

Le fichier Fidéli

Centre d'Accès Sécurisé aux
Données



Pierre Lamarche, chef de la division Logement

- 01** Le principe en quelques mots
- 02** Quelle exploitation des données ?
- 03** Développements et perspectives

01

Le principe en quelques mots

Les données d'origine fiscale

- Le fichier d'imposition des personnes (FIP)
- Le fichier de la taxe d'habitation (PLFC)
- Le fichier de l'impôt sur le revenu (POTE)
- Le fichier des propriétés bâties (Majic)

Les données complémentaires

- Le répertoire des communautés
- Les résidences hôtelières
- Le répertoire des logements locatifs des bailleurs sociaux (RPLS)

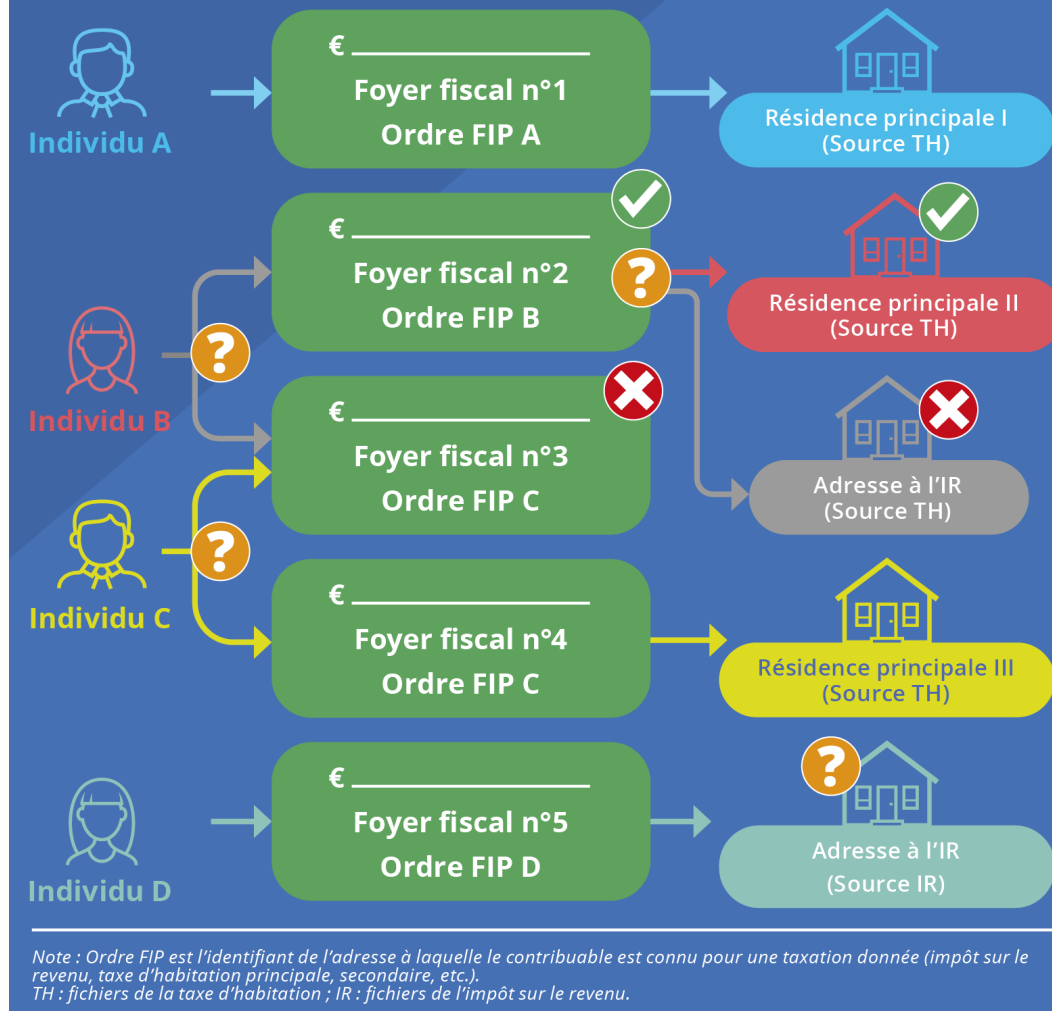
Objectif

- Justifier le **dé** de Fidéli : éviter les doubles-comptes, se rapprocher des concepts du Recensement de Population

Plusieurs types de traitements

- Correction d'erreurs manifestes (*editing*) et imputations de valeurs
- Pondérations (variable **poids**)
 - Cas des gardes partagées ou des bilocalisation
- Repérage des résidences principales
 - Essentiellement sur la base de l'information TH, mais pas seulement
- Repérage des doublons
 - Objectifs antagonistes de l'administration fiscale (recouvrement de l'impôt) et de la statistique
 - Mise en cohérence sur la base des "traits d'identité"

Figure 2. Le processus de localisation des individus dans un logement



Construction de variables statistiques

- Mobilités résidentielles (information de localisation en [AAAA-1](#) et en [AAAA](#))
- Description du parc de logements sociaux et de ses habitants (*via* l'appariement avec le RPLS)
- Composition socio-démographique des ménages vivant sur le territoire national

Géoréférencement de l'information (logement/individu) sur la base de l'adresse et de la référence cadastrale

Information sur les revenus

- Au niveau individuel, l'information de l'Impôt sur le Revenu
- Au niveau du ménage, un appariement avec le [dispositif Filosofi](#) sur l'identifiant fiscal du logement

Un gros processus de production

- Une chaîne initialement prototypée en SAS et intégrée dans un environnement classique de production



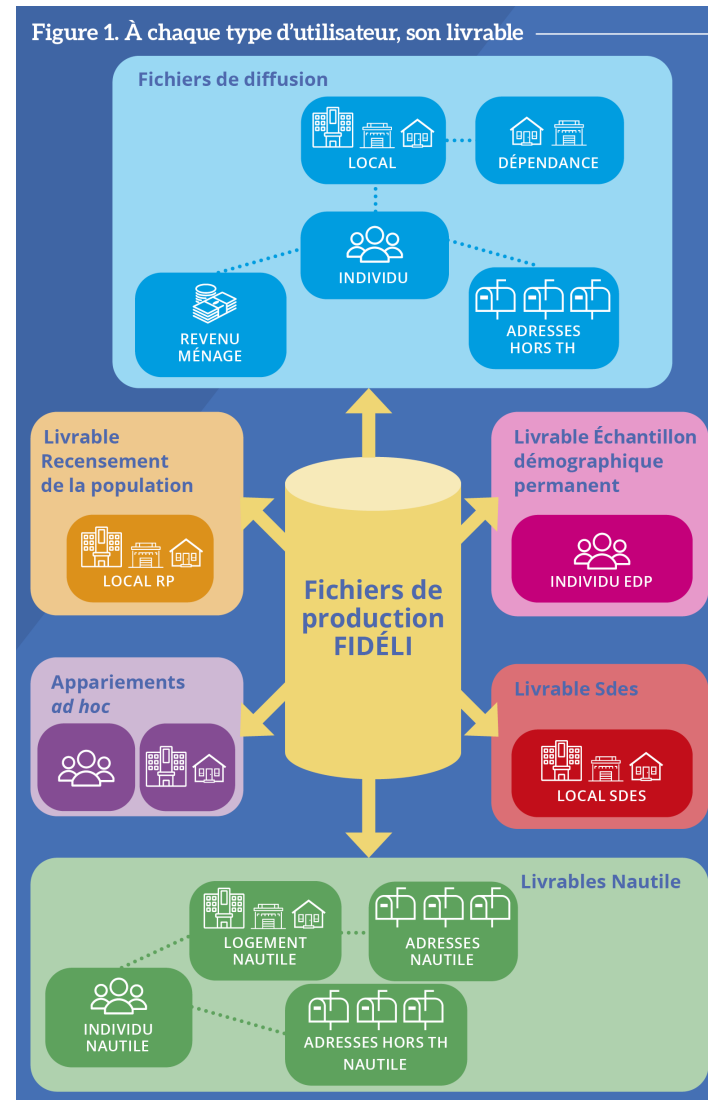
- Avec des "rustines" développées en self pour répondre aux besoins (plus ou moins) ponctuels



Répondre à plusieurs besoins :

- Tirage des échantillon des enquêtes ménages
- Connaissance locale des logements et de leurs occupants
- ...

Figure 1. À chaque type d'utilisateur, son livrable



Essentiellement sur les tables `local` et `dependance`

- identification unique d'un local au travers des variables `id_log` et `id_loc`
- deux définitions du logement (important par exemple pour les variables `type_res1` et `type_res2`)

```
## définition au sens du RP  
logement == '1'  
  
## définition adhérente à la nature du local  
natloc %in% c('MA', 'AP', 'ME')
```

- les lignes dans `local` pour lesquelles le couple (`id_log`, `id_loc`) est manquant sont des logements dits "fictifs" :
 - l'identifiant est l'adresse, c'est-à-dire le n-uplet (`csdep`, `cne1`, `rue1`, `noimm1`, `btq1`)
 - logements qui sont inconnus à la TH, mais pour lesquels on identifie l'existence d'un foyer fiscal à l'adresse indiquée

La table `individu (!!)` et `revenus_menage`

- identification unique d'un individu au travers des variables `id_foy` et `typ_fisc` (ou `id_num` et `biloc_final`)
- association d'un individu à un logement :
 - via l'identifiant `id_log` pour les logements assujettis à la TH
 - via l'adresse pour les logements hors TH
 - identifiant du logement en `AAAA-1` via `id_log_n_1` (qui est l'identifiant du logement dans le millésime `AAAA`, mais pas celui de ce logement dans le millésime `AAAA-1`)
- association d'un revenu du ménage à un logement :
 - via le couple (`id_log`, `id_log_n_1`)
 - uniquement pour les logements connus à la TH et dans le champ de Filosofi `AAAA-1`
 - information issue de Filosofi, non nécessairement cohérente avec celle de Fidéli

Le chaînage des logements et des individus n'est pas possible d'un millésime sur l'autre :

- `id_log` et `id_num` ne sont pas invariants dans le temps
- seules les variables rétrospectives (suffixées `_n_1`) permettent de panéliser l'information

Il y a deux types de variables de revenu :

- les variables de revenus individuels, localisées dans la table `individu`, issues du POTE
- les variables de revenus au niveau du ménage (ou du logement), obtenues par appariement entre Filosofi et Fidéli, qui contiennent de l'information fiscale et sociale

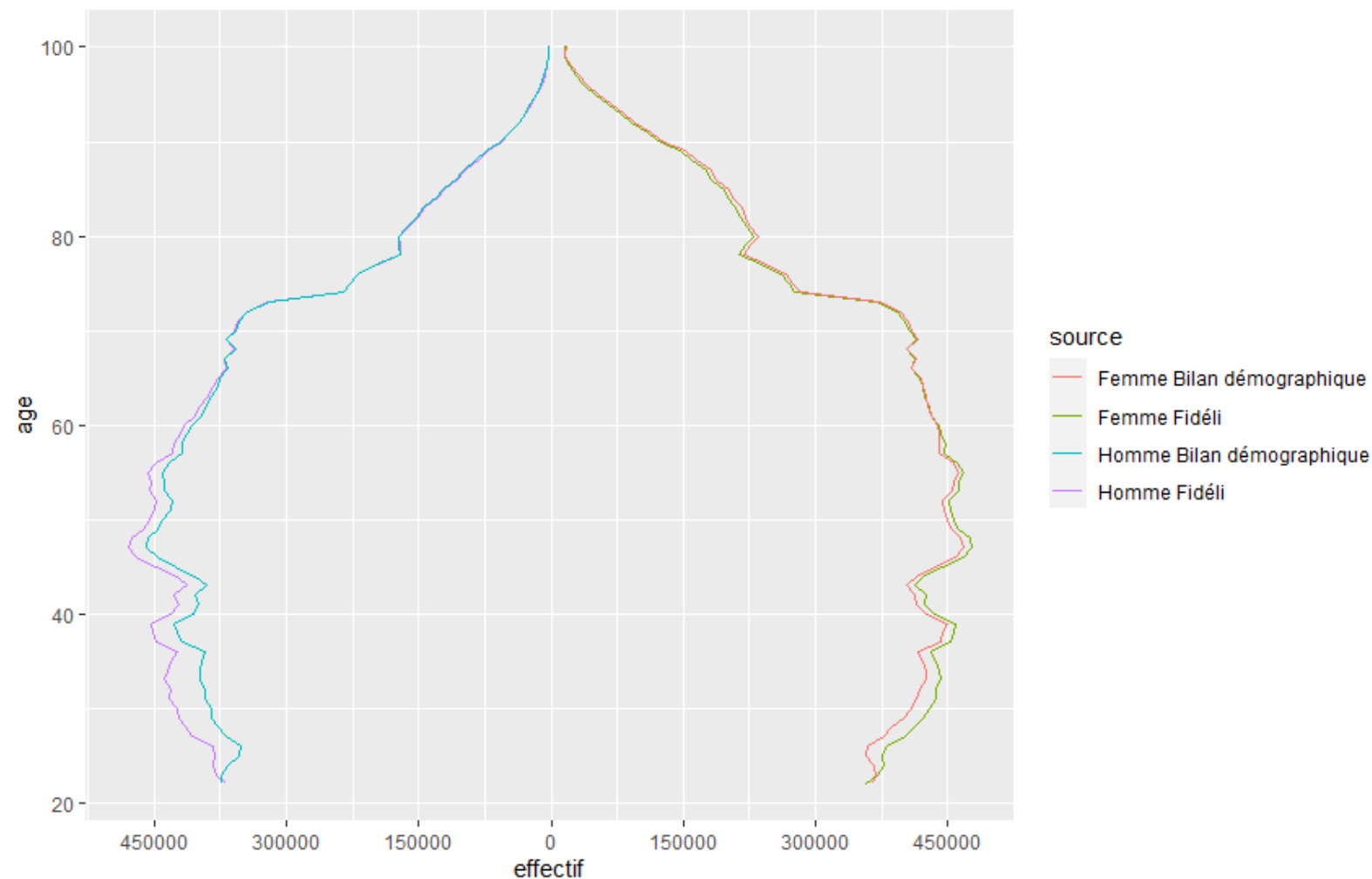
Comparaison des sources sur 2018, 2019 et 2020

<u>Année</u>	<u>Effectif Fidéli</u>	<u>Effectif bilan démographique</u>	<u>Écart absolu</u>	<u>Écart relatif</u>
2018	66 999 333	66 890 699	100 634	0,15 %
2019	67 487 616	66 977 703	509 913	0,76 %
2020	68 677 594	67 063 703	1 613 891	2,41 %

Champ : France entière

Écart entre Fidéli et le bilan démographique qui se creuse

Campagne active en 2020 de l'administration fiscale auprès des individus n'ayant pas déclaré leur impôt sur le revenu et déclaration tacite



Champ : France entière

Comparaison entre sources des effectifs par département en 2020

Écart relatif	< -10 %	-10 à -5 %	-5 à -2 %	-2 à 0 %	0 à 2 %	2 à 5 %	5 à 10 %	> 10 %
Nombre de départements	2	2	2	19	43	24	7	2

Champ : France entière

Écarts les plus importants :

- sur-représentation dans Fidéli en Guadeloupe et dans les Alpes-Maritimes
- sous-représentation dans Fidéli en Guyane et à Mayotte

02

Quelle exploitation des données ?

- Pour les **communes de 5 000 habitants ou plus**
 - la maille de diffusion est **la commune ou l'IRIS**
 - Pour les IRIS, à la condition que la somme des habitants dans les IRIS de moins de 2 000 habitants soit supérieure à 2 000
- Pour les **communes de moins de 5 000 habitants**
 - la maille de diffusion est **l'EPCI**
 - À la condition que la somme des habitants des communes de moins de 5 000 habitants dans l'EPCI soit supérieure à 2 000
- Pour les **QPV**
 - La maille de diffusion est la **région**
- Pour tout autre type de publication (non géographique) :
 - chaque case doit comporter plus de 11 individus

Fidéli est un objet dont la diffusion est décentralisé, ce qui pose de nombreux risques

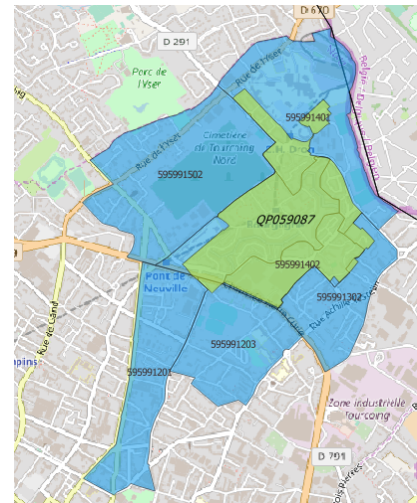
Le secret primaire est largement maîtrisable, mais c'est le secret secondaire qui pose problème

- la grande variété de variables présentes dans Fidéli démultiplie d'autant les possibilités d'analyses
- la diffusion d'informations au niveau local peut accroître le risque lié au secret secondaire...
- d'autant que la géolocalisation des individus rend plus facile la construction de zonages à façon

Cas typique : la superposition d'un QPV et d'un IRIS

La question de la diffusion de statistiques sur le QPV est conditionnée par :

- La taille (en termes de résidences principales) de la zone verte
- Mais également la taille de la zone bleue (par différenciation)
- Les choses se complexifient avec les chevauchements de plus de deux zonages



⇒ les règles de diffusion doivent tenir compte des cas les plus problématiques

La cartographie est assimilable à la diffusion d'un tableau

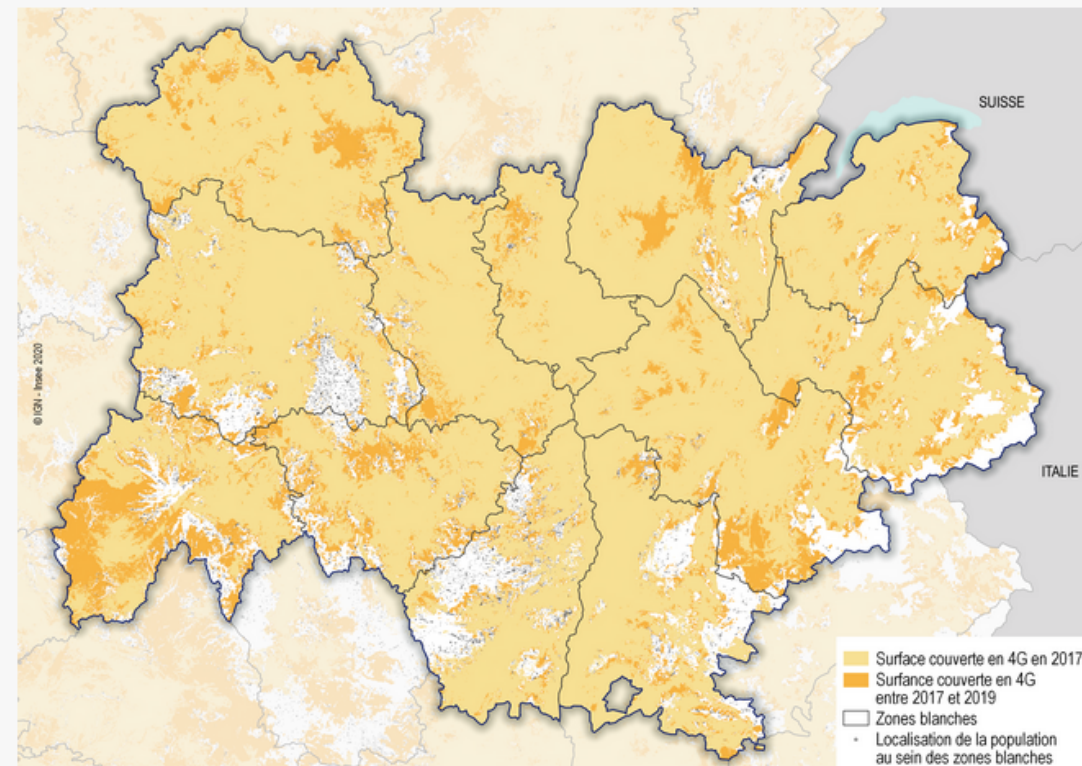
Sauf preuve du contraire, la diffusion d'une carte équivaut à la publication des données sous-jacente :

- La carte n'est qu'un mode de représentation de la donnée
- La notion de "floutage" n'a de sens que si elle procède d'une agrégation des données sous-jacente
 - la question de la résolution de la carte s'apparente à une diffusion au carreau
 - les données carroyées sont la meilleure réponse à cette question

⇒ la présence de données de géolocalisation dans Fidéli ne peut être vue comme un blanc-seing à la diffusion de cartes à partir de ces données. Elles sont d'abord utiles à **l'enrichissement de la donnée**, ainsi qu'à une prise en compte de la **dimension locale dans une analyse à plus grande échelle**.

Se poser la question de la source à utiliser

Figure 2 – Vers une couverture plus homogène du territoire
Extension de la couverture numérique en 4G entre 2017 et 2019



Note : couverture du territoire par au moins l'un des quatre opérateurs Orange, Bouygues, SFR/SFR-Numéricable et Free

Sources : France Très Haut Débit et Arcep, données au 01/04/2017 et 01/04/2019 ; Insee, *filosofi* 2015

Les points à avoir en tête :

- Les données de revenu disponible sont le résultat d'un appariement avec Filosofi N-1, et la composition du ménage Fidéli n'est pas (nécessairement) identique à celle du ménage Filosofi
- Les descripteurs du logement (nombre de pièces par exemple) sont issues des sources fiscales et peuvent souffrir d'un défaut d'actualisation
- La vacance résidentielle est mal appréhendée par les sources fiscales. Le RP est la référence en la matière (y compris pour les autres catégories de logements)
- **La plus-value de Fidéli** : l'éclairage des mobilités résidentielles, en lien avec des descripteurs sociaux, économiques et géographiques. Toute autre exploitation doit être soigneusement analysée en comparaison des autres sources disponibles.

03

Développements et perspectives



Un enrichissement de la source fiscale par la source fiscale

- **1^{er} étage de la fusée** : mise en évidence du lien logement-propriétaire(s)
 - Utilisation de l'information disponible dans les données du cadastre
 - Pour les SCI, mise en transparence du lien de propriété avec les données du RCS
 - un premier prototype développé sur le millésime 2017, qui a donné lieu à **de premières publications**
- **2^{ème} étage de la fusée** : appariement avec les **Demandes de Valeurs Foncières (DVF)**
 - Obtention d'une valeur de marché pour tous les logements ayant fait l'objet d'une transaction dans les années récentes
 - Prédiction d'une valeur de marché pour l'ensemble des logements de la base à partir d'un modèle basé sur les caractéristiques observables de ces logements
 - À partir de l'information chaînée de Fidéli, descripteurs précis des acheteurs et des vendeurs pour les transactions identifiées dans les DVF

Les sources fiscales offrent beaucoup de potentiel

- Apparier les données du RPLS au niveau logement
 - Associer les loyers aux logements du parc social et à leurs occupants
 - Conditionnel à l'inclusion de l'identifiant fiscal du logement dans le RPLS
- Analyse des données issues de la déclaration 2044
 - Question de la représentativité de ces données
 - **Défi de l'identification du logement** dans la déclaration 2044
 - **Défi du calcul d'un loyer mensuel** à partir de l'information sur les flux de revenus perçus par les bailleurs

Un enjeu de taille : la disparition de la taxe d'habitation

- Fidéli construit en grande partie la localisation des individus et le contour des ménages avec les informations de la TH
- La disparition de cette dernière en 2023 constitue un défi pour la pérennité de la source
 - Aménagement de court terme de l'application de production Fidéli pour accommoder la disparition de la source
 - À moyen terme, les projets de la DGFIP de collecte de l'information auprès des propriétaires de logements doit permettre de reconstituer l'information manquante
 - À plus long terme, l'Insee construit un projet d'intégration de différentes sources administratives (pas seulement fiscales)

Si vous souhaitez nous contacter, nous vous recommandons d'utiliser la boîte fonctionnelle de l'équipe de production de Fidéli :

dg75-administration-fideli@insee.fr

Retrouvez-nous sur :

insee.fr



Pierre LAMARCHE
Division Logement
01 87 69 63 67
pierre.lamarche@insee.fr