre une variable d'intérêt de l'enquête et les variables d'équilibrage : si celles-ci sont explicatives des variables d'intérêt, au sens d'une régression linéaire, alors on montre que la variance des estimateurs issus du sondage est de l'ordre du niveau de variabilité des résidus de la régression, et non de celle de la variable d'intérêt. Il convient donc de choisir des variables d'équilibrage

qui expliquent au mieux les futures variables d'intérêt des enquêtes. Néanmoins, pour l'élaboration du plan, il convient de limiter le jeu de variables d'équilibrage qui, de facto, limitent le champ de l'aléatoire dans la sélection des unités primaires.

Or le nombre d'unités primaires est relativement faible (5 128), de même que le nombre d'unités sélectionnées (environ 500). Par conséquent, le caractère surdéterminant des contraintes d'équilibrage, essentiellement lié à leur nombre, est assez vite atteint. Par exemple, introduire trop de variables d'équilibrage conduirait à ne plus pouvoir respecter le principe d'une taille d'échantillon-maître fixe en termes de nombre d'unités primaires.

Une analyse en composantes principales (ACP) a permis de réduire les variables d'équilibrage aux 15 premiers axes, représentant 99 % de l'inertie de l'ensemble des variables disponibles. Cette représentation de la quasi-totalité de l'information sur la population par l'échantillon obtenu se fait cependant au détriment de l'équilibrage parfait de celui-ci sur un nombre restreint de variables. C'est pourquoi, on complète ces axes de quelques variables clés, comme la population, ou celles permettant d'assurer un nombre fixe d'unités de l'échantillon-maître dans chaque région. Le plan est finalement équilibré au niveau régional, avec une phase d'atterrissage au niveau national permettant de libérer les contraintes qu'on ne peut pas respecter au niveau régional (Chauvet, 2009).

La sélection du plan, donc des variables retenues pour l'équilibrage, se fonde sur l'optimisation empirique du coefficient de variation des variables d'intérêt. Techniquement, celui-ci peut-être approché en simulant, par une méthode de Monte-Carlo¹⁵, un grand nombre de fois le plan de sondage retenu, la moyenne et la variance empiriques des estimations produites lors de chaque simulation convergeant vers la valeur vraie des estimateurs. Ces estimateurs sont calculés aux niveaux national et régional, du fait des contraintes du règlement IESS à ces deux niveaux géographiques. Les allocations retenues, ainsi que le plan de sondage de premier degré, visent donc à respecter les contraintes européennes.

La méthode d'équilibrage retenue *in fine* est une variante de l'algorithme du Cube, dite « échantillonnage doublement équilibré », adaptée aux situations dans lesquelles les variables d'intérêt sont spatialement corrélées, ce qui est le cas de la plupart des variables socio-économiques. En effet, dans ce type de situation, toutes choses égales par ailleurs, deux unités proches vont apporter plutôt moins d'information statistique que deux unités plus éloignées (Loonis et De Bellefon, 2018). On aura donc intérêt à sélectionner les unités modulo une légère force de répulsion entre unités proches, engendrant un échantillon mieux réparti spatialement (**figure 4**). C'est l'objet de l'algorithme d'équilibrage spatial, mis au point par (Grafström et Tillé, 2013), et appliqué dans le cas présent. Des simulations montrent que, pour des variables spatialement corrélées ne contribuant pas à l'équilibrage, comme le nombre de chômeurs ou de cadres, les gains de l'équilibrage spatial peuvent représenter jusqu'à 20 %, en termes d'erreur quadratique moyenne des estimations, par rapport à l'équilibrage simple (Favre-Martinoz et Merly-Alpa, 2016).

L'échantillon-maître fixe les zones géographiques sur lesquelles sont collectées la plupart des enquêtes auprès des ménages de l'Insee. Toutefois, pour des raisons liées principalement au protocole d'enquête, l'enquête Emploi est collectée sur des zones distinctes de celles

15. Le terme désigne une famille de méthodes algorithmiques visant à calculer une valeur numérique approchée en utilisant des techniques probabilistes.

de l'échantillon-maître. Il paraissait judicieux de rapprocher ces zones de collecte pour faciliter l'activité des enquêteurs. C'est ce qui a été réalisé pour la première fois lors du présent exercice.

LE RENOUVELLEMENT DE L'ÉCHANTILLON DE L'ENQUÊTE EMPLOI

Parmi les enquêtes ménages, l'enquête Emploi se distingue par un temps de collecte très court (trois semaines dans le trimestre) imposant à l'enquêteur sur le terrain un parcours rapide *in situ*. C'est ce qui justifie, entre autres, de tirer son échantillon par grappes de logements, géographiquement proches.

« En 2009, l'échantillon de l'enquête Emploi a été tiré de manière indépendante de l'échantillon-maître. »

En 2009, l'échantillon de l'enquête Emploi a été tiré de manière indépendante de l'échantillon-maître. Le parti retenu lors de ce renouvellement (Loonis, 2009) a été de s'appuyer sur les sources fiscales pour constituer des **secteurs de collecte**, chacun composé de six **grappes contenant chacune environ 20 logements** quasiment contigus. La France métropolitaine est découpée en 232 000

secteurs Emploi, composés de 1,4 million de grappes, regroupant 28,3 millions de résidences principales. Un échantillon de secteur a été tiré. Tous les logements d'une grappe échantillonnée sont interrogés, pendant six trimestres consécutifs (pour plus de détail, voir **encadré 1**), puis remplacés par les logements d'une autre grappe du même secteur. L'échantillon de l'enquête Emploi a ainsi une durée de vie de neuf ans.

En 2019, la volonté d'optimiser l'activité de collecte des enquêteurs de l'Insee a conduit à envisager une coordination des échantillons de l'EEC et des autres enquêtes auprès des ménages.

Figure 4. Effet de l'équilibrage spatial sur la sélection des unités primaires

Tirage équilibré



Tirage spatialement équilibré



Encadré 1. La constitution des grappes et des secteurs de l'enquête Emploi en 2019

L'enquête Emploi en continu (EEC) interroge les ménages échantillonnés pendant six trimestres consécutifs : on peut ainsi apprécier l'évolution de la situation d'emploi des personnes, tout en limitant l'attrition que provoquerait un nombre d'interrogations plus important. Chaque interrogation se déroule sur une période relativement courte (3 semaines), ce qui nécessite de réduire au maximum l'étendue de la zone de collecte pour un enquêteur donné. D'où l'intérêt de constituer des **grappes d'une vingtaine de logements** géographiquement très proches, et d'interroger tous les logements d'une grappe en même temps. À la fin de la période de six trimestres, une grappe « entrante » s'y substitue : partant de l'idée que la situation face à l'emploi dans deux grappes voisines est proche, dès la constitution de l'échantillon, on regroupe les grappes dans des **secteurs**, au sein desquels s'opère le renouvellement de l'échantillon.

Pour constituer les grappes, on mobilise une logique de proximité entre les logements, analogue à celle retenue entre les communes lors de la constitution de l'échantillon-maître. Sauf qu'ici on s'appuie sur une localisation très précise, disponible indirectement dans les sources fiscales : les coordonnées géographiques des logements, à partir des parcelles cadastrales.

Dans les immeubles collectifs, on facilite le travail des enquêteurs, en constituant des grappes comprenant soit toutes les résidences principales d'un même étage, soit aucune. Un « chemin » est ensuite créé entre chaque étage d'immeuble ou maison individuelle, au sein d'un même Iris*. Pour garantir le principe de faible étendue géographique, l'étendue moyenne des grappes est pénalisée dans une fonction de coût. Une fois le chemin parcouru et les grappes associées constituées** dans un Iris, le processus d'association est rejoué, en sélectionnant d'autres points de départ du chemin : pour chaque Iris considéré, on retient *in fine* le chemin qui aboutit à la plus faible étendue des grappes.

Une fois les grappes constituées, celles-ci sont rassemblées, toujours au sein des Iris, en **secteurs** de six à sept grappes, sur un principe similaire à celui déjà appliqué pour les communes dans les unités primaires ou pour les logements dans les grappes***. On détermine donc différents chemins optimaux pour différents points de départ parmi les barycentres des grappes et **retient, parmi les chemins calculés, celui qui confère la plus petite étendue géographique aux secteurs.**

L'échantillon devant être utilisé pendant 9 ans, il est rafraîchi chaque année en associant les nouveaux logements aux grappes ainsi constituées et en mettant à jour le statut des logements (résidences principales, secondaires ou vacantes). Là aussi, la source fiscale est la seule à pouvoir offrir ce service.

Ce schéma aboutit à constituer 1,4 millions de grappes, regroupant 28,3 millions de résidences principales en 232 000 secteurs. L'étendue géographique étant pénalisée dans l'algorithme de sélection, elle est nettement diminuée par rapport à l'échantillon de 2009. Par exemple, la longueur médiane passe de 1 km pour l'ancien échantillon à 0,6 km pour le nouveau.

Note : le détail des opérations sera l'objet d'un document de travail de l'Insee, à paraître en 2020.

* Îlots regroupés pour l'information statistique, maille élémentaire de diffusion des résultats du recensement correspondant à des zones de 2 000 logements en moyenne. Ce découpage concerne les communes de 5 000 habitants ou plus, en dessous de ce seuil, la commune est assimilée à un Iris.

** Une phase de consolidation finale permet d'affecter les résidences principales résiduelles.

*** Afin de limiter les secteurs comprenant 7 grappes, la contrainte d'appartenance des secteurs à l'Iris est assouplie, ce qui conduit, dans certains cas, à constituer des secteurs à cheval sur deux Iris.

Une première idée pouvait être de sélectionner les secteurs emploi au sein des unités primaires de l'échantillon-maître. Mais d'une part ceci aurait conduit à l'épuisement rapide de certaines unités primaires trop petites ; d'autre part, le niveau de précision sur les estimateurs issus de l'enquête Emploi aurait été insuffisant au regard de ce qui est attendu, notamment dans le règlement européen IESS. Cette solution a donc été vite abandonnée au profit d'une solution consistant à définir des voisinages des unités primaires sélectionnées au sein desquels les secteurs de l'EEC ont vocation à être échantillonnés. Ces zones voisines constituent des **unités de coordination**.

Une analyse détaillée des différentes options a été menée, notamment s'agissant de la taille et du nombre d'unités primaires constituant les unités de coordination, et s'appuyant sur des considérations liées à la précision des deux types d'enquêtes. Elle a conduit à retenir, par rapport à la taille minimale des unités primaires qui est de 2 500 logements, une taille plancher des unités de coordination à 10 000 logements-résidences principales.

Sur le plan du mécanisme de sélection, les unités de coordination apparaissent comme indirectement sélectionnées, une fois les unités primaires de l'échantillon-maître tirées. Plusieurs unités primaires peuvent être à l'origine d'une même unité de coordination. On est donc dans le contexte du sondage indirect (Deville et Lavallée, 2006) qui nécessite de procéder à un partage des poids.

“ Une fois les unités de coordination définies, celles-ci contiennent les secteurs emploi qui font l'objet, comme les unités primaires de l'échantillon-maître, d'un échantillonnage spatialement équilibré. ”

Une fois les unités de coordination définies, celles-ci contiennent les secteurs emploi qui font l'objet, comme les unités primaires de l'échantillon-maître, d'un échantillonnage spatialement équilibré. Les secteurs sont sélectionnés indépendamment d'une région à l'autre (stratification

régionale) ce qui permet d'ajuster convenablement les allocations régionales en fonction des contraintes du règlement européen IESS.

Comme pour l'échantillon-maître, des simulations de Monte-Carlo permettent d'éclairer le paramétrage du plan de sondage, par analyse des précisions obtenues sur des *proxies* des variables d'intérêt de l'enquête Emploi, à partir des variables de la base de sondage. C'est le cas par exemple de la variable d'activité dont un *proxy*, fondé sur les données fiscales, est l'indicatrice de somme non nulle des revenus d'activité.

FINALISATION ET OPTIMISATION CONJOINTE DES DEUX ÉCHANTILLONS

Le schéma d'ensemble requiert une optimisation, puisque les deux échantillons sont fortement connectés du fait de la coordination instaurée : une variation du nombre d'unités primaires retenues a une incidence sur la précision des enquêtes ménages, comme sur celle de l'enquête Emploi. Le nombre des secteurs a lui aussi une incidence sur la précision de cette dernière. Ainsi, plusieurs leviers sont possibles pour, par exemple, jouer sur la précision de l'enquête Emploi : le nombre d'unités primaires sélectionnées, la taille des unités de coordination, le nombre de secteurs. Des simulations sont mobilisées pour évaluer l'impact des différents paramètres en termes de précision. Ces résultats sont comparés aux niveaux de précisions à atteindre au regard du règlement cadre IESS.

Tous les moyens mis en œuvre pour améliorer la précision permettent *in fine* d'améliorer de manière spectaculaire, d'un facteur 5 à 6, le niveau de précision du premier degré du nouvel échantillon-maître, par rapport à celui fondé sur les enquêtes annuelles de recensement.

En rejouant les tirages simultanés des deux échantillons, on peut procéder à des estimations de précision convergeant vers les vraies valeurs, connues, pour des variables de la base de sondage. Des dizaines de millions d'échantillons d'unités primaires et de secteurs emploi ont été tirés, dans de nombreuses configurations différentes, pour parvenir à un choix de paramétrage. Celui-ci portait sur la configuration des zones, en particulier sur les seuils plancher de nombre de logements par unités primaires et unités de coordination, sur les variables d'équilibrage, et sur les allocations, c'est-à-dire le nombre de zones sélectionnées dans les échantillons retenus.

Les derniers paramètres cible ont été les allocations régionales : elles ont permis d'affiner les *scenarii* présentés au comité de direction de l'Insee, ainsi qu'aux directions régionales qui gèrent les enquêteurs. Initialement proportionnelles à la taille de la région (en termes de nombre de résidences principales), elles ont été dans un second temps renforcées dans les petites régions, au détriment des plus grandes, afin d'éviter une concentration excessive des secteurs emploi au voisinage d'un nombre trop restreint d'unités primaires. Des simulations de tirage ont permis de vérifier que la précision attendue sur les variables phares de l'enquête Emploi apparaissait compatible avec les besoins de diffusion. L'exercice de validation finale a débouché sur un échantillon de 541 unités primaires, desquelles découlent la sélection de 524 unités de coordinations et finalement 2 944 secteurs pour l'enquête Emploi¹⁶.

Naturellement, la variance totale dépend aussi du second degré qui dépend lui-même des allocations (nombre de ménages tirés dans chaque unité primaire). La répartition de variance entre premier et second degré étant très diverse, il convient de ne pas trop réduire le nombre d'unités primaires au motif que la précision du premier degré serait très favorable. La prudence nécessaire a conduit à retenir un nombre d'unités primaires en légère baisse par rapport à l'échantillon antérieur, passant de 567 à 541.

Spontanément, à nombre de secteurs sélectionnés identiques, les gains attendus des évolutions de méthode sur la précision de l'EEC sont modestes, puisque l'échantillon antérieur avait déjà été construit sur les fichiers fiscaux (Loonis, 2009). Toutefois la possibilité, nouvelle en raison de la disponibilité de la source Fidéli, d'équilibrer le tirage des secteurs sur des *proxies* de variables de chômage, par exemple, a constitué une amélioration substantielle. À l'opposé, la coordination entre les deux échantillons génère un surcroît de variance sur l'enquête Emploi. Compte tenu des gains attendus sur le premier degré des enquêtes ménages, il a été possible d'ajuster les allocations d'unités primaires par région de façon à assurer un niveau de précision de l'EEC par région compatible avec les contraintes du règlement IESS, malgré la perte de précision spontanée liée à la coordination. En outre, il a été possible de réduire un peu le nombre de secteurs de l'EEC, car dans le bilan de précision de l'enquête, le gain marginal associé aux unités primaires est nettement plus favorable¹⁷ que celui associé aux secteurs, en raison de l'effet de grappe, caractéristique de ces derniers.

16. Les allocations régionales seront présentées dans un document de travail de l'Insee, à paraître en 2020.

17. Le ratio des élasticités des coefficients de variation de moyennes, estimées par l'EEC, au nombre d'unités primaires d'une part, et de secteurs d'autre part, vaut 4.

UN PROJET SERVANT UNE STRATÉGIE D'EXTENSION DES MODES DE COLLECTE

La rénovation des échantillons des enquêtes auprès des ménages et de l'enquête Emploi est un projet stratégique pour l'Insee. Elle répond à la volonté de pouvoir mener des enquêtes, sur la base d'une méthodologie solide et éprouvée, adaptée au mode de collecte en face-à-face.

Les modes alternatifs (internet, téléphone, papier) se développent, en raison de leur souplesse, de leur faible coût et parce qu'ils permettent d'atteindre mieux certaines catégories de ménages : ils ne requièrent pas, *stricto sensu*, d'échantillon-maître. Du reste, l'Insee renforce activement sa capacité à administrer les enquêtes en multimode (Cotton et Dubois, 2019 ; Koumarios et Sigaud, 2019). Mais, dans un pays étendu comme la France¹⁸, le principe d'un échantillon-maître reste le plus adapté à une collecte en face-à-face, que ce soit en mode unique ou comme composante d'une enquête multimodale.

Le projet a aussi été exigeant. Mené en mode agile, avec une interaction forte entre les concepteurs et les développeurs, il a mobilisé trois experts de la direction de la Méthodologie à plein temps pendant 18 mois. Le déploiement des enquêteurs sur les nouvelles zones a été étudié et déterminé par les directions régionales et la direction des statistiques démographiques et sociales, en concertation avec les organisations syndicales. Au-delà des aspects méthodologiques et organisationnels, une application spécifique a été développée : Nautille¹⁹ standardise et sécurise les tirages dans l'échantillon-maître, et marque les logements sélectionnés pour une enquête. Cette application est compatible avec les outils de gestion des métadonnées actives²⁰. Elle alimentera par ses échantillons les applications de gestion d'enquête, de calcul de charge des enquêteurs et d'administration des contacts avec les ménages, pour les dix ans qui viennent.

18. D'autres pays, comme l'Australie et le Canada, disposent d'un échantillon-maître. Et pour leurs enquêtes en face à-face, la plupart des pays procèdent par sélection initiale de zones géographiques, puis de ménages au sein de ces zones (OECD, 2019).

19. Nautille (Nouvelle application utilisée pour le tirage des individus et des logements des enquêtes) succède ainsi à Octopusse (Organisation coordonnée de tirages optimisés pour une utilisation statistique des échantillons) qui a servi aux échantillonnages des enquêtes de 2009 à 2019.

20. RMéS, le référentiel de métadonnées statistiques de l'Insee, a fait l'objet d'un article dans le numéro N2 du Courrier des statistiques (Bonnans, 2019).

BIBLIOGRAPHIE

APPLEGATE, David L., BIXBY, Robert E., CHVÁTAL, Vašek et COOK, William J., 2006. *The traveling salesman problem: a computational study*. Princeton University Press, Princeton Series in Applied Mathematics. ISBN : 978-0691129938.

BONNANS, Dominique, 2019. RMÉS, le référentiel de métadonnées statistiques de l'Insee. In : *Courrier des statistiques*. [en ligne]. 27 juin 2019. N°N2, pp. 46-57. [Consulté le 16 juin 2020]. Disponible à l'adresse : <https://www.insee.fr/fr/statistiques/fichier/4168396/courstat-2-6.pdf>.

CASES, Chantal, 2019. IESS : l'Europe harmonise ses statistiques sociales pour mieux éclairer les politiques. In : *Courrier des statistiques*. [en ligne]. 19 décembre 2019. N°N3, pp. 127-139. [Consulté le 16 juin 2020]. Disponible à l'adresse : <https://www.insee.fr/fr/statistiques/fichier/4254233/courstat-3-10.pdf>.

CHAUVET, Guillaume, 2009. Stratified balanced sampling. In : *Survey Methodology*. [en ligne]. Juin 2009. Statistique Canada, Catalogue n°12-001-X, vol. 35, n°1, pp. 115-119. [Consulté le 16 juin 2020]. Disponible à l'adresse : <https://www150.statcan.gc.ca/n1/pub/12-001-x/12-001-x2009001-eng.pdf>.

CHRISTINE, Marc et FAIVRE, Sébastien, 2009. OCTOPUSSE : un système d'Échantillon-Maître pour le tirage des échantillons dans la dernière Enquête Annuelle de Recensement. In : *X^e Journées de Méthodologie Statistique de l'Insee*. [en ligne]. 23-25 mars 2009. [Consulté le 16 juin 2020]. Disponible à l'adresse : http://www.jms-insee.fr/2009/S14_3_ACTE_FAIVRE_JMS2009.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 16 juin 2020]. Disponible à l'adresse : <https://www.insee.fr/fr/statistiques/fichier/4254216/courstat-3-3.pdf>.

DEVILLE, Jean-Claude, 1991. Une théorie des enquêtes par quotas. In : *Techniques d'Enquête*. [en ligne]. 16 décembre 1991. Statistique Canada, vol. 17, pp. 177-195. [Consulté le 16 juin 2020]. Disponible à l'adresse : <https://www150.statcan.gc.ca/n1/fr/pub/12-001-x/1991002/article/14504-fra.pdf?st=1bW9pHnv>.

DEVILLE, Jean-Claude et LAVALLÉE, Pierre, 2006. Indirect sampling : The foundations of the generalized weight share method. In : *Survey Methodology*. [en ligne]. Décembre 2006. Statistique Canada, catalogue n°12-001, vol. 32, n° 2, pp. 165-176. [Consulté le 16 juin 2020]. Disponible à l'adresse : <https://www150.statcan.gc.ca/n1/en/pub/12-001-x/2006002/article/9551-eng.pdf?st=i92lk5a4>.

DEVILLE, Jean-Claude et TILLÉ, Yves, 2004. Efficient balanced sampling : The cube method. In : *Biometrika*. [en ligne]. 1^{er} décembre 2004. Vol. 91, n°4, pp. 893-912. [Consulté le 16 juin 2020]. Disponible à l'adresse : <https://academic.oup.com/biomet/article-pdf/91/4/893/1249543/914893.pdf>.

FAVRE-MARTINOZ, Cyril, 2015. *Étude préliminaire de l'impact de la prise en compte des groupes de rotation du RP pour l'échantillonnage des enquêtes ménages*. Insee, note interne n°776/DG75-L110/DE/CFM/SF.

FAVRE-MARTINOZ, Cyril et MERLY-ALPA, Thomas, 2016. Utilisation des méthodes d'échantillonnage spatialement équilibré pour le tirage des unités primaires des enquêtes ménages de l'Insee. In : *9^e Colloque Francophone sur les Sondages*. [en ligne]. 14 octobre 2016. Société française de Statistique (SFDS) et Université du Québec en Outaouais (UQO), Gatineau, Canada. [Consulté le 16 juin 2020]. Disponible à l'adresse : <http://sondages2016.sfds.asso.fr/wp-content/uploads/2016/10/Session21-MERLYALPA.pdf>.

GRAFSTRÖM, Anton et TILLÉ, Yves, 2013. Doubly balanced spatial sampling with spreading and restitution of auxiliary totals. In : *Environmetrics*. [en ligne]. Vol. 24, n°2, pp. 120-131. [Consulté le 16 juin 2020]. Disponible à l'adresse : <https://core.ac.uk/download/pdf/43674147.pdf>.

KOUMARIANOS, Heïdi et SIGAUD, Eric, 2019. Eno, un générateur d'instruments de collecte. In : *Courrier des statistiques*. [en ligne]. 19 décembre 2019. N°N3, pp. 17-28. [Consulté le 16 juin 2020]. Disponible à l'adresse : <https://www.insee.fr/fr/statistiques/fichier/4254218/courstat-3-4.pdf>.

LOONIS, Vincent, 2009. La construction du nouvel échantillon de l'Enquête Emploi en Continu à partir des fichiers de la Taxe d'Habitation. In : *Journées de Méthodologie Statistique*. [en ligne]. 23 mars 2009. Insee, X^e journées, Paris, Session 2 – Autour de l'Enquête Emploi. [Consulté le 16 juin 2020]. Disponible à l'adresse : http://www.jms-insee.fr/2009/S02_5_PRESENTATION_LOONIS_JMS2009.PDF.

LOONIS, Vincent et DE BELLEFON, Marie-Pierre, 2018. *Manuel d'analyse spatiale – Théorie et mise en œuvre pratique avec R*. [en ligne]. 29 octobre 2018. Insee. Eurostat. Collection Insee Méthodes, n°131, [Consulté le 16 juin 2020]. Disponible à l'adresse : <https://www.insee.fr/fr/statistiques/fichier/version-html/3635442/imet131.pdf>.

MERLY-ALPA, Thomas et SILLARD, Patrick, 2019. *The use of the French administrative dataset Fidéli as Sampling Frame for Household Surveys*. [en ligne]. 4-5 juin 2019, ESS workshop Administrative data for social statistics, Valence, Espagne. [Consulté le 16 juin 2020]. Disponible à l'adresse : https://ec.europa.eu/eurostat/cros/system/files/mr-patrick-sillard_french-admin-dataset-fideli_insee.pdf.

OECD, 2019. Sampling and Weighting. In : *Technical Report of the Survey of Adult Skills (PIAAC) (3rd Edition)*. [en ligne]. Section 4, chapitres 14-16. [Consulté le 16 juin 2020]. Disponible à l'adresse : http://www.oecd.org/skills/piaac/publications/PIAAC_Technical_Report_3rd_Edition_2019_Section4_Chapters14-16.pdf.

TILLÉ, Yves, 2019. *Théorie des sondages : Échantillonnage et estimation en populations finies*. [en ligne]. Dunod, 2^e édition, ISBN : 978-2-10-079355-6. [Consulté le 16 juin 2020]. Disponible à l'adresse : <https://www.insee.fr/fr/statistiques/fichier/4254233/courstat-3-10.pdf>.

LA TROISIÈME REFONTE DU RÉPERTOIRE SIRENE

TROP AMBITIEUSE OU PAS ASSEZ ?

Christophe Alviset*

Longtemps au cœur du système statistique sur les entreprises, Sirene, le répertoire d'entreprises de l'Insee, a évolué en profondeur depuis 2005, date de sa dernière grande refonte : les grandes fonctionnalités ont été clarifiées, avec un répertoire des unités statistiques distinct (Sirus) et une mise à disposition déportée sur une interface (API). Sirene est maintenant consacré entièrement à la gestion administrative et aux échanges avec les partenaires de cette gestion.

La troisième refonte de Sirene est concomitante à d'importantes réformes administratives : la création d'un « guichet unique » et d'un « registre général » pour les formalités des entreprises, ainsi que le service d'identification électronique des personnes représentant les unités légales. Le programme vise d'abord à réduire les coûts de gestion du répertoire, dans un contexte d'augmentation régulière du volume de déclarations à traiter. Il doit aussi permettre d'améliorer l'évolutivité du répertoire, pour intégrer plus rapidement les modifications des formalités d'entreprise, et rendre un meilleur service aux partenaires et aux déclarants. Il s'appuie sur une analyse de l'urbanisation du système d'information, et sur une démarche qualité : la conduite agile des projets sur la période 2019-2023 facilite l'adaptation aux changements organisationnels, fonctionnels et techniques, susceptibles de se produire pendant leur déroulement tout en répondant aux besoins des gestionnaires de l'Insee.

 *Having long been at the heart of the business statistics system, SIRENE, INSEE's business register, has undergone major changes since 2005, when it had its last major overhaul: the main functions have been clarified, with a separate register for statistical units (SIRUS) and dissemination through an API (Application Programming Interface). SIRENE is now fully dedicated to administrative management and electronic exchanges with the partners involved in this management.*

The third overhaul of SIRENE is taking place alongside other significant administrative reforms: the creation of a "one-stop shop" and a "general register" for business procedures, together with the electronic identity for persons representing legal units. The programme first aims to cut the costs of managing the register, against a backdrop of steady increase in the volume of declarations to be processed. It should also ease the evolutions of the register, so that changes in business procedures can be incorporated more quickly as well as to provide better service to partners and businesses. The programme is founded on an analysis of the information system architecture and on quality management: the agile management of the projects over the 2019-2023 period facilitates the adaptation to the various organisational, functional and technical changes likely to occur during their progress, while also meeting the needs of INSEE end-users.

* Directeur du programme Sirene 4, direction des Statistiques d'Entreprises (DSE), Insee, christophe.alviset@insee.fr

Le répertoire Sirene identifie les unités légales et les établissements en France depuis 1970 (Létoquart et Reffet, 1969). En près de 50 ans, le répertoire a connu deux refontes importantes : comme ses prédécesseurs Sirene 2 en 1986 et Sirene 3 en 2005 (Picard et Chazelas, 2003), le programme Sirene 4 modernisera la gestion du répertoire, cette fois-ci à l'horizon 2022-2023.

Sirene aujourd'hui, c'est 31 millions d'établissements, 4 millions d'événements traités chaque année, et plus de 10 000 utilisateurs réguliers. Faire évoluer un référentiel de cette importance suppose un certain nombre de précautions, que nous allons aborder dans cet article. Mais si la troisième refonte de Sirene répond à des préoccupations identiques à celles qui ont précédé – diminution des coûts de gestion, simplification administrative pour les entreprises, rationalisation des traitements et du système d'information – elle s'en démarque par son contexte : l'environnement administratif est traversé d'évolutions majeures, impulsées par le politique ; les institutions accélèrent leur adaptation à la révolution numérique ; les exigences des utilisateurs s'amplifient. Enfin, la statistique publique met en œuvre de nouvelles méthodes de développement de ses projets.

LE RÉPERTOIRE SIRENE EST UN DES RÉFÉRENTIELS MAJEURS DE L'ÉTAT

Le répertoire national d'identification des entreprises et des établissements concerne toutes les unités légales implantées sur le territoire national. Il a été créé par le décret du 14 mars 1973¹ et sa gestion confiée dès ce moment à l'Insee : l'institut a conçu et met à jour depuis près de 50 ans, le système Sirene (Système Informatique pour le Répertoire des ENTREPRISES et des Etablissements). À ce titre, il est un des rares instituts de statistique, avec celui de la Suisse, qui tient un répertoire administratif des entreprises.

« L'usage généralisé de ce numéro unique par l'administration et par la statistique publique place le répertoire Sirene au cœur du dispositif d'information sur les entreprises. »

L'Insee attribue un numéro unique d'identification, le numéro Siren, aux personnes morales et aux entrepreneurs individuels et le numéro Siret à chacun de leurs établissements, c'est-à-dire à chaque implantation géographiquement distincte. L'usage généralisé de ce numéro unique par l'administration et par la statistique publique place le répertoire Sirene au cœur du dispositif d'information sur les entreprises.

Le répertoire concerne toutes les entreprises, quelle que soit leur forme juridique et quel que soit leur secteur d'activité, situées en métropole, ainsi que dans les tous les départements² et certaines collectivités³ d'outre-mer. Les organismes publics y sont également répertoriés ainsi que les entreprises étrangères qui ont une représentation ou une activité en France⁴. Sont donc inscrits au répertoire tous les entrepreneurs individuels ou les personnes morales immatriculées au Registre du commerce et des sociétés, ou au Répertoire des métiers, ou employant du personnel salarié, ou soumises à des obligations fiscales, ou bénéficiaires de transferts financiers publics.

1. Décret repris par les articles R. 123-220 à R. 123-234 du code de commerce.

2. Guadeloupe, Guyane, Martinique, la Réunion et Mayotte.

3. Saint-Barthélemy, Saint-Martin et Saint-Pierre-et-Miquelon.

4. Les entreprises et établissements de Monaco ou en Andorre peuvent également demander à y figurer.

Pour mettre à jour Sirene, l'Insee reçoit des informations du réseau constitué des 1 400 centres de formalités d'entreprises (CFE) : implantés au sein des institutions en charge des fichiers et registres cités plus haut⁵, ces centres permettent aux entreprises de déclarer en un seul lieu les formalités nécessaires à leur prise en compte dans les systèmes d'information de l'ensemble des administrations concernées.

Pour que le principe à la base de la création des CFE puisse fonctionner, il faut non seulement réceptionner les informations, mais également les retransmettre. L'Insee communique en permanence les évolutions enregistrées dans le répertoire Sirene (créations, modifications, cessations d'activité, appelées aussi « mouvements ») aux associés et partenaires⁶ du répertoire : ces flux d'informations placent depuis longtemps Sirene au cœur d'un écosystème complexe (*figure 1*).

Le dispositif (identification unique, écosystème autour du répertoire) permet de constituer avec Sirene le référentiel français des entreprises et des établissements. Celui-ci irrigue les systèmes d'information de l'administration, mais également du secteur privé, évitant ainsi à chacun d'assurer son propre suivi de la vie des entreprises.

« Un référentiel n'est pas une base de données comme une autre. »

Toutefois, un référentiel n'est pas une base de données comme une autre. Le responsable de la tenue d'un référentiel se doit de suivre un certain nombre de bonnes pratiques (Bizingre, Paumier et Rivière, 2013). En France, l'État a confié à la direction interministérielle du Numérique (Dinum) la mission de définir ses propres critères (voir *encadré 1* et (Disic, 2013)). Si la gestion du répertoire Sirene suit déjà la majorité des

préconisations de la Dinum, le programme Sirene 4 sera l'occasion d'un pas supplémentaire dans cette direction, par exemple en matière d'historisation systématique des données d'identification (règle RA6).

Référentiel officiel pour servir une large palette d'utilisateurs institutionnels ou privés, Sirene est également un outil au cœur du système statistique public sur les entreprises. Or au fil du temps, il est apparu que la description purement juridique ne suffisait pas à représenter correctement l'activité des entreprises du point de vue économique.

LA MISSION STATISTIQUE DE SIRENE : CONFIÉE À UN RÉPERTOIRE DÉDIÉ

Si l'Insee gère depuis les années soixante-dix le répertoire Sirene, c'est en partie parce que l'institut y avait un intérêt fort : la gestion du répertoire lui permet de se constituer la base de sondage indispensable à la conduite de ses enquêtes auprès des entreprises.

Mais l'extension progressive des sources d'informations a aussi alourdi le répertoire de tout un ensemble d'unités qui n'intéressent pas le statisticien, parce que dépourvues de consistance économique. De plus, l'évolution de l'organisation interne aux entreprises a rendu l'unité légale seule de moins en moins adaptée à la statistique économique.

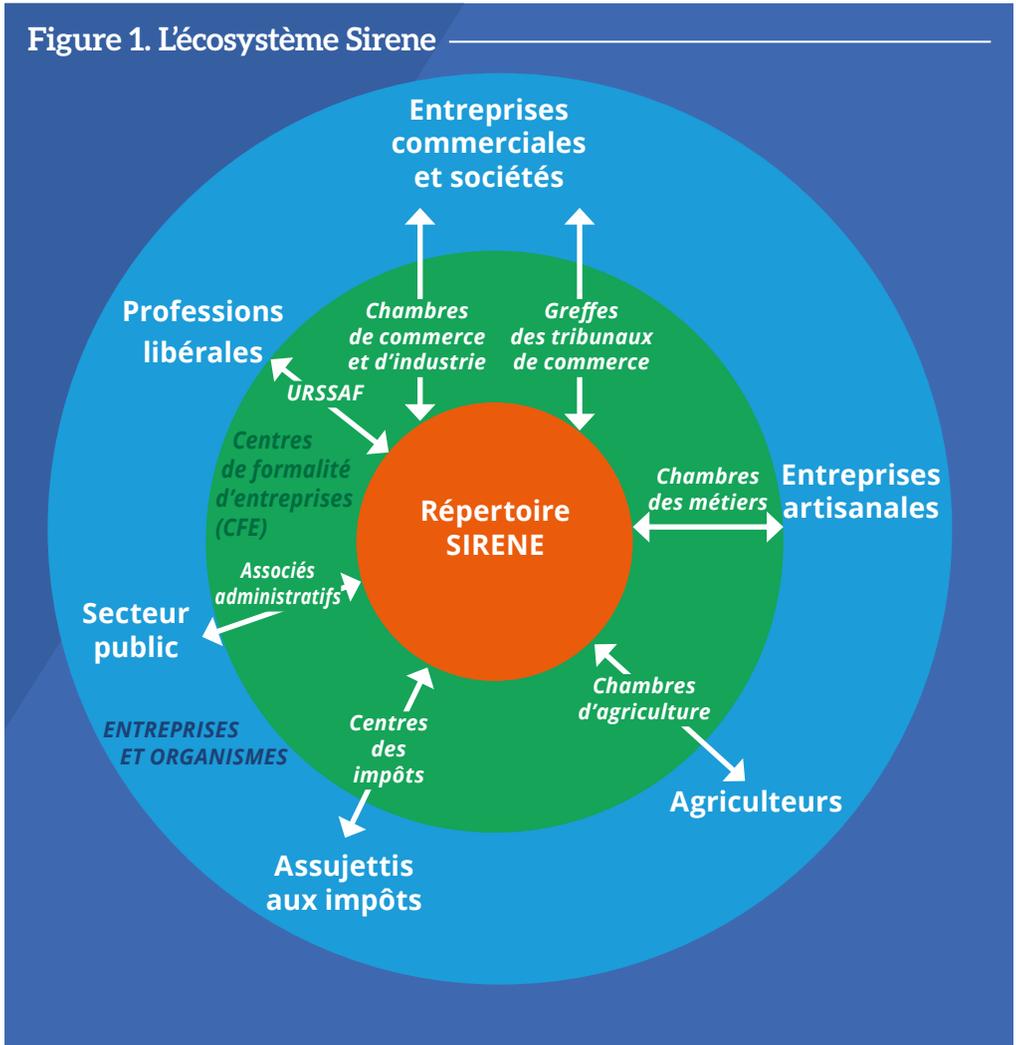
5. Greffes des tribunaux de commerce et analogues, unions de recouvrement des cotisations de sécurité sociale et d'allocations familiales (URSSAF), chambres de commerce et d'industrie, chambres des métiers et de l'artisanat, chambres d'agriculture, services des impôts.

6. Il s'agit par exemple du ministère des Solidarités et de la Santé, des directions générales des Finances publiques et des Douanes.

Depuis 2012, l'Insee utilise le répertoire Sirene pour construire Sirius, son répertoire des unités statistiques⁷ (Béguin et Haag, 2017). Ces unités statistiques permettent de mieux décrire le tissu économique que les unités légales, en incluant les groupes, les réseaux et les entreprises profilées (Haag, 2019), mais aussi en mettant à disposition du service statistique public (SSP)⁸ un grand nombre de variables statistiques absentes de Sirene.

Sirus est aujourd'hui l'outil de référence pour vérifier les données d'identification dans les opérations statistiques impliquant la connaissance de variables relatives à des entreprises (au sens statistique du terme, voir (Haag, 2019)) ou des établissements. C'est également une base de données pour les études structurelles sur des zones géographiques plus ou moins fines ou sur certaines catégories d'entreprises.

Figure 1. L'écosystème Sirene



7. Sirius ou Système d'immatriculation au répertoire des unités statistiques.

8. L'Insee et les services statistiques ministériels (SSM).

Base de sondage des enquêtes de la statistique publique, le répertoire Sirius est utilisé notamment pour tirer les échantillons des enquêtes annuelles d'entreprises dans l'industrie, la construction, le commerce et les services. Sources privilégiées du suivi de la démographie des entreprises, les mises à jour enregistrées dans le répertoire Sirius permettent également d'élaborer les statistiques mensuelles de création d'entreprises publiées par l'Insee.

Pour les utilisateurs n'appartenant pas à la Statistique publique, l'accès au répertoire Sirene relève d'une approche spécifique, qui elle aussi a évolué au cours des années récentes.

📍 LA DIFFUSION DE SIRENE : MODERNISÉE ET DISTINCTE DE LA GESTION

L'administration française est traversée régulièrement par des réformes portant des enjeux de simplification administrative pour les entreprises ; à leur création, Sirene et les CFE avaient d'ailleurs constitué des étapes importantes de ce mouvement. Plus récemment, plusieurs événements ont contribué à l'évolution de la diffusion du répertoire Sirene.

Lancé en 2013, le programme « Dites-le-nous une fois⁹ » prévoit que les données d'entreprises soient échangées entre administrations au lieu d'être demandées plusieurs fois aux entreprises. Dans un autre ordre d'idée, l'orientation de « l'État-plateforme » cherche à favoriser les services numériques pour accéder aux bases de données publiques, plutôt que la compilation de ces bases dans les environnements de leurs utilisateurs (SGMAP, 2020 ; Colin et Verdier, 2015 ; Dinsic, 2016 ; Biehl, 2016). *In fine*, ces deux actions de la puissance publique ont incité l'Insee à conduire en 2017 son propre projet de modernisation, celui de la diffusion de Sirene.

Dans le même temps, la direction de la Sécurité sociale¹⁰ demandait d'accéder directement à la totalité des données historiques du répertoire Sirene ; ceci a également contribué à restructurer fortement l'offre de l'Insee concernant Sirene.

En choisissant de développer une « API » (*Application Programming Interface*) pour mettre Sirene à la disposition de ses utilisateurs (**encadré 2**), l'Insee a contribué à séparer les deux fonctions autrefois intimement liées : la gestion et la diffusion du répertoire.

Ce faisant, on assiste à une petite révolution culturelle : du côté Insee, la technique des API diminue fortement la charge de conception des produits de diffusion ; du côté des utilisateurs des données, qu'ils soient ou pas partie prenante de la gestion des formalités d'entreprises, Sirius et l'API Sirene rationalisent la partie du système d'information consacrée à la mise à disposition des données Sirene (**figure 2**).

Avant même de lancer sa troisième refonte de Sirene, l'Insee avait donc déjà fait évoluer fortement le répertoire : en isolant dans Sirius la partie du référentiel utile aux statisticiens, en séparant la gestion du répertoire administratif de sa diffusion, en modernisant la technologie de la mise à disposition de Sirius et de Sirene. Dans ces conditions, on est conduit à s'interroger sur ce qui a justifié le lancement d'une nouvelle refonte, le programme Sirene 4.

9. Voir la plaquette de présentation (SGMAP, 2013).

10. La demande a été formulée dans le cadre de la mise en place du Répertoire général des carrières unique.

SIRENE 4 A BÉNÉFICIE (INVOLONTAIREMENT) D'UNE PRÉPARATION APPROFONDIE

Sirene est une des opérations majeures de l'Insee, par sa finalité d'abord, mais aussi par l'ampleur des moyens que l'Insee consacre à sa gestion. L'institut revisite régulièrement les coûts engagés dans les opérations dont il a la charge : une revue concernant spécifiquement Sirene a débouché en 2013 sur des recommandations, dont certaines ont pu être mises en œuvre rapidement : rationalisation de l'accueil téléphonique, refonte de l'automate de traitement des déclarations, en vue de le rendre plus performant, etc.

Encadré 1. Cinq obligations d'un référentiel, parmi les règles préconisées par la Dinum

(Cadre commun d'architecture des référentiels de données de la direction interministérielle du Numérique)*

(Règle RS3) Tout référentiel de données dispose d'un **cadre de gouvernance à jour**, accessible librement à tous, et aligné ou en cours d'alignement :

- sur le présent corpus réglementaire interministériel (Cadre Commun d'Urbanisation, Cadre Commun d'Architecture des Référentiels de Données, Référentiel Général d'Interopérabilité),
- sur la stratégie du système d'information de l'État,
- sur les besoins opérationnels des utilisateurs des données de référence gérées,
- sur la réglementation métier correspondante en vigueur. [...]

(Règle RS5) **La qualité des données d'un référentiel est mesurée, suivie et publiée.** Une démarche d'amélioration continue est en place pour atteindre le niveau de qualité attendu par les utilisateurs directs pour la performance de leurs processus métiers.

(Règle RM4) Chaque référentiel doit intégrer un **dispositif d'alerte ou de signalement** permettant à un utilisateur du référentiel de faire remonter au responsable du référentiel toutes anomalies sur les données détectées en aval (incomplétude, incohérence, doublon, amalgame, problèmes d'intégrité, etc.). Le processus de traitement des signalements doit être également décrit, à jour et publié. Il est recommandé également de rechercher son automatisation et donc son outillage.

(Règle RA2) Les données de référence encapsulées dans un référentiel ne sont accessibles que par **des services documentés et catalogués**, disposant de SLA** partagés et alignés sur les besoins des consommateurs des données. Les services réalisent les fonctionnalités définies (RF7). **Le mode d'accès synchrone est recommandé** ; le mode asynchrone n'est pas interdit, mais doit être limité et encadré.

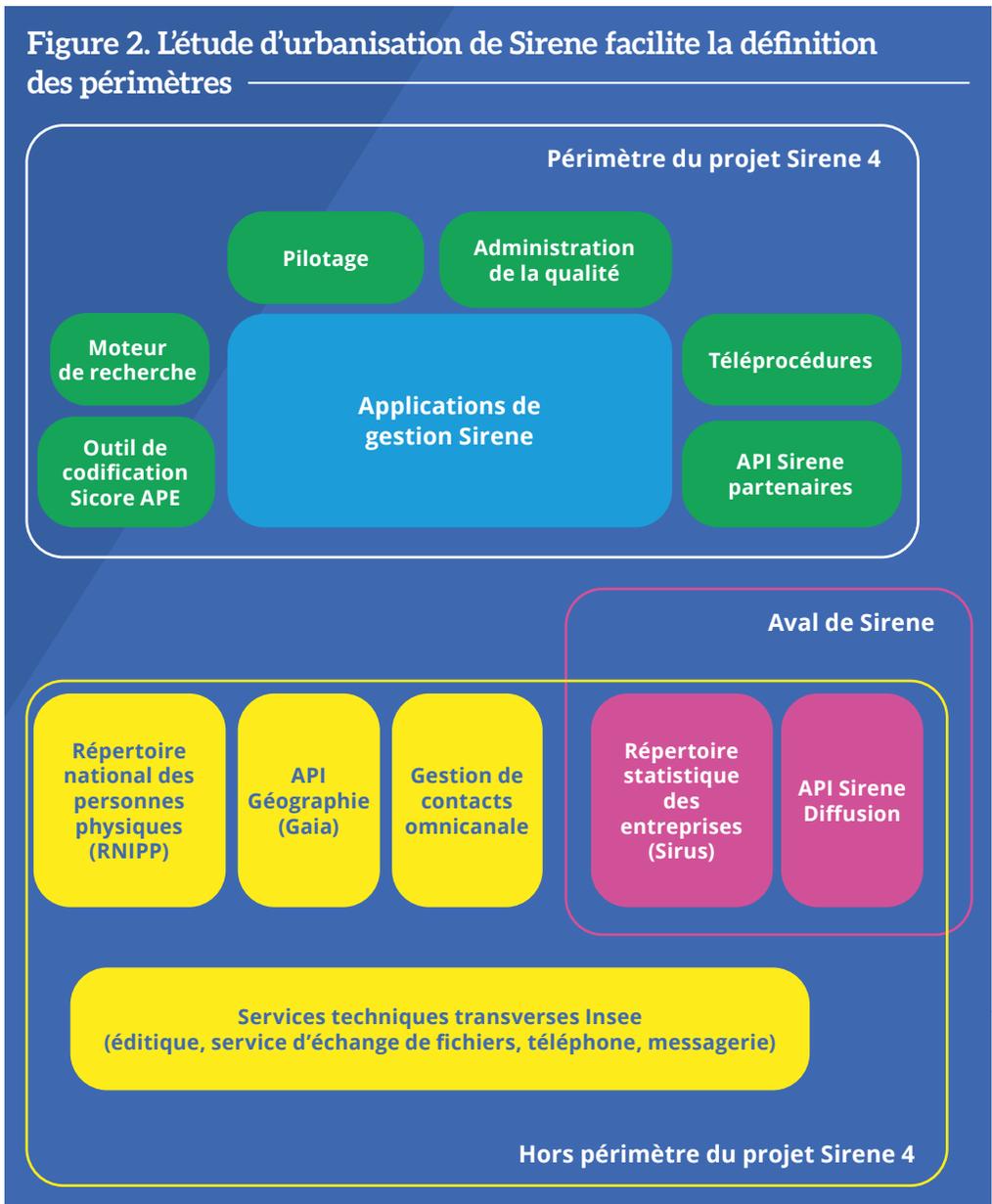
(Règle RA6) Toute action sur les données d'un référentiel doit être tracée et journalisée. **Les données sont historisées.** Les instances des objets métiers sont versionnées en fonction de leur cycle de vie. Les données ne sont pas détruites, mais marquées comme non actives. La politique de conservation et d'archivage des données est définie.

* (Disic, 2013) La direction interministérielle du Numérique (Dinum), par le décret du 25 octobre 2019, prend la suite de la direction interministérielle du Numérique et du système d'information et de communication de l'État (Dinsic, auparavant Disic).

** Le *Service Level Agreement*, ou SLA est un contrat par lequel un prestataire informatique s'engage à fournir un ensemble de services à un ou plusieurs clients.

Mais d'autres demandes nécessitaient une forte évolution applicative. Or, côté métier comme côté informatique, les équipes étaient largement mobilisées par un autre chantier, celui de la rationalisation de l'exploitation informatique de l'Insee, lequel s'est achevé fin 2017.

Ce délai, involontaire, a permis de conduire le projet d'API Sirene cité plus haut, mais également deux études qui se sont révélées importantes pour la préparation du programme Sirene 4.



- ❶ L'étude d'urbanisation du système d'information Sirene : menée sur l'ensemble des activités de gestion et de diffusion, l'étude a permis de formaliser les principaux objectifs de la refonte et les a traduits en architecture fonctionnelle, applicative (*figure 2*) et technique. Elle s'inscrit dans les objectifs d'urbanisation du système d'information de l'État (Disic, 2012 ; Disic, 2013 ; Weill, Robertson et Ross, 2006).
- ❷ Les nouveaux développements de la démarche Qualité Sirene, portant sur l'ensemble de la gestion et de la diffusion, sont venus compléter la démarche de 2005 (Rivière, 2005). Pour répondre aux préconisations de la Dinum en matière de tenue de référentiels (*encadré 1*), des objectifs chiffrés sont en cours de finalisation. Les indicateurs devraient à terme se focaliser sur la qualité de l'adresse, celle du code APE¹¹, ou de l'état administratif des unités référencées (fermé/actif). Ainsi, la mesure de la qualité pourra servir de contrepois à la recherche d'économie de moyens.

Ces travaux préparatoires ont bien évidemment intégré les objectifs déjà cités¹² du gouvernement de l'époque, en matière de simplification et de sécurisation de la vie des entreprises, avec notamment la décision du comité interministériel de la Modernisation de l'action publique du 18 décembre 2013 faisant du répertoire Sirene la référence en matière d'identité des entreprises.

Leurs résultats devront être adaptés pour mettre en œuvre les décisions plus récentes, notamment la loi pour la croissance et la transformation des entreprises (PACTE), très structurante pour les entreprises (*encadré 3*). Mais à très court terme, un autre plan d'actions gouvernemental a un impact déjà mesurable sur le programme Sirene 4.

Encadré 2. La diffusion de Sirene radicalement modernisée

- ❶ 2017 : la loi pour République numérique se traduit par la mise en *open data* des données sur <https://data.gouv.fr> ;
- ❷ 2018 : ouverture de l'API Sirene via <https://api.insee.fr> donnant accès à la totalité des données courantes et historiques du répertoire (22 millions d'unités légales, 31 millions d'établissements). Pour éviter la recopie de tout ou partie des données et des aléas dans la tenue à jour de ces copies, les API (interfaces de programmation applicative) permettent, avec quelques compétences de programmation, d'utiliser directement et quasi-immédiatement les données Sirene hébergées par l'Insee ;
- ❸ la Base Sirene des entreprises et de leurs établissements (SIREN, SIRET) est un des neuf **jeux de données de référence du Service Public de la Donnée** (<https://www.data.gouv.fr/fr/referance>), avec le répertoire national des associations, la base adresse nationale, le référentiel de l'organisation administrative de l'État, le plan cadastral informatisé, le registre parcellaire graphique, le référentiel à grande échelle, le répertoire opérationnel des métiers et des emplois, et enfin le code officiel géographique de l'Insee ;
- ❹ pour un usage simple ou épisodique, le site <https://sirene.fr> permet de faire des recherches et d'obtenir des listes suivant les principales caractéristiques des établissements.

11. Le code d'activité principale exercée (APE) est attribué par l'Insee à chaque unité légale et à chaque établissement de ces unités légales, en référence à la nomenclature d'activité française (NAF).

12. Dites-le nous une fois, État-plateforme, loi pour une République numérique (*encadré 2*).

1 CONTRIBUTUER À LA TRANSFORMATION NUMÉRIQUE DU SERVICE PUBLIC

L'État a lancé en avril 2019 TECH.GOUV, un programme pour accélérer la transformation numérique du service public. Sa feuille de route distingue plusieurs missions, notamment « Idnum » dont un des projets est le « *Développement et déploiement de ProConnect et de l'écosystème associé* ». ProConnect est le pendant pour les entreprises de FranceConnect pour les particuliers¹³ ; ce sera un dispositif d'authentification des personnes physiques agissant pour le compte des unités légales.

« ProConnect est le pendant pour les entreprises de FranceConnect pour les particuliers ; ce sera un dispositif d'authentification des personnes physiques agissant pour le compte des unités légales. »

Les premiers travaux sur ce projet 2019-2020 positionnent l'Insee, via le répertoire Sirene, comme un contributeur majeur. L'architecture fonctionnelle prévoit d'inclure les dirigeants des entreprises dans le répertoire (sans pour autant en faire une diffusion publique) et

d'y distinguer les habilitateurs principaux qui, au sein des entreprises, sont responsables de l'attribution des droits d'accès aux différents portails des services publics, comme www.impot.gouv.fr ou www.net-entreprises.fr. Pour les habilitateurs principaux, le répertoire Sirene contiendra leur courriel et éventuellement leur téléphone mobile, qui seront mis à disposition de ProConnect. Ces données permettront de plus à toute administration de contacter l'entreprise par internet, évitant ainsi un courrier papier au siège social.

Après les orientations politiques, intéressons-nous maintenant aux utilisateurs opérationnels du répertoire Sirene.

1 CERNER LES BESOINS DES UTILISATEURS QUI SERONT SATISFAITS

Dans la doctrine traditionnelle de conduite des projets, une bonne pratique consiste à synthétiser l'ensemble des besoins utilisateurs pour permettre la meilleure définition du projet.

Plusieurs opérations ont permis d'illustrer concrètement certaines attentes.

En 2014, avec l'aide du secrétariat général à la Modernisation de l'action publique (SGMAP), dans le cadre du programme « Dites-le-nous une fois », l'Insee a rencontré plusieurs administrations pour mieux comprendre leur usage des données de Sirene et leurs attentes.

13. FranceConnect <https://franceconnect.gouv.fr/> permet d'utiliser un compte des impôts ou de l'assurance maladie, par exemple, pour utiliser une téléprocédure sans devoir à chaque fois créer un compte spécifique. Lancé en 2014, il compte plus de 15 millions d'utilisateurs. FranceConnect relie une dizaine de fournisseurs d'identité et plus de 600 fournisseurs de services.

En voici trois illustrations :

- ① fournir les coordonnées géographiques (dites « xy ») des établissements : retenue dans Sirene 4, la géolocalisation aujourd'hui devient indissociable de l'adresse géographique ;
- ① identifier des locaux, des bâtiments ou des terrains et pas uniquement des établissements : la notion actuelle d'établissement convient pour les besoins de la sphère fiscale et de la sphère sociale ainsi que pour ceux de la statistique ; une extension du concept ou l'ajout d'un nouveau type d'unités nécessiteraient des moyens de gestion qui ne sont pas envisageables à l'Insee, en plus d'avoir un impact difficile à mesurer pour les déclarants et tous les utilisateurs des données ;
- ① étendre le champ de Sirene à toutes les associations (Blein, 2014) : source de simplification pour les associations et les utilisateurs des données des associations, cette extension ne pourra être réalisée que lorsque les formalités administratives pour les associations seront suffisamment automatisées pour que le surcoût de gestion pour l'Insee soit acceptable ; une première étape a été franchie avec le portail <https://lecompteasso.associations.gouv.fr/> ; il n'y a actuellement pas d'horizon temporel déterminé pour que Sirene contienne toutes les associations.

« Source de simplification pour les associations et les utilisateurs des données des associations, cette extension ne pourra être réalisée que lorsque les formalités administratives pour les associations seront suffisamment automatisées. »

Dans le cadre de *hackathons*¹⁴ organisé par le SGMAP¹⁵ en 2016 ou l'Insee en 2018, d'autres besoins ont été notés en plus de ceux recueillis par l'enquête de satisfaction des utilisateurs des données réalisée dans le cadre de la démarche qualité en 2017 (Insee, 2017 ; Insee, 2018). Trois exemples peuvent être présentés :

- ① l'amélioration de la qualité de l'information en matière de cessation d'activité : cette information n'arrive que très tardivement à l'Insee, quand elle arrive ; un garde-fou est mis en place par l'article 38 de la loi PACTE (« *Si le travailleur indépendant est entrepreneur individuel, la radiation [...] emporte de plein droit celle des fichiers, registres ou répertoires tenus par les autres administrations* »), mais ne suffit pas à épuiser ce sujet complexe ;
- ① l'amélioration de la qualité des adresses : ceci est bien retenu dans le cadre de Sirene 4, sous réserve que les moyens de gestion nécessaires pour la résorption des écarts éventuels avec la BAN, base adresse nationale, à la fois pour le rattrapage du stock et le traitement courant des déclarations soient compatibles avec ceux de l'Insee (Faure-Muntian, 2018) ; à défaut, l'amélioration sera lissée sur plusieurs années ;
- ① l'amélioration de la qualité des informations sur le secteur public : ce sujet est retenu dans Sirene 4, mais dépend également de la capacité des administrations partenaires à améliorer et moderniser leur contribution au répertoire Sirene.

Une fois passés en revue l'environnement du répertoire, le contexte politique du programme et certains besoins utilisateurs, il est temps d'aborder les objectifs de cette troisième refonte.

14. Le *hackathon* désigne le rassemblement d'informaticiens durant plusieurs jours et au minimum une nuit (souvent organisé sur un week-end) en vue de collaborer sur des sujets de programmation informatique pointus et innovants.
15. Le secrétariat général pour la Modernisation de l'action publique (SGMAP) accompagne le gouvernement pour transformer les administrations et moderniser les services publics. Il est rattaché au Premier ministre.

📍 LES TROIS OBJECTIFS PRINCIPAUX DU PROGRAMME SIRENE 4

Comme indiqué ci-dessus, tout usage des données en dehors de la sphère de gestion est renvoyé à la diffusion du répertoire Sirene (API Sirene <https://api.insee.fr>, <https://data.gouv.fr>, <https://sirene.fr>). Pour les utilisateurs des données, les seuls changements envisagés à ce stade sont d'augmenter le nombre de variables historisées, de détailler un peu plus les composantes de l'adresse et d'être plus strict sur les valeurs admissibles.

Les objectifs de la refonte, par ordre décroissant de priorité, sont les suivants :

- 📍 les gains en gestion : le travail quotidien des agents de l'Insee qui traitent les formalités n'ayant pu être traitées automatiquement ;
- 📍 les gains en administration du répertoire : le pilotage du répertoire dans son ensemble et des échanges avec les partenaires ;
- 📍 les gains pour les partenaires et les déclarants : le service rendu *in fine* par l'Insee.

📍 LES GAINS EN GESTION RELÈVENT MAJORITAIREMENT DE L'INSEE

L'Insee traite chaque jour 18 000 déclarations. Plus de 60 % sont traitées automatiquement. Le traitement manuel des déclarations est consacré à trois activités bien déterminées : l'identification de l'unité visée par la déclaration, la codification de l'activité principale de l'entreprise et le traitement d'anomalies, d'incohérences ou d'imprécisions dans la déclaration.

- 📍 La question de **l'identification** est un sujet d'étude depuis longtemps (Fellegi et Sunter, 1969 ; Winkler, 2006). Pour Sirene, des études techniques ont montré en 2018 que des gains substantiels étaient possibles en matière d'identification, d'une part avec l'utilisation d'outils modernes de recherche (en l'occurrence, le moteur Solr), d'autre part en améliorant la qualité des données, y compris les données historiques. L'amélioration des données peut être faite par confrontation avec d'autres sources (base adresse nationale, répertoire national d'identification des personnes physiques, registre national du commerce et des sociétés, par exemple) ou par des enquêtes d'amélioration du répertoire ciblées sur des questions précises.
- 📍 Il n'a pas été possible de mettre en évidence des gains accessibles en matière de **codification** de l'activité principale ; ce sujet nécessite encore des travaux méthodologiques pour profiter des avancées en matière d'apprentissage automatique (*machine learning*) ou d'autres techniques d'intelligence artificielle. Ces travaux ne pourront pas démarrer avant 2023, après la fin du programme Sirene 4.
- 📍 Enfin, des simplifications en matière de **traitements** seront obtenues en éliminant ceux qui sont repris par le répertoire statistique ou la diffusion du répertoire. Ceci nécessite de passer au peigne fin la totalité des traitements.

Pour des gains en gestion supplémentaires, l'Insee dépend de ses partenaires. Le premier sujet est la qualité des données entrantes ; le deuxième est la capacité de confronter

le répertoire Sirene aux autres registres et répertoire¹⁶ ; le troisième est la capacité à dématérialiser les relations avec les entreprises.

La qualité des données contenues dans les déclarations ou, pour le secteur public, fournies par les administrations, ne peut pas s'améliorer spontanément : avec des objectifs différents (la qualité des données du répertoire par opposition au traitement immédiat des déclarations) et des moyens en baisse dans toutes les organisations, il est nécessaire d'instaurer un dialogue régulier pour convenir petit à petit de contrôles plus stricts sur le contenu des déclarations, qui permettra la diminution du délai global de traitement. Le bouleversement des acteurs en matière de déclarations d'entreprises (par exemple les conséquences de la loi PACTE, voir **encadré 3**) peut être un accélérateur de cette évolution, ou un frein.

Deuxième sujet, la confrontation avec d'autres registres et répertoires est nécessaire pour rectifier certaines données divergentes (en général pour des raisons matérielles) et, au-delà, vérifier les procédures d'échanges et de traitements des données dans un environnement

Encadré 3. La loi PACTE rationalise l'organisation des partenaires de la gestion de Sirene

La loi pour la croissance et la transformation des entreprises (PACTE), publiée en mai 2019, prévoit des changements substantiels dans le traitement des déclarations des entreprises :

- mise en place d'un guichet unique électronique (Article 1^{er}) ;
- constitution d'un registre général des entreprises (Article 2), en maintenant distinct le répertoire Sirene ;
- dématérialisation à terme de toutes les déclarations des entreprises (seulement 20 % aujourd'hui) ;
- l'Inpi (Institut national de la propriété intellectuelle) unique opérateur pour gérer à la fois le guichet unique et le registre général.

Pour le programme Sirene 4, la mise en œuvre de cette loi pose quelques questions clés :

- quel positionnement du répertoire Sirene par rapport au guichet unique et au registre général ?
- pour satisfaire la demande sociale, quels engagements de délais de traitement des formalités par l'Insee (et ses partenaires) ?
- l'Insee doit-il, en plus de l'activité principale*, fournir pour chaque entreprise tout ou partie de ses activités secondaires éventuelles et sous quelle forme ?

En termes d'échéance, le démarrage du guichet unique est envisagé à compter de janvier 2021, et la mise en œuvre du registre général des entreprises pour une échéance plus lointaine, estimée à fin 2022.

* Outre sa fonction d'identification des entreprises, le répertoire Sirene confie à l'Insee le soin de coder une activité principale exercée (APE), dans la Nomenclature d'activité française (NAF). L'attribution de ce code répond à des finalités exclusivement statistiques, même si le code APE se révèle à l'usage structurant dans la mise en œuvre de réglementations pour les entreprises et leurs salariés.

16. Registre du commerce et des sociétés, répertoire des métiers, registre des actifs agricoles, notamment.

administratif complexe. Les travaux préparatoires à la mise en place du registre général des entreprises pourraient être l'occasion d'améliorer la qualité d'ensemble, si tous disposent des moyens d'instruire et de corriger les écarts.

Enfin, plus classiquement, certaines procédures spécifiques de l'Insee, comme la demande de modification de l'activité principale, pourraient être dématérialisées avec ProConnect, comme la demande de modification du statut de diffusion <https://statut-diffusion-sirene.insee.fr/> l'a été avec FranceConnect. Ceci se fera progressivement, en fonction du calendrier de migration des entreprises à ProConnect.

L'objectif le plus important pour l'Insee, donc indirectement pour les partenaires et les entreprises, est de faire des gains en gestion. Cependant, les évolutions de l'organisation des partenaires et du contenu des formalités sont récurrentes, et parfois urgentes : il faut également être capable de les traiter de façon plus efficace.

L'INSEE A UN RÔLE DE COORDINATION DES CFE

La commission de coordination des centres de formalités des entreprises (CCCCE)¹⁷ élabore et fait évoluer la norme fonctionnelle d'échanges automatisés d'information des CFE (EDI CFE) ; cette norme constitue la traduction en langage informatique des renseignements recueillis par les CFE et transmis à leurs divers partenaires. L'Insee préside cette commission, dont

le secrétariat est assuré par la direction générale des Entreprises, et gère cette norme d'échanges EDI CFE avec l'agence centrale des organismes de Sécurité sociale (ACOSS)¹⁸ à travers le groupe Norme de la CCCCE.

« Les gains en administration du répertoire seront apportés principalement par la mise en place d'un outil de gestion de la norme : il permettra de vérifier la cohérence interne de la norme et d'en gérer les différentes versions. »

Les gains en administration du répertoire seront apportés principalement par la mise en place d'un outil de gestion : il permettra de vérifier la cohérence interne de la norme et d'en gérer les différentes versions ; cela concerne également la partie de la norme qui s'applique aux échanges de l'Insee avec les

destinataires. Ce sera l'occasion de mettre en place des bonnes pratiques après concertation avec les parties prenantes.

En ce qui concerne le rôle de tenue du répertoire, l'objectif du programme Sirene 4 est de gérer plus facilement les évolutions en mettant en place un **moteur de règles**, ce qui, en complément de la gestion de la norme, permettra de standardiser les traitements et, ainsi, de faciliter leurs évolutions.

Enfin, il sera nécessaire de mettre en œuvre des études d'impact : pour ce faire, il faudra développer la capacité à faire tout ou partie des traitements en double, afin de mesurer la qualité des processus. De même, des environnements de simulation seront nécessaires pour étudier l'impact des changements sur les traitements et le poste de gestion.

17. <https://www.entreprises.gouv.fr/politique-et-enjeux/la-commission-coordination-des-centres-formalites-des-entreprises-cccce>.

18. https://www.cfe.urssaf.fr/static_cfe/NormesEDI-CFE/index.htm.

Sirene 4 devrait ainsi permettre des gains réguliers en efficacité, seule réponse possible à l'accroissement annuel de la charge de travail, qui peut être assez fort : on l'a constaté récemment lors des changements juridiques concernant les micro-entrepreneurs, qui ont eu des conséquences importantes sur le volume des traitements dans Sirene.

Le recours aux techniques d'automatisation du développement et de l'exploitation informatique (DevOps) va également dans ce sens (Doerrfeld, van der Mersch, Sandoval et Wood, 2016). Outre la nouvelle application de gestion, c'est une véritable usine de développement logiciel qui est mise en place et servira pour les maintenances ultérieures. Ce sont aussi les nouvelles architectures, sur la base de services indépendants, qui permettront de gérer plus facilement les évolutions.

🕒 LES GAINS EN DÉLAI : UNE QUESTION D'ARCHITECTURE D'ENSEMBLE

La loi PACTE prévoit un guichet unique, mais derrière le guichet, que va-t-il se passer ? Le fonctionnement actuel des échanges entre les partenaires repose sur des traitements par lots en temps différé. Un fonctionnement typique est d'envoyer le lot des déclarations de la journée J à l'Insee le soir ; l'Insee l'intègre le lendemain et renvoie le soir même ce qui a été traité dans la journée, automatiquement ou par les gestionnaires ; le destinataire le reçoit donc à J+2 alors que le traitement pourrait être instantané pour les déclarations qui ne requièrent pas de traitement manuel.

En plus du traitement par lots, le programme Sirene 4 prévoit de permettre le traitement au fil de l'eau : le partenaire envoie une déclaration par service web, l'Insee renvoie immédiatement le résultat du traitement s'il peut être traité automatiquement. Même pour les déclarations qui doivent être reprises par un gestionnaire, tout le monde gagne deux jours de délai par suppression des temps morts.

La mise à disposition des déclarants de leur avis d'immatriculation au répertoire Sirene par support électronique permettra également de gagner le délai d'acheminement du courrier papier, dès que les entreprises auront des données de contact électronique sûres.

D'autres gains pour les partenaires et les déclarants seront également possibles *via* un bouquet de services fourni par l'API Sirene (pour le pré-remplissage de formulaires de modification), le répertoire national d'identification des personnes physiques pour la vérification de l'identité ainsi que par des API d'assistance à la saisie de données géographiques, par l'institut national de l'information géographique et forestière (IGN).

Après avoir décrit les trois objectifs de Sirene 4, quelle trajectoire pour les atteindre ?

🕒 LA SUPERPOSITION DES CALENDRIERS ACCROÎT LES INCERTITUDES

On a vu que certains objectifs de Sirene 4 ne dépendaient que des moyens de l'Insee tandis que d'autres nécessitaient la coopération d'acteurs externes qui ont leurs propres projets et leurs propres priorités. De plus, les missions de certains de ces acteurs vont être redéfinies par la mise en œuvre de la loi PACTE. Le biais de l'approche projet nous fait croire, ou nous fait espérer, que tous les changements interviendront dans le bon calendrier, celui du programme Sirene 4.

En fait, on peut distinguer plusieurs calendriers :

- ❶ la convention avec la direction du Budget, qui nous demande d'obtenir des gains dès 2021 et de réaliser l'essentiel du programme pour 2022 ;
- ❶ la mise en place de ProConnect, 2020-2021 ;
- ❶ le guichet unique électronique et la suppression des centres de formalités des entreprises, 2021-2022 ;
- ❶ le registre général des entreprises, 2022 ;
- ❶ les simplifications administratives, notamment la directive sur l'utilisation d'outils et de processus numériques en droit des sociétés, qui doit être transposée en 2021.

Il n'est pas possible de connaître le détail de tous ces calendriers, et de leurs répercussions sur nos partenaires, avant d'organiser le programme Sirene 4. La conduite du programme devra s'adapter aux circonstances. Les méthodes actuelles de développement prônent l'agilité. Toutefois, l'agilité pour les grands projets n'est pas encore aussi bien établie.

❶ LE FONCTIONNEMENT AGILE DOIT ÊTRE ADAPTÉ À LA TAILLE DU PROGRAMME

Face aux incertitudes, la méthode agile (Schwaber et Sutherland, 2017 ; Reinertsen, 2009 ; Sutherland et Sutherland, 2014) suivie par la très grande majorité des projets de l'Insee permettra plus facilement de s'adapter aux circonstances ainsi qu'aux imprévus techniques (**encadré 4**). Les méthodes agiles privilégient l'adaptation aux changements de l'environnement, des besoins ou des technologies au suivi d'un plan « immuable ».

Toutefois, la taille de l'équipe (12 personnes en septembre 2019, une vingtaine en septembre 2020) ne permet pas d'utiliser telle quelle la méthode Scrum, bien adaptée à une taille d'équipe de trois à neuf membres. Une adaptation pragmatique sera nécessaire en s'appuyant sur différents éléments des méthodes de passage de l'agilité à l'échelle, comme Less (Larman et Vodde, 2016), Safe (Leffingwell, 2018) ou Scrum@Scale (<https://www.scrumatscale.com/>).

Encadré 4. Sirene 4 utilise Scrum, la méthode agile la plus répandue

Scrum est un cadre de travail pour le développement et la maintenance de produits complexes (Schwaber et Sutherland, 2017). Plutôt que de suivre un plan détaillé sur quatre ans, Scrum utilise une approche itérative et incrémentale pour optimiser la prédictibilité et pour contrôler le risque.

Trois piliers soutiennent la méthode : la transparence, l'inspection et l'adaptation.

Quatre types d'« événements Scrum » sont mis en œuvre : l'itération (3 semaines pour Sirene 4), la mêlée quotidienne (15'), la revue d'itération et la rétrospective (1 journée).

Pendant la réunion de **revue d'itération**, l'équipe Scrum et les parties prenantes (clients, encadrement, utilisateurs ou leurs représentants) échangent sur ce qui a été fait durant l'itération. Les participants collaborent pour déterminer les prochains items ayant le plus de valeur qui pourraient être réalisés.

La **rétrospective** est une occasion pour l'équipe d'inspecter son fonctionnement et de créer un plan d'amélioration qui sera mis en place au cours de l'itération suivante.

Les **travaux sont priorisés par la valeur** qu'ils apportent aux utilisateurs en produisant à chaque itération un résultat démontrable.

Ainsi, le programme pourrait être configuré en sous-équipes, le temps de quelques itérations, ou mettre en place une revue et une planification trimestrielle, pour mieux coordonner les projets entre eux.

En plus de l'association des utilisateurs internes aux démonstrations de l'avancement des travaux, l'Insee a mis au point une méthode originale de prise en compte de leurs besoins.

🕒 L'APPLICATION SIRENE ADAPTÉE AUX GESTIONNAIRES DE L'INSEE

Pour les projets ayant un impact organisationnel potentiellement fort, l'Insee a recours à une méthode de détermination des rôles et des compétences (**encadré 5**). Les travaux préparatoires au programme Sirene 4 se sont largement concentrés sur les gains de productivité demandés, les besoins des gestionnaires, des experts et des responsables de site ou de pôle de compétence doivent être cartographiés.

Cette démarche en organisation du travail ne permet pas de faire l'économie des travaux d'ergonomie et d'expérience utilisateur, pour aborder plus en détail tel ou tel aspect de l'application.

Évidemment, un aspect important du changement pour les utilisateurs est de savoir quand et comment il aura lieu.

🕒 LA STRATÉGIE DE MIGRATION ET DE TRANSITION DOIT ÊTRE FINALISÉE

En ce qui concerne le premier objectif du programme, des gains en gestion, la stratégie de mise en place a été définie : des premiers changements seront réalisés en 2020 pour des gains, toutes choses égales par ailleurs, en année pleine dès 2021. Les autres gains mis en place en 2022 pour un effet en 2023.

Encadré 5. Quand Sirene inspire un Maiol

Initialisée en 1998, la **maîtrise d'œuvre locale en organisation du travail** renvoie aux utilisateurs la responsabilité de concevoir l'organisation du travail, lorsqu'un projet impacte leur processus de production.

Maiol est une démarche « sur mesure », adaptée aux spécificités de l'Insee, qui s'insère dans le déroulement des projets, en harmonie avec la méthode de conduite de projet en vigueur. Elle doit permettre un atterrissage plus satisfaisant des applications dans les unités concernées, et garantit la cohérence avec la politique générale de l'Insee en matière d'organisation du travail dans les directions régionales.

La coopération des trois équipes statistique, informatique et organisation du travail responsabilise les acteurs, leur permet de confronter leurs points de vue et de choisir de nouvelles organisations performantes, motivantes et adaptées. L'évaluation des tâches à accomplir et des objectifs à atteindre contribue à créer les conditions de l'efficacité tant individuelle que collective.

Assez classiquement, un Maiol s'organise en trois phases : l'analyse de l'existant, l'élaboration des organisations cibles et l'accompagnement du changement (formations, communication, définition des fiches de poste, reconversion éventuelle).

La démarche est mise en œuvre pour Sirene 4 (de même qu'elle l'a été en 2002-2004 pour Sirene 3).

Pour le reste du programme, des services périphériques seront disponibles régulièrement chaque année à partir de 2020, mais pour l'essentiel de la transformation du système d'information, et notamment le travail des gestionnaires, la stratégie n'est pas arrêtée sachant qu'en aucun cas il n'est possible de suspendre la gestion pour changer d'outil.

Un travail d'analyse conjoint Sirene 3 – Sirene 4 doit être réalisé pour la mi-2020 pour examiner la façon dont certains changements pourraient être mis en place progressivement et limiter, autant que faire se peut, les risques liés à une approche en *big bang*.

📍 AU-DELÀ DE SIRENE 4, UNE DYNAMIQUE DE CHANGEMENTS RÉGULIERS

Pour ce qui est de la question posée dans le titre, un programme trop ambitieux ou pas assez, le lecteur jugera.

Au-delà de Sirene 4, plusieurs chantiers apparaissent déjà : la modernisation de la codification de l'activité principale, par exemple avec des techniques d'intelligence artificielle, éventuellement la mise à disposition des activités secondaires des entreprises, sans compter la poursuite régulière de l'amélioration de l'identification, l'impact de la dématérialisation des formalités des entreprises.

BIBLIOGRAPHIE

BÉGUIN, Jean-Marc et HAAG, Olivier, 2017. *Méthodologie de la statistique annuelle d'entreprises. Description du système « É sane »*. [en ligne]. Octobre 2017. Insee Méthodes n°130. [Consulté le 26 mai 2020]. Disponible à l'adresse : <https://www.insee.fr/fr/information/3056089>.

BIEHL Matthias, 2016. *RESTful API Design. APIs your consumers will love*. [en ligne]. Août 2016. API-University Press, série API-University, Vol. 3. [Consulté le 26 mai 2020]. Disponible à l'adresse : <https://api-university.com/wp-content/uploads/2015/05/api-design-toc.pdf>.

BIZINGRE Joël, PAUMIER, Joseph et RIVIÈRE, Pascal, 2013. *Les référentiels du système d'information : Données de référence et architectures d'entreprise*. 3 juillet 2013. Édition Dunod. ISBN : 978-2-10-060124-0.

BLEIN, Yves, 2014. *Simplifications pour les associations*. [en ligne]. 9 octobre 2014. Rapport au Premier ministre. [Consulté le 26 mai 2020]. Disponible à l'adresse : <https://www.vie-publique.fr/var/storage/rapports-publics/144000663.pdf>.

COLIN, Nicolas et VERDIER, Henri, 2015. *L'âge de la multitude : Entreprendre et gouverner après la révolution numérique*. 13 mai 2015. Édition Armand Colin. 2^e édition. ISBN : 978-2-200-60343-4.

DIRECTION INTERMINISTÉRIELLE DU NUMÉRIQUE ET DU SYSTÈME D'INFORMATION ET DE COMMUNICATION (DINSIC), 2016. *Plan d'Occupation des Sols du Système d'Information des Services Publics Français*. [en ligne]. 25 juillet 2016. Version 3.1.1. [Consulté le 26 mai 2020]. Disponible à l'adresse : <http://references.modernisation.gouv.fr/sites/default/files/Plan%20d%27Occupation%20des%20Sols%20du%20SI%20des%20Services%20Publics%20Fran%3%A7ais%20V3.1.1.pdf>.

DIRECTION INTERMINISTÉRIELLE DES SYSTÈMES D'INFORMATION ET DE COMMUNICATION (DISIC), 2012. *Cadre Commun d'Urbanisation du Système d'Information de l'État*. [en ligne]. 26 octobre 2012. Version 1.0. [Consulté le 26 mai 2020]. Disponible à l'adresse : http://references.modernisation.gouv.fr/sites/default/files/Cadre%20Commun%20d%27Urbanisation%20du%20SI%20de%20l%27Etat%20v1.0_0.pdf.

DIRECTION INTERMINISTÉRIELLE DES SYSTÈMES D'INFORMATION ET DE COMMUNICATION (DISIC), 2013. *Cadre commun d'architecture des référentiels de données. Complément n°2 au Cadre Commun d'Urbanisation du Système d'Information de l'État version 1.0*. [en ligne]. 18 décembre 2013. Version 1.0. [Consulté le 26 mai 2020]. Disponible à l'adresse : http://references.modernisation.gouv.fr/sites/default/files/Cadre%20Commun%20d%27Architecture%20des%20R%3%A9f%3%A9rentiel%20de%20donn%3%A9%20v1.0_0.pdf.

DOERRFELD, Bill, VAN DER MERSCH, Vassili, SANDOVAL, Kristopher et WOOD, Chris, 2016. *API driven DevOps : Strategies for Continuous Deployment*. 6 juin 2016. Nordic APIs, Lean publication.

FAURE-MUNTIAN, Valéria, 2018. *Les données géographiques souveraines*. [en ligne]. Juillet 2018. Rapport au Gouvernement. [Consulté le 26 mai 2020]. Disponible à l'adresse : https://www.ecologique-solidaire.gouv.fr/sites/default/files/Rapport_DonneesGeographiquesSouveraines.pdf.

- FELLEGI, Ivan et SUNTER, Alan, 1969. A Theory for Record Linkage. In : *The Journal of the American Statistical Association*. [en ligne]. Décembre 1969. Vol. 64, n°328, pp. 1183-1210. [Consulté le 26 mai 2020]. Disponible à l'adresse : <https://courses.cs.washington.edu/courses/cse590q/04au/papers/Felligi69.pdf>.
- HAAG, Olivier, 2019. Le profilage à l'Insee : une identification plus pertinente des acteurs économiques. In : *Courrier des statistiques* [en ligne]. 27 juin 2019. N°N2, pp. 86-100. [Consulté le 26 mai 2020]. Disponible à l'adresse : <https://www.insee.fr/fr/statistiques/fichier/4168409/courstat-2-9.pdf>.
- INSEE, 2017. *Qualité du service rendu par les sites de gestion Sirene. Résultats de l'enquête menée en 2016*. [en ligne]. 17 août 2017. [Consulté le 26 mai 2020]. Disponible à l'adresse : <https://www.insee.fr/fr/information/2965145>.
- INSEE, 2018. *Enquête Qualité Sirene. Résultats de l'enquête menée en 2017*. [en ligne]. 20 décembre 2018. [Consulté le 26 mai 2020]. Disponible à l'adresse : <https://www.insee.fr/fr/information/3645526>.
- LARMAN, Craig et VODDE, Bas, 2016. *Large-Scale Scrum : More with LeSS*. [en ligne]. Août 2016. Édition Addison-Wesley. [Consulté le 26 mai 2020]. Disponible à l'adresse : <http://ptgmedia.pearsoncmg.com/images/9780321985712/samplepages/9780321985712.pdf>.
- LÉTOQUART, Philippe et REFFET, Jean-Michel, 1969. Le projet « SIRENE » vise à unifier les procédures d'immatriculation des entreprises. In : *Économie et Statistique*. [en ligne]. Novembre 1969. N°6, pp. 67-68. [Consulté le 26 mai 2020]. Disponible à l'adresse : https://www.epsilon.insee.fr/jspui/bitstream/1/123269/1/estat_1969_6.pdf.
- LEFFINGWELL, Dean, 2018. *SAFe 4.5 Reference Guide : Scaled Agile Framework for Lean Enterprises*. Mai 2018. Édition Addison Wesley-Professional. 2^e édition. ISBN : 978-0-13-489286-3.
- PICARD, Hugues et CHAZELAS, Nicole, 2003. Identification interadministrative des entreprises et des établissements. De nouvelles avancées avec Sirene 3. In : *Courrier des Statistiques*. [en ligne]. Juin 2003. N°105-106, pp. 13-17. [Consulté le 26 mai 2020]. Disponible à l'adresse : <https://www.epsilon.insee.fr/jspui/bitstream/1/122507/1/cs105-106.pdf>.
- REINERTSEN, Donald G., 2009. *The Principles of Product Development Flow : Second Generation Lean Product Development*. Mai 2009. Édition Celeritas. ISBN : 9781935401001.
- RIVIÈRE, Pascal, 2005. Indicateurs de qualité en matière de production de données : quelques éléments de réflexion. In : *Courrier des statistiques*. [en ligne]. Septembre 2005. N°115, pp. 34-40. [Consulté le 26 mai 2020]. Disponible à l'adresse : <https://www.epsilon.insee.fr/jspui/bitstream/1/122525/1/cs115.pdf>.
- SECRÉTARIAT GÉNÉRAL POUR LA MODERNISATION DE L'ACTION PUBLIQUE (SGMAP), 2013. « *Dites-le nous une fois* » : un programme pour simplifier la vie des entreprises. [en ligne]. 26 septembre 2013. Site internet. [Consulté le 26 mai 2020]. Disponible à l'adresse : https://www.modernisation.gouv.fr/sites/default/files/fichiers-attaches/sgmap_plaquette_dlnuf.pdf.
- SECRÉTARIAT GÉNÉRAL POUR LA MODERNISATION DE L'ACTION PUBLIQUE (SGMAP), 2020. *État Plateforme. Concevoir autrement les services publics numériques*. [en ligne]. Site internet. [Consulté le 26 mai 2020]. Disponible à l'adresse : <http://etatplateforme.modernisation.gouv.fr/>.

SCHWABER, Ken et SUTHERLAND, Jeff, 2017. *Le guide SCRUM™ – Le Guide de Référence de SCRUM : Les Règles de Jeu*. [en ligne]. Novembre 2017. [Consulté le 26 mai 2020]. Disponible à l'adresse : <https://www.scrumguides.org/docs/scrumguide/v2017/2017-Scrum-Guide-French.pdf>.

SUTHERLAND, Jeff et SUTHERLAND, J. J., 2014. *Scrum : The Art of Doing Twice the Work in Half the Time*. 30 septembre 2014. édition Penguin Random House. ISBN : 978-0-385-34645-0.

WEILL, Peter, ROBERTSON, David C. et ROSS Jeanne W., 2006. *Enterprise Architecture As Strategy. Creating a Foundation for Business Execution*. [en ligne]. Mai 2006. Harvard Business School Press. [Consulté le 26 mai 2020]. Disponible à l'adresse : https://www.researchgate.net/profile/Peter_Weill/publication/236972734_Enterprise_Architecture_as_Strategy_-_Creating_a_Foundation_for_Business_Execution/links/00b7d5335b0b48cccb000000/Enterprise-Architecture-as-Strategy-Creating-a-Foundation-for-Business-Execution.pdf.

WINKLER, William E, 2006. *Overview of Record Linkage and Current Research Directions*. [en ligne]. 8 février 2006. Research Report Series, Statistical Research Division, U.S. Census Bureau. [Consulté le 26 mai 2020]. Disponible à l'adresse : <https://www.census.gov/srd/papers/pdf/rrs2006-02.pdf>.

● FONDEMENTS JURIDIQUES

Code de commerce. In : *Légifrance* [en ligne]. [Consulté le 26 mai 2020]. Disponible à l'adresse : <https://www.legifrance.gouv.fr/affichCode.do?cidTexte=LEGITEXT000005634379>.

Code des relations entre le public et l'administration. In : *Légifrance* [en ligne]. [Consulté le 26 mai 2020]. Disponible à l'adresse : <https://www.legifrance.gouv.fr/affichCode.do?cidTexte=LEGITEXT000031366350>.

Décret n°2019-31 du 18 janvier 2019 relatif aux échanges d'informations et de données entre administrations dans le cadre des démarches administratives. In : *Légifrance* [en ligne]. [Consulté le 26 mai 2020]. Disponible à l'adresse : <https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000038029589&categorieLien=id>.

Directive (UE) n°2019/1151 du Parlement européen et du Conseil du 20 juin 2019 modifiant la directive (UE) 2017/1132 en ce qui concerne l'utilisation d'outils et de processus numériques en droit des sociétés. In : *Journal officiel de l'Union européenne*. [en ligne]. [Consulté le 26 mai 2020]. Disponible à l'adresse : <https://eur-lex.europa.eu/legal-content/FR/TXT/PDF/?uri=CELEX:32019L1151>.

Loi n°2014-1545 du 20 décembre 2014 relative à la simplification de la vie des entreprises et portant diverses dispositions de simplification et de clarification du droit et des procédures administratives. In : *Légifrance* [en ligne]. [Consulté le 26 mai 2020]. Disponible à l'adresse : <https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000029926655&categorieLien=id>.

Loi n°2015-990 du 6 août 2015 pour la croissance, l'activité et l'égalité des chances économiques. In : *Légifrance* [en ligne]. [Consulté le 26 mai 2020]. Disponible à l'adresse : <https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000030978561&categorieLien=id>.

Loi n°2015-1779 du 28 décembre 2015 relative à la gratuité et aux modalités de la réutilisation des informations du secteur public. In : *Légifrance* [en ligne]. [Consulté le 26 mai 2020]. Disponible à l'adresse : <https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000031701525&categorieLien=id>.

Loi n°2016-1321 du 7 octobre 2016 pour une République numérique. In : *Légifrance* [en ligne]. [Consulté le 26 mai 2020]. Disponible à l'adresse : <https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000033202746&categorieLien=id>.

Décret n°2017-331 du 14 mars 2017 relatif au service public de mise à disposition des données de référence. In : *Légifrance* [en ligne]. [Consulté le 26 mai 2020]. Disponible à l'adresse : <https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000034194946&categorieLien=id>.

Loi n°2018-727 du 10 août 2018 pour un État au service d'une société de confiance. In : *Légifrance* [en ligne]. [Consulté le 26 mai 2020]. Disponible à l'adresse : <https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000037307624&categorieLien=id>.

Loi n°2019-486 du 22 mai 2019 relative à la croissance et la transformation des entreprises. In : *Légifrance* [en ligne]. [Consulté le 26 mai 2020]. Disponible à l'adresse : <https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000038496102&categorieLien=id>.

Règlement (UE) n°910/2014 du parlement européen et du conseil du 23 juillet 2014 sur l'identification électronique et les services de confiance pour les transactions électroniques au sein du marché intérieur. In : *Journal officiel de l'Union européenne*. [en ligne]. [Consulté le 26 mai 2020]. Disponible à l'adresse : <https://eur-lex.europa.eu/legal-content/FR/TXT/PDF/?uri=CELEX:32014R0910&from=FR>.

Règlement (UE) n°016/679 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 : *Journal officiel de l'Union européenne*. [en ligne]. [Consulté le 26 mai 2020]. Disponible à l'adresse : <https://eur-lex.europa.eu/legal-content/FR/TXT/PDF/?uri=CELEX:32016R0679&from=FR>.

LE MODÈLE DE PROCESSUS STATISTIQUE EN SUÈDE

MISE EN ŒUVRE, EXPÉRIENCES ET ENSEIGNEMENTS

Johan Erikson*

L'institut national statistique suédois (SCB) a développé son modèle de processus statistique en 2007. Comme le GSBPM, modèle générique de description des processus de production statistique diffusé à grande échelle en 2009, il se fonde sur les travaux de Statistics New Zealand. Les deux modèles ont été développés en parallèle : malgré certaines différences, ils sont très proches l'un de l'autre. Des normes, des lignes directrices, des modèles et des listes de contrôle, ainsi que des outils informatiques ont été décrits et diffusés au sein de l'institut suédois, par le biais d'une infrastructure servant de support à la démarche. On a ainsi posé les bases d'une approche de la production statistique standardisée et « orientée processus ».

À l'œuvre depuis 2008, la démarche a radicalement changé l'organisation du SCB. En outre, depuis 2014, l'institut est certifié selon la norme internationale ISO 20252, dont tous les critères ont été incorporés à l'infrastructure associée au modèle. L'ensemble est maintenant largement implanté au sein du SCB, non sans avoir dû surmonter quelques difficultés : l'acceptation par les parties prenantes, l'identification des informations pertinentes et le maintien à jour des éléments constitutifs. Utilisé depuis plus de dix ans, le dispositif pourra se développer, par exemple en créant une version adaptée à chaque produit statistique, ou en étendant son usage à d'autres autorités responsables de la statistique publique en Suède.

 *The Swedish Process Model was developed in 2007. Like the Generic Statistical Business Process Model (GSBPM, released widely in 2009) it was based on work done by Statistics New Zealand. There are some differences between the models since they were developed in parallel, but they are very closely aligned. Standards, guidelines, templates and checklists as well as IT tools are described and disseminated through the Process Support System. The whole formed the basis for a standardised and process oriented approach to statistical production.*

That approach has been used by Statistics Sweden since 2008, affecting widely its organisation. Since 2014, the Office has been also certified according to the international standard ISO 20252. All the criteria from this standard have been included in the Process Support System. The Process Model and the Process Support System has reached a broad use in the organisation. That does not mean there haven't been challenges along the way, like acceptance, finding the relevant information and keeping everything up to date. While the systems have now been used for over ten years, there are still possibilities for further development. These include creating a more tailored view for each statistical product, as well as spreading the use of the Process Model and the Process Support System to other authorities responsible for official statistics in Sweden.

* Propriétaire de processus, Département de développement des processus et des méthodes, Statistiska centralbyrån (SCB), johan.erikson@scb.se

Cet article décrit la mise en œuvre du modèle de processus suédois et de l'infrastructure qui supporte le modèle (*Process Support System*), dans une démarche d'« orientation processus » de la production statistique (*Process Oriented Production*). Nous décrirons en premier lieu le modèle suédois, ses similitudes et ses légères différences avec le GSBPM, modèle générique international pour les processus statistiques. Nous présenterons ensuite brièvement le système de la statistique publique en Suède : en nous remettant dans le contexte de la création et de la mise en œuvre du modèle et de l'infrastructure associée, nous identifierons les facteurs de réussite qui ont abouti à la généralisation de leur utilisation au sein de l'institut national statistique suédois *Statistiska centralbyrån* (SCB). Nous étudierons ensuite les changements intervenus depuis que les dispositifs ont été créés, la conformité aux systèmes et certains des principaux défis auxquels nous avons été confrontés au fil du temps. Pour finir, nous présenterons les principales idées quant à l'évolution future.

LE GSBPM,...

Le modèle générique de description des processus de production statistique (GSBPM pour *Generic Statistical Business Process Model*) décrit les différentes étapes à suivre pour produire des statistiques publiques. Développé en 2008 par le groupe de travail commun Unece/Eurostat/OCDE sur les métadonnées statistiques (METIS¹), il se fondait initialement sur le modèle conçu par *Statistics New Zealand*. La première version utilisée à grande échelle a été publiée en 2009 (version 4.0) et largement adoptée par la communauté mondiale de la statistique publique.

Le GSBPM structure le processus de production statistique en le divisant en plusieurs phases et sous-processus, avec à chaque fois une description des éléments inclus dans le processus ou le sous-processus et des travaux effectués en son sein (*figure 1*). Le GSBPM est un modèle de référence. Il peut donc, et est conçu pour, être utilisé de différentes façons par différents services, selon ce qui leur semble approprié.

Il est lié à d'autres modèles développés afin de décrire les travaux des services statistiques, à savoir :

- ① le « modèle générique d'activité des organismes statistiques » (GAMSO pour *Generic Activity Model for Statistical Organizations*), qui décrit les activités de haut niveau des services statistiques ;
- ① et le « modèle générique d'informations statistiques » (GSIM pour *Generic Statistical Information Model*), un cadre de référence pour les informations statistiques.

Le GAMSO et le GSIM ont tous les deux été développés après le GSBPM et s'inscrivent dans sa continuité.

Le modèle GSBPM complet est accessible sur le site de l'Unece (Unece, 2019). Une bonne description de sa première version, ainsi que de la façon dont il a été développé, est disponible dans (Vale, 2009).

1. METIS (*METadata Information System*) a réuni l'Unece (*United Nations Economic Commission for Europe* ou Commission économique pour l'Europe des Nations unies), Eurostat et l'OCDE (Organisation de coopération et de développement économiques), cf. <https://stats.oecd.org/glossary/detail.asp?ID=5494>.

Figure 1. Niveaux 1 et 2 du modèle générique de description des processus de



Source : version 5.1 sur le site de l'Unece (<https://statswiki.unece.org/display/GSBPM/GSBPM+v5.1>).

Gérer les métadonnées

Traiter les données

5.1
Intégrer les données

5.2
Classer et coder les données

5.3
Examiner et valider les données

5.4
Redresser et imputer les données

5.5
Dériver de nouvelles variables et unités statistiques

5.6
Calculer les pondérations

5.7
Calculer les agrégats

5.8
Finaliser les fichiers de données

Analyser

6.1
Préparer la version préliminaire des résultats

6.2
Valider les résultats

6.3
Interpréter et expliquer les résultats

6.4
Appliquer les contrôles de confidentialité

6.5
Finaliser les résultats

Diffuser

7.1
Mettre à jour les systèmes de diffusion

7.2
Produire les produits de diffusion

7.3
Gérer la sortie des produits de diffusion

7.4
Promouvoir les produits de diffusion

7.5
Gérer l'assistance aux clients

Évaluer

8.1
Recueillir les contributions d'évaluation

8.2
Effectuer l'évaluation

8.3
Convenir d'un plan d'action

📍 ... LE MODÈLE DE PROCESSUS SUÉDOIS,...

Le modèle de processus suédois a été développé en 2007, indépendamment du GSBPM mais en parallèle avec lui. Il se fondait également sur les travaux de *Statistics New Zealand* (Pearson et Savage, 2007). Il a été créé dans le cadre d'un projet de modernisation et de standardisation de grande envergure, appelé *Lotta* (voir *infra*), et mis en œuvre par le *Statistiska centralbyrån* (SCB) entre 2006 et le début 2008.

Le modèle de processus standardisé jouait un rôle important dans cette approche, car il définissait la liste des processus et sous-processus pour lesquels il fallait rassembler et développer des normes. La première version du modèle de processus suédois a été établie en octobre 2007 et n'a subi que des révisions mineures depuis lors. La *figure 2* présente sa version actuelle.

📍 ... ET LA DIFFÉRENCE ENTRE LES DEUX

Si l'on examine les deux modèles, on peut identifier quelques différences, les principales étant les suivantes :

- ① le GSBPM comporte huit phases, assorties de processus transverses, tandis que le modèle suédois comprend une neuvième phase (Support et infrastructure) qui remplace les processus transverses. En outre, la phase transversale d'Évaluation ne contient pas de sous-processus dans le modèle, alors qu'elle se subdivise en trois dans le GSBPM ;
- ① la phase de Définition des besoins est abordée différemment dans les deux modèles ;
- ① dans la phase de Conception, le modèle suédois comprend des sous-processus spécifiques pour la diffusion et la communication, et sépare la conception du traitement et de l'analyse en deux sous-processus distincts, tandis que le GSBPM établit un sous-processus séparé pour concevoir la description des variables. De plus, le modèle suédois ajoute un sous-processus pour planifier le cycle de production ;
- ① le GSBPM fait la différence entre l'élaboration d'outils à des fins de traitement et d'analyse d'une part, et à des fins de diffusion d'autre part, établissant pour cela deux sous-processus, tandis que le modèle suédois rassemble ces deux activités dans le même sous-processus ;
- ① dans la phase de Collecte, le modèle suédois distingue la création de la base de sondage et le tirage de l'échantillon, le GSBPM rassemblant ces deux activités dans le même sous-processus ;
- ① les sous-processus de la phase de Traitement diffèrent légèrement dans les deux modèles ;
- ① dans les phases de Construction, de Collecte et de Traitement, le GSBPM comprend des sous-processus pour la finalisation de la phase, mais ces sous-processus ne sont pas explicites dans le modèle suédois.

Si l'on regarde de plus près, en examinant la description détaillée de chaque phase et de ses subdivisions, on s'aperçoit que, au niveau du contenu, les différences sont minimales. Les deux modèles couvrent les mêmes étapes de travail, mais les attribuent dans certains cas à des sous-processus différents. C'est la raison pour laquelle on peut affirmer que, malgré quelques différences subtiles, les deux modèles sont très proches. Lorsque le GSBPM a été présenté au SCB, comme lorsque des révisions majeures ont été apportées au modèle suédois, des discussions ont eu lieu en interne afin de décider s'il fallait s'en remettre entièrement au GSBPM ; mais l'institut a jusqu'à présent préféré conserver son propre modèle. Celui-ci est bien connu au sein du SCB, le *Process Support System* est structuré en conséquence et somme toute, les différences sont minimales.

L'institut suédois a joué un rôle actif dans le développement du GSBPM au fil du temps, et a participé à chaque révision depuis 2009. Il a également pris part au projet MEMOBUST (méthodologie pour la modernisation des statistiques d'entreprises²) d'Eurostat, qui, de fait, traite beaucoup de ce que le *Process Support System* gère déjà : la collecte des méthodes et des guides relatifs à la production des statistiques d'entreprises, selon la grille des sous-processus du GSBPM. Le SCB a beaucoup œuvré dans ce projet pour fonder ses résultats sur le GSBPM.

« L'institut suédois a joué un rôle actif dans le développement du GSBPM au fil du temps, et a participé à chaque révision depuis 2009. »

LE SYSTÈME DE LA STATISTIQUE PUBLIQUE EN SUÈDE

En Suède, la statistique publique est régie par une loi sur les statistiques officielles (« *Lag (2001:99) om den officiella statistiken* ») et par un décret (« *Förordning (2001:100) om den officiella statistiken* »). Depuis 1994, la Suède applique un système décentralisé pour la production des statistiques publiques. Aujourd'hui, 28 agences gouvernementales sont responsables de la statistique publique relative à leur domaine respectif.

Les statistiques sont elles-mêmes réparties entre plusieurs « sujets », « domaines » et produits. À l'heure actuelle, il existe 22 sujets, 112 domaines et 356 produits. Chaque agence étant responsable d'un ou de plusieurs domaines statistiques, il peut y avoir plus d'une agence responsable d'un sujet donné.

Le SCB joue deux rôles au sein du système : en premier lieu, il fait partie des 28 agences responsables de la statistique (pour 45 domaines inclus dans 13 sujets) ; en second lieu, en tant qu'institut national de statistique, il coordonne la globalité du système statistique public et évalue la qualité de la statistique publique.

Un conseil de la statistique publique (« *Rådet för den officiella statistiken* »), rassemblant les dirigeants des 28 agences, a été créé en 2002. Le Conseil compte douze délégués, dont six permanents et six tournants, avec une rotation tous les trois ans. Deux nouveaux délégués sont élus chaque année. Il est présidé par le directeur général du SCB.

Aujourd'hui, le *Statistiska centralbyrån* est organisé comme suit :

- ① un département dédié à la collecte ;
- ① quatre départements dédiés à différents « sujets » ;
- ① un département dédié au développement des processus et des méthodes ;
- ① un département dédié à la communication ;
- ① un département informatique.

Il existe également un service des ressources humaines et un service administratif, ainsi qu'un cabinet exécutif central assurant la gestion de l'institut. Ce cabinet exécutif gère aussi ce qui relève des missions d'un employeur et d'une agence gouvernementale administrative (*figure 3*).

2. Voir https://ec.europa.eu/eurostat/cros/content/memobust_en.

ET RÉAGIR

5

Traiter

6

Analyser

7

Diffuser et
communiquer

5.1
Codifier
et classifier
les micro-données

6.1
Produire
les statistiques

7.1
Préparer
la diffusion

5.2
Éditer les
micro-données

6.2
Éditer
les macro-données

7.2
Compiler
le produit final

5.3
Imputer les
non-réponses

6.3
Effectuer
un contrôle
de la divulgation

7.3
Diffuser
le produit final
au client

5.4
Compléter les
micro-données

6.4
Finaliser
le registre
des observations

7.4
Communiquer
sur le produit final

5.5
Calculer les
pondérations

6.5
Interpréter
et expliquer

7.5
Éliminer
et conserver

6.6
Finaliser
le produit
de diffusion

STRUCTURES

❶ POURQUOI LE SCB A DÉCIDÉ DE METTRE EN ŒUVRE UN MODÈLE DE PROCESSUS

En 2005, l'environnement de production du SCB était largement décentralisé. Cela découlait pour partie du passage d'une unité centrale à des ordinateurs individuels vers la fin des années quatre-vingt-dix³, et pour une autre partie, de la nature extrêmement décentralisée de notre institution : chaque département dédié à un thème (sujet) était responsable de ses propres processus de production, appliquait ses propres méthodes et avait ses propres experts informatiques. Dans ce contexte, chacun avait plus ou moins son propre système informatique pour la production de ses statistiques, selon ses propres développements. Les différences étaient grandes dans la manière d'exécuter chaque tâche des processus de production. En outre, les systèmes informatiques étaient déjà anciens et avaient besoin d'être reconstruits, sans parler du fait qu'ils ne faisaient l'objet d'aucune documentation et que leur maintenance était coûteuse. Enfin, il était devenu également coûteux et difficile d'introduire de nouvelles méthodes et solutions techniques pour ces systèmes dans un tel environnement informatique.

Deux exemples en témoignent : la collecte de données sur internet et la nouvelle méthode de contrôle sélectif. Dans les deux cas, il semblait inefficace de développer des solutions spécifiques à chaque enquête, et il était donc naturel de développer des solutions centrales communes. Mais pour les mettre en œuvre dans l'ensemble du SCB, il était nécessaire de procéder à de nombreuses intégrations dans le système spécifique à chaque enquête. Cela avait conduit à une « architecture spaghetti », exigeant de consacrer beaucoup de temps à la maintenance de toutes les intégrations.

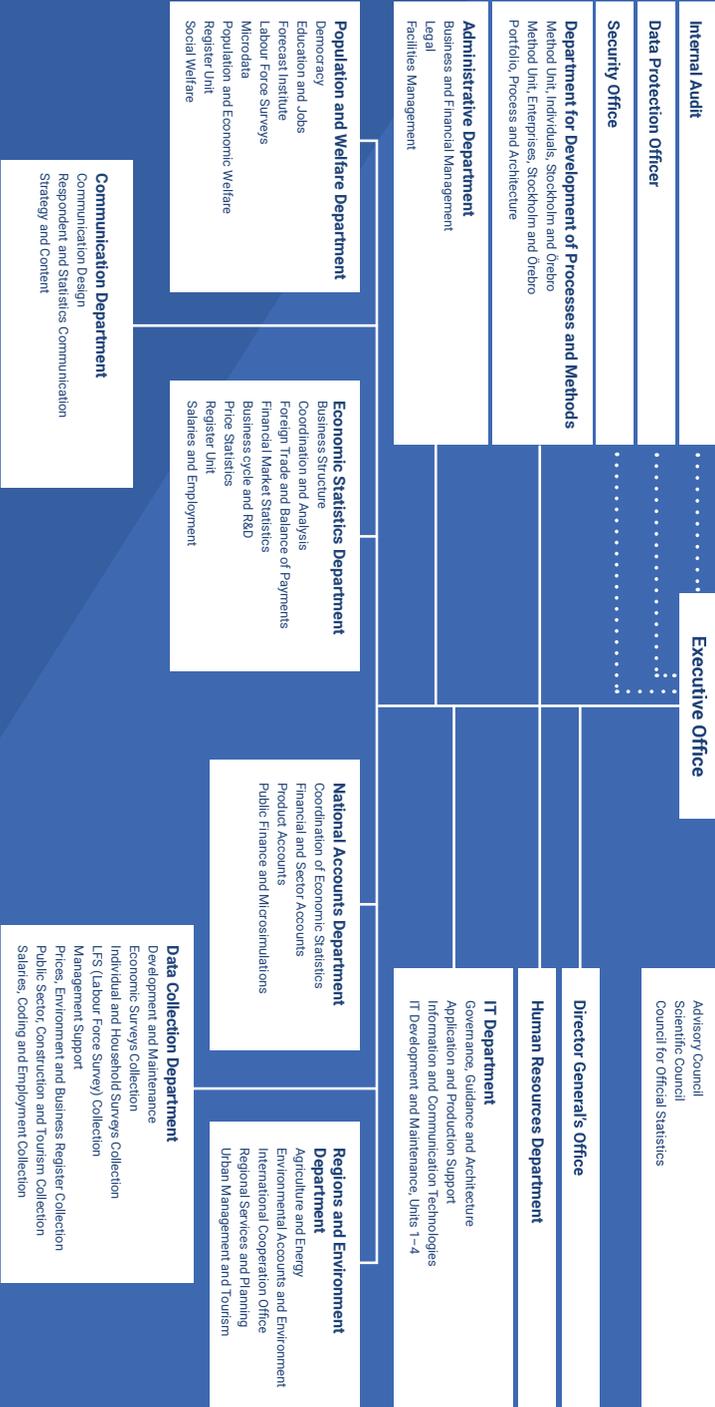
En plus de tout cela, de nombreux experts du SCB allaient partir en retraite au cours des années suivantes, des experts qui connaissaient parfaitement les processus de production statistique « locaux ». Cela était également le cas pour de nombreux informaticiens. Il aurait alors fallu transmettre à d'autres personnes les connaissances relatives au fonctionnement de chaque enquête et de son système informatique spécifique.

Le SCB a donc pris la décision stratégique de lancer un projet de modernisation et de standardisation, le *projet Lotta*⁴. Son but était de réduire le nombre de systèmes de production et d'améliorer l'efficacité de la production statistique. Afin d'atteindre cet objectif, il a été décidé d'adopter une approche « orientée processus » de la production statistique (*Process Oriented Approach*) et de rassembler, développer et définir des méthodes, traitements automatisés et outils standardisés pour tous les sous-processus. On considérait que cela permettrait également « d'institutionnaliser » les connaissances sur les procédures ainsi standardisées.

Dans le cadre de ce projet, le SCB a développé et défini un modèle de processus permettant de décrire tous les processus de production. S'appuyant sur les travaux effectués par *Statistics New Zealand*, cela a abouti en octobre 2007 au modèle de processus suédois (*figure 2*).

-
3. Un grand nombre de ces systèmes avaient été rapidement développés ou modifiés à cette époque en raison de craintes quant à l'impact potentiel du « bug de l'an 2000 ».
 4. Le *projet Lotta* est nommé ainsi car c'était le prénom fêté le jour du lancement, le 13 mai. Le projet et l'organisation des processus sont décrits dans (Jorner, 2008).

Figure 3. Organigramme du *Statistiska centralbyrån* (SCB)



Janvier 2020

Disponible à l'adresse : <https://www.scb.se/en/About-us/main-activity/organisation/>.

IMPACT SUR L'ORGANISATION

La décision d'adopter une perspective axée sur les processus pour rationaliser la production statistique allait avoir un impact sur l'organisation du SCB, et placer le nouveau modèle standardisé de processus sur le devant de la scène.

« Développer toute une infrastructure (Process Support System) rassemblant toutes les normes ainsi définies, et de fonder ce système sur le modèle de processus suédois. »

Pour stimuler la progression vers la standardisation et vers des outils informatiques communs, il a été décidé de créer un département des processus, comprenant des propriétaires de processus (ou *Process Owner*) pour tous les sous-processus inclus dans la production statistique de l'institut. Les propriétaires de processus allaient être responsables du développement et de

la maintenance des méthodes, traitements automatisés et outils standardisés. En outre, tous les agents traitant de la méthodologie ou de l'informatique ont été regroupés au sein de ce département des processus, qui a vu le jour le 1^{er} janvier 2008.

Il a également été décidé de développer toute une infrastructure (*Process Support System*) rassemblant toutes les normes ainsi définies, et de fonder ce système sur le modèle de processus suédois.

Des propriétaires de processus ont été nommés pour cinq domaines définis à partir des sous-processus du modèle :

- ① Définition des besoins, Diffusion et communication ;
- ① Conception et planification, Construction et essai ;
- ① Collecte ;
- ① Traitement et analyse ;
- ① Évaluation et retour d'information, Support et infrastructure.

Pour chaque domaine, un propriétaire de processus, et un adjoint, ont été nommés. Chaque domaine a également reçu un budget pour la maintenance des normes et outils existants et pour le développement de nouvelles normes. Il était prévu que les grands projets de développement informatiques soient initiés par les propriétaires de processus, mais financés par les ressources propres du SCB qui devaient également couvrir d'autres types de projets – tels que ceux consacrés au développement du contenu statistique – mais se concentrer en priorité sur le développement de processus.

Les personnes travaillant à la production statistique restaient dans les départements dédiés à leurs sujets d'étude, mais devaient également se consacrer à la collecte des données et à la communication. Cela a donné lieu *in fine* à l'établissement d'une structure « matricielle » contenant à la fois les processus/sous-processus et les produits statistiques.

Deux des dernières initiatives du *projet Lotta* ont consisté à préparer d'une part une première décision sur les normes, d'autre part une décision sur la façon dont l'infrastructure commune serait conçue :

- ① sur le premier point, il a fallu examiner les outils, méthodes et traitements automatisés déjà utilisés par le SCB, puis faire la liste de ceux qui deviendraient des normes dans la nouvelle approche de travail orientée processus. Un grand nombre de sous-projets dans le *projet Lotta* avaient établi des normes et des outils. On estimait que cela serait très utile pour identifier les sujets dans lesquels les normes faisaient encore défaut. Le directeur général décidait des normes, et les propriétaires de processus allaient être chargés de leur maintenance et de leur développement ;
- ① pour le deuxième aspect, il a été décidé de fonder le *Process Support System* sur MS SharePoint, où toutes les normes et tous les outils devaient être décrits et rendus accessibles aux utilisateurs. Hormis les outils informatiques, il était prévu qu'il contienne des traitements automatiques standardisés, ainsi que des outils de support tels que les listes de contrôles (*checklists*), les modèles et d'autres éléments nécessaires à l'application des normes. Les propriétaires de processus ont été chargés d'alimenter toute cette infrastructure logicielle et documentaire. Ce fut un travail considérable, mobilisant de nombreuses personnes dans l'institut, et une première version a été publiée dès septembre 2008. Le *Process Support System* avait - et a toujours - une architecture simple, construite à partir d'une version « cliquable » du modèle de processus suédois. Chaque phase et chaque sous-processus (qui peut être divisé en pages supplémentaires ou en sous-processus secondaires plus détaillés selon les besoins) comprend une description sur une page, divisée en quatre parties (*figure 4*) :
 - une courte description du processus ou du sous-processus et de son objectif ;
 - les intrants (*input*) : ce qui doit avoir été fait avant de démarrer le processus, et source des informations requises (avec des liens vers les sous-processus précédents, le cas échéant) ;
 - la réalisation : une description détaillée de la façon dont le processus doit être exécuté, y compris des liens vers les outils appropriés et une description des normes à appliquer (dont modèles, listes de contrôle, etc.) ;
 - les extrants (*output*) : les résultats en sortie du processus.

L'infrastructure logicielle que constitue le *Process Support System* est toujours pleinement utilisée aujourd'hui. Elle joue un rôle légèrement différent de celui du modèle standardisé de processus : son principal objectif est de présenter les normes définies par le SCB qui doivent être appliquées par tous les utilisateurs des processus (il existe un processus spécifique à appliquer en cas d'exception, si, pour une raison ou pour une autre, on ne peut pas respecter une norme donnée). Si la description des phases et des sous-processus est générique dans le GSBPM, les descriptions incluses dans le *Process Support System* sont en revanche détaillées, précises et spécifiques au SCB.

L'utilisation d'un modèle général et standardisé de processus a permis d'une part de créer de nouveaux outils informatiques standardisés, et d'autre part, de faciliter la communication entre ces outils puisqu'ils sont reliés aux processus. Parallèlement, ces outils ont rendu la production plus efficace, notamment à l'étape de la collecte des données. Grâce à la standardisation, les coûts liés à l'apprentissage de nouveaux outils informatiques ont diminué et il est plus facile de travailler sur différentes productions statistiques. Bien que la standardisation puisse engendrer une certaine perte de flexibilité, globalement l'institut estime avoir atteint un grand nombre des objectifs fixés avant de débiter la démarche d'« orientation processus » de la production statistique.

CERTIFICATION ISO 20252 ET MISE EN ŒUVRE DANS LE PROCESS SUPPORT SYSTEM

Cette infrastructure commune, et en conséquence le modèle standardisé de processus suédois sur lequel elle s'appuie, ont pris de plus en plus d'importance entre 2008 et 2014. Dès mars 2008, le directeur général avait décidé que le SCB demanderait à être certifié suivant la norme internationale ISO 20252 (destinée aux services menant des études de marché, des sondages d'opinion et des recherches sociales⁵).

Des recherches préliminaires ont souligné que le SCB devrait fournir de gros efforts pour parvenir à la conformité avec cette norme. Il était notamment nécessaire d'en interpréter

Figure 4. Exemple de structuration sur l'infrastructure logicielle et documentaire

Menu base sur le modèle de processus

Entrant ou input - Ce qui doit être fait avant

Realisations : description et exemples

Modèles à utiliser selon le type d'enquête

5. L'ISO 20252, norme internationale publiée en 2006, définit les termes et les exigences pour un système de management de la qualité, applicables aux entreprises et aux professionnels réalisant des études de marché, des études sociales et des études d'opinion.

certaines parties de façon plus approfondie, et surtout de développer de nouvelles normes. Ces travaux ont été réalisés dans le cadre d'un projet mené de décembre 2008 à décembre 2009, visant à aboutir pour le début de l'année 2010 au plus tard.

Lors d'un processus de certification ISO, le système d'assurance-qualité doit avoir été mis en place et utilisé pendant une période assez longue avant que la certification puisse être accordée. Pour chaque critère défini dans la norme, SCB s'est demandé s'il fallait établir une solution commune ou des instructions spécifiques à chaque production quant à la façon dont la norme pouvait être respectée. Il a été décidé d'inclure dans le *Process Support System* à la fois des solutions communes et des instructions, afin que, si une production est conforme aux normes définies dans le système du SCB, elle soit automatiquement jugée conforme aux exigences de la norme ISO 20252. Cette intégration des exigences de l'ISO 20252 a également permis à SCB d'en imposer de plus strictes lorsque cela semblait nécessaire, et de combiner les standards internes avec les critères externes de la norme ISO.

Le *Process Support System* était quasiment à jour à la fin 2010 ; mais la norme ISO exige une solution technique permettant de surveiller les enquêtes téléphoniques, que le SCB ne possédait pas encore, et la certification était donc impossible à obtenir avant d'avoir installé une telle solution. Cela a pris un certain temps. Le SCB a fini par obtenir la certification en mars 2014. Le parcours vers la certification est décrit dans (Hoff *et alii*, 2010) et (Bergdahl *et alii*, 2014).

❶ FACTEURS DE RÉUSSITE : INTENSITÉ, SPONSOR ET MOTEURS D'UNE APPROCHE GLOBALE

Le *projet Lotta*, la mise en œuvre d'un modèle de processus très proche du GSBPM, la mise au point d'une vaste infrastructure opérationnelle commune (*Process Support System*), la création du département des processus et la certification à la norme ISO 20252 – autant d'initiatives étroitement liées et mises en œuvre durant une période intense (2006-2010). Ces changements ont indéniablement été contraignants pendant ces quelques années, mais constituent cependant nos principaux facteurs de réussite.

« Ces changements ont indéniablement été contraignants pendant ces quelques années, mais constituent cependant nos principaux facteurs de réussite. »

Un profond sentiment de l'urgence prévalait, à tous les niveaux de l'institut. Les changements étaient conduits et appuyés par le *top management*, et traités de façon cohérente. Les collaborateurs qui n'étaient pas directement concernés n'étaient peut-être pas de cet avis, pris

au beau milieu d'un ouragan de changements : mais si toutes les évolutions avaient été introduites une par une, le projet n'aurait sans doute pas réussi.

La standardisation et le *Process Support System* n'auraient probablement pas abouti sans qu'un certain groupe de personnes ne s'organise pour faire avancer le développement et la mise en œuvre (les propriétaires de processus). Si on n'avait pas fixé l'objectif de certification ISO 20252, le système ne serait peut-être pas devenu un outil central de support et de définition des orientations pour les différentes productions. Et, sans modèle réellement standardisé, il aurait été beaucoup plus difficile d'étendre les normes et les orientations à l'ensemble du SCB. Il a en effet permis d'établir une interprétation commune des différentes phases de la production statistique, et a fourni un cadre permettant à la fois de décrire et

d'identifier les normes, les orientations, les outils et les supports. Le modèle standardisé de processus et l'infrastructure qui supporte ce modèle ont défini un cadre approprié pour vérifier la conformité aux normes existantes.

● CHANGEMENTS AU FIL DU TEMPS

Entre 2010 et 2016, l'organisation interne et les systèmes sont restés relativement stables. L'institut était concentré sur le développement d'outils informatiques communs pour différents processus ; les fonctions informatiques ont quitté le département des processus pour intégrer leur propre unité en 2010. À noter toutefois que le développement informatique était toujours initié par les propriétaires de processus. Le département des processus et le département informatique travaillaient alors en étroite collaboration. Les procédures automatisées et les normes internes étaient toujours sous la responsabilité du département des processus et étaient mises à jour en continu. Le système constitué d'un département dédié aux processus en général et de propriétaires de processus a très peu évolué. Les responsabilités et les sous-processus attribués à chaque propriétaire ont légèrement changé, mais l'organisation principale est restée la même. Une description plus détaillée de l'environnement de production commun est fournie dans (Erikson et Odenkrants, 2016).

Des changements organisationnels plus importants ont été apportés à partir de 2017. Des discussions avaient déjà eu lieu sur les frontières et responsabilités, entre département des processus et département R&D. Il devenait nécessaire de renforcer nos capacités

“ Des discussions avaient déjà eu lieu sur les frontières et responsabilités, entre département des processus et département R&D. ”

de développement et d'améliorer l'efficacité des projets, ainsi que d'assurer une mise en œuvre plus rapide des résultats des projets. Dans cette optique, les deux départements ont fusionné en un nouveau « département du développement des processus et des méthodes ». L'institut a également décidé de mettre en œuvre un nouveau modèle de maintenance. Ces deux décisions ont engendré des changements dans les rôles et les responsabilités.

Les anciens propriétaires de processus ont été remplacés par des responsables de la maintenance, avec un seul propriétaire pour le processus de production statistique dans son ensemble. Puisque les responsables de la maintenance étaient dorénavant situés dans des départements différents au sein de l'institut, la maintenance du *Process Support System* s'est trouvée quelque peu décentralisée. Aujourd'hui, c'est à la fois aux responsables de la maintenance et au propriétaire de processus qu'il appartient d'assurer la cohérence et la mise à jour. Mais le rôle du *Process Support System* et du modèle standardisé de processus reste le même. Le système contient toujours les normes à appliquer, et le principe est toujours valide, selon lequel, si un produit suit le standard, il est automatiquement jugé conforme à la norme ISO 20252. C'est l'organisation autour de laquelle le système s'articule qui a changé, et non pas le système lui-même ou son objectif.

● CONFORMITÉ AU CADRE DE RÉFÉRENCE

Dans la mesure où les exigences de la norme ISO 20252 sont intégrées au *Process Support System*, les utilisateurs des processus sont fortement incités à utiliser ce système et ses normes.

La conformité est vérifiée par un auditeur externe qui mène chaque année un audit de conformité à la norme ISO 20252 (grâce à un échantillon aléatoire de produits), ainsi que par des contrôleurs-qualité internes qui vérifient à la fois la conformité à l'infrastructure et la qualité de l'infrastructure elle-même et des orientations qu'elle contient. L'audit qualité interne est organisé de façon à examiner l'utilisation de tous les sous-processus sur une période glissante de trois ans.

Bien que les audits identifient parfois des écarts par rapport aux normes existantes, on peut dire que globalement la conformité est plutôt satisfaisante dans l'ensemble du SCB. Ce système est utilisé depuis dix ans et l'institut y est maintenant habitué. Les propriétaires de produits savent qu'ils peuvent faire – et feront assurément – l'objet d'un audit de conformité. Les auditeurs externes vérifiant la conformité à la norme ISO se sont dits très impressionnés par le *Process Support System* et l'ont jugé excellent en termes de conformité à la norme et de qualité du processus de production statistique.

Ce qui précède montre que la création et la mise en œuvre d'une approche de la production statistique orientée processus a globalement réussi en Suède. Toutefois cela n'a pas été sans difficulté. Les trois principaux défis rencontrés durant cette période ont été l'acceptation, la facilité pour les utilisateurs des processus d'identifier les informations pertinentes et la mise à jour de ces informations.

TROIS NOUVEAUX DÉFIS : ACCEPTER,...

La standardisation pose toujours des problèmes en termes d'acceptation. Certaines personnes se demandent pourquoi elles doivent changer ce qui fonctionne bien. Une procédure standardisée peut être pertinente pour le SCB dans son ensemble, mais pas nécessairement pour chaque utilisateur. La nécessité d'une standardisation est remise en question par des personnes expérimentées qui connaissent les batchs actuels par cœur grâce à une longue expérience. La standardisation peut être considérée comme un frein à la créativité. Les normes elles-mêmes, ainsi que la pertinence de leur mise en œuvre

“ La nécessité d'une standardisation est remise en question par des personnes expérimentées qui connaissent les batchs actuels par cœur grâce à une longue expérience. La standardisation peut être considérée comme un frein à la créativité. ”

(par exemple afin d'obtenir une certification), peuvent également être remises en question. Toutes ces situations ont été rencontrées lors de l'implémentation.

Le modèle de processus a été accusé de trop simplifier, de ne tenir compte ni des différences entre les types de produits, ni

des complexités de la production statistique dans la pratique. Les normes, ainsi que la décision de demander la certification, ont été remises en question. Des collaborateurs expérimentés ont craint de devenir les « esclaves » des listes de contrôle, d'être contraints dans leur libre arbitre.

Ce type de discussion survient encore de temps à autre, si quelqu'un pense qu'il y a de meilleures manières de faire que les normes existantes, et qu'il est empêché de les utiliser comme bon lui semble. Pour répondre à ces arguments, il est important de débattre ouvertement des avantages de la standardisation, de créer un sentiment d'implication personnelle des utilisateurs des processus dans le développement des normes et d'obtenir

le ferme support de la direction en vue des changements. Lors du développement, la direction a apporté un support inébranlable, ce qui a joué le rôle le plus important pour surmonter cette difficulté. On a également tenté de traiter ce problème en continu et, s'il persiste dans une certaine mesure aujourd'hui, il concerne des normes, modèles et outils informatiques particuliers et leur mode d'utilisation, plutôt que le concept global de l'orientation vers les processus et de la standardisation. C'est positif et beaucoup plus facile à gérer. Rappelons également que, aux alentours de 2010, il s'est avéré que certaines des statistiques du SCB contenaient des erreurs importantes, et il a fallu fournir de gros efforts pour améliorer la qualité de la production. Pour ce faire, les normes ont été très utiles, ce qui a renforcé l'acceptation de l'approche de travail axée sur les processus.

... IDENTIFIER LES INFORMATIONS PERTINENTES,...

Le modèle de processus et l'infrastructure qui le supporte contiennent de gros volumes d'informations. Et toutes les normes ne s'appliquent pas à tous les types d'enquêtes. Par exemple, certaines normes ne concernent que les enquêtes par sondage avec collecte de données primaires, tandis que d'autres s'appliquent aux collectes fondées sur des données administratives. De très nombreux traitements automatiques standardisés sont liés à différents modes de collecte des données, et les enquêtes utilisent des modes différents. Les enquêtes obligatoires sont assorties de règles très strictes qui ne concernent pas les enquêtes facultatives. Il existe également des différences selon que l'on s'intéresse à des particuliers ou à des entreprises, etc.

« *Beaucoup de collaborateurs estiment toujours qu'il est difficile d'identifier toutes les informations nécessaires, et cela affecte également l'acceptation des normes en général.* »

Pour ces raisons, s'agissant de nombreux sous-processus, il est nécessaire d'identifier les normes applicables aux produits concernés, et uniquement ces normes. Cela peut parfois être difficile. En outre, on risque de passer à côté d'informations importantes. Certains de ces problèmes ont été abordés en divisant davantage les sous-processus et

en créant des documents globaux, par exemple pour la production statistique fondée sur des répertoires et pour la vérification et la validation dans l'ensemble du processus de production. Mais cela n'a résolu qu'une partie de la difficulté. Beaucoup de collaborateurs estiment toujours qu'il est difficile d'identifier toutes les informations nécessaires, et cela affecte également l'acceptation des normes en général. Ceux qui ne trouvent pas ce qu'ils cherchent sont plus susceptibles de rejeter le système dans son ensemble. C'est là le défi le plus important à relever et gérer en continu.

... ET MAINTENIR LES INFORMATIONS À JOUR

Comme on a vu plus haut, le *Process Support System* contient de gros volumes d'informations. Or, pour qu'il soit accepté au sein du SCB, il est impératif que tout soit à jour et pertinent. Cela a deux conséquences. Premièrement, lorsque les traitements automatisés ou les normes changent, il est important que les informations incluses dans l'infrastructure opérationnelle commune soient mises à jour en conséquence. Ce peut être une difficulté, surtout si les personnes qui mettent à jour ou développent une norme ne sont pas celles qui ont la responsabilité des informations incluses dans le système. C'est une chose de développer une norme ou un outil, mais c'en est une autre d'en faire la description dans des routines pouvant être intégrées à l'infrastructure. Pour cette raison, il arrive parfois que la nouvelle

norme ne soit pas décrite immédiatement dans le système, ce qui pose problème. Il peut aussi arriver que les changements affectent plusieurs sous-processus dont la maintenance est assurée par différents responsables, ce qui peut engendrer des incohérences si toutes les parties concernées ne sont pas mises à jour au même moment. Dans d'autres cas encore, il peut être difficile de décider de la partie du modèle standardisé la plus appropriée pour y affecter les informations : doivent-elles être intégrées à un processus de conception, à un processus de réalisation ou bien à la phase Support et infrastructure ?

S'agissant de la mise à jour des informations, une autre complication découle du fait que toutes les normes ne changent pas au fil du temps, ou du moins pas très souvent. Ainsi, certaines normes développées il y a cinq ou dix ans sont encore utilisées de la même

« Certaines normes développées il y a cinq ou dix ans sont encore utilisées de la même façon aujourd'hui. »

façon aujourd'hui. Lorsqu'il est confronté à un processus créé et mis à jour pour la dernière fois en 2013, l'utilisateur n'est pas assuré que ces informations soient toujours pertinentes. Dans ce cas, il a du mal à accepter le système dans son ensemble. Pour résoudre ce problème, l'idée a été d'attribuer à chaque sous-processus non seulement la date de

la dernière mise à jour mais aussi la dernière date à laquelle les informations ont été vérifiées et confirmées comme étant pertinentes. Cela permet de dire, par exemple, que les informations ont été mises à jour en 2013 mais vérifiées en 2020 et sont par conséquent à jour. Cette approche a amélioré les choses et semble permettre de surmonter la difficulté.

ÉVOLUTION FUTURE : QUELQUES VOIES À EXPLORER

La mise en œuvre d'une production statistique « orientée processus », fondée sur un modèle commun, a été pour l'essentiel une réussite, malgré les difficultés évoquées plus haut. Le modèle standardisé est désormais bien établi en tant que moyen de description du processus de production statistique, et l'infrastructure fondée sur ce modèle est utilisée dans l'ensemble du SCB. Mais cela ne veut pas dire qu'il ne reste plus rien à améliorer ou à développer. Trois évolutions potentielles prioritaires ont été identifiées :

- ❶ **une meilleure ergonomie** : comme nous l'avons vu, le plus gros défi repose sur la capacité de nos collaborateurs à identifier facilement des informations pertinentes et à jour, afin d'assurer un bon niveau d'acceptation. Plusieurs idées ont été avancées pour optimiser l'ergonomie du *Process Support System*, notamment des approches sur mesure pour les différents types de statistiques, ainsi que des descriptions plus globales pour les sujets couvrant plusieurs parties du processus de production. Il a également été suggéré d'inclure une plus grande partie du cadre juridique relatif à la production statistique, et d'établir un lien avec les évolutions stratégiques du SCB. Pour inclure de plus gros volumes d'informations, il faudra réfléchir de façon plus approfondie à la façon dont elles peuvent être présentées, afin que l'ergonomie ne se dégrade pas en raison de volumes trop importants, mais s'améliore au contraire grâce à l'inclusion d'informations plus pertinentes ;
- ❷ **une utilisation plus adaptée aux produits** : dès le début, des idées ont été suggérées pour créer une perspective radicalement différente sur le *Process Support System* et pour l'adapter de sorte que chaque produit reçoive les informations nécessaires, mais pas plus. Il devrait être possible d'identifier les informations nécessaires à un produit donné si l'on a accès à des informations sur ses choix de conception de base. On a tenté de mettre ces idées en pratique dans le cadre de la collecte des données avec un système

« basique » d'assurance-qualité : chaque produit fournit des informations quant à sa conception, à partir desquelles une liste d'opérations d'assurance-qualité est établie pour le produit concerné. Cela consiste à filtrer les opérations pertinentes à partir d'une liste brute d'opérations potentielles, en fonction des choix de conception du produit. Cette version simple est déjà prometteuse (*figure 5*) mais, pour fonctionner pleinement, elle devra couvrir la totalité du processus de production statistique. Elle devra devenir plus « intelligente » qu'elle ne l'est aujourd'hui, pour présenter un outillage (tels que les modèles et les listes de contrôle) adaptés à des choix de conception spécifiques. Pour être vraiment utile aux personnes chargées de la production statistique, elle devra également permettre d'ajouter des informations spécifiques aux enquêtes, ainsi qu'une description de la façon dont un produit utilise une norme donnée. Par ailleurs, les choix de conception servent à créer un domaine pour chaque produit, où il est possible d'accéder aux outils pertinents et où les données du processus relatives à la production peuvent être présentées. À l'heure actuelle cela se limite à la collecte des données, mais il est prévu d'étendre cette utilisation à toutes les phases du processus de production statistique.

① **un même système pour l'ensemble de la statistique publique en Suède** : comme on l'a décrit *supra*, la responsabilité de la statistique publique en Suède est répartie entre 28 agences différentes. Pour l'instant, le *Process Support System* n'est utilisé que par le SCB. Mais le débat est ouvert : les autres agences pourraient ou devraient-elles l'utiliser ? Dans l'affirmative, il devrait probablement y avoir deux versions du système, l'une spécifique au SCB et l'autre plus générique pour toutes les agences. À ces fins, il sera essentiel de définir un terrain d'entente, sous la forme du modèle de processus, sans quoi il serait extrêmement difficile d'établir une interprétation commune des éléments à relier aux normes et aux traitements automatisés. Si les autres agences ont leur propre modèle de processus, cela pourrait également nécessiter des ajustements, et il faudrait faire des compromis pour établir un modèle commun. Un test à portée limitée sera mené au printemps 2020 afin de déterminer si les informations sont utiles à d'autres agences.

Compte tenu de ces potentielles évolutions futures, le modèle standardisé de processus et tous les éléments qui y sont rattachés continueront indéniablement de jouer un rôle important dans l'infrastructure du SCB.



BIBLIOGRAPHIE

UNECE, 2019. *Generic Statistical Business Process Model GSBPM*. [en ligne]. Janvier 2019. Version 5.1. [Consulté le 19 juin 2020]. Disponible à l'adresse : <https://statswiki.unece.org/display/GSBPM/GSBPM+v5.1>.

BERGDAHL, Heather, COLLIN, Marie, ELLTING, Pernilla, LISAI, Dan, PETERSSON, Åke et HOFF, Sara, 2014. *The final steps of the journey towards an ISO certification. Implementing ISO 20252 for Market, Opinion and Social Research at Statistics Sweden*. [en ligne]. 2-5 juin 2014. European Conference on Quality in Statistics Q2014, Vienne. [Consulté le 19 juin 2020]. Disponible à l'adresse : http://www.q2014.at/fileadmin/user_upload/The_final_steps_of_the_journey_towards_an_ISO_certification_Q2014.pdf.

ERIKSON, Johan et ODENCRANTS, Martin, 2016. *Building a common production environment – the Swedish experience*. [en ligne]. 20-23 juin 2016. American Statistical Association (ASA), International Conference on Establishment Statistics ICES-V, Genève. [Consulté le 19 juin 2020]. Disponible à l'adresse : https://ww2.amstat.org/meetings/ices/2016/proceedings/137_ices15Final00203.pdf.

HOFF, Sara, JAPEC, Lilli, LISAI, Dan et PETERSSON, Åke, 2010. *The journey towards an ISO certification – Implementing ISO 20252 for Market, Opinion and Social Research at Statistics Sweden*. [en ligne]. 4-6 mai 2010. European Conference on Quality in Statistics Q2010, Helsinki. [Consulté le 19 juin 2020]. Disponible à l'adresse : https://q2010.stat.fi/media//presentations/Hoff_Q2010_paper_ISO_paper.pdf.

JORNER, Ulf, 2008. *Summa summarum. SCBs första 150 år*. [en ligne]. Septembre 2008. Statistics Sweden (SCB), Stockholm. [Consulté le 19 juin 2020]. Disponible à l'adresse : http://share.scb.se/OV9993/Data/Publikationer/statistik/publikationer/OV9999_2008A01_BR_X05BR0801.pdf.

PEARSON, John et SAVAGE, Tracey, 2007. *Statistics New Zealand's Business Model Transformation Strategy : Creating a new Business Model for the 21st Century National Statistical Office*. [en ligne]. 18-21 juin 2007. American Statistical Association (ASA), International Conference on Establishment Statistics ICES-III, Montréal, Québec, Canada, pp. 1150-1157. [Consulté le 19 juin 2020]. Disponible à l'adresse : https://ww2.amstat.org/meetings/ices/2007/proceedings/ICES2007-000185_PDF.

VALE, Steven, 2009. *Generic Statistical Business Process Model*. [en ligne]. Avril 2009. Joint UNECE, Eurostat, OECD Work Session on Statistical Metadata (METIS) Version 4.0. [Consulté le 19 juin 2020]. Disponible à l'adresse : <https://www.unece.org/fileadmin/DAM/stats/documents/ece/ces/ge.61/2009/mtg1/zip.32.e.pdf>.

LE SYSTÈME STATISTIQUE DU LOGEMENT

ÉTENDUE ET PERSPECTIVES

Jérôme Harnois* et Pierre Lamarche**

Le système statistique français décrit et quantifie une large part des événements concernant le logement : depuis la construction ou la rénovation de locaux à usage d'habitation, jusqu'au mode d'occupation par les ménages, en passant par les transactions et autres opérations économiques relatives à leur usage.

Le secteur de la construction constitue un élément majeur de l'activité économique ; la dynamique et les conditions d'acquisition des logements sont de ce fait scrutées de près par les acteurs concernés. Le logement est porteur par ailleurs d'un enjeu social fort, à travers les questions de son coût pour les ménages, des difficultés d'accès, ou du mal-logement. La qualité de l'information statistique est donc essentielle, que ce soit en termes de précision, de fraîcheur et de pertinence.

En France, le spectre des dispositifs statistiques sur le logement est large : enquêtes thématiques auprès d'un échantillon de ménages, données administratives exhaustives produites par l'appareil administratif qui régule l'activité immobilière du pays.

Ce système d'information reste perfectible : il doit s'adapter aux évolutions de la société et intégrer les sources nouvelles. Au cours des prochaines années, la statistique publique relèvera deux défis en particulier : mieux exploiter la granularité toujours plus fine des données pour ancrer les possibilités d'analyse dans le territoire, et intégrer les nouveaux enjeux du développement durable.

 *The French statistical system makes it possible to describe and quantify many events related to housing, ranging from construction or refurbishment of housing units to tenure status for dwellings, also transactions and other economical operations involving real estate assets.*

Regarding economics, the construction sector is at the core of activity. As a consequence, dynamics and conditions for real estate transactions are closely monitored by the economical actors. They also constitute an aspect of the social issue, along with housing costs and housing difficulties.

In France, all these events and phenomenons are measured thanks to a large spectrum of tools, ranging from household surveys to administrative data, also administrative procedures for data collection related to the regulation of real estate activities in the country.

The information system has to evolve and be enhanced, so as to account for the changes in the society and the economy, as well as achieving the integration of new data sources. Better exploiting the always decreasing data granularity and improving the analyses on the territorial side are the challenges that the statistical system is now facing. Obviously it will also have to address more and more issues related to sustainable development as a result of the coming times.

* Sous-directeur des statistiques sur le logement et la construction, Sdes,
jerome.harnois@developpement-durable.gouv.fr

** Chef de la division Logement, Insee,
pierre.lamarche@insee.fr

📍 L'ACTION PUBLIQUE EN FAVEUR DU LOGEMENT : UNE LONGUE HISTOIRE

Jusqu'à la fin du XIX^e siècle, il était admis que le logement relevait de la sphère privée. L'industrialisation a confronté les classes populaires des territoires urbains à un manque structurel de logements peu onéreux, secteur fui par les investisseurs. Le surpeuplement et l'insalubrité de grande ampleur ont poussé les dirigeants de la III^e République à prendre des initiatives en faveur du logement. Deux conflits mondiaux, l'apparition du chômage de masse à la fin des Trente Glorieuses ou encore la crise initiée en 2008 à la suite de celle des *subprimes* ont rendu nécessaire le maintien d'un solide interventionnisme d'État en matière de logement et de construction. En ce début de XXI^e siècle, de nouveaux enjeux se font jour, désormais en lien avec ceux du développement durable.

Si le sujet du logement et de la construction est inscrit depuis longtemps dans l'action gouvernementale, il en est de même dans la statistique publique. Cet article dresse le panorama des différents outils d'observation du Service statistique public (SSP) en la matière, chacun utilisant une focale spécifique. Ces dispositifs mobilisent l'Insee, le Sdes¹ mais aussi le réseau des notaires ; ils s'appuient sur des enquêtes statistiques et sur des données d'origine administrative. La richesse de l'appareil français n'évitera pas des évolutions à court ou moyen terme. Tout en modernisant ses outils traditionnels sur les thématiques économiques et sociales, la statistique publique se doit désormais d'éclairer le débat sur la maîtrise de l'artificialisation, la rénovation énergétique des bâtiments, ou la régulation de l'offre au regard de la demande de logements. Il lui faudra aussi répondre à une demande très soutenue de données localisées, qui émane des acteurs du secteur comme du grand public.

📍 AUTOUR DU LOGEMENT, UNE MULTIPLICITÉ D'ÉVÈNEMENTS —

Offrir une vision statistique complète du domaine du logement est loin d'être une sinécure. La notion de bâti recouvre un vaste domaine, marqué dans le temps comme dans l'espace par une très grande diversité d'états, de situations physiques ou de conditions d'occupation (**figure 1**).

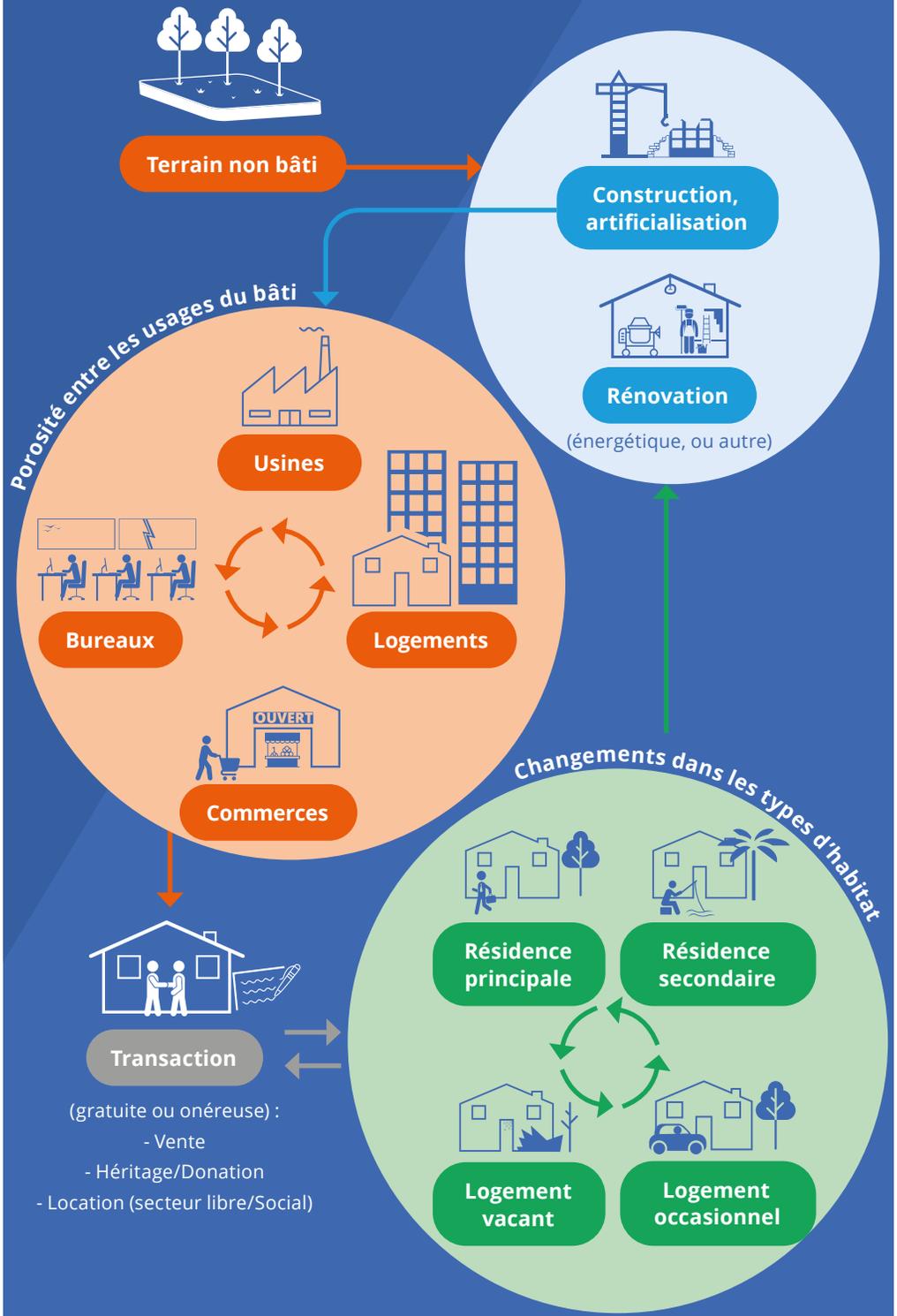
« L'histoire d'un bâtiment, quel qu'en soit le premier usage, habitation ou non résidentiel, débute généralement par l'acquisition, onéreuse ou gratuite, d'une parcelle de terrain nue de toute construction. »

L'histoire d'un bâtiment, quel qu'en soit le premier usage, habitation ou non résidentiel, débute généralement par l'acquisition, onéreuse ou gratuite, d'une parcelle de terrain nue de toute construction. La superficie de cette parcelle, sa localisation géographique et son emplacement par rapport aux autres

zones bâties ou aux réseaux de transport déterminent son prix d'acquisition : celui-ci représente une part non négligeable du coût total d'un projet de construction immobilière (près du tiers dans le cas d'une maison individuelle).

1. Le Service des données et des études statistiques est le service statistique des ministères chargés de l'Environnement, de l'énergie, de la construction, du logement et des transports.

Figure 1. L'immobilier résidentiel dans tous ses états



L'acte de construction, mais aussi celui d'agrandir ou de transformer, est très réglementé en France : il est régi selon des principes généraux, comme les codes de l'urbanisme ou de l'habitation, ou spécifiques comme les schémas de cohérence territoriale ou les plans locaux d'urbanisme. Avant toute mise en chantier, le porteur de projet de construction, une personne physique ou morale appelée « pétitionnaire », doit formuler une demande d'autorisation de construction auprès de la mairie de la commune. Celle-ci en délègue l'instruction à un centre instructeur afin qu'il lui propose une décision. Cette étape est importante, car elle préfigure le plus souvent une construction effective, appelée « mise en chantier ».

Bien qu'il puisse y avoir des exceptions, un projet impliquant plusieurs logements, immeuble collectif ou maisons individuelles groupées, est le plus souvent porté par une personne morale, publique ou privée (bailleur social ou promoteur immobilier). Les logements correspondants sont vendus selon divers degrés de maturité, allant de la réservation sur plan, à l'achat après achèvement et raccordement aux réseaux d'énergie.

« Dès l'installation du premier occupant, le local perd son statut de « logement neuf ». C'est alors qu'il est l'objet d'évènements successifs de natures très diverses. »

du premier occupant, le local perd son statut de « logement neuf »². C'est alors qu'il est l'objet d'évènements successifs de natures très diverses.

Un propriétaire peut occuper son local : il est qualifié d'« accédant à la propriété » lorsqu'il en supporte encore les frais d'emprunt

immobilier. Tous les occupants ne possèdent pas le bien qu'ils occupent : une part non négligeable est locataire (**encadré 1**) et verse à ce titre un loyer à un bailleur qu'on dira du secteur libre pour le parc privé, et social pour les habitations à loyer modéré et régulier³.

Un bâtiment est un bien durable. Toutefois il arrivera toujours au cours de son « cycle de vie » un moment où, pour des raisons physiques ou réglementaires, il sera nécessaire de le rénover : on cherchera à en améliorer les performances énergétiques, les caractéristiques physiques (motifs de salubrité) ou les fonctionnalités, par exemple pour l'adapter aux seniors ou aux personnes atteintes d'un handicap physique.

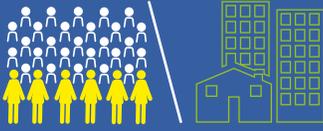
Les usages d'un local donné ne sont pas figés au cours du temps. Ainsi, nombre de professions libérales exercent dans un appartement originellement à vocation d'habitation. Inversement, les aspirations de certains ménages en matière de logement les conduisent à s'engager dans des projets de reconversion de locaux d'usine en appartement spacieux et ou lumineux.

De même, un logement peut être occupé à divers titres. Bien que la majeure partie du parc existant le soit comme résidence principale, une autre est recensée comme vacante par défaut d'occupant. D'autres logements accueillent des occupants de façon plus occasionnelle ou en tant que résidence secondaire.

2. Le statut dont il est question ici est celui déterminé par la Statistique publique : il ne doit pas être confondu avec celui établi par les services fiscaux.

3. Il existe d'autres statuts d'occupation, beaucoup plus anecdotiques, tels que l'occupation à titre gratuit, ou encore l'usufruit qui se caractérise juridiquement par la possession du droit d'usage du logement uniquement.

Encadré 1. Chiffres-clés du logement en France



La France hors Mayotte compte 66,7 millions d'habitants et 36,6 millions de logements dont 82% de résidences principales.

Sources : Estimations annuelles de population légales et Estimations annuelles du parc de logements, 2019, Insee



D'ici 2070, la population française devrait augmenter de 9,5 millions d'habitants (si les tendances démographiques récentes se poursuivent).

Source : Projections de population à horizon 2070, Insee



35 % des occupants sont locataires, soit + 4 points p/r à l'Union européenne, - 13,5 points p/r à l'Allemagne.

Source : EU Survey on Income and Living Conditions*, 2018, Eurostat



40 % des résidences principales en location relèvent de bailleurs sociaux.

Source : Fidéli, 2016, Insee



Le logement, le chauffage et l'éclairage représentent 20 % de la consommation des ménages et un patrimoine de 3 800 milliards d'€.

Source : Comptes nationaux, 2018, Insee



Les administrations publiques ont versé 40 milliards d'€ d'aides dont la moitié en prestations sociales.

Source : Compte satellite du logement, 2018, SDES



8 % des ménages vivent dans un logement surpeuplé (16 % dans l'habitat collectif).

Source : Enquête logement, 2013, Insee



21 % des logements sont considérés par leurs occupants comme difficile ou trop coûteux à bien chauffer.

Source : SRCV*, 2017, Insee



11 % des résidences principales sont mal isolées (fuites dans la toiture, humidité ou moisissure dans les cadres de fenêtre ou le sol).

Source : SRCV*, 2017, Insee

* L'enquête statistique sur les ressources et conditions de vie (Dispositif SRCV) est la partie française du système communautaire EU-SILC (European union-Statistics on income and living conditions).

Faisant écho à la multiplicité des états d'un local à vocation potentiellement résidentielle, l'observation statistique du logement ne se réduit pas – loin s'en faut – à celle de l'habitat. Au cours de son histoire encore récente, la statistique publique française, en l'occurrence l'Insee et le Service des données et études statistiques (Sdes), s'est progressivement dotée d'un système d'information permettant de se placer à des points de vue nécessairement distincts mais complémentaires.

📍 IL ÉTAIT UNE FOIS UN TERRAIN...

Disposer d'un terrain à bâtir est une condition nécessaire et originelle à l'acte de construction neuve. L'enquête sur les prix des terrains et du bâti (EPTB), mensuelle et exhaustive, est menée auprès des ménages pétitionnaires. L'enquête recueille des informations sur les terrains destinés à la construction d'une maison individuelle : mode d'acquisition, superficie, date d'achat, prix ; la maison est aussi l'objet de questionnements sur son prix prévisionnel⁴, le degré de finition des travaux et leur coordinateur, mais aussi sur sa future performance thermique. Enfin l'enquête s'intéresse aux caractéristiques du détenteur du permis de construire : sa catégorie socioprofessionnelle, son âge, la taille de son ménage, son statut de primo-accédant, etc.

Il n'existe aucune autre source que l'EPTB fournissant des informations sur les prix des terrains et sur leur rapport avec les prix des maisons qui y sont construites. Les sources

« Il n'existe aucune autre source que l'EPTB fournissant des informations sur les prix des terrains et sur leur rapport avec les prix des maisons qui y sont construites. »

notariales peuvent apporter un éclairage sur les autres transactions foncières, mais les données contenues sont parfois lacunaires, disponibles plus tardivement et ne donnent pas d'information sur le projet de construction. De ce fait, le système statistique français capte mal les spécificités des mutations foncières⁵ préalables à la construction d'immeubles collectifs.

📍 SIT@DEL : UN RECENSEMENT « EN TEMPS RÉEL » DES LOCAUX NEUFS

Depuis 1972, l'administration française dispose d'un répertoire des demandes d'autorisation d'urbanisme, désormais appelé Sit@del. Sa vocation ne se limite pas à la production de statistiques publiques : il sert aussi à réaliser des contrôles, à établir les taxes d'urbanisme, à suivre des changements relatifs aux propriétés bâties dans le cadre de l'assiette de la fiscalité directe locale ou à suivre certaines politiques publiques basées sur la construction neuve. Sit@del est néanmoins administré par le service statistique ministériel en charge du logement (Sdes).

4. Le prix collecté au moment de l'enquête peut être encore un approximatif et être l'objet de révisions jusqu'au moment de la livraison de la maison individuelle.

5. La mutation d'un bien immobilier désigne un changement de propriétaire (transfert de la propriété), à titre gratuit ou à titre onéreux.

Pour une très large part, les données qui alimentent Sit@del proviennent des formulaires de permis de construire de locaux à vocation résidentielle ou autre : elles sont donc en principe exhaustives et concernent potentiellement des logements et des locaux professionnels⁶. Tout projet de construction neuve ou de transformation d'une construction nécessite le dépôt d'un permis de construire auprès de la mairie de la commune où se situe le terrain ; le pétitionnaire renseigne un formulaire Cerfa, lequel est traité par les services instructeurs (État, collectivité territoriale) dont relève la commune. Après autorisation, le pétitionnaire pourra soit démarrer les travaux et déclarer l'ouverture de son chantier, soit abandonner son projet et demander l'annulation. La fin des travaux est signalée par une déclaration spécifique, à partir de laquelle est vérifiée la conformité au projet initial. Sans être très fréquents, le refus par la collectivité ou l'annulation – après autorisation – d'une demande de permis de construire ne sont pas des événements rares. D'ailleurs, le conjoncturiste, les professionnels de la construction neuve ou encore les décideurs nationaux suivent toujours le nombre d'annulations par le pétitionnaire avec intérêt parce qu'ils renseignent sur le « climat » dans le secteur de la construction.

Les mouvements relatifs à la vie du permis (dépôts, autorisations, annulations, modificatifs, mises en chantier, achèvements et conformité des travaux) sont exploités à des fins statistiques. Le Sdes réceptionne environ 200 000 déclarations par mois. Les autorisations sont transmises par les services instructeurs dans les six mois après le prononcé. Les déclarations de mises en chantier ou d'achèvement et de conformité des travaux sont faites par les pétitionnaires ; leur remontée est beaucoup plus tardive et intervient généralement dans les dix-huit mois après l'ouverture de chantier, bien qu'au bout de 12 mois, encore 30 % d'information manquent sur les logements commencés.

La conjoncture de la construction neuve de logements est appréciée au travers d'estimations en date réelle ventilées par échelon géographique. Les séries statistiques correspondantes visent à retracer dès le mois suivant les autorisations et les mises en chantier. Majoritairement constituées de données collectées, elles intègrent une estimation de l'information non remontée. Ces séries sont actualisées chaque mois pour prendre en compte les nouvelles informations.

En parallèle, l'enquête sur la commercialisation des logements neufs (ECLN) est une opération d'ampleur réalisée chaque trimestre par téléphone auprès des promoteurs immobiliers pour le suivi de l'activité du secteur. L'ECLN permet de connaître les volumes des mises en vente et des réservations, les prix à la réservation⁷, les stocks disponibles à la vente et leur délai d'écoulement. L'enquête est exhaustive sur son champ, celui des programmes de plus de cinq logements, maisons individuelles sur permis groupés ou immeubles collectifs.

Sit@del et ECLN ont en commun de permettre un suivi fiable des flux de construction des logements neufs puis de leur commercialisation, quels qu'en soient les acquéreurs (particuliers, investisseurs, bailleurs sociaux⁸).

6. Pour ne parler que des deux principales destinations desdits locaux.

7. Hors frais de notaire et frais divers.

8. Il s'agit de la vente en bloc, qui désigne l'acquisition d'un immeuble entier ou d'un programme immobilier entier par un seul acquéreur. Celle-ci est effectuée au profit de bailleurs sociaux ou de gros investisseurs privés comme des banques ou des assureurs.

ESTIMER LE PARC DE LOGEMENTS...

Dans un contexte d'augmentation tendancielle de la population française, apprécier les capacités de logement suppose d'être au moins en mesure de dénombrer les logements, neufs ou anciens, bâtis sur le territoire. Combien y a-t-il de logements en France ? Combien d'entre eux sont occupés ? Pour répondre à ces questions fondamentales, les estimations annuelles du parc de logements font la synthèse et la mise en cohérence de plusieurs sources statistiques, la principale étant actuellement le recensement de la population.

En France métropolitaine, l'estimation mobilise également d'autres sources (voir *infra*) : les fichiers de la taxe d'habitation, le répertoire d'immeubles localisés du recensement (RIL) et l'enquête Logement (ENL), etc. Pour les départements et régions d'outre-mer, l'estimation est principalement réalisée à l'aide du recensement et des enquêtes sur le logement de 2006 et 2013.

L'estimation intègre l'information pour laquelle chaque source est jugée la plus pertinente et précise, et l'exercice consiste à rendre cohérent l'ensemble des sources utilisées. Elle se déroule en plusieurs étapes, selon un protocole de révision similaire à celui des comptes nationaux. Les données relatives aux trois dernières années sont provisoires ; elles sont ensuite révisées année après année, jusqu'au chiffrage définitif établi au travers de la compilation des enquêtes annuelles de recensement⁹.

« Cet exercice d'estimation est crucial, car il permet en particulier d'alimenter le compte satellite du logement, qui est un élément important des comptes nationaux. »

L'estimation établit un nombre de logements total, déclinable par **catégorie** (résidences principales, secondaires, occasionnelles ou vacantes), par **type de logement** (habitat collectif ou individuel), par **statut d'occupation** (propriétaires, locataires du secteur libre ou social, autres) et par tranche d'unité urbaines (rural ou urbain). Cet exercice d'estimation est crucial, car il permet en particulier d'alimenter le compte

satellite du logement, qui est un élément important des comptes nationaux (voir *infra*). De nombreuses enquêtes du SSP se calent sur l'estimation du parc de logement, de sorte que les enquêtes produites par les services de la statistique publique sont cohérentes en termes de nombre de logements pour une période de référence donnée.

... AVEC LE RECENSEMENT, L'ENQUÊTE LOGEMENT ET BIEN D'AUTRES SOURCES

Le recensement de la population (RP) permet d'établir la population légale de chaque commune en France, et fournit également des informations sur leurs caractéristiques : âge, profession, moyens de transport utilisés, conditions de logement, etc. À travers une meilleure connaissance de la population et de son évolution, le RP permet par exemple de quantifier les besoins en logements et se présente à ce titre comme un outil central de l'aménagement du territoire. Il aide en outre à caractériser les logements recensés selon

9. Ce qui est la conséquence de la méthodologie très spécifique de la France pour le recensement de population, conduit tous les ans sur un échantillon rotatif de logements.

plusieurs critères : catégorie de logement, type de construction (bâtiment d'habitation isolé, groupé, à usage autre que d'habitation, etc.), type de logement, période d'achèvement de la construction, taille, desserte par un ascenseur, statut d'occupation, ancienneté d'emménagement des occupants.

L'enquête Logement (ENL) est la grande enquête statistique auprès des ménages consacrée à la problématique du logement en France. Elle est la seule source suffisamment détaillée pour ventiler les résidences principales par filière et par type de logement. Ainsi, la connaissance d'une statistique clé telle que le taux de propriétaires de leur résidence principale repose pour beaucoup sur l'information collectée par l'ENL.

Le volet social du parc de logements est également très finement observé *via* le répertoire des logements locatifs des bailleurs sociaux (RPLS). Succédant à un dispositif de collecte de données agrégées auprès d'un nombre restreint de bailleurs sociaux, le RPLS a été mis en place en 2011 sur un champ plus large : les données collectées, logement par logement, permettent de connaître la structure au 1^{er} janvier du parc locatif social en France, ainsi que les modifications intervenues au cours de l'année écoulée : nouvelles constructions, réhabilitation, mouvements, etc. Les organismes interrogés déclarent l'ensemble des logements dont ils ont la pleine propriété, ou sur lesquels ils ont un bail de longue durée, à construction ou à réhabilitation, ou dont ils ont l'usufruit¹⁰, que ceux-ci ouvrent ou pas à leurs occupants un droit à l'aide personnalisée au logement (APL)¹¹.

Par ailleurs, les sources administratives, et en particulier fiscales constituent des éléments structurants pour l'estimation du parc de logements en France de par leur caractère exhaustif et la qualité des informations qu'elles contiennent. Elles sont en effet propres à fournir une information dense et précise sur les caractéristiques des logements et celles de leurs occupants.

Historiquement, le fichier Filocom¹², produit sous la maîtrise d'ouvrage du service statistique ministériel en charge du logement, a rempli ce rôle et permis nombre d'analyses aux niveaux national et local. Depuis 2016, l'Insee diffuse le fichier démographique des logements et des individus (Fidéli), qui vise à mettre en cohérence les grandes sources fiscales sur les logements et les individus. Fidéli peut être vu comme un registre statistique fiable et complet associant logements et occupants ; il alimente de nombreux points névralgiques du système d'information statistique français : base de sondage pour les enquêtes auprès des ménages¹³, source alimentant l'Échantillon Démographique Permanent ou les indicateurs de mobilité résidentielle¹⁴, etc.

LE LOGEMENT, ÉLÉMENT DU PATRIMOINE DES MÉNAGES

Une fois que l'on a établi une image complète du stock de logements en France, on va chercher à comprendre dans le détail le mécanisme d'acquisition de ces logements par les ménages. La dynamique des prix des logements est cruciale de ce point de vue.

10. Ils ne déclarent pas les logements qu'ils gèrent sans détenir de droit réel immobilier ou sans en avoir l'usufruit.

11. Les loyers de travailleurs, les résidences sociales et les foyers logement sont inclus dans le RPLS depuis la collecte 2020.

12. Le Fichier des Logements par Communes est un fichier construit par la DGFIP (Direction Générale des Finances Publiques) pour les besoins du Ministère en charge du Logement.

13. Voir l'article de Sébastien Faivre, Nicolas Paliot, Patrick Sillard et Ludovic Vincent sur la refonte de l'échantillon-maître dans ce même numéro.

14. Entre parcs social et privé notamment.

Le logement a la particularité d'être à la fois un bien de consommation, mais également, pour les ménages propriétaires de leur logement, un élément constitutif de leur patrimoine (Trannoy, 2018). Le niveau des prix des logements est non seulement révélateur du dynamisme ou des tensions du marché immobilier, mais il a également un impact sur l'accès à la propriété et potentiellement sur le niveau des loyers. Les prix de l'immobilier sont donc particulièrement suivis dans le débat public, non sans multiplier les sources et les concepts : prix prévisionnel (EPTB), prix à la réservation (ECLN), prix de réalisation, prix de transaction, etc. Les statistiques produites sont conçues pour retracer l'évolution des prix à qualité comparable, pas leur niveau.

📍 COMMENT ÉVOLUENT LES « PRIX » DES LOGEMENTS NEUFS ? —

Parce que ce n'est pas prévu dans les documents administratifs qui en définissent le contenu, le répertoire Sit@del ne comporte aucune information sur le prix d'un projet immobilier. L'objectif quasiment exclusif de l'enquête sur le prix de revient des logements neufs (PRLN) est de permettre le calcul de l'indice du coût de la construction (ICC). Bien que consacrée par l'usage, la dénomination de cet indice est impropre puisqu'il s'agit d'un indice de prix à la production. Il est fondé sur l'observation des marchés conclus entre les maîtres d'ouvrage et les entreprises de construction, à l'exclusion des autres composantes du prix de revient des logements (terrain, frais annexes de promotion, frais financiers, etc.). Le coût de la construction proprement dit est appréhendé par d'autres indicateurs, notamment les « index BT » utilisés pour les révisions des prix des marchés de construction.

L'ICC a de nombreux usages : calcul de l'indice des loyers commerciaux, indexation des baux commerciaux, suivi de l'évolution des prix dans le secteur de la construction. Cet indice est également indispensable aux comptes nationaux comme déflateur pour évaluer l'activité de construction, pour mesurer l'effet qualité dans l'estimation de la progression du volume de la FBCF¹⁵ en logement (compte satellite du logement) ou encore pour l'établissement des comptes de patrimoine.

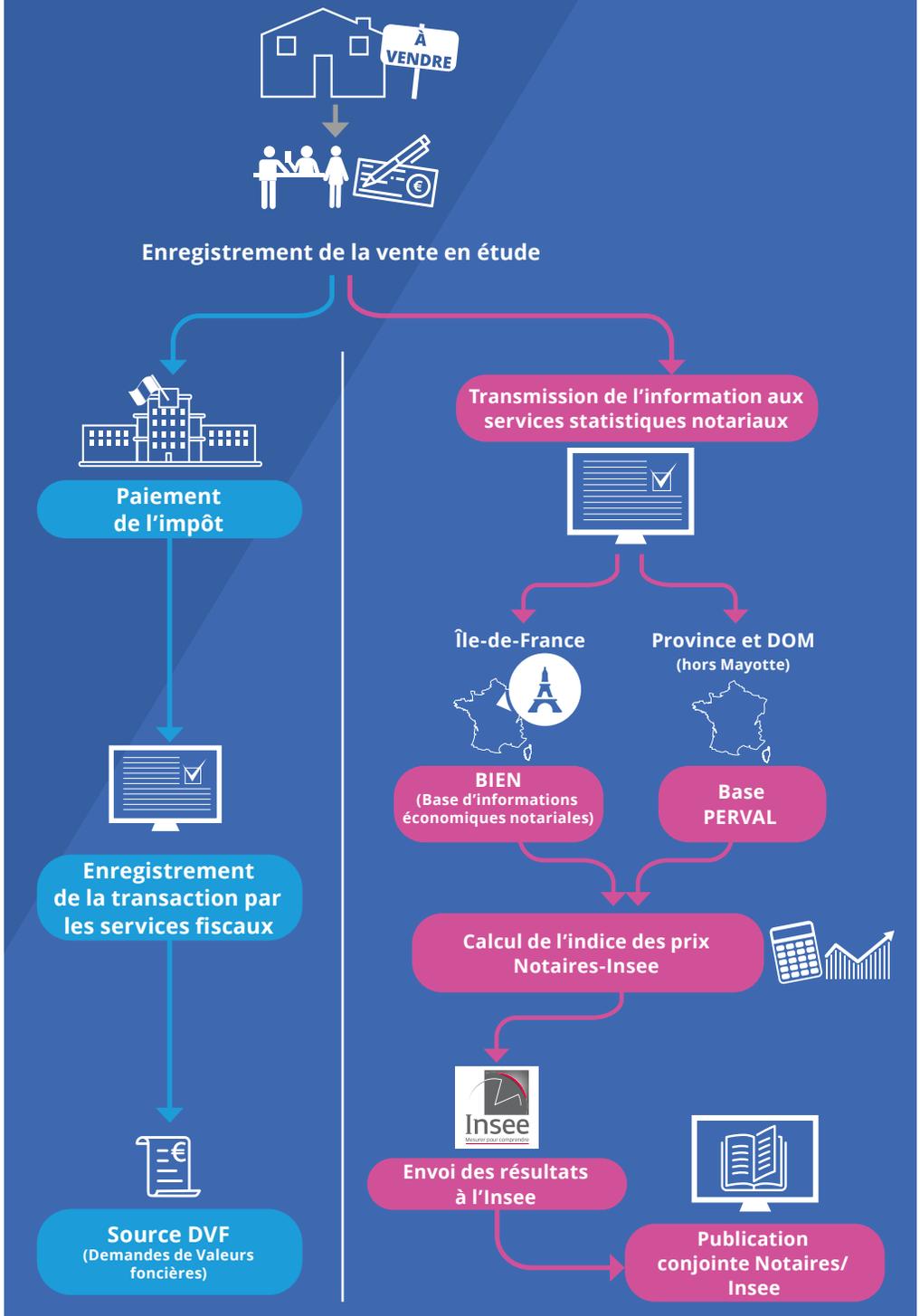
L'enquête PRLN collecte des informations techniques et financières sur les marchés signés : surface construite, nature du contrat de construction, dates du marché, du devis et des travaux, et éventuelles indexations de prix. Elle vise aussi à décrire de la manière la plus détaillée les prestations prévues dans les marchés ainsi que les prix associés. L'enquête renseigne également sur la forme et le nombre de niveaux habitables de la construction, le type de chauffage (source d'énergie, émetteur et producteur de chaleur, système de distribution) et les labels associés. Elle est à ce dernier titre utilisée pour des statistiques spécifiques sur les consommations d'énergie des ménages menées par le Centre d'études et de recherches économiques sur l'énergie (Ceren).

📍 ET CEUX DES LOGEMENTS ANCIENS ? —

L'Office des Notaires constitue des bases de données dans le cadre de sa mission de service public. Ces bases sont alimentées par des informations descriptives très fraîches et détaillées, relatives aux transactions immobilières : le suivi trimestriel du montant des transactions, mais également des caractéristiques des biens, permettent de calculer l'indice des prix des logements anciens Notaires-Insee (*figure 2*).

15. FBCF : formation brute de capital fixe.

Figure 2. Deux systèmes parallèles pour décrire le prix de vente des logements anciens



La méthode des prix hédoniques a été élaborée conjointement par l'Insee et le notariat dans le cadre d'un conseil scientifique. Elle vise en particulier à tenir compte de l'hétérogénéité des biens échangés, et ainsi neutraliser les évolutions liées à des différences dans la structure des échanges.

La principale fragilité de la source notariale est la relative faiblesse de son taux de couverture, puisque l'ensemble des transactions n'est *de facto* pas remonté aux offices statistiques des notaires : cela concerne en particulier les territoires ruraux.

« D'autres sources, de nature proche mais quasi-exhaustives celles-ci, renseignent également sur l'ensemble des transactions immobilières enregistrées en France dans les cinq dernières années. »

D'autres sources, de nature proche mais quasi-exhaustives celles-ci, renseignent également sur l'ensemble des transactions immobilières enregistrées en France dans les cinq dernières années. Les **Demandes de Valeurs Foncières (DVF)** produites par la direction générale des Finances publiques (DGFIP) décrivent les transactions immobilières intervenues au cours des cinq dernières années¹⁶. Les données contenues

sont issues des actes notariés et des informations cadastrales et résultent du recouvrement de l'impôt sur les mutations à titre onéreux : les mutations gratuites en sont donc exclues. Ces données sont accessibles au grand public en *open data*¹⁷ depuis avril 2019. Elles permettent par exemple de confirmer les volumes de transactions estimés tous les trimestres au travers des données notariales combinées avec les agrégats du recouvrement de l'impôt transmis par la DGFIP. Bien que théoriquement exhaustives, elles arrivent beaucoup plus tardivement que les sources notariales, et, en dehors de l'enrichissement au travers des données fiscales sur les locaux, elles ne fournissent qu'une description très frustrante des biens échangés. Cette source fait l'objet d'un intérêt légitime de la part des services producteurs d'information sur le marché de l'immobilier ; du côté du Service statistique public, si les DVF peuvent servir pour décrire les volumes de transactions, il reste à évaluer la fiabilité des montants de ces transactions.

Le prix de l'immobilier résidentiel est une donnée d'analyse essentielle, non seulement parce qu'elle synthétise beaucoup de facteurs macroéconomiques – y compris à l'échelle locale – mais aussi parce qu'elle est en lien fort avec un enjeu majeur de société : l'accès au logement. Ainsi, le prix des logements a un impact direct sur le budget des ménages accédant à la propriété¹⁸, mais également sur celui des locataires, car les loyers, bien que largement régulés en France, sont corrélés avec les prix des logements neufs et anciens.

COMBIEN SE LOGER COÛTE-T-IL ?

De manière générale, le coût du logement est une donnée fondamentale pour l'évaluation des conditions de vie des ménages français¹⁹. Mesurer correctement les loyers que doivent acquitter les locataires est essentiel du point de vue du budget des ménages.

16. À l'exception de l'Alsace-Moselle et de Mayotte.

17. Voir le site <https://www.data.gouv.fr/fr/datasets/demandes-de-valeurs-foncieres/>.

18. C'est-à-dire les ménages endettés pour l'acquisition de leur résidence principale.

19. Voir par exemple à ce sujet le post de blog de l'Insee :

<https://blog.insee.fr/mais-si-linsee-prend-bien-en-compte-le-logement-dans-linflation-et-au-bon-niveau/>.

Pour ce faire, on distingue, parmi les bailleurs de logements, deux grandes catégories : les bailleurs sociaux, qui mettent en location des logements dans des conditions réglementées, en particulier en matière de loyers (voir *supra* le RPLS), et les autres qui appliquent en général les loyers de marché. Pour cela, deux enquêtes statistiques trimestrielles de l'Insee permettent de mesurer non pas les niveaux mais l'évolution des loyers au cours du temps dans les secteurs libre et social.

L'**enquête Loyers et Charges** interroge, cinq trimestres d'affilée un échantillon de logements du parc privé susceptibles d'être loués vides. Sur un principe similaire à celui de l'enquête Emploi, les ménages sont interrogés en face-à-face au premier et au dernier de ces trimestres, et par téléphone pour les trimestres intermédiaires. Le renouvellement de l'échantillon, d'environ 3 000 logements, se fait sur le principe du panel rotatif.

Dans l'enquête conjoncturelle sur les loyers auprès des bailleurs sociaux (ELBS), il est demandé de renseigner, tous les trimestres, l'évolution des loyers d'un échantillon de logements sélectionnés dans le RPLS (voir *supra*)²⁰. Le renouvellement de l'échantillon se fait tous les cinq ans.

Dans le cadre de sa mission de service public, l'Agence nationale pour l'information sur le logement (Anil) assure le traitement des données de loyers du parc privé collectées par des observatoires locaux agréés et les met à la disposition du Sdes chaque année depuis 2019.

D'autres enquêtes de l'Insee auprès des ménages permettent de mesurer ponctuellement le coût du logement, telles que le dispositif de statistiques sur les revenus et les conditions de vie des ménages (SRCV), qui se préoccupe de cette question sous l'angle de la pauvreté en conditions de vie, mais surtout l'enquête Logement.

L'HABITAT, UN SUJET COUVERT PAR DES DISPOSITIFS ÉPROUVÉS

Les questions structurelles liées au logement sont éclairées par l'enquête Logement, dispositif central dans la connaissance du logement en France. Comme évoqué précédemment, elle aide à dénombrer le parc de logements et contribue à l'évaluation du coût du logement.

« Il est donc possible, grâce aux données collectées au fil du temps, de constituer un panorama de l'évolution des conditions de logement en France et de constater les tendances sur longue période en matière d'amélioration de la qualité de l'habitat. »

L'enquête Logement permet aussi de préciser la qualité de l'habitat, que ce soit au travers d'une description objective ou d'un ressenti, et intègre depuis quelques millésimes des préoccupations environnementales.

L'enquête Logement a une longue histoire : depuis sa première édition en 1955, le questionnaire a assez peu évolué. Il est donc possible, grâce aux données collectées au fil du temps, de constituer un panorama de l'évolution des conditions de logement en France et de constater les tendances sur longue période en matière d'amélioration de la qualité de l'habitat.

20. La collecte annuelle visant à actualiser le RPLS au 1^{er} janvier inclut, si le logement est occupé, des informations de loyer et de surface retenues pour son calcul.

L'enquête s'inscrit dans la tradition des grandes enquêtes statistiques de l'Insee, avec un questionnaire très détaillé, abordant les caractéristiques physiques du logement, la qualité de l'habitat, le statut d'occupation, les difficultés d'accès au logement, les questions de mobilité résidentielle, etc. Le questionnement est long, avec des durées d'entretien en face-à-face au domicile des personnes enquêtées excédant assez fréquemment une heure.

Contrairement à de nombreuses enquêtes de la statistique publique française, l'enquête Logement n'est adossée à aucun référentiel européen. Il n'y a donc pas eu d'effort d'harmonisation et de coordination entre États-membres, et ce parce que les équivalents sont rares. Seuls la Grande-Bretagne (au travers de l'*English Housing Survey*²¹) et les Pays-Bas (avec le WoON²²) réalisent une enquête auprès des ménages similaire à l'enquête française. De manière générale, s'agissant particulièrement des conditions de logement vues sous l'angle des privations, la source européenne de référence passe par le dispositif EU-SILC (SRCV en France) ; les données sur le logement qui y sont collectées permettent d'alimenter des indicateurs de pauvreté matérielle²³. La taille de l'échantillon de SRCV n'est toutefois pas comparable à celle de l'enquête Logement, et offre peu de possibilités d'analyses par sous-populations.

Dans les années récentes, la saturation du réseau d'enquêteurs de l'Insee a progressivement conduit le SSP à ajuster la périodicité de l'enquête Logement, qui était d'environ 5 ans jusqu'à l'édition 2006. Les millésimes suivants sont ceux de 2013 et de 2020, le champ du dernier étant de surcroît limité à la France métropolitaine.

Les prochaines éditions devraient rompre avec l'exclusivité de l'enquête en face-à-face pour ouvrir la voie à l'Insee de la collecte multimode (internet, téléphone, et ultimement en face-à-face). Tout en intégrant une filière technique standardisée²⁴, les responsables de l'enquête Logement devront relever de nouveaux défis, en termes de méthodologie de collecte (assurer l'intelligibilité et l'adéquation d'un long questionnaire avec internet et le téléphone) comme de méthodologie de traitements statistiques (corriger les effets de mode²⁵, pour calculer les indicateurs traditionnellement dérivés de l'ENL sans rupture de série). Ce passage au multimode est une opportunité pour l'enquête Logement, en ce qu'elle doit permettre de collecter de l'information auprès de publics traditionnellement difficiles à joindre en face-à-face.

Par ailleurs, recourir à des sources de données alternatives devrait permettre d'alléger le questionnaire. De ce point de vue, l'enquête néerlandaise sur le logement donne des indications sur la voie à suivre : elle dispense en effet les ménages de questions sur leur consommation en énergie en mobilisant par ailleurs les données détenues par les fournisseurs d'énergie.

21. <https://www.gov.uk/guidance/english-housing-survey-guidance-and-methodology>.

22. <https://www.cbs.nl/en-gb/our-services/methods/surveys/korte-onderzoeksbeschrijvingen/netherlands-housing-survey-woon->.

23. Voir le règlement européen n°1177/2003 dans les références juridiques.

24. Voir les articles parus dans le numéro N3 du Courrier des statistiques, portant sur l'outil de conception de questionnaire (Cotton et Dubois, 2019) et sur le générateur d'instruments de collecte (Koumarianos et Sigaud, 2019).

25. L'information collectée par internet ou par téléphone peut différer de celle collectée habituellement en face-à-face : les répondants n'ont pas le même profil, et un même individu ne fournit pas nécessairement la même information selon le canal de réponse.

COMMENT VALORISER ET METTRE EN COHÉRENCE TOUTES CES INFORMATIONS ?

Toutes les sources d'information statistique sur le logement et la construction présentées précédemment sont l'objet d'une politique de diffusion désormais assez classique au sein du SSP : mise en ligne – le cas échéant sous embargo – de séries chronologiques, notes trimestrielles de conjoncture, support de primo-valorisation de nouvelles données, tableaux standards, analyses à forte valeur ajoutée, etc. Certains producteurs de données ont même développé des outils spécifiques de *datavisualisation* ou de cartographie pour favoriser l'appropriation d'indicateurs clefs par les utilisateurs. La mise à disposition des chercheurs et du monde académique de données individuelles est aussi favorisée par l'Insee, le Sdes et même depuis quelques années par le service statistique ministériel de l'administration fiscale, *via* le Centre d'accès sécurisé aux données (CASD)²⁶.

« Compte tenu du poids du logement dans l'économie française, les utilisateurs des informations produites par le SSP expriment un fort besoin de disposer d'indicateurs macroéconomiques de référence. »

Compte tenu du poids du logement dans l'économie française, les utilisateurs des informations produites par le SSP expriment un fort besoin de disposer d'indicateurs macroéconomiques de référence. L'une des réponses à cette demande est la constitution de ce qu'on appelle un « compte satellite ». Est ainsi désigné « un cadre de présentation des données

de l'économie d'un domaine particulier en relation avec l'analyse économique globale du cadre central de la comptabilité nationale. ». Cela concerne par exemple le tourisme, l'éducation, la santé, l'environnement ou encore le logement.

Le Compte satellite sur le logement (CSL) consiste plus particulièrement à expliciter chaque année la dépense totale de logements, aussi bien courante que d'investissement, ainsi que la place du logement dans l'économie nationale. Les éléments publiés, généralement dans le courant de l'été suivant la dernière année sous revue, contiennent également un chiffrage de l'activité immobilière ainsi que des différentes aides au logement (prestations sociales, aides à la production ou au propriétaire). Enfin, les estimations annuelles du parc de logements (voir *supra*) sont également rappelées et déclinées selon les différents modes d'occupation.

La constitution du Compte satellite du logement ne se limite pas à un travail technique de calcul ou de rédaction. Elle passe aussi par une large consultation des différents acteurs publics ou privés du logement et de la construction neuve, au sein du SSP mais également au-delà. Cette concertation, qui prend généralement la forme d'une conférence technique annuelle (longtemps appelée « commission logement »), permet de mieux partager les messages et indicateurs du CSL, au sein de la société française, médias compris.

Autre exemple d'indicateur synthétique, les **indices des prix des logements**, qui mesurent l'évolution des prix des transactions sur les logements, effectuées par les ménages,

26. Voir à ce sujet l'article paru sur le CASD dans le numéro N3 du Courrier des statistiques (Gadouche, 2019).

et qui sont transmis à Eurostat selon les modalités prévues par un règlement européen²⁷, intègrent à la fois les données issues de l'ECLN et des sources notariales, données qui fournissent les prix de ces transactions respectivement dans le neuf et l'ancien.

Mais les données produites par le système statistique du logement ont une utilité plus large.

Les dépenses associées au logement, et en particulier les loyers, sont incluses dans la consommation des ménages et à ce titre participent à l'évaluation du pouvoir d'achat des ménages. Ainsi, les chiffres sur l'évolution des loyers sont intégrés à l'indice des prix à la consommation (IPC). Ces données servent également à l'indice des prix à la consommation harmonisé défini au niveau européen (IPCH) et dont Eurostat agrège les données de chaque pays de l'Union européenne. Par ailleurs, Eurostat a également défini d'autres indices de manière à tenir compte du fait que les propriétaires engagé de façon implicite des dépenses de logement en habitant le logement qu'ils possèdent : il s'agit de l'*Owner-Occupied Housing Price Index* (OOH). L'OOH permet de compléter la mesure fournie par l'IPCH sur le poids des dépenses de logement dans la consommation totale et de tenir compte des propriétaires occupants de leur logement.

TIRER PARTI DE LA GRANULARITÉ DES DONNÉES, DE PLUS EN PLUS FINE

L'ère numérique voit l'émergence de données plus détaillées et plus complètes sur le logement. Ceci constitue une opportunité d'améliorer la mesure de certains phénomènes, et en particulier ceux pour lesquels les notions associées sont directement quantifiables à l'échelle du logement : consommation d'énergie et d'eau, prix et loyers, etc. En voici quelques exemples :

- ❶ on peut mettre directement en relation une valeur de marché observée dans les DVF, avec les caractéristiques du logement, celles de ses habitants et de ses propriétaires avant et après la transaction, données qu'on observe dans Fidéli ;
- ❶ l'évaluation du patrimoine immobilier des ménages avec Fidéli est également un élément prometteur, y compris à l'échelle locale, en ce qu'il permet de comprendre les logiques d'investissement dans les actifs immobiliers des ménages, ainsi que sa répartition sur le territoire national. Il faudra pour cela auparavant approfondir la méthodologie employée, et améliorer l'utilisation par Fidéli des données du foncier bâti.

« L'éclairage des conditions de logement et des mobilités résidentielles, en particulier au niveau local, requiert la recherche continue de sources alternatives. »

L'éclairage des conditions de logement et des mobilités résidentielles, en particulier au niveau local, requiert la recherche continue de sources alternatives : de ce point de vue, Fidéli permet de compléter les

données du recensement, en particulier pour les territoires identifiés en tension²⁸. Fidéli permet de suivre sur deux années successives les logements et leurs occupants et d'analyser plus finement des migrations résidentielles.

27. Voir le règlement européen n° 93/2013 du 1^{er} février 2013 dans les références juridiques.

28. Depuis 2003, la tension du marché immobilier local constitue le critère de classement dans une des zones des dispositifs successifs d'investissement locatif défiscalisé : « Robien », « Duflot », « Pinel », etc.

Cette même source permet d'investiguer également des questions liées aux conditions de logement qui font, au même titre que l'emploi ou le revenu, l'objet d'un besoin croissant d'analyses fines à l'échelle locale.

Toutefois, l'exploitation des sources administratives à des fins statistiques est limitée à la seule information réglementaire, rarement alignée avec les concepts d'intérêt statistique. En particulier, si le besoin de statistiques localisées est fort en matière de mal-logement (Join Lambert, Labarthe, Marpsat et Rougerie, 2011 ; Benvéniste et Calame, 1996), l'apport des données administratives telles que Fidéli reste limité par la difficulté à appréhender la notion au travers des variables disponibles (**encadré 2**), que ce soit au niveau national ou local.

Enfin, l'analyse du logement (et en particulier du marché immobilier) à l'échelle locale nécessite de construire un ensemble de strates locales. C'est le travail qu'a accompli le Sdes en construisant la maille habitat, qui regroupe les communes les plus similaires en termes de logement, sous une contrainte de contiguïté. Ce maillage permet de faire ressortir au mieux les disparités à l'échelle locale, et de caractériser les territoires pour ce qui relève des difficultés d'accès au logement et de l'attractivité. Il constitue en soi une grille d'analyse commune et partagée, aisément mobilisable. La promotion et la poursuite de l'effort méthodologique sont des éléments clés dans la mise en place d'une plus grande territorialisation des statistiques sur le logement.

À TERME, MIEUX ÉCLAIRER LES PROBLÉMATIQUES DE DÉVELOPPEMENT DURABLE

La statistique publique française doit aujourd'hui encore tenir compte de l'émergence de nouveaux enjeux en lien avec le développement durable : lutte contre le réchauffement climatique, soutien des ménages modestes pour la transition énergétique, maîtrise de l'artificialisation.

En 2012, le service statistique des ministères du Logement et de l'Écologie avait réalisé l'enquête sur la performance de l'habitat, équipements, besoins et usages de l'énergie (Phébus) afin de fournir une photographie des performances énergétiques du parc des résidences principales et de les analyser en fonction des caractéristiques des ménages occupants, de leurs équipements ménagers et automobiles et de leurs usages et comportements énergétiques. Phébus avait alors aussi permis d'étudier la problématique de la précarité énergétique.

Ces dernières années, le réchauffement climatique s'est inscrit (et probablement installé) parmi les principales préoccupations quotidiennes des Français. L'enquête sur les travaux de rénovation énergétique des maisons individuelles » (Trémi) réalisée en 2020 sous copilotage du Sdes et de l'Ademe²⁹ renseignera prochainement sur les travaux de rénovation achevés en 2019 dans le parc résidentiel individuel (coût, durée des travaux, mode de réalisation et de financement, etc.), et d'évaluer leur niveau de performance énergétique. Elle permet aussi aux pouvoirs publics d'identifier les leviers et les freins des ménages pour réaliser ce genre de travaux (motivations, éléments déclencheurs, etc.), de mesurer la notoriété et l'usage des dispositifs d'aides à la rénovation en fonction de caractéristiques du logement et de la personne de référence du ménage occupant.

29. Agence de l'environnement et de la maîtrise de l'énergie.

Le dispositif devrait être étendu aux immeubles collectifs (**Tréco**) au cours des prochaines années, ce qui n'est pas sans poser de délicates difficultés méthodologiques : dynamique de rénovation spécifique, distinction entre travaux effectués entre parties privatives et partie collective, accès aux informations adéquates propres aux copropriétés, etc.

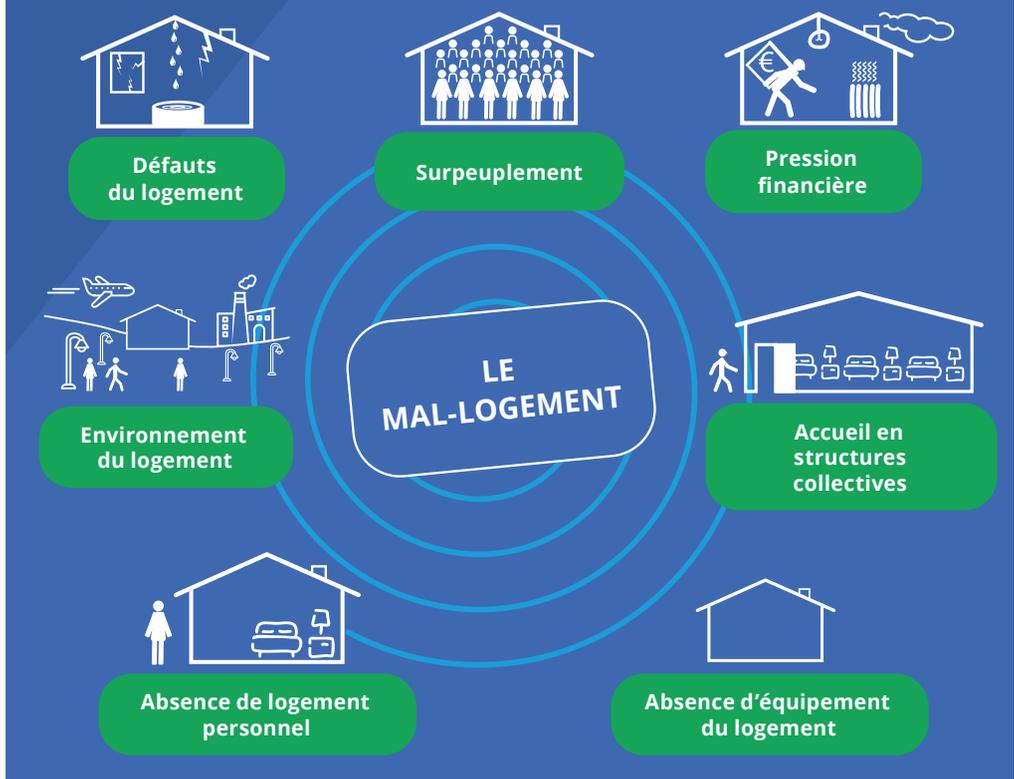
L'exploitation des données d'origine administrative ou fiscale (Crédit d'impôt pour la transition énergétique, Certificats d'économie d'énergie, etc.) sera aussi certainement approfondie. Mais elle pose d'une part de redoutables questions de dédoublement

Encadré 2. Le mal-logement, un phénomène aux formes multiples

Au-delà de la situation des sans-abris, la notion de mal-logement recouvre également :

- les difficultés d'accès au logement : absence de logement personnel, par exemple individus occupant une chambre d'hôtel comme résidence principale ou logeant de façon permanente et contrainte chez des proches ;
- l'absence de confort voire la dégradation de l'habitat ;
- l'inadaptation de l'habitat, en particulier le surpeuplement ;
- les difficultés financières que peuvent engendrer les dépenses propres à l'habitat ;
- la dégradation de l'environnement : nuisances, pollution, problème de voisinage, etc.

Voir (Join Lambert, Labarthe, Marpsat et Rougerie, 2011 ; Benvéniste et Calame, 1996).



des ménages cumulant les dispositifs de soutien public, en l'absence d'un identifiant inter-administratif du logement ou du local. Et d'autre part, il faudra savoir définir la conversion des gestes de rénovation en quantité d'énergie effectivement économisée.

Construire des logements ou des bâtiments à usage non résidentiel a pour conséquence d'artificialiser voire d'imperméabiliser des terrains qui, dans un passé plus ou moins

« L'exploitation des données d'origine administrative ou fiscale sera aussi certainement approfondie. Mais elle pose d'une part de redoutables questions de dédoublonnage des ménages cumulant les dispositifs de soutien public, en l'absence d'un identifiant inter-administratif du logement ou du local. »

lointain, avaient une vocation agricole. Cet acte s'avère rarement réversible et engage sur très long terme la collectivité qui l'autorise : raccordement aux réseaux d'eau, de gaz, d'électricité voire de chauffage urbain, réseau internet ou fibre optique, desserte des transports en commun, construction de nouveaux équipements publics ou privés, etc. Le plan stratégique de juillet 2018 portant sur la biodiversité prévoit d'enrayer l'augmentation des surfaces des bâtiments, infrastructures de transports, parkings, etc. et d'atteindre l'objectif de zéro artificialisation nette d'ici 2050.

L'artificialisation est aujourd'hui encore un concept assez mal défini. Les statisticiens, géographes et chargés d'étude sont engagés, sous l'égide du ministère de la Transition écologique et solidaire, dans une réflexion et des expérimentations pour produire des indicateurs, fondés tantôt sur des informations du SSP (construction neuve, Teruti-Lucas³⁰...), tantôt satellitaires. Le chemin à parcourir paraît aujourd'hui encore assez long avant que les acteurs du débat public puissent disposer de données de référence sur l'artificialisation, à la méthodologie faisant consensus.

30. Enquête annuelle réalisée par les services statistiques du ministère en charge de l'agriculture ayant pour objectif de suivre l'évolution de l'occupation et de l'usage des sols sur tout le territoire national.

BIBLIOGRAPHIE

ALBIZZATI, Colin, POULHES, Mathilde et SULTAN PARRAUD, Joyce, 2017. Caractérisation des espaces consommés par le bâti en France métropolitaine entre 2005 et 2013. In : *Les acteurs économiques et l'environnement*. [en ligne]. 5 décembre 2017. Collection Insee Références. pp. 73-85. [Consulté le 17 juin 2020]. Disponible à l'adresse : https://www.insee.fr/fr/statistiques/fichier/3280938/Enviro17f_D5_Artificialisation.pdf.

BENVÉNISTE, Corinne et CALAME, Pierre, 1996. *Pour une meilleure connaissance des sans-abri et de l'exclusion du logement*. CNIS, Rapport de groupe de travail n°29. ISBN 978-2-11-066393-1.

CGDD (COMMISSARIAT GÉNÉRAL AU DÉVELOPPEMENT DURABLE), 2018. *Comparer le poids du logement en France et en Allemagne : le taux d'effort moyen ne suffit pas*. [en ligne]. Janvier 2018. Édité par le service de la donnée et des études statistiques (Sdes). [Consulté le 17 juin 2020]. Disponible à l'adresse : <https://www.statistiques.developpement-durable.gouv.fr/sites/default/files/2019-03/datalab-29-comparer-poids-logement-france-allemande-janvier2018.pdf>.

CGDD (COMMISSARIAT GÉNÉRAL AU DÉVELOPPEMENT DURABLE), 2019. *Rapport de la Commission des comptes du logement*. [en ligne]. 30 septembre 2019. Édité par le service de la donnée et des études statistiques (Sdes). [Consulté le 17 juin 2020]. Disponible à l'adresse : <https://www.statistiques.developpement-durable.gouv.fr/sites/default/files/2019-10/datalab60-compte-du-logement-2018-septembre2019.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 17 juin 2020]. Disponible à l'adresse : <https://www.insee.fr/fr/statistiques/fichier/4254216/courstat-3-3.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. N°N3, pp. 76-92. [Consulté le 17 juin 2020]. Disponible à l'adresse : <https://www.insee.fr/fr/statistiques/fichier/4254227/courstat-3-7.pdf>.

INSEE, 2017. *Les conditions de logement en France*. [en ligne]. 21 février 2017. Collection Insee Références. Édition 2017. [Consulté le 17 juin 2020]. Disponible à l'adresse : <https://www.insee.fr/fr/statistiques/2586377>.

JOIN-LAMBERT, Marie-Thérèse, LABARTHE, Julie, MARPSAT, Maryse et ROUGERIE, Catherine, 2011. *Le mal-logement*. [en ligne]. Juillet 2011. Rapport d'un groupe de travail n°126, CNIS. [Consulté le 17 juin 2020]. Disponible à l'adresse : http://www.epsilon.insee.fr/jspui/bitstream/1/56513/1/CNIS_rapports_2011_126.pdf.

KOUMARIANOS, Heïdi et SIGAUD, ERIC, 2019. Eno, un générateur d'instruments de collecte. In : *Courrier des statistiques*. [en ligne]. 19 décembre 2019. N°N3, pp. 29-44. [Consulté le 17 juin 2020]. Disponible à l'adresse : <https://www.insee.fr/fr/statistiques/fichier/4254218/courstat-3-4.pdf>.

TRANNOY, Alain, 2018. Le logement : un bien espace-temps. In : *Économie et Statistique*. [en ligne]. 29 octobre 2018. N°500-501-502, pp.5-11. [Consulté le 17 juin 2020]. Disponible à l'adresse : https://insee.fr/fr/statistiques/fichier/3621959/500-501-502_introduction-Trannoy_FR.pdf.

❶ FONDEMENTS JURIDIQUES

Règlement (CE) n°1177/2003 du Parlement européen et du Conseil du 16 juin 2003 relatif aux statistiques communautaires sur le revenu et les conditions de vie (EU-SILC). In : *Journal officiel de l'Union européenne*. [en ligne]. 3 juillet 2003. L165. [Consulté le 25 mai 2020]. Disponible à l'adresse :

<https://eur-lex.europa.eu/legal-content/FR/TXT/PDF/?uri=CELEX:02003R1177-20081211&from=ES>.

Règlement (UE) n°1157/2010 de la Commission européenne du 9 décembre 2010 mettant en oeuvre le règlement n°1177/2003 en ce qui concerne la liste 2012 des variables cibles secondaires relatives aux conditions de logement. In : *Journal officiel de l'Union européenne*. [en ligne]. 10 décembre 2010. L326/3. [Consulté le 25 mai 2020]. Disponible à l'adresse :

<https://eur-lex.europa.eu/legal-content/FR/TXT/PDF/?uri=CELEX:32010R1157&from=PT>.

Règlement (UE) n°93/2013 de la Commission européenne du 1^{er} février 2013 portant modalités d'application du règlement n°2494/95 relatif aux indices des prix à la consommation harmonisés, en ce qui concerne l'établissement d'indices des prix des logements occupés par leur propriétaire. In : *Journal officiel de l'Union européenne*. [en ligne]. 2 février 2013. L33/14. [Consulté le 25 mai 2020]. Disponible à l'adresse :

<http://eur-lex.europa.eu/LexUriServ/LexUriServ.do?uri=OJ:L:2013:033:0014:0016:FR:PDF>.

Règlement (UE) n°792/2016 de la Commission européenne du 11 mai 2016 relatif aux indices des prix à la consommation harmonisés et à l'indice des prix des logements, et abrogeant le règlement (CE) n°2494/95 du Conseil. In : *Journal officiel de l'Union européenne*. [en ligne]. 24 mai 2016. L135/11. [Consulté le 22 avril 2020]. Disponible à l'adresse :

<https://eur-lex.europa.eu/legal-content/FR/TXT/PDF/?uri=CELEX:32016R0792&from=fr>.



Présentation du numéro N4

Poursuivant son exploration des métiers et méthodes de la statistique publique, ce numéro N4 s'intéresse d'abord à une pratique qu'on ne range pas habituellement dans ce domaine : la microsimulation, dynamique ou statique, pratiquée dans un INS ou dans des services ministériels. Deux modèles spécifiques sont détaillés : TRAJECTOIRE, sur le système de retraite et Ines, autour des politiques sociales et fiscales. Quatre papiers sont ensuite consacrés à des références pour le statisticien : en France d'abord, à travers la refonte de la nomenclature des PCS, le programme de refonte du répertoire d'entreprises Sirene et la mise au point du nouvel échantillon-maître. L'expérience suédoise ensuite, sur la modélisation des processus statistiques et son impact organisationnel, apporte l'éclairage externe que nous affectionnons. Enfin, un panorama sur le système d'information du logement en France, dans sa globalité, vient conclure ce numéro.

ISSN 2107-0903
ISBN 978-2-11-151264-1

