PATRIMOINE IMMOBILIER DES MÉNAGES

ENSEIGNEMENTS D'UNE EXPLOITATION DE SOURCES ADMINISTRATIVES EXHAUSTIVES

*Mathias André et Olivier Meslin**

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

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

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

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

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

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

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

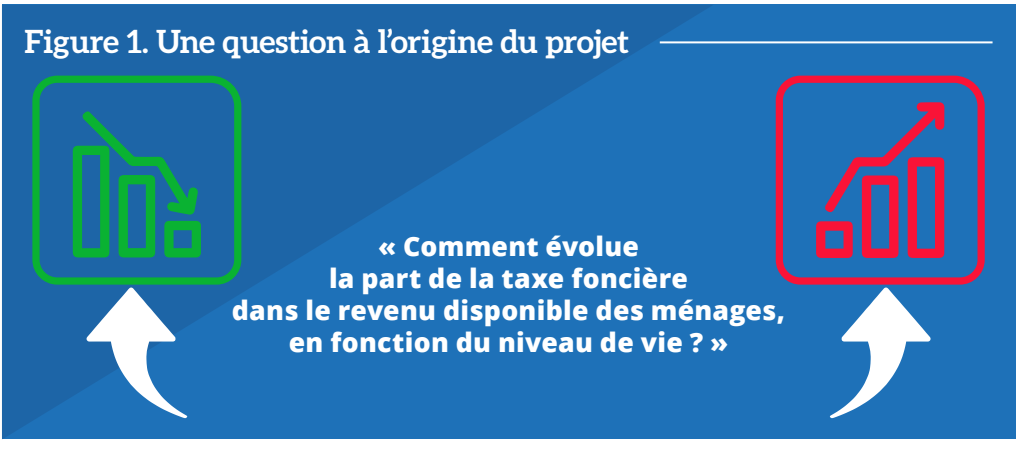
Deux questions sont à l'origine des travaux[1] présentés dans cet article. Une question économique d'abord ***(figure 1)*** : quel est le profil redistributif de la taxe foncière sur les locaux d'habitation ? Porte-t-il davantage sur les ménages aisés, médians ou modestes ? Peu d'études ont abordé cette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

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

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

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

Figure 1. Une question à l'origine du projet



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

DONNÉES ADMINISTRATIVES: UN PAYSAGE QUI A RÉCEMMENT ÉVOLUÉ

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

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

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

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

UN ACCÈS AUX DONNÉES, PROGRESSIVEMENT ÉLARGI ET FACILITÉ

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

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

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

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

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

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

UNE UTILISATION PAR LES STATISTICIENS, FACILITÉE PAR LES ÉVOLUTIONS TECHNIQUES

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

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

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

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

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

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

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

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

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

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

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

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

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

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

(PHASE 1) RÉCUPÉRER LES DONNÉES PERTINENTES

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

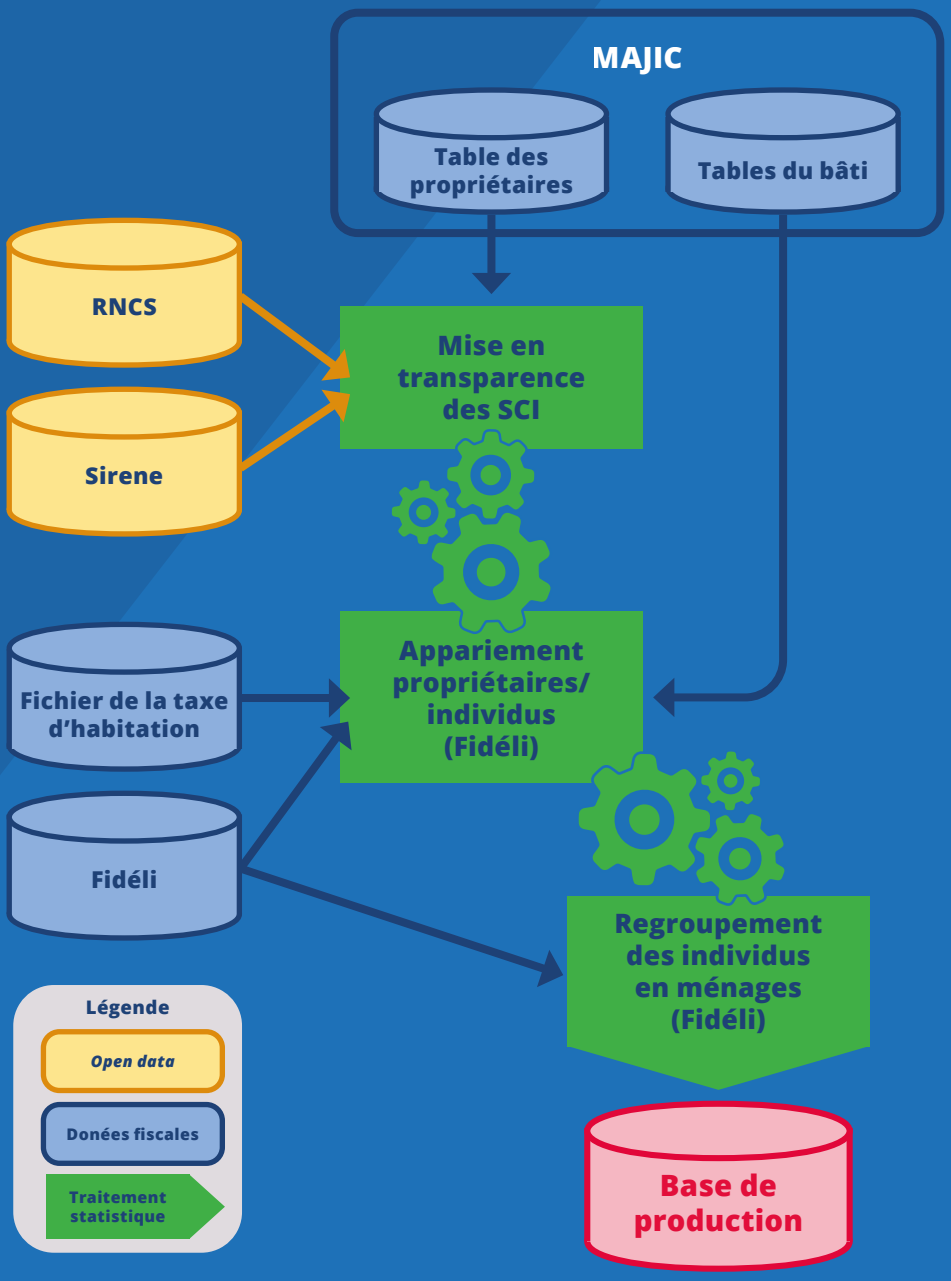
> *« Les difficultés ont plutôt consisté à avoir connaissance de l'existence de données potentiellement utiles. »*

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

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

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

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



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

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

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

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

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

Encadré 1. Les cinq sources du projet

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

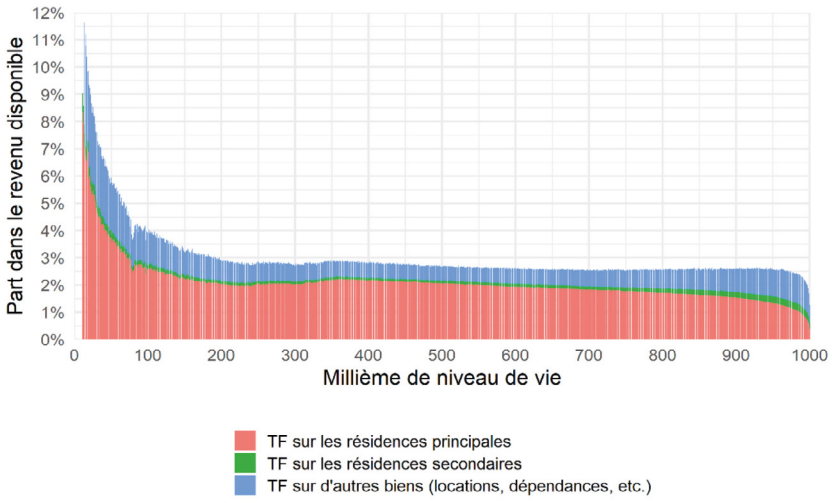
Forts de cette expérience, nous avons souhaité en tirer des enseignements pour le statisticien qui souhaiterait se lancer dans une entreprise similaire. Ces enseignements peuvent se résumer en trois caractéristiques distinctives des sources administratives, et quatre grandes questions que l'on doit se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

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

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

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

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

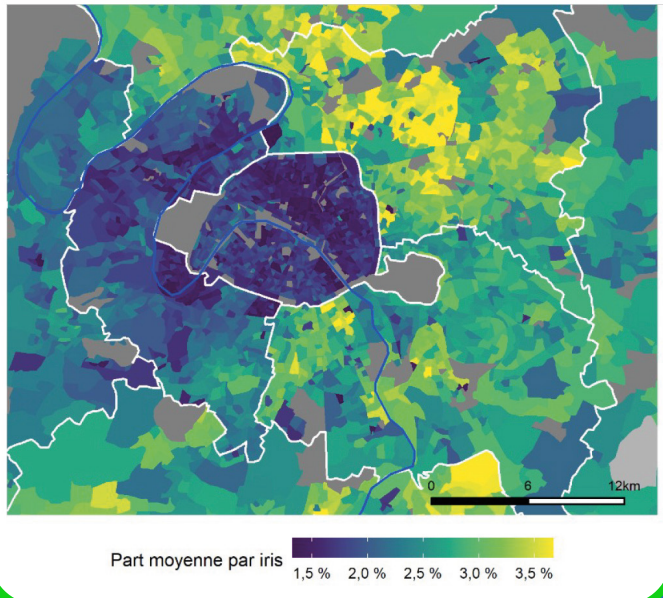


« Le graphique met en évidence que la part de la taxe foncière dans le revenu disponible des ménages imposables à cet impôt décroît en fonction du niveau de vie pour la première moitié de la distribution, puis est stable à environ 2,5 % du revenu disponible pour la seconde moitié de la distribution,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

(André, Arnold et Meslin, 2021)

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

(André et Meslin, 2021)



À QUEL MOMENT OBSERVE-T-ON LES INFORMATIONS?

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

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

DE QUELLE EXHAUSTIVITÉ PARLE-T-ON?

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

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

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

18. Caisse nationale des allocations familiales.

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

QUELLE EST LA FIABILITÉ DES INFORMATIONS ADMINISTRATIVES?

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

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

Ce travail d'expertise de la fiabilité implique des questions essentielles pour le statisticien : quelle est la définition administrative de cette variable ? Qui fournit l'information ? Est-elle à jour ? L'administration en a-t-elle contrôlé la qualité ? Y répondre requiert une compréhension fine du processus de gestion qui est à l'origine des données et un dialogue fréquent avec les administrations productrices de données. Enfin, une dernière question va se poser au statisticien qui rapproche ou s'appuie sur des bases administratives.

QUELS SONT LE CHAMP ET L'UNITÉ D'ANALYSE PERTINENTS POUR L'EXPLOITATION STATISTIQUE ?

Il s'agit là de l'étape à laquelle le statisticien n'échappe jamais, car c'est en répondant à cette question qu'il transforme les bases administratives en bases à usage proprement statistique. Dans un processus d'enquête, cette réflexion se situe en amont, au moment de la construction du questionnaire et de la constitution de l'échantillon[19]. Pour l'exploitation des sources administratives, elle est présente à chaque étape, tant au moment de la recherche des sources qu'en filigrane à toutes les étapes de production de la base finale. Dans le présent projet, il a fallu notamment répondre aux interrogations suivantes :

- quels types de biens immobiliers étudier ? Uniquement les logements, ou aussi les dépendances (garages, caves, etc.), voire les locaux industriels et commerciaux ?
- quelles modalités de détention retenir ? Uniquement les biens possédés en nom propre, ou également ceux possédés *via* une société (SCI, SA, SARL, SAS, etc.[20]) ?
- quels droits de propriété considérer ? Uniquement la pleine-propriété ? Que faire des usufruitiers et des nus-propriétaires[21] ?
- quelle unité de patrimoine immobilier étudier ? Au niveau des individus, des foyers fiscaux ou des ménages ?

Ainsi, les différentes « nuances d'exhaustivité », l'absence de production à visée statistique et les différences de définitions des unités d'observation portent à souligner l'importance de mener avec précaution les appariements entre les données d'enquête et les bases administratives ou entre bases administratives[22].

QUELS ENSEIGNEMENTS RETENIR DU PROJET ?

Trois principaux enseignements peuvent être tirés de ces travaux :

- le simple fait d'**avoir accès aux sources administratives n'est en aucun cas suffisant pour en faire des exploitations statistiques rigoureuses.** À partir d'une idée initiale, de nombreuses étapes sont nécessaires pour passer de bases initiales, riches mais hétérogènes et éparses, à une base statistique cohérente. Comme les autres travaux statistiques, les différentes difficultés qu'il a fallu surmonter ont correspondu à un cheminement vers toujours plus d'homogénéité, car les données administratives peuvent adopter des conventions, des définitions et des nomenclatures distinctes. Il revient alors au statisticien de les rapprocher et de les harmoniser avec patience et rigueur ;
- **l'exploitation systématique de données administratives correspond bien à un mode de collecte à part entière pour la statistique publique.** Celui-ci induit néanmoins un changement majeur par rapport aux autres modes : le statisticien ne maîtrise plus la production des données. Cette évolution implique le développement d'une expertise propre : comprendre le fonctionnement des administrations qui produisent les données, échanger avec les producteurs, anticiper les changements. Mais cela souligne également

19. Dans le modèle générique de processus de production statistique (GSBPM), ces aspects sont traités dans les phases de conception, au tout début du process.
20. Société civile immobilière, société anonyme, société à responsabilité limitée ou société par actions simplifiée.
21. La nue-propriété est le droit donnant à son titulaire, appelé nu-propriétaire, la faculté de disposer d'une chose mobilière ou immobilière (en la vendant, la donnant, la léguant, etc.) alors que l'usufruitier dispose seulement du droit d'en avoir l'usage.
22. Voir également l'expérience d'appariement relatée dans (Midy, 2021).

les limites imposées par la nature même de ces données, auxquelles il peut manquer des variables importantes pour l'étude, ou dont le contenu peut être d'une fiabilité relative et fonction de son importance dans les activités quotidiennes des administrations ;

- l'exploitation des données administratives ouvre des perspectives nouvelles à la statistique publique pour au moins deux raisons. D'une part, **l'exhaustivité rend possible l'étude de phénomènes rares**, tels que les patrimoines immobiliers les plus importants ou le croisement fin de plusieurs variables, **et la réalisation d'analyses à un niveau géographique très fin**, ce qui est difficilement réalisable avec des enquêtes. D'autre part, **des productions innovantes deviennent envisageables grâce aux appariements :** l'appariement entre les données cadastrales et le répertoire Fidéli permet de reconstituer la distribution conjointe des revenus et de la propriété immobilière, qui n'existait dans aucune source avec ce niveau de détail. De même, la mise en transparence des SCI permet de dépasser la distinction habituelle entre données sur les ménages et données sur les entreprises et donc d'étudier des phénomènes économiques peu documentés, comme le comportement de recours aux SCI en fonction de la composition du patrimoine et du niveau de revenu des ménages. Ces travaux pourraient aboutir à l'introduction d'un module sur les patrimoines immobiliers dans les fichiers Fidéli.

L'exploitation des données administratives est à la fois un défi et une opportunité. Un défi en ceci qu'elle exige des statisticiens qu'ils acquièrent de nouvelles compétences et renouvellent leurs méthodes. Une opportunité parce que la statistique publique élargit ainsi les informations qu'elle mobilise, source d'une grande richesse pour accomplir sa mission : mesurer et comprendre les phénomènes économiques et sociaux.

BIBLIOGRAPHIE

ANDRÉ, Mathias, ARNOLD, Céline et MESLIN, Olivier, 2021. 24 % des ménages détiennent 68 % des logements possédés par des particuliers. In : *France, Portrait Social.* [en ligne]. 25 novembre 2021. Insee références, édition 2021, pp. 91-104. [Consulté le 15 décembre 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/5432517?sommaire=5435421*.

ANXIONNAZ, Isabelle et MAUREL, Françoise, 2021. Le Conseil national de l'information statistique – La qualité des statistiques passe aussi par la concertation. In : *Courrier des statistiques.* [en ligne]. 8 juillet 2021. Insee. N° N6, pp. 123-142. [Consulté le 15 décembre 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/5398693/courstat-6-art-7.pdf*.

BONNET, Odran, LOISEL, Tristan et OLIVIA, Tom, 2021. *Impact de la crise sanitaire sur un panel anonymisé de clients de La Banque Postale. Les revenus de la plupart des clients ont été affectés de manière limitée et temporaire.* [en ligne]. 3 novembre 2021. Insee Analyses n° 69. [Consulté le 15 décembre 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/5760458*.

CARBONNIER, Clément, 2019. L'impact distributif de la fiscalité locale sur les ménages en France. In : *Économie et Statistique / Economics and Statistics.* [en ligne]. 11 juillet 2019. Insee. N° 507-508, pp. 31–52. [Consulté le 15 décembre 2021]. Disponible à l'adresse : *https://doi.org/10.24187/ecostat.2019.507d.1977*.

COUSIN, Guillaume et HILLAIREAU, Fabrice, 2019. Les données de téléphonie mobile peuvent-elles améliorer la mesure du tourisme international en France ? In : *Économie et Statistique / Economics and Statistics.* [en ligne]. 11 avril 2019. Insee. N° 505-506, pp. 89–107. [Consulté le 15 décembre 2021]. Disponible à l'adresse : *https://www.insee.fr/en/statistiques/fichier/3706269/ES_505-506_EN.pdf*.

FREDON, Simon et SICSIC, Michaël, 2020. Ines, le modèle qui simule l'impact des politiques sociales et fiscales. In : *Courrier des statistiques.* [en ligne]. 29 juin 2020. Insee. N° N4, pp. 42-61. [Consulté le 15 décembre 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/4497070/courstat-4-4.pdf*.

GADOUCHE, Kamel, 2019. Le Centre d'accès sécurisé aux données (CASD), un service pour la data science et la recherche scientifique. In : *Courrier des statistiques.* [en ligne]. 19 décembre 2019. Insee. N° N3, pp. 76-92. [Consulté le 15 décembre 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/4254227/courstat-3-7.pdf*.

HUMBERT-BOTTIN, Élisabeth, 2018. La déclaration sociale nominative. Nouvelle référence pour les échanges de données sociales des entreprises vers les administrations. In : *Courrier des statistiques.* [en ligne]. 6 décembre 2018. Insee. N° N1, pp. 25-34. [Consulté le 15 décembre 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/3647025/courstat-1-6.pdf*.

LAGARDE, Sylvie, 2008. La nouvelle exploitation exhaustive des DADS. In : *Courrier des statistiques*. [en ligne]. Juin 2008. N° 85-86, pp. 65-69. [Consulté le 15 décembre 2021]. Disponible à l'adresse : *https://gallica.bnf.fr/ark:/12148/bc6p06z99f7/f1.pdf*.

LAMARCHE, Pierre et LOLLIVIER, Stéfan, 2021. Fidéli, l'intégration des sources fiscales dans les données sociales. In : *Courrier des statistiques.* [en ligne]. 8 juillet 2021. Insee. N° N6, pp. 28-46. [Consulté le 15 décembre 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/5398683/courstat-6-art-2.pdf*.

LECLAIR, Marie, 2019. Utiliser les données de caisses pour le calcul de l'indice des prix à la consommation. In : *Courrier des statistiques.* [en ligne]. 19 décembre 2019. Insee. N° N3, pp. 61-75. [Consulté le 15 décembre 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/4254225/courstat-3-6.pdf*.

MIDY, Loïc, 2021. Un outil d'appariement sur identifiants indirects : l'exemple sur le système d'information des jeunes. In : *Courrier des statistiques.* [en ligne]. 8 juillet 2021. Insee. N° N6, pp. 82-99. [Consulté le 15 décembre 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/5398689/courstat-6-art-5.pdf*.

RIVIÈRE, Pascal, 2020. Qu'est-ce qu'une donnée ? Impact des données externes sur la statistique publique. In : *Courrier des statistiques.* [en ligne]. 31 décembre 2020. Insee. N° N5, pp. 114-131. [Consulté le 15 décembre 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/5008707/courstat-5-8.pdf*.

SILLARD, Patrick, FAIVRE, Sébastien, PALIOD, Nicolas et VINCENT, Ludovic, 2020. Pour les enquêtes auprès des ménages, l'Insee rénove ses échantillons. In : *Courrier des statistiques*. [en ligne]. 29 juin 2020. Insee. N° N4, pp. 81-100. [Consulté le 15 décembre 2021]. Disponible à l'adresse : *https://www.insee.fr/fr/statistiques/fichier/4497081/courstat-4-6.pdf*.

FONDEMENTS JURIDIQUES

Loi n° 2016-1321 du 7 octobre 2016 pour une République numérique. In : *site de Légifrance*. [en ligne]. Modifié le 23 août 2017. [Consulté le 15 décembre 2021]. Disponible à l'adresse : *https://www.legifrance.gouv.fr/dossierlegislatif/JORFDOLE000031589829/*.

Règlement (UE) 2016/679 du Parlement européen et du Conseil du 27 avril 2016 relatif à la protection des personnes physiques à l'égard du traitement des données à caractère personnel et à la libre circulation de ces données. In : *site EUR-Lex*. [en ligne]. [Consulté le 15 décembre 2021]. Disponible à l'adresse : *https://eur-lex.europa.eu/legal-content/FR/TXT/HTML/?uri=CELEX:32016R0679&from=FR*.