Insee Méthodes

N° 132
Juin 2019

Les indices Notaires-Insee des prix des logements anciens Méthodologie v4

Ont participé à cet ouvrage : Cédric Cailly, Jean-François Côte, Alain David, Jacques Friggit, Stéphane Gregoir, Adélia Nobre, François Proost, Catherine Rougerie, Stéphane Schoffit, Nelly Tauzin, Hélène Thélot

Insee
Mesurer pour comprendre

Insee Méthodes

Les indices Notaires-Insee des prix des logements anciens Méthodologie v4

Sommaire

Résumé

Connaître l'évolution des prix des logements est un préalable au bon fonctionnement et à la transparence de leurs marchés. Une méthode a été mise au point par l'Insee en liaison avec le notariat pour produire des indices de prix réguliers et fiables des logements anciens, les indices Notaires-Insee. Elle a été exposée dans sa première version dans le numéro 98 d'*Insee Méthodes*, (David *et al.*, 2002), puis dans des versions révisées : la « version 2 », dans le numéro 111 d'*Insee Méthodes* (Beauvois *et al.*, 2005), et la « version 3 », dans le numéro 128 d'*Insee Méthodes* (Clarenc *et al.*, 2014). Le présent numéro d'*Insee Méthodes* présente une actualisation de cette version 3, la « version 4 », des indices Notaires-Insee, qui a été mise en production à partir de 2018.

Cette « version 4 » ne remet pas en question les grands principes de la méthode. Cette dernière repose toujours sur des modèles économétriques (« hédoniques »), qui permettent d'estimer le prix d'un bien en fonction de ses caractéristiques. Les modèles sont estimés par type de bien (appartements/maisons) et par zone géographique, car la formation des prix des logements dépend fortement de la localisation du bien. Cette méthode est appliquée afin de répondre aux deux difficultés majeures rencontrées lorsqu'on souhaite estimer des prix pour les logements : d'abord, le prix d'un logement donné est rarement observé, ensuite la qualité des logements évolue au cours du temps. Ainsi, il est impossible pour une analyse conjoncturelle de suivre les prix des mêmes biens au cours du temps ; on ne peut également pas se contenter de comparer la moyenne des prix des transactions courantes entre deux périodes, car cela mélangerait un effet prix et un effet qualité.

Les données sont fournies par les notaires, qui rédigent les contrats pour la quasi-totalité des ventes de logements anciens en France. Les caractéristiques des logements comportent leur superficie, leur nombre de pièces, de salles de bains, l'étage, la présence d'un ascenseur, d'un garage, etc.

Comme prévu initialement, les modèles initiaux sont révisés tous les 5 à 8 ans, l'objectif étant de s'assurer que la méthode est toujours pertinente et adaptée à la période récente. Pour cette nouvelle version, la stratification a été révisée et enrichie en Île-de-France, et les modèles économétriques ont été revus et améliorés, notamment en ajoutant des variables explicatives du prix (étiquette énergétique, distance à la commune de plus de 50 000 habitants la plus proche en province, distance à Paris pour l'Île-de-France). Enfin, pour cette nouvelle version, le champ de calcul a été élargi aux Drom hors Mayotte et les indices régionaux diffusés le sont également par nouvelle région.

Les indices décrits dans ce document sont validés sous l'appellation « indices Notaires-Insee » ; leur conception est surveillée par un Conseil scientifique. Ils sont publiés chaque trimestre dans une première version provisoire deux mois après la fin du trimestre, puis dans une version définitive six mois après la fin du trimestre (cinq mois pour les indices d'Île-de-France). Les indices trimestriels des prix des logements anciens ont été labellisés par l'Autorité de la statistique publique (ASP), à la suite d'une analyse des dispositifs de collecte et de calcul.

Le présent document décrit la méthode de calcul et ses enrichissements apportés en version 4.

Ont participé à cet ouvrage :
Cédric Cailly, Jean-François Côte, Alain David, Jacques Friggit, Stéphane Gregoir, Adélia Nobre, François Proost, Catherine Rougerie, Stéphane Schoffit, Nelly Tauzin, Hélène Thélot

1 – Introduction

Connaître l'évolution des prix des logements est un préalable au bon fonctionnement et à la transparence des marchés immobiliers. La production d'indicateurs statistiques sur le sujet peut bénéficier à de multiples acteurs : aux particuliers comme élément de décision sur le choix d'acheter ou vendre un bien ou pour l'estimation d'un patrimoine pour des successions ou dans un cadre fiscal, aux pouvoirs publics et collectivités qui souhaitent veiller au bon fonctionnement du secteur du logement, aux banques qui financent les particuliers dans leurs achats de logements et qui valorisent des garanties hypothécaires, aux banques centrales qui effectuent des analyses de stabilité financière…

C'est dans ce contexte que sont produits les indices Notaires-Insee des prix des logements anciens. Comme leur nom l'indique, ces indices sont le produit d'une collaboration entre les notaires et l'Insee. Les premiers enregistrent les mutations immobilières en France et traitent les informations liées aux transactions dans des bases de données. L'Insee apporte quant à lui son expertise et son expérience en matière d'exploitation et de méthodologie statistiques.
Ce partenariat est officialisé à travers deux conventions : l'une entre l'Insee, les Chambres des notaires d'Île-de-France et l'Association des notaires du Châtelet (Paris Notaires Services – PNS), l'autre entre l'Insee, ADNOV[1] et le Conseil supérieur du notariat. PNS calcule chaque trimestre les indices relatifs à l'Île-de-France, tandis que la société ADNOV, entité du groupe ADSN (Association pour le développement du service notarial), fait de même pour la province. L'Insee calcule tous les ans des coefficients de correction des variations saisonnières (CVS) et valide les indices avant chaque publication trimestrielle. Afin d'organiser au mieux cette coopération, les acteurs se retrouvent au sein d'un Conseil scientifique des indices Notaires-Insee (CSIN). Celui-ci se réunit au minimum une fois par trimestre et exerce un rôle de veille méthodologique et de suivi des indices publiés.
Les indices ont été labellisés par l'Autorité de la statistique publique (ASP), à la suite d'une analyse des dispositifs de collecte et de calcul.

Le calcul de ces indices de prix des logements anciens doit tenir compte de deux difficultés majeures :

- le prix d'un logement est rarement observé −le prix n'est en effet connu que lors d'une transaction −et un logement change rarement de main, en général après de nombreuses années de propriété. On ne peut donc pas suivre le prix d'un même bien régulièrement au cours du temps ;
- la qualité des logements évolue au cours du temps : cette propriété est commune à de nombreux biens lorsqu'on cherche à en calculer des indices de prix. Lorsque la qualité des biens évolue au cours du temps, les prix enregistrés ne seront pas relatifs à des biens identiques. Par exemple, si le prix d'une ampoule électrique passe de 2 à 4 euros, alors que sa durée de vie augmente, son prix a apparemment doublé. Mais, si le service rendu au consommateur d'ampoule est le même pour une ampoule de durée de vie longue que pour deux ampoules de durée de vie courte, on dira que la qualité a aussi été multipliée par deux et donc que le prix « pur » est resté constant. Au niveau du logement, la qualité va se définir à travers des caractéristiques multiples telles que la localisation, la surface, le

[1] À compter du 1er juin 2019, Min.not est devenue ADNOV.

nombre de salles de bains, la présence d'un garage… Mais le domaine du logement est aussi plus simple à aborder pour le statisticien parce que le nombre de révolutions techniques est limité ; les changements de qualité, dans le sens de l'apparition de nouvelles caractéristiques, sont très rares, contrairement à certains domaines comme l'informatique où de nouveaux composants plus performants entrent régulièrement sur le marché. Sur une période donnée de quelques années, on peut donc considérer la qualité des logements à travers un nombre fini et constant de caractéristiques.

Pour suivre les tendances immobilières, on ne peut donc pas suivre directement le prix de mêmes biens au cours du temps, car ces prix ne sont pas disponibles. On ne peut pas non plus se contenter de faire la moyenne des prix des transactions courantes pour la comparer à celle de la période précédente, car une telle comparaison mélangerait l'évolution « pure » des prix (que l'on souhaite mesurer) à celle de la qualité des biens échangés. Pour traiter ces questions, la méthode retenue repose sur des modèles économétriques dits « modèles hédoniques », fondés sur l'estimation des prix en fonction des caractéristiques définissant la qualité des logements.
Cette méthode est passée entièrement en revue tous les 5 à 8 ans et actualisée au besoin afin de s'assurer qu'elle est toujours pertinente et adaptée aux périodes récentes. C'est à cette occasion qu'a été mise en place la version 4 de la méthode de calcul des indices décrite dans ce document et effective à partir des évolutions de prix de 2018[2].

[2] La méthode de calcul de la version 3 des indices est décrite dans l'*Insee Méthodes* n° 128 de juillet 2014 (https://insee.fr/fr/information/2569926). Cet ouvrage détaille notamment dans son chapitre 2 une approche plus théorique des méthodes de calcul appliquées aux indices.

2 – Données et champ

2.1 – Le cadre institutionnel

Lors de la transaction d'un bien immobilier en France, un acte de vente doit être réalisé chez un notaire. Dans le cas d'un logement existant, cet acte doit obligatoirement mentionner les informations suivantes :

- concernant les parties :
  - les coordonnées du vendeur et de l'acquéreur ;
- concernant le bien :
  - l'adresse du bien ;
  - l'origine du bien (date du précédent acte de vente, nom du précédent propriétaire, acte notarié…) ;
  - le descriptif détaillé du bien, de ses équipements et annexes ;
  - l'existence d'une hypothèque et/ou d'une servitude ;
- concernant la vente :
  - le montant des honoraires du professionnel chargé de la vente et à qui en incombe le paiement ;
  - le prix de vente et ses modalités de paiement (avec ou sans l'aide d'un prêt immobilier) ;
  - la date de disponibilité du bien ;
  - les conditions suspensives lorsqu'il en existe.

Le notaire est un officier public, intervenant dans l'ensemble des domaines du droit incluant l'immobilier, mais également la famille, la fiscalité et le patrimoine, les entreprises, les zones rurales, les collectivités locales… Agissant pour le compte de l'État et nommé par arrêté du ministre de la justice, il possède de véritables prérogatives de puissance publique et confère aux actes qu'il rédige, par sa signature, l'authenticité.

Légalement, la mission de service public du notariat a été définie par l'article premier de l'ordonnance n° 45-2590 du 2 novembre 1945 : « Les notaires sont les officiers publics, établis pour recevoir tous les actes et contrats auxquels les parties doivent ou veulent faire donner le caractère d'authenticité attaché aux actes de l'autorité publique et pour en assurer la date, en conserver le dépôt, en délivrer des grosses et expéditions. »

Depuis la loi n° 2011-331 du 28 mars 2011, complétée par le décret d'application n° 2013-803 du 3 septembre 2013 et de ses arrêtés du 30 septembre 2016, la collecte d'informations statistiques sur les transactions de logement devient également partie de la mission de service public des notaires qui « contribuent à la diffusion des informations relatives aux mutations d'immeubles à

titre onéreux. Ils transmettent au Conseil supérieur du notariat[3] les données nécessaires à l'exercice de cette mission de service public ».
Cette même loi reconnaît aussi comme mission de service public l'activité de centralisation et de diffusion des données collectées par le Conseil supérieur du notariat :

- « Le Conseil supérieur du notariat met gratuitement des résultats statistiques à la disposition du public sur le réseau internet. »
- « Le Conseil supérieur du notariat, ou son délégataire, transmet à toute personne qui le demande, pour au moins vingt mutations, les informations relatives à la transaction, au prix et aux caractéristiques essentielles de chaque bien. »
- « Le Conseil supérieur du notariat, ou son délégataire, transmet à toute personne qui le demande, moyennant le paiement d'une rémunération, un ou plusieurs tableaux de résultats statistiques (…). »

2.2 – Présentation des bases

Les indices Notaires-Insee sont calculés à partir des informations enregistrées dans les bases de données notariales, qui sont au nombre de deux :

- La base BIEN, gérée par PNS (Paris Notaires Services), couvre l'Île-de-France. Elle a été mise en place à partir de 1989 pour Paris, 1991 pour la petite couronne et 1996 pour la grande couronne ; elle est exploitable pour les indices depuis 1991.
- La base Perval, gérée par l'ADSN (Association pour le développement du service notarial, dépendant du notariat), couvre la province et les Drom. Elle a été mise en place et exploitée à partir de 1994.

Ces bases enregistrent les transactions de tous les types de biens : appartements et maisons, mais aussi immeubles, locaux d'activités, terrains, garages, vignobles et autres biens agricoles. Les informations sont récupérées dans les actes de ventes, permettant de recueillir des informations détaillées pour les biens échangés.
Les actes de vente de biens immobiliers sont transmis directement par les offices notariaux situés sur l'ensemble du territoire national. Les biens situés en province et dans les Drom hors Mayotte alimentent la base Perval, tandis que ceux situés en Île-de-France sont enregistrés dans la base BIEN.
Il existe deux filières d'alimentation de la base de données :

- l'envoi en version papier était le seul en vigueur jusqu'en 2009 : les offices notariaux transmettent aux gestionnaires des bases des copies des actes de vente complétées de fiches descriptives des biens. Les informations sont ensuite contrôlées et saisies par des codificateurs ;
- la télétransmission des actes est à la disposition des offices depuis 2010 : elle se fait à partir du logiciel de rédaction d'actes lors de la saisie de la vente dans l'office notarial. L'acte de vente est ensuite télétransmis *via* l'intranet de la profession pour intégrer la base de données immobilières, où sont effectués des contrôles qualité par les codificateurs du côté des gestionnaires des bases.

[3] Le Conseil supérieur du notariat est un établissement d'utilité publique créé par l'ordonnance du 2 novembre 1945 ; il représente la profession auprès des pouvoirs publics, détermine sa politique générale, contribue à l'évolution du notariat et fournit des services collectifs aux notaires.

Le délai de réception, conditionnant la réactivité des indices, dépend du mode de transmission : en 2016, le délai est en moyenne de 18 jours pour les actes télétransmis et 59 jours pour les actes papiers en province (respectivement 19 jours et 47 jours en Île-de-France). La télétransmission permet ainsi un gain notable sur les délais. Elle se développe progressivement depuis 2010 et a vocation à devenir à terme le seul vecteur de transmission. Ce processus est déjà bien avancé : en 2016, 67 % des actes de province ont été télétransmis et 54 % en Île-de-France.

L'envoi des données par les notaires s'effectuait jusqu'à fin 2016 sur une base volontaire, le taux de couverture de la base de données n'était donc pas de 100 %. Depuis le 1er janvier 2017, la loi[4]oblige les notaires à alimenter les bases, aussi le taux de couverture a progressé, même si la phase de montée en charge n'est pas terminée. Il se situait autour de 60 % début 2017. Ce taux de couverture est estimé en comparant le montant des droits de mutation figurant dans les bases de données avec celui perçu par la direction générale des Finances publiques (DGFiP) (voir partie 4.4.1).
Fin 2018, les bases contenaient 18 millions d'enregistrements correspondant aux mutations des années 1990 à 2017 (25 % en Île-de-France et 75 % en province), dont 11 millions portaient sur des logements anciens au sens fiscal (voir partie 2.3). Pour l'année 2017, elles contenaient 910 000 mutations, dont 610 000 portaient sur des logements anciens au sens fiscal.

2.3 – Champ des indices

Les indices couvrent aujourd'hui l'ensemble de la France excepté Mayotte. Ils portent sur les maisons et appartements anciens au sens fiscal (c'est-à-dire non soumis en partie ou en totalité à la TVA) : les logements neufs ou ayant subi une rénovation lourde sont donc exclus. Sont également exclus les logements atypiques tels que chambres, greniers, lofts, ateliers, loges de gardien, châteaux, grandes propriétés, hôtels particuliers. De plus, les logements doivent être vendus de gré à gré, libres d'occupation au moment de la vente[5], destinés à un usage strict d'habitation et acquis en pleine propriété par un particulier ou une société civile immobilière. Sur le parc des logements anciens, environ 12 % des appartements et 13 % des maisons en province (respectivement 11 % et 7 % en Île-de-France) sont exclus du champ des indices en appliquant ces critères (calcul effectué sur les données de 2012 à 2016).
En outre, afin d'éviter les enregistrements trop incomplets ou atypiques, voire faux, les transactions retenues doivent respecter les caractéristiques suivantes :

- Pour les appartements :
  - le nombre de pièces est renseigné et strictement inférieur à 9 ;
  - la surface habitable réelle ou estimée[6] est comprise entre 10 m² et 200 m² (inclus) ;
  - le prix de vente est renseigné et compris entre 1 500 euros et 5 000 000 euros (exclus) ;

[4] Les arrêtés du décret n°2013-803 du 3 septembre 2013 sont parus le 30 septembre 2016. Avec leur publication, la loi n°2011-331 du 28 mars 2011 relative à la modernisation des professions judiciaires et juridiques est entrée en vigueur, conférant aux notaires une nouvelle mission de service public au travers de la collecte et de la diffusion des statistiques immobilières issues de leurs contrats et avant-contrats de vente.

[5] Les logements sont retirés lorsque la période d'occupation par un tiers ou par le vendeur excède six mois en considérant que, compte tenu de la réglementation en matière de baux locatifs, ces logements subissent en général une décote.

[6] Lorsque la surface habitable n'est pas disponible, elle est estimée à partir d'autres informations. Voir partie 2.4.

- le prix au m² est strictement inférieur à 25 000 euros ;
- la date de transaction est correctement renseignée ;
- la commune de la transaction est bien présente dans le référentiel géographique.

- Pour les maisons :
  - le nombre de pièces est renseigné et strictement inférieur à 13 ;
  - la surface habitable réelle ou estimée est comprise entre 20 m² et 300 m² (inclus) ;
  - la surface du terrain est renseignée ;
  - le prix de vente est renseigné et compris entre 1 500 euros et 15 000 000 euros (exclus) ;
  - la date de transaction est correctement renseignée ;
  - la commune de la transaction est bien présente dans le référentiel géographique.

Le prix retenu est le prix net vendeur, hors droits de mutation, frais de notaire et commission d'agence.
Au total, après ces deux étapes, on conserve, dans le champ des indices, 87 % des appartements et 80 % des maisons en province (86 % pour les appartements et maisons d'Île-de-France).

2.4 – Variables utilisées et traitement des valeurs manquantes

Parmi les variables collectées dans les bases notariales, celles retenues dans le cadre des indices sont, au-delà du prix de vente et de la date de mutation, celles qui ont un impact sur le prix du bien. Elles doivent en outre être suffisamment bien renseignées pour permettre une estimation pertinente de l'équation des prix.
Les caractéristiques ainsi retenues sont :

- pour les appartements :
  - la localisation[7] ;
  - la distance en kilomètres à la commune de 50 000 habitants la plus proche pour les strates de province, la distance à Paris pour les strates d'Île-de-France ;
  - la surface habitable ;
  - le nombre de pièces ;
  - l'époque de construction ;
  - le nombre de salles de bains ;
  - le nombre de garages ou parkings ;
  - l'état du bien[8] ;
  - l'étiquette énergétique ;
  - l'étage ;
  - la présence d'un ascenseur (à partir du 4e étage) ;
  - la présence d'une terrasse (ou balcon ou loggia) ;
  - la présence d'une cave.

[7] La localisation, ayant un fort impact sur la détermination des prix, est prise en compte de plusieurs manières : *via* la stratification, puis au sein de chaque modèle. Ces points sont détaillés dans les chapitres suivants.
[8] L'état, quand il est renseigné, est le résultat d'une appréciation faite par les acteurs de la transaction. Il peut prendre trois valeurs : bon état, travaux à prévoir, à rénover.

- pour les maisons :
    - la localisation ;
    - la distance en kilomètres à la commune de 50 000 habitants la plus proche pour les strates de province, la distance à Paris pour les strates d'Île-de-France ;
    - la surface habitable ;
    - la surface du terrain ;
    - le nombre de pièces ;
    - l'époque de construction ;
    - le nombre de salles de bains ;
    - le nombre de garages ou parkings ;
    - l'état du bien ;
    - l'étiquette énergétique ;
    - la présence d'une cave ou d'un sous-sol ;
    - le nombre de niveaux.

Les caractéristiques présentes dans les bases notariales peuvent parfois être incomplètes. Lorsqu'une transaction est enregistrée dans les bases mais qu'une ou plusieurs variables relatives au bien échangé sont manquantes[9], un traitement est effectué sur ces dernières afin de pouvoir utiliser la transaction dans le calcul des indices. Selon les cas, les variables peuvent être imputées (par estimation ou recodage) ou bien la valeur manquante est considérée en tant que telle dans les modèles (ajout d'une modalité « inconnu »). Les règles retenues sont synthétisées dans les figures 2-1 et 2-2.

Figure 2-1 : Règles de traitement de la non-réponse pour les valeurs manquantes - Appartements

Appartements				
Variable	*Modalité attendue*	*Action en cas de non-réponse*	*Pourcentage de non-réponse en 2017*	
			Île-de-France	*Province*
Surface habitable	Valeur numérique	Estimation linéaire en fonction du nombre de salles de bains, du nombre de pièces, de l'époque de construction, de l'étage et du nombre de garages. Les coefficients sont remis à jour tous les 2 ans.	2 % (avant imputation)	2 % (avant imputation)
Nombre de salles de bains	Valeur numérique	Recodage à 1 salle de bains.	2 %	2 %

[9] Les informations essentielles, à savoir la nature du bien (appartement ou maison), le prix, la date et le lieu de mutation, le nombre de pièces et (pour les maisons) la surface du terrain doivent au minimum être présentes. Les observations ne respectant pas ces critères ont été exclues lors de la sélection du champ (cf. partie 2.3).

Appartements				
Époque de construction	10 tranches	Si possible, imputation à partir des ventes réalisées à la même adresse sur les 10 dernières années.[10] Sinon, ajout d'une modalité « inconnu ».	25 % (après imputation)	17 % (après imputation)
Étage	Valeur numérique	Recodage à 0 (rez-de-chaussée).	1 %	2 %
Présence d'un ascenseur (à partir du 4e étage)	Oui/non	Si possible, imputation à partir des ventes réalisées à la même adresse sur les 10 dernières années[11]. Sinon, recodage à « oui ».	37 % (après imputation)	31 % (après imputation)
Présence d'une cave	Oui/non	Recodage à « non »[12].	Proche de 0 %	9 %
Nombre de garages ou parkings	Valeur numérique	Recodage à 0 (pas de garage ni parking)[13].	3 %	8 %
Présence d'une terrasse (ou balcon ou loggia)	Oui/non	Ajout d'une modalité « inconnu ».	23 %	13 %
État du bien	3 modalités	Ajout d'une modalité « inconnu ».	82 %	70 %
Étiquette énergie	8 modalités	Ajout d'une modalité « inconnu ».	41 %	32 %

10 Le principe est de redresser ces variables en prenant les informations fournies par les autres ventes disponibles dans le même bâtiment au cours des 10 dernières années, lorsqu'il y en a, dans les bases immobilières notariales. Afin de s'assurer que l'information est suffisamment fiable, ce traitement est effectué uniquement s'il y a au moins cinq ventes à la même adresse, et dont au moins 60 % indiquent la même valeur. Ce traitement n'est fait qu'en appartements, en raison du peu de maisons dans les bases qui ont fait l'objet d'au moins deux ventes.

11 Avec les mêmes contraintes que pour l'époque de construction.

12 Un numéro de lot doit être indiqué sur l'acte s'il y a une cave.

13 Un numéro de lot doit être indiqué sur l'acte s'il y a un garage.

Figure 2-2 : Règles de traitement de la non-réponse pour les valeurs manquantes - Maisons

Maisons				
Variable	*Modalité attendue*	*Action en cas de non-réponse*	*Pourcentage de non-réponse en 2017*	
			Île-de-France	*Province*
Surface habitable[14]	Valeur numérique	Estimation linéaire en fonction du nombre de salles de bains, du nombre de pièces, de l'époque de construction et du nombre de niveaux. Les coefficients sont remis à jour tous les 2 ans.	23 %	8 %
Nombre de salles de bains	Valeur numérique	Recodage à 1 salle de bains.	1 %	2 %
Époque de construction	10 tranches	Ajout d'une modalité « inconnu ».	32 %	13 %
Nombre de niveaux	2 modalités	Recodage à « 2 niveaux ou plus ».	0 %	2 %
Présence d'une cave ou d'un sous-sol	Oui/non	Recodage à « non ».	7 %	18 %
Nombre de garages ou parkings	Valeur numérique	Ajout d'une modalité « inconnu ».	15 %	12 %
État du bien	3 modalités	Ajout d'une modalité « inconnu ».	82 %	57 %
Étiquette énergie	8 modalités	Ajout d'une modalité « inconnu ».	32 %	31 %

[14] La méthode utilisée pour estimer la surface habitable des maisons anciennes a fait l'objet d'une attention particulière, car le taux de non-réponse est assez élevé sur ce marché. En effet, la loi Carrez (entrée en vigueur le 19 juin 1997), obligeant le vendeur à déclarer la surface sur l'acte de vente, n'est valable que pour les appartements. Le modèle retenu est finalement proche de celui utilisé en version 3, à l'exception de l'ajout dans les modèles de quelques variables ainsi que de la mise à jour plus fréquente des coefficients.

3 – Stratification

La formation des prix des logements dépend fortement de la localisation du bien. Pour prendre en compte cet aspect dans le calcul des indices, le territoire a été divisé en strates pour chaque type de bien (appartements et maisons), l'idée étant de distinguer différents marchés immobiliers : dans chaque strate sera ensuite calculé un modèle hédonique propre.

L'objectif recherché à travers ce découpage géographique est d'obtenir des strates homogènes en matière de valorisation des logements. Un équilibre a cependant dû être trouvé. Il était d'une part recherché une stratification très fine séparant au mieux la variabilité des prix sur le territoire, mais il fallait d'autre part que chaque strate possède un nombre suffisant d'observations afin d'effectuer des traitements statistiques de bonne qualité. Pour cela, il a été imposé que chaque strate contienne un nombre minimum de transactions, arbitré à 110 par trimestre en moyenne sur la période d'étude[15].

Étant donné cette taille minimale des strates et l'importance de la localisation sur les prix, il reste des effets locaux non identifiés par la stratification. Afin de les capter au mieux, un zonage détaillé est parfois défini au sein des strates, qui rentre dans l'étape de modélisation (cf. partie 4-2).

La stratification a été définie différemment en Île-de-France et en province, compte tenu des contextes différents des marchés immobiliers :

- en Île-de-France, où les prix de l'immobilier s'établissent en bonne partie selon le positionnement par rapport à Paris, la stratification a été construite par une approche statistique. Celle-ci a été déclinée sur trois zones ayant été distinguées au préalable : deux pour le marché des appartements, Paris composant une zone et le reste de l'Île-de-France l'autre zone, et une seule pour les maisons, car distinguer Paris du reste de l'Île-de-France n'aurait pas d'intérêt du fait du trop faible effectif des maisons à Paris. La démarche statistique consiste, à partir du zonage géographique le plus fin disponible[16], d'effectuer des regroupements successifs de ces zones selon leurs similarités en fonction de différentes variables d'intérêt[17]. Ces dernières, relatives à la zone (il peut s'agir alors de valeurs moyennes sur la commune ou de la répartition des modalités d'une variable), sont listées dans la figure 3-1.

[15] Soit les années 2003 à 2013 en Île-de-France et 1998 à 2007 pour la province.

[16] Il s'agit des quartiers administratifs à Paris, qui sont au nombre de 80, et des communes pour le reste de l'Île-de-France.

[17] Outre les variables issues de la base BIEN, des variables communales produites par la statistique publique ont également été utilisées.

Figure 3-1 : Zonage et variables de classification en Île-de-France

Appartements		Maisons
Paris	**Reste de l'Île-de-France**	
Nombre de pièces	Nombre de pièces	Nombre de pièces
Surface habitable	Surface habitable	Surface habitable
Époque de construction	Époque de construction	Surface du terrain
Étage et présence d'un ascenseur (au-delà du 4e étage)	Étage et présence d'un ascenseur (au-delà du 4e étage)	Époque de construction
Nombre de garages ou parkings	Nombre de garages ou parkings	Présence d'une dépendance
Nombre de salles de bains	Nombre de salles de bains	Nombre de bâtiments
Présence d'une cave	Présence d'une cave	Nombre de garages ou parkings
	Distance à Paris (calculée à partir des coordonnées des centres des communes)	Nombre de salles de bains
	Part des résidences principales occupées par des locataires	Nombre de niveaux
	Nombre d'habitants	Distance à Paris (calculée à partir des coordonnées des centres des communes)
	Revenu médian	Revenu médian

Dans le détail, une analyse en composantes principales (ACP) a premièrement été effectuée sur les communes ou quartiers parisiens avec ces variables. Cette méthode permet de réduire la dimension de l'espace de données étudié, tout en préservant le maximum d'informations possible : de nouvelles variables sont obtenues, appelées « composantes principales ». Ce sont des combinaisons linéaires des variables initiales, non corrélées deux à deux, et de variance maximale[18].

Une classification ascendante hiérarchique (CAH) est ensuite appliquée sur ces composantes principales. Cette deuxième méthode, algorithmique, permet de créer un arbre de classification[19] en partant des « individus » (dans ce cas les communes, ou les quartiers parisiens) qui sont regroupés pas à pas. Différents critères existent pour effectuer

[18] La méthode crée initialement autant de nouvelles variables que d'anciennes, mais non corrélées et triées par leur part d'inertie (généralisation de la variance dans le cas multidimensionnel) expliquée par rapport à l'inertie totale du nuage de points (correspondant à la somme des variances de toutes les variables). Cela permet alors, en appliquant un seuil, de retenir moins de variables (les n premières, selon le seuil choisi) tout en conservant le maximum de variabilité entre les différents individus.

[19] La classification est effectuée en partant des individus jusqu'à ce qu'ils soient tous regroupés en une seule classe. Le résultat peut être présenté sous forme d'arbre, permettant de visualiser tous les regroupements effectués successivement et de choisir à quelle étape on souhaite s'arrêter pour déterminer les classes définitives.

les regroupements ; c'est la méthode de Ward[20], optimale du point de vue de l'homogénéité des agrégats construits, qui a été utilisée ici. Dans l'arbre ainsi calculé, la classification retenue est la plus fine respectant le critère des 110 transactions minimales par trimestre pour chaque agrégat.

Afin d'affiner une dernière fois le résultat obtenu, on vérifie s'il y a de « grands agrégats » (plus de 450 transactions par trimestre) qui seraient le fruit d'un regroupement ayant eu lieu avant l'étape finale, que l'on scinde alors à nouveau sous réserve que toutes les subdivisions respectent le critère des 110 transactions trimestrielles.

La constitution des strates est alors terminée ; il est à noter que les communes (ou quartiers pour les appartements parisiens) composant une strate ne sont pas forcément contiguës, l'important étant qu'elles possèdent des caractéristiques proches concernant leurs marchés immobiliers.

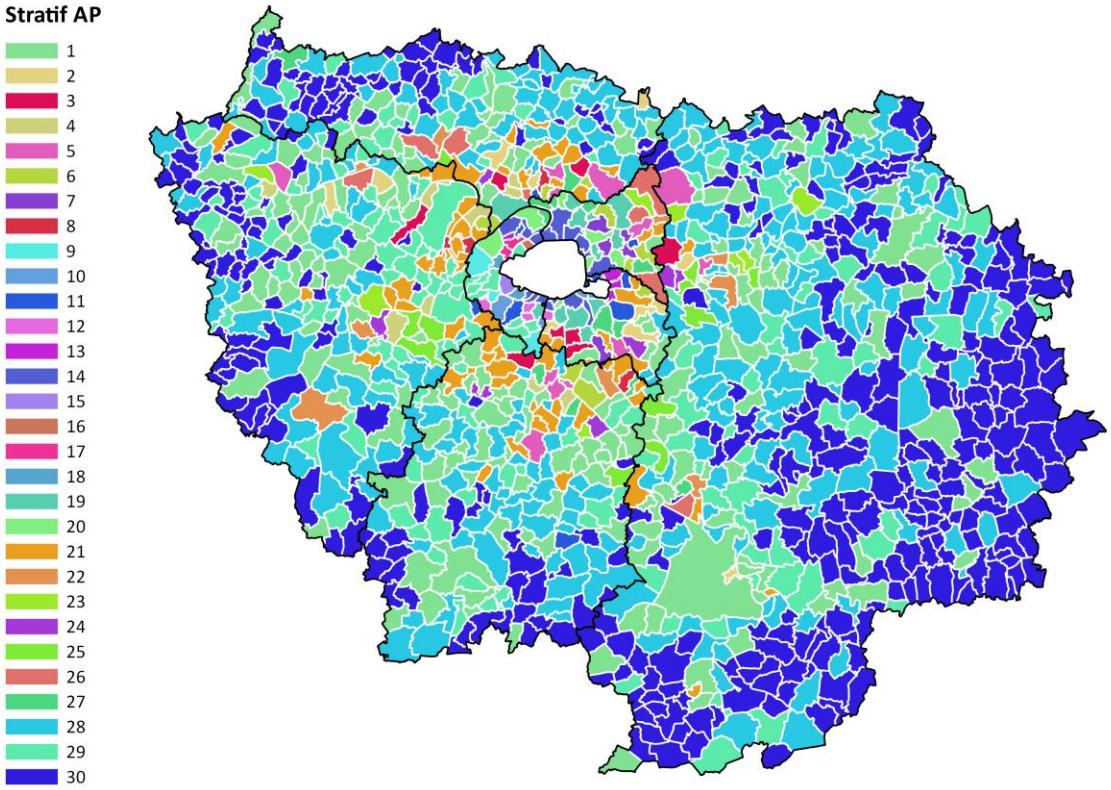
Au total, on obtient 18 strates pour les appartements parisiens (figure 3-2), 29 strates pour les appartements d'Île-de-France hors Paris (figure 3-3) et 22 strates pour les maisons (figure 3-4).

Figure 3-2 : Stratification retenue pour les appartements parisiens

Exemples de strates : la strate 1 regroupe les quartiers centraux de Paris, composée de 33 % d'appartements datant d'avant 1850 et de 38 % de la période hausmanienne (1850-1913). La strate 18, composée de quartiers du Nord parisien, est composée de seulement 2 % d'appartements datant d'avant 1850, mais de 54 % de la période hausmanienne.

[20] On peut décomposer, dans le cadre d'une classification, l'inertie totale en deux parties : l'inertie intra-classe (cumul des variabilités au sein de chaque classe) et l'inertie inter-classe (variabilité entre les classes). Avec la méthode de Ward, le regroupement effectué à chaque itération est celui qui conserve l'inertie inter-classe la plus grande possible, et inversement, qui minimise l'inertie intra-classe.

Figure 3-3 : Stratification retenue pour les appartements en Île-de-France hors Paris

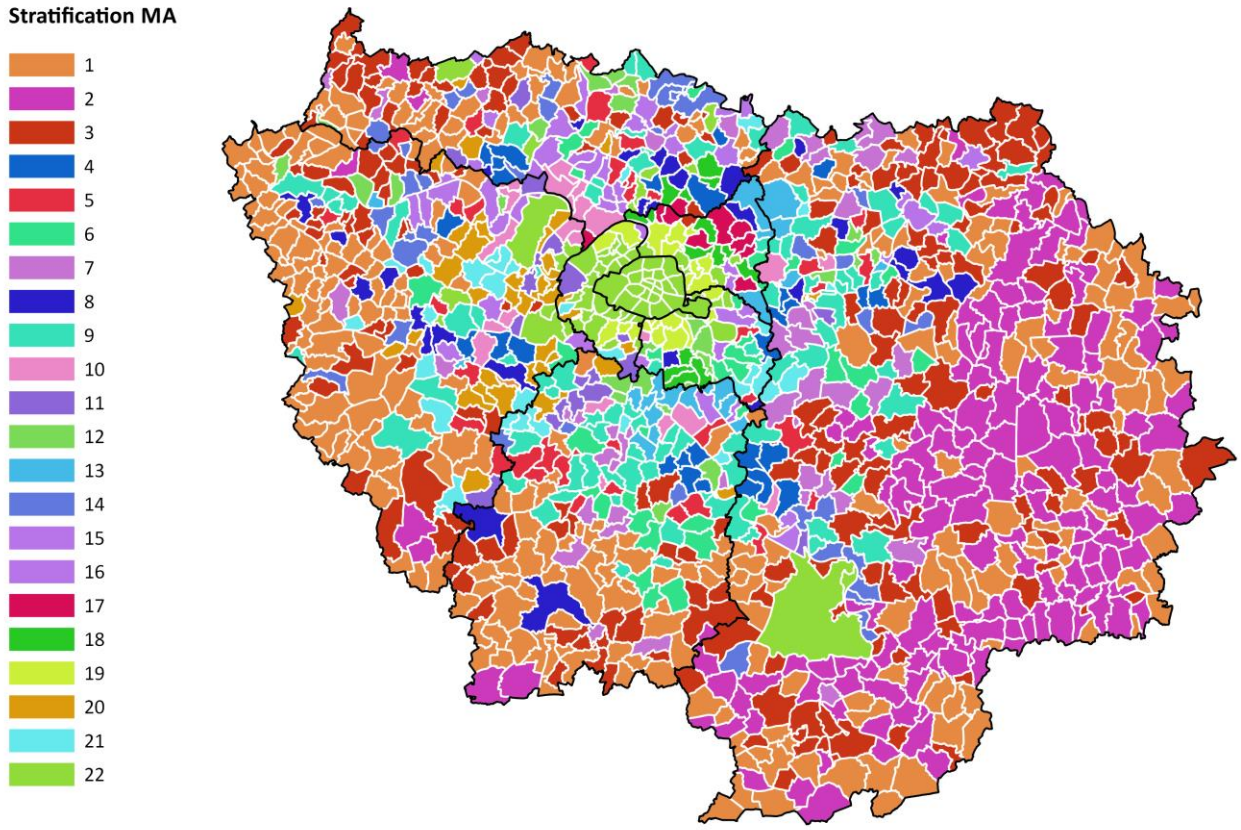


Note : la strate 30 représente les communes « orphelines », c'est-à-dire les communes (au nombre de 382) n'enregistrant aucune vente d'appartements sur la période 2003-2013 dans la base BIEN. Elles sont toutes rattachées à la strate 28, qui est la strate la plus proche en matière de typologie de commune basée sur les variables d'aménités du SDES et du recensement Insee (distance à Paris, population, part des résidences principales occupées par des locataires, revenu par ménage médian).

Exemples de strates : la strate 3 regroupe des communes à la périphérie de la petite couronne (Massy, Chelles, Poissy, Chevilly-Larue, Orly, Thiais, Franconville, Villiers-Le-Bel), situées en moyenne à 16 km du centre de Paris. Ce sont plutôt des grands appartements (8 % sont des studios, 34 % ont 4 pièces ou plus), construits sur la période après-guerre (74 % ont été construits entre 1948 et 1980, 24 % après 1980).

La strate 28 est surtout composée de communes de la grande couronne, situées en moyenne à 35 km du centre de Paris. Ce sont plutôt des petits appartements (22 % sont des studios, 13 % ont 4 pièces ou plus), construits récemment (17 % ont été construits entre 1948 et 1980, 55 % après 1980).

Figure 3-4 : Stratification retenue pour les maisons en Île-de-France



Note : 3 communes sont des communes « orphelines » pour les maisons d'Île-de-France. Elles ont été rattachées à la strate 1, qui est la strate la plus proche en matière de typologie de commune basée sur les variables d'aménités du SDES et du recensement Insee (distance à Paris, population, part des résidences principales occupées par des locataires, revenu par ménage médian).
Exemples de strates : la strate 1 regroupe 302 communes « rurales » de 2 000 habitants en moyenne, se situant en périphérie de la grande couronne, en moyenne à 50 km du centre de Paris. Cette strate est caractérisée par des maisons avec de grands terrains (62 % des maisons de cette strate ayant une surface de terrain supérieure au 3e quartile de l'Île-de-France), et relativement récentes (plus de 50 % des maisons construites après 1970). La part des résidences principales occupées par des locataires dans cette strate est de 16 %.
À l'inverse, la strate 22 regroupe 54 communes « urbaines » de 11 000 habitants en moyenne (dont Paris et 36 communes de la petite couronne), en moyenne à 13 km du centre de Paris. Cette strate concentre des maisons avec des petits terrains (43 % des maisons de cette strate ont une surface de terrain inférieure au 1er quartile de l'Île-de-France), et plus anciennes (seules 24 % ont été construites après 1970). La part des résidences principales occupées par des locataires dans cette strate est de 47 %. Ces maisons sont construites plus en hauteur, car elles possèdent dans un tiers des cas 2 étages ou plus, alors que seules 12 % des maisons d'Île-de-France ont plus d'un étage.

En province, on ne retrouve pas un effet central unique comme celui de Paris en Île-de-France. Il peut de plus exister des marchés locaux très cloisonnés telles les zones touristiques (stations balnéaires, stations de ski). Lors de la mise en place de la version 4 des indices, de nombreux tests ont été réalisés dans l'idée d'établir une stratification pour la province *via* des méthodes de classification, sur le même principe que pour l'Île-de-France, en ajoutant des variables distinguant les communes littorales et les stations de sport d'hiver et en utilisant non plus la distance à Paris mais la distance à la commune de plus de 50 000 habitants la plus proche. Cependant, la stratification obtenue par ce biais permettait une séparation des prix légèrement moins bonne que celle définie

précédemment et n'a donc pas été retenue. La stratification en province a ainsi été reconduite telle qu'elle était en version 3 des indices. La méthodologie est basée sur une démarche de regroupement selon la proximité géographique. Une première sélection a été effectuée à partir des unités de territoire suivantes :

- la commune : maille la plus fine. On isole dans un premier temps les communes qui dépassent le seuil minimum de 110 transactions par trimestre, qui définiront une strate à elles seules. Ceci n'est réalisé dans la pratique que pour les plus grandes communes (il y en a 33 pour les appartements et 9 pour les maisons) ;
- l'unité urbaine : on isole les unités urbaines qui dépassent le seuil minimum de 110 transactions par trimestre, en retirant au préalable les communes qui ont déjà été sélectionnées en tant que strate. Il peut donc y avoir des strates formées par une banlieue seule ou bien par l'agglomération tout entière quand une grande ville n'était pas suffisamment importante pour constituer une strate à elle seule ;
- les stations de sports d'hiver alpines : les communes ayant sur leur territoire une station de ski sont regroupées en strates par massif montagneux (massif du Mont Blanc, etc.) ;
- les communes littorales : les strates sont obtenues par regroupement géographique des communes littorales de même département ou région. Elles se distinguent des autres communes limitrophes par un prix moyen des logements plus élevé.

Toutes les autres communes, soit celles qui n'ont ni un nombre de ventes suffisant par trimestre pour être distinguées à ce titre, ni de caractéristiques particulières, ont été réparties en strates d'après les quantiles de revenu moyen par habitant de l'année 2006. Ce travail a été réalisé au cas par cas à partir d'un processus descendant en commençant au niveau des zones économiques et d'aménagement du territoire (ZEAT).

Si le seuil limite de transactions par trimestre pour une ZEAT était largement dépassé, on descendait au niveau régional. Il en va de même pour une région donnée, pour laquelle on a pu être amené à descendre à un niveau encore plus fin, le département. Dans d'autres cas, deux départements d'une même région ont été regroupés pour former une strate.

Depuis la mise en place de la V4, les 4 Drom et la Corse ont été introduits dans le champ des indices : la Corse forme à elle seule une strate et les 4 Drom ont été réunis dans la même strate faute d'activité suffisante pour pouvoir les étudier séparément.

Au total, on obtient en province 99 strates pour les appartements et 176 strates en maisons. Les cartes de la stratification « province » pour la métropole hors Corse sont fournies dans l'annexe 2 de l'*Insee Méthodes* V3 de juillet 2014 (https://insee.fr/fr/information/2569926).

4 – Calculs mis en œuvre

4.1 – Les modèles hédoniques

Le calcul des indices des prix des logements anciens repose sur la construction de modèles hédoniques : ceux-ci visent à lier mathématiquement le prix des biens à leurs caractéristiques.

Un modèle distinct est estimé pour chaque strate, le lien entre le prix et les caractéristiques pouvant varier d'un marché à un autre.

Les modèles sont de la forme :

$$\log(p_{s,i}) = \log(p_{0,s}) + \mu_{A-1,s}.Y_{A-1,s,i} + \sum_{m=1}^{11} \theta_{m,s}.M_{m,s,i} + \sum_{k=1}^{K_s} \beta_{k,s}.X_{k,s,i} + \epsilon_{s,i} \quad (1)$$

avec les notations suivantes :
s : la strate (relative à un type de bien donné) sur laquelle on considère le modèle
$p_{s,i}$: le prix au m² (pour les appartements) ou le prix total (pour les maisons) du bien i de la strate s
$p_{0,s}$: le prix de référence de la strate s : il s'agit du prix moyen qu'auraient les biens de la strate pour un jeu de caractéristiques fixé ; on cherchera à suivre l'évolution de ce prix pour calculer un indice sur la strate.
$Y_{A-1,s,i}$: valeur de l'indicatrice de l'année de mutation pour le bien i de la strate s (on estime en pratique les modèles sur deux années de transactions consécutives notées A-1 et A ; il y a une seule indicatrice pour l'année A-1, car A est l'année de référence)
$M_{m,s,i}$: valeur de l'indicatrice du mois de mutation pour le bien i de la strate s, indexée par m allant de 1 à 11 (le mois de décembre est le mois de référence et ne dispose donc pas d'indicatrice associée)
$X_{k,s,i}$: valeur de la k-ième variable caractéristique du modèle pour le bien i de la strate s, indexée par k allant de 1 à K_s (K_s étant le nombre de variables caractéristiques utilisées dans le modèle associé à la strate s)
$\mu_{A-1,s}$: coefficient associé à l'indicatrice d'année $Y_{A-1,s}$
$\theta_{m,s}$: coefficient associé à l'indicatrice de mois $M_{m,s}$
$\beta_{k,s}$: coefficient associé à la variable $X_{k,s}$
$\epsilon_{s,i}$: terme d'erreur pour le bien i de la strate s

Ils sont estimés sur une période passée, appelée « période de référence », qui correspond au flux des transactions observées sur deux années. C'est la raison pour laquelle sont introduites les

variables $Y_{A-1,s}$ et $M_{m,s}$, pour prendre en compte les variations de prix au cours de ces deux ans.
La période de référence est constituée des années *n-3* et *n-2* pour les indices des années *n* et *n+1*[21] (avec *n* pair), le parc étant mis à jour tous les deux ans.

La méthode hédonique implémentée repose sur l'hypothèse que la valorisation des caractéristiques des biens (les coefficients $\beta_{k,s}$) est stable tout au long d'une période comprenant la période de référence sur laquelle les modèles sont estimés (*années n-3 et n-2*) ainsi que la période associée sur laquelle on va calculer les indices (*années n et n+1*), appelée « période courante »[22].

On réalise deux phases d'estimation, avec la méthode des moindres carrés ordinaires qui fournit la meilleure estimation non biaisée des coefficients $\mu_{A-1,s}$, $\theta_{m,s}$ et $\beta_{k,s}$.

La première phase a pour but d'exclure les observations atypiques, c'est-à-dire celles pour lesquelles la valeur estimée s'écarte trop de la valeur réelle, qui correspondent souvent à des logements avec des caractéristiques très particulières et auxquels la valorisation décrite par l'équation hédonique s'applique mal. L'objectif poursuivi est de produire une estimation plus robuste des coefficients, donc moins sensible aux caractéristiques particulières de l'échantillon de données exploité.

En notant $\log(\hat{p_{0,s,temp}})$, $\hat{\mu_{A-1,s,temp}}$, $\hat{\theta_{m,s,temp}}$, et $\hat{\beta_{k,s,temp}}$ les estimateurs obtenus à l'issue de cette première phase d'estimation, la relation suivante est vérifiée :

$$\log(p_{s,i}) = \log(\hat{p_{0,s,temp}}) + \hat{\mu_{A-1,s,temp}} \cdot Y_{A-1,s,i} + \sum_{m=1}^{11} \hat{\theta_{m,s,temp}} \cdot M_{m,s,i} + \sum_{k=1}^{K_s} \hat{\beta_{k,s,temp}} \cdot X_{k,s,i} + \hat{\epsilon_{s,i,temp}}$$

où les $\hat{\epsilon_{s,i,temp}}$ sont appelés résidus et correspondent à la différence entre la valeur observée et la valeur estimée de $\log(p_{s,i})$.

Il existe plusieurs critères de détection des points atypiques, on les définit ici comme ceux dont la valeur estimée par le modèle s'écarte de la valeur réelle de plus de deux écarts-types.
Plus précisément, on calcule les résidus standardisés $\hat{r_{s,i,temp}}$:

$$\hat{r_{s,i,temp}} = \frac{\hat{\epsilon_{s,i,temp}}}{\hat{\sigma}_s \cdot \sqrt{1 - h_{s,ii}}}$$

---

[21] On cherche à travailler sur une période de référence proche de celle des indices ; cependant, les données de l'année n-1 ne sont pas encore toutes enregistrées lorsqu'on calcule les modèles utilisés lors de l'année n, ce pourquoi on s'en tient aux années n-3 et n-2 qui elles sont complètes en matière de capacité de collecte et d'atteinte d'un taux de couverture maximal.

[22] Une autre méthode hédonique, dite par « périodes adjacentes », a été testée dans le cadre de la mise en place de la v4 des indices. Chaque trimestre, un nouveau modèle est estimé sur les données des deux derniers trimestres en incluant une indicatrice temporelle comme variable explicative. L'estimateur associé à cette indicatrice permet alors de rendre compte de l'évolution des prix entre les deux trimestres, toutes choses égales par ailleurs. En se basant uniquement sur les deux derniers trimestres, cette méthode présente l'avantage de s'adapter rapidement si le lien entre le prix et les caractéristiques évolue au cours du temps mais, *a contrario*, la précision des estimateurs est plus faible, car l'échantillon est plus restreint. Au total, les résultats obtenus étaient similaires à ceux produits avec la méthode à qualité fixée, ce qui étaye la robustesse des chiffres produits.

Avec :

$\hat{\sigma}_s$: la racine carrée de la variance estimée de $\epsilon_{s,i}$ égale à $\hat{\sigma}_s^2 = \frac{\sum_{1}^{n_s} \hat{\epsilon_{s,i,temp}}}{n_s - (p_s + 1)}$

n_s : le nombre d'observations dans la strate

p_s : le nombre total de variables (caractéristiques du bien et indicatrices d'année ou de mois) dans le modèle associé à la strate

$h_{s,ii} = x_{s,i}' (X_s' X_s)^{-1} x_{s,i}$: l'effet levier de l'observation i où X_s est la matrice de taille n_s *(p_s +1) représentant les valeurs des variables du modèle (ainsi que la constante) pour l'ensemble des observations de la strate s et où $x_{s,i}$ désigne le vecteur de taille 1*(p_s +1) regroupant les valeurs des variables pour l'observation i de la strate s.

On supprime alors toutes les observations ayant un résidu standardisé sortant de l'intervalle]-2 ; 2[. Sur le parc 2015-2016, 4,6 % des appartements anciens et 4,9 % des maisons anciennes ont été supprimés à cette étape aussi bien en Île-de-France qu'en province.

À l'issue de cette étape, l'ensemble des observations encore présentes compose ce que l'on appelle le parc de référence : il s'agit donc de l'ensemble des transactions du champ sur la période de référence et n'ayant pas été détectées atypiques. C'est sur ce parc qu'on va effectuer la seconde phase d'estimation pour produire les modèles finaux.

Ce parc de référence constitue le « panier de biens » ou « portefeuille » dont on va mesurer la variation de prix pour calculer les indices. On s'assure ainsi que les indices retracent l'évolution de prix des mêmes logements et qu'ils ne sont pas sensibles aux variations de la structure du marché (il s'agira alors de valoriser les logements du parc de référence lors de la période courante grâce aux modèles, bien qu'ils n'aient pas été vendus à nouveau).

Finalement, on effectue la seconde phase d'estimation pour obtenir les estimateurs définitifs $\log(\hat{p}_{0,s})$, $\hat{\mu_{A-1,s}}$, $\hat{\theta_{m,s}}$ et $\hat{\beta_{k,s}}$ caractérisant le modèle :

$$\log(p_{s,i}) = \log(\hat{p}_{0,s}) + \hat{\mu_{A-1,s}} . Y_{A-1,s,i} + \sum_{m=1}^{11} \hat{\theta_{m,s}} . M_{m,s,i} + \sum_{k=1}^{K_s} \hat{\beta_{k,s}} . X_{k,s,i} + \hat{\epsilon_{s,i}} \quad (2)$$

4.2 – Les variables caractéristiques

Les variables caractéristiques des biens X_k sont calculées à partir des caractéristiques initiales disponibles dans les bases (voir partie 2.4). Elles peuvent inclure des effets croisés[23] ou des regroupements de valeurs au sein d'une même modalité[24]. Elles sont par ailleurs codifiées dans un format adapté à leur inclusion dans les modèles de régression : elles peuvent être des variables indicatrices (prenant les valeurs 0 ou 1) pour les caractéristiques avec un nombre fini de modalités ou des variables continues.

On définit une valeur de référence pour chaque caractéristique et on appelle « logement de référence » le logement possédant l'ensemble des valeurs de référence. Les indices calculés seront les mêmes quelles que soient les valeurs fixées, mais le choix d'une référence est impératif lorsqu'on souhaite implémenter l'ensemble des modalités d'une même caractéristique au sein d'un modèle afin d'éviter la colinéarité des variables explicatives. Ainsi, pour une caractéristique possédant m modalités, on définira m-1 variables indicatrices.

Par exemple, la caractéristique *nombre de salles de bains* est scindée en trois modalités : *0*, *1* et *2 ou plus*. On choisit *1* comme référence et on définit alors deux variables indicatrices pour les autres modalités :

- $I_{0\ salle\ de\ bain}$ qui prendra la valeur 1 si le logement ne possède pas de salle de bains et 0 sinon ;
- $I_{2\ salles\ de\ bain\ ou\ plus}$ qui prendra la valeur 1 si le logement possède deux salles de bains ou plus et 0 sinon.

Si ces deux variables sont utilisées dans un modèle, alors on calculera les estimateurs des effets associés :

- $\hat{\beta}_{0\ salle\ de\ bain}$ qui va chiffrer l'impact sur le logarithme du prix de ne pas avoir de salle de bains plutôt que d'en avoir une (une salle de bains étant la référence) ;
- $\hat{\beta}_{2\ salles\ de\ bain\ ou\ plus}$ qui va chiffrer l'impact sur le logarithme du prix d'avoir deux salles de bains ou plus plutôt que d'en avoir une.

L'ensemble des caractéristiques introduites dans les modèles sont listées dans les figures 4-1 et 4-2.

[23] On croise des variables quand celles-ci sont très corrélées entre elles pour éviter les problèmes de colinéarité.

[24] On regroupe des valeurs pour deux raisons : d'une part pour avoir un nombre suffisant d'observations pour chaque modalité et que l'estimation du coefficient associé soit suffisamment précise, d'autre part pour avoir des modalités ayant un sens économique sur la formation du prix. On ne dénote par exemple pas de différence de prix entre le 5e et 6e étage, donc il n'est pas nécessaire de les considérer séparément ; en revanche, ce sera le cas entre le RDC et le 1er étage.

Figure 4-1 : Variables introduites dans les modèles – Appartements

Variables caractéristiques des appartements	
Caractéristiques	**Modalités (les références sont signalées en gras)**
Époque de construction	Époque AB : avant 1913 Époque C : de 1914 à 1947 **Époque D : de 1948 à 1969** Époque E : de 1970 à 1980 Époque F : de 1981 à 1991 Époque G : de 1992 à 2000 Époque HI : depuis 2001 Époque X : non renseigné
Nombre de salles de bains	0 **1** 2 ou plus
Nombre de garages ou parkings	**0** 1 2 ou plus
État du bien	**B : bon état** M : travaux à prévoir V : à rénover XX : non renseigné
Étage croisé avec la présence d'un ascenseur	**Rez-de-chaussée** 1er étage 2e étage 3e étage Plus que 3e étage avec ascenseur Plus que 3e étage sans ascenseur
Présence d'une cave	**Pas de cave** Au moins une cave
Présence d'une terrasse ou balcon ou loggia	**Sans terrasse ni balcon ni loggia** Avec terrasse ou balcon ou loggia Non renseigné
Nombre de pièces croisé avec la surface (sous forme de la surface moyenne par pièce)	Studio - petit : <20 m² **Studio - moyen : entre 20 et 30 m²** Studio - grand : >30 m² 2 pièces - petit : <17 m² **2 pièces - moyen : entre 17 et 24 m²** 2 pièces - grand : >24 m² 3 pièces - petit : <18 m² **3 pièces - moyen : entre 18 et 23 m²** 3 pièces - grand : >23 m² 4 pièces - petit : <17 m² **4 pièces - moyen : entre 17 et 21 m²** 4 pièces - grand : >21 m² 5 pièces ou + - petit : <16 m² **5 pièces ou + - moyen : entre 16 et 22 m²** 5 pièces ou + - grand : >22 m²

Variables caractéristiques des appartements	
Étiquette du diagnostic de performance énergétique	A ou B : <=90 Kw/h C : 91 à 150 Kw/h D : 151 à 230 Kw/h E : 231 à 330 Kw/h F ou G : >330 Kw/h **X : Non renseigné**
Distance de la commune d'appartenance à la grande commune[25] la plus proche (en kilomètres, calculée de centre à centre)	variable numérique continue - référence 0 km en province, 15 km en Île-de-France
Zonage détaillé croisé avec le nombre de pièces (en 5 classes : 1, 2, 3, 4 et 5 pièces ou plus)	Il existe 5 modalités par zone (pour chaque modalité du nombre de pièces). **Référence : zone 1 pour les 3 pièces**

Figure 4-2 : Variables introduites dans les modèles - Maisons

Variables caractéristiques des maisons	
Caractéristiques	**Modalités (les références sont signalées en gras)**
Nombre de niveaux	1 **2 ou plus**
Nombre de salles de bains	0 **1** 2 ou plus
Nombre de garages ou parkings	0 **1** 2 ou plus Non renseigné
Époque de construction	Époque AB : avant 1913 (inclus) Époque C : de 1914 à 1947 Époque D : de 1948 à 1969 Époque E : de 1970 à 1980 Époque F : de 1981 à 1991 Époque GHI : depuis 1992 **Époque X : non renseigné**
Logarithme de la surface habitable	*variable numérique continue - référence :* ***ln(100)***
Logarithme de la surface du terrain	*variable numérique continue - référence :* ***ln(610)***
État du bien	B : bon état M : travaux à prévoir V : à rénover **XX : non renseigné**
Présence d'un sous-sol / cave	**pas de sous-sol / cave** présence d'un sous-sol / cave

[25] Il s'agit de la distance à la commune de 50 000 habitants la plus proche pour les strates de province et de la distance à Paris pour les strates d'Île-de-France.

Variables caractéristiques des maisons	
Étiquette du diagnostic de performance énergétique	A ou B : <=90 Kw/h C : 91 à 150 Kw/h D : 151 à 230 Kw/h E : 231 à 330 Kw/h F ou G : >330 Kw/h **X : Non renseigné**
Distance de la commune d'appartenance à la grande commune la plus proche[26] (en kilomètres, calculée de centre à centre)	variable numérique continue - référence 0 km en province, 27 km en Île-de-France
Zonage détaillé croisé avec le nombre de pièces (en 5 classes : moins de 4, 4, 5, 6 et 7 pièces ou plus)	Il existe 5 modalités par zone (pour chaque modalité du nombre de pièces). **Référence : zone 1 pour les 4 pièces**

Le zonage détaillé croisé avec le nombre de pièces dans l'étape de modélisation[27] n'a pas le même rôle que les strates pour prendre en compte l'effet de la localisation sur les prix : une strate correspond à un périmètre sur lequel on estime un modèle hédonique (cf. partie 3), tandis que la zone (croisée avec le nombre de pièces) correspond à une variable explicative du modèle. En d'autres termes, on suppose qu'au sein d'une strate, un logement peut être valorisé différemment selon la zone détaillée dans laquelle il se trouve et son nombre de pièces, mais que les effets des autres caractéristiques (époque de construction, nombre de salles de bains…) sont les mêmes partout à l'intérieur de la strate (ils peuvent en revanche varier d'une strate à une autre).

Ces zones ont été construites de la manière suivante :

- elles doivent respecter un seuil minimum de 110 transactions en moyenne par an ;
- au sein des strates des appartements parisiens, le zonage détaillé correspond simplement au quartier administratif. Pour les strates des maisons et celles des appartements du reste de l'Île-de-France, le zonage détaillé correspond au département. En effet, comme les strates peuvent regrouper des communes éloignées géographiquement, l'utilisation du département permet de prendre en compte, le cas échéant, des niveaux de prix différents liés à la situation géographique ;
- en province, le zonage détaillé a été construit sur le même principe que celui des strates, à l'exception du seuil minimal de transactions qui a été porté à 110 transactions annuelles. Pour les communes qui forment une strate à elles seules, la segmentation a été établie par dires d'experts : les notaires locaux, interrogés sur la question, ont défini des unités de quartiers d'après leur connaissance du marché immobilier. Pour la strate regroupant les 4 Drom, ainsi que celle comprenant la Corse, le zonage détaillé correspond au département.

[26] Il s'agit de la distance à la commune de 50 000 habitants la plus proche pour les strates de province et de la distance à Paris pour les strates d'Île-de-France.

[27] L'examen des données montre que la variation de prix d'une zone à une autre se différencie selon le nombre de pièces des logements. Par exemple, passer d'un quartier résidentiel à un centre-ville pourra se répercuter différemment sur les prix des studios ou sur les prix des quatre-pièces pour lesquels la demande n'est pas la même.

Toutes les variables fournies dans les figures 4-1 et 4-2 ne feront pas forcément partie du modèle retenu pour une strate donnée. Une caractéristique peut en effet ne pas avoir le même impact d'un marché à un autre et notamment n'est pas toujours explicative du prix dans une strate particulière. On peut par exemple penser que la présence d'un garage ne sera pas déterminante sur le prix des logements dans une zone où il est facile de se garer à l'extérieur tandis que cela sera fortement valorisé dans une zone où il existe très peu de places de stationnement.

On souhaite donc, pour chaque strate, faire entrer dans l'équation associée uniquement les variables les plus corrélées avec le prix. Limiter le nombre de variables explicatives permet d'accroître la validité et la robustesse du modèle[28] et réduit aussi par la même occasion le risque de colinéarité entre les variables. Pour cela, on a recours à une procédure de sélection de variables[29]. Il existe diverses méthodes de sélection ; on utilise celle dite ascendante (*méthode Forward*)[30].

La méthode Forward est une méthode de pas à pas, où l'on injecte une variable après l'autre. On initialise un modèle sans variables, puis à chaque itération on retient, parmi l'ensemble des variables proposées, celle qui améliore le plus le modèle selon un critère donné. Le critère utilisé ici, le plus classique, est un test de Fischer. Lors de l'ajout de la j-ème variable, la statistique de test pour une variable v est :

$$F_v = \frac{SSE_{M(j-1)} - SSE_{M(j-1,v)}}{\frac{SSE_{M(j-1,v)}}{n-j-1}}$$

avec $SSE_{M(j-1)}$: la somme des carrés des résidus du modèle avec les j-1 variables précédemment retenues ;

$SSE_{M(j-1,v)}$: la somme des carrés des résidus du modèle avec les j-1 variables précédemment retenues et la variable v ;

n : le nombre d'observations.

On retient la variable pour laquelle on obtient la plus faible p-value associée au test. La procédure se termine lorsqu'on atteint une itération où aucune variable n'est jugée suffisamment pertinente pour intégrer le modèle selon un critère d'entrée ; dans notre cas, la procédure s'arrête lorsqu'aucune variable ne produit de p-value inférieure à 0,05.

---

[28] En incluant des variables peu explicatives du prix, on risque de sur-ajuster le modèle, c'est-à-dire d'obtenir un modèle avec un grand nombre d'estimateurs qui s'adaptera trop spécifiquement aux données de l'échantillon utilisé pour la régression (ici le parc de référence) et qui s'appliquera alors beaucoup moins bien pour tous les logements n'ayant pas fait partie de cet échantillon. Pour jauger de la qualité générale d'un modèle dans ce cadre, on cherche à obtenir à la fois un R^2 élevé, représentant la qualité de l'estimation sur l'échantillon, et un PRESS faible, représentant la qualité de prévision hors échantillon par une forme de validation croisée.

[29] Il est parfois préféré de tester manuellement l'inclusion des variables. Le recours à une procédure de sélection est ici très avantageux par son caractère automatique, étant donné le grand nombre de modèles à estimer.

[30] Différentes méthodes de sélection et différents critères ont été testés dans le cadre de la mise en place de la version 4 des indices, aboutissant à des résultats très similaires, ce qui est positif quant à la robustesse des variables sélectionnées.

4.3 – Les indices élémentaires pour un parc donné

Les indices élémentaires sont les indices calculés au niveau géographique le plus fin, c'est-à-dire au niveau des strates : ils représentent le rapport entre la valeur courante d'un parc de logements de référence et sa valeur sur une période de base.

On note $I_{t/b}(s)$ l'indice de la strate s pour la période t par rapport à la base b. Cette période de base, que l'on va noter 0, est fixée au dernier trimestre de l'année n-1 pour les indices des années *n* et *n+1* (avec *n* pair). La valeur des indices en t représente l'évolution de la valorisation du parc de référence entre 0 et t (ils seront ensuite chaînés pour établir les séries longues et représenteront l'évolution de prix par rapport à une base 100 fixée par ailleurs).

Les logements du parc de référence peuvent être valorisés en période courante à travers la relation hédonique. On décline pour cela le modèle (1) (voir partie 4-1), appliqué uniquement sur la période de référence, afin de l'adapter à l'ensemble de la période d'application du parc ; le prix d'un bien j de la strate s au cours d'une période t peut alors s'exprimer par :

$$\log(p_{s,j,t}) = \log(p_{0,s,t}) + \sum_{k=1}^{K_s} \beta_{k,s}.X_{k,s,j} + \epsilon_{s,j,t} \quad (3)$$

On remarquera que le modèle (1) utilisé uniquement dans la phase d'estimation sur la période de référence est compatible avec ce dernier modèle. L'introduction des indicatrices d'années et de mois vient du fait que le parc de référence couvre une période longue sur laquelle peuvent avoir lieu des variations de prix. De façon équivalente, le prix de référence de la strate s pour le mois *(a,m)* de la période de référence serait :

$$\log(p_{0,s,a,m}) = \log(p_{0,s}) + \mu_{a,s} + \theta_{m,s}$$

Avec $\mu_{a,s}=0$ si a est l'année de référence et $\theta_{m,s}=0$ si m est le mois de référence.

On cherche maintenant à estimer $\log(p_{0,s,t})$ en période courante. On ne va pas estimer le modèle (3) par les moindres carrés ordinaires, comme on l'a fait précédemment pour le modèle (1), car on va travailler en période courante sur des périodes plus courtes (un trimestre), qui de plus sont récentes par rapport à la date du calcul et donc pour lesquelles toutes les transactions n'auront pas encore été enregistrées. On a alors peu d'observations, ce qui engendrerait des estimateurs peu précis.

On va en revanche pouvoir réutiliser les estimateurs $\hat{\beta}_{k,s}$, plus fiables, obtenus lors de l'estimation (2) sur le parc de référence tronqué des valeurs atypiques, vu qu'on suppose que les effets des caractéristiques $\beta_{k,s}$ sont stables tout au long de la période.

Tout d'abord, on introduit $\tilde{p_{s,j,t}}$ le « prix équivalent bien de référence » du bien j de la strate s au temps t, tel que :

$$\log(\tilde{p_{s,j,t}}) = \log(p_{0,s,t}) + \epsilon_{s,j,t}$$

D'après la relation (3), cette quantité vaut également :

$$\log(\tilde{p_{s,j,t}}) = \log(p_{s,j,t}) - \sum_{k=1}^{K_s} \beta_{k,s}.X_{k,s,j} = \log(p_{0,s,t}) + \epsilon_{s,j,t}$$

En passant à la moyenne sur l'ensemble des $J_{s,t}$ biens échangés dans la strate s au cours de la période t, on obtient[31] :

$$\frac{\sum_{j=1}^{J_{s,t}} \log(\tilde{p_{s,j,t}})}{J_{s,t}} = \frac{\sum_{j=1}^{J_{s,t}} \left[\log(p_{s,j,t}) - \sum_{k=1}^{K_s} \beta_{k,s}.X_{k,s,j} \right]}{J_{s,t}} = \frac{\sum_{j=1}^{J_{s,t}} \left[\log(p_{0,s,t}) + \epsilon_{s,j,t} \right]}{J_{s,t}} = \log(p_{0,s,t})$$

En utilisant les estimateurs $\hat{\beta}_{k,s}$, on va pouvoir estimer $\log(\tilde{p_{s,j,t}})$ par :

$$\log(\hat{\tilde{p}}_{s,j,t}) = \log(p_{s,j,t}) - \sum_{k=1}^{K_s} \hat{\beta}_{k,s}.X_{k,s,j}$$

puis $\log(p_{0,s,t})$ par la moyenne des $\log(\hat{\tilde{p}}_{s,j,t})$. Par souci de robustesse, on n'utilise pas pour cela l'ensemble des $J_{s,t}$ biens : on retire, au sein de chaque zone détaillée de la strate, les 2 % des observations avec $\log(\hat{\tilde{p}}_{s,j,t})$ le plus élevé et les 2 % avec $\log(\hat{\tilde{p}}_{s,j,t})$ le plus faible. En notant $J'_{s,t}$ le nombre d'observations restantes, on obtient :

$$\log(\hat{p}_{0,s,t}) = \frac{\sum_{j=1}^{J'_{s,t}} \log(\hat{\tilde{p}}_{s,j,t})}{J'_{s,t}} = \frac{\sum_{j=1}^{J'_{s,t}} \left[\log(p_{s,j,t}) - \sum_{k=1}^{K_s} \hat{\beta}_{k,s}.X_{k,s,j} \right]}{J'_{s,t}}$$

On note $\hat{p_{0,s,t}} = \exp(\log(\hat{p}_{0,s,t}))$ l'estimation du prix de référence de la strate s à la période t.

On peut maintenant estimer la valeur du parc de référence pendant une période t.

On considère un logement i du parc de référence sur la strate s. On va noter $\log(\hat{p}_{s,i,t})$ l'estimation du logarithme de son prix à la période t calculé à partir des précédents estimateurs :

$$\log(\hat{p}_{s,i,t}) = \log(\hat{p}_{0,s,t}) + \sum_{k=1}^{K_s} \hat{\beta}_{k,s}.X_{k,s,i}$$

et $\hat{p_{s,i,t}} = \exp(\log(\hat{p}_{s,i,t}))$ l'estimation de son prix (au mètre carré pour les appartements, total pour les maisons) à la période t :

$$\hat{p_{s,i,t}} = \exp(\log(\hat{p}_{0,s,t}) + \sum_{k=1}^{K_s} \hat{\beta}_{k,s}.X_{k,s,i}) = \hat{p_{0,s,t}}.\exp(\sum_{k=1}^{K_s} \hat{\beta}_{k,s}.X_{k,s,i})$$

Pour les appartements, on multiplie de plus le prix au mètre carré estimé par la surface pour obtenir l'estimation du prix total.

Afin d'homogénéiser les notations, on introduit $\hat{P_{s,i,t}}$ l'estimation du prix total :

$$\hat{P_{s,i,t}} = \hat{p_{s,i,t}}.A_{s,i} = \hat{p_{0,s,t}}.A_{s,i}.\exp(\sum_{k=1}^{K_s} \hat{\beta}_{k,s}.X_{k,s,i})$$

[31] Les termes d'erreur $\epsilon_{s,j,t}$ sont de moyenne nulle par définition.

Avec $A_{s,i}$ valant la surface du bien i en mètres carrés pour les strates d'appartements et $A_{s,i}=1$ pour les maisons.

L'estimation de l'évolution du prix d'un logement, entre deux périodes t_1 et t_2, est la même pour tous les logements de la strate :

$$\frac{\hat{P_{s,i,t_2}}}{\hat{P_{s,i,t_1}}}=\frac{\hat{p_{0,s,t_2}}.A_{s,i}.\exp(\sum_{k=1}^{K_s}\hat{\beta_{k,s}}.X_{k,s,i})}{\hat{p_{0,s,t_1}}.A_{s,i}.\exp(\sum_{k=1}^{K_s}\hat{\beta_{k,s}}.X_{k,s,i})}=\frac{\hat{p_{0,s,t_2}}}{\hat{p_{0,s,t_1}}}$$

La valorisation du parc de référence en t pour une strate s donnée, notée $\hat{W_{s,t}}$, correspond à la somme des prix estimés en t de tous les logements composant le parc de référence pour cette strate :

$$\hat{W_{s,t}}=\sum_{i=1}^{N_s}\hat{P_{s,i,t}}=\sum_{i=1}^{N_s}\left(\hat{p_{0,s,t}}.A_{s,i}.\exp(\sum_{k=1}^{K_s}\hat{\beta_{k,s}}.X_{k,s,i})\right)=\hat{p_{0,s,t}}.\sum_{i=1}^{N_s}\left(A_{s,i}.\exp(\sum_{k=1}^{K_s}\hat{\beta_{k,s}}.X_{k,s,i})\right) \quad (4)$$

avec N_s le nombre de logements composant le parc de référence pour la strate s.

Ainsi, en calculant l'évolution de la valorisation du parc de référence entre deux périodes t_1 et t_2, ou, en d'autres termes, l'évolution de l'indice élémentaire de la strate s entre t_1 et t_2, on retrouve :

$$\frac{\hat{W_{s,t_2}}}{\hat{W_{s,t_1}}}=\frac{\hat{p_{0,s,t_2}}.\sum_{i=1}^{N_s}\left(A_{s,i}.\exp(\sum_{k=1}^{K_s}\hat{\beta_{k,s}}.X_{k,s,i})\right)}{\hat{p_{0,s,t_1}}.\sum_{i=1}^{N_s}\left(A_{s,i}.\exp(\sum_{k=1}^{K_s}\hat{\beta_{k,s}}.X_{k,s,i})\right)}=\frac{\hat{p_{0,s,t_2}}}{\hat{p_{0,s,t_1}}}$$

De là, on peut déduire la valeur des indices élémentaires : il reste à choisir une période de base par rapport à laquelle on mesure la variation de la valeur du parc de référence[32]. Pour les indices trimestriels des années *n* et *n+1*, calculés sur le parc de référence contenant les transactions des années *n-3* et *n-2*, on place la période de base *au 4e trimestre de l'année n-1* que l'on appelle t_0 [33]

[32] Le choix de la période de base n'a pas d'impact sur les évolutions de prix calculées, il s'agit simplement d'une référence. Il est très simple de changer la base une fois un indice calculé, on divise pour cela toutes les valeurs de l'indice par la valeur qu'il prend à la nouvelle période de base souhaitée.

[33] L'Insee publie uniquement les indices trimestriels (les seuls ayant l'appellation « Notaires-Insee »). ADNOV et PNS calculent et publient également des indices mensuels sur trimestre glissants calculés selon la même méthodologie et dénommés par le dernier mois de la période. Par exemple, l'indice de février d'une année A sera calculé sur une période comprenant décembre A-1, janvier A et février A. Les indices de mars, juin, septembre et décembre correspondent alors aux indices trimestriels. Dans ce cadre et par souci de continuité, les indices de janvier, avril, juillet et octobre des années *n* et *n+1* sont alors en base t_0' (août, septembre et octobre de l'année *n-1*). De même les indices de février, mai, août et novembre des années *n* et *n+1* sont en base t_0'' (septembre, octobre et novembre de l'année *n-1*).

On obtient alors les indices élémentaires :

$$I_{t/t_0}(s) = \frac{\hat{W}_{s,t}}{\hat{W}_{s,t_0}} = \frac{\hat{p}_{0,s,t}}{\hat{p}_{0,s,t_0}}$$

4.4 – Les indices agrégés pour un parc donné

Une fois les indices élémentaires connus, ceux-ci sont agrégés à différents échelons pour obtenir les indices sur les niveaux géographiques souhaités. Une agrégation des indices des prix des appartements et de ceux des maisons a aussi lieu selon le même procédé pour obtenir les indices sur l'ensemble des logements.

4.4.1 Correction de la non-exhaustivité des bases

Les bases notariales ne couvrent pas la totalité des biens échangés (cf. partie 2.2). Pour corriger cette non-exhaustivité, on introduit des coefficients de redressement. Ces coefficients sont estimés à partir de données fiscales (montants des droits de mutation établis par la Direction générale des Finances publiques, DGFiP). Comme celles-ci donnent peu de précisions sur les biens échangés, on estime les coefficients sur des agrégats assez larges : par département et par année de mutation, appartements et maisons confondus. On obtient le coefficient $\delta_{a,d}$ en divisant le montant des assiettes des droits de mutation des transactions dans le département d pour l'année a, d'après les données fiscales, par ce même montant enregistré dans la base notariale. On estime un décalage de 2 mois entre la date de l'acte de vente et la comptabilisation des droits par la DGFiP.

Chaque coefficient est appliqué à tous les biens du croisement (département x année de mutation) dans le calcul des termes qui pondèrent les indices.

4.4.2 Calcul des poids

Le poids des différents indices d'un zonage géographique au sein d'un agrégat plus large représente leur part de la valeur totale du parc de référence sur l'agrégat. Cette valeur du parc est calculée sur le dernier trimestre du parc de référence, que l'on appelle période de pondération q, en prenant en compte les coefficients de correction de la non-exhaustivité[34]. Par exemple, pour une strate s :

$$\hat{W}^{\delta}_{s,q} = \sum_{i=1}^{N_s} \delta_{a_i,d_i} . \hat{P}_{s,i,q} = \sum_{i=1}^{N_s} \left(\delta_{a_i,d_i} . \hat{p_0}_{,s,q} . A_{s,i} . \exp\left(\sum_{k=1}^{K_s} \hat{\beta}_{k,s} . X_{k,s,i} \right) \right)$$

avec a_i et d_i l'année et le département du bien i.

[34] La prise en compte ou non de ces coefficients dans la phase de calcul des indices élémentaires n'a pas d'impact sur les résultats obtenus. Elle est cependant nécessaire dans la phase de pondération pour que celle-ci ne soit pas faussée par des taux de couverture inégaux sur différents territoires.

Toujours en prenant l'exemple d'une strate s, son poids au sein d'un agrégat A composé de n_A strates vaut alors :

$$\frac{\hat{W}^{\delta}_{s,q}}{\sum_{u=1}^{n_A} \hat{W}^{\delta}_{u,q}}$$

4.4.3 Indices agrégés aux niveaux infra-départementaux et départementaux

Pour obtenir des indices agrégés à des niveaux infra-départementaux ou départementaux, on calcule une moyenne géométrique[35] sur les indices élémentaires[36]. Pour un agrégat A, infra-départemental ou départemental, composé de n_A strates, l'indice agrégé vaut :

$$I_{t/t_0}(A) = \prod_{s=1}^{n_A} I_{t/t_0}(s)^{\left(\frac{\hat{w}^{\delta}_{s,q}}{\sum_{u=1}^{n_A} \hat{w}^{\delta}_{u,q}}\right)}$$

4.4.4 Indices agrégés aux niveaux supra-départementaux

Pour obtenir des indices agrégés à des niveaux supra-départementaux, on calcule une moyenne arithmétique. Pour un agrégat B supra-départemental, composé de n_B départements, l'indice agrégé vaut :

$$I_{t/t_0}(B) = \sum_{d=1}^{n_B} \left(\frac{\hat{W}^{\delta}_{d,q}}{\sum_{e=1}^{n_B} \hat{W}^{\delta}_{e,q}} \right) . I_{t/t_0}(d)$$

---

[35] Ce mode de calcul est utilisé dans la construction des indices statistiques lorsque l'on pense qu'il peut exister un arbitrage entre différents biens de nature voisine. Dans notre cas, il peut y avoir un choix d'acquisition entre des biens appartenant à des strates géographiques voisines. On utilise des moyennes arithmétiques pour les niveaux plus agrégés que les niveaux départementaux, car l'acheteur a moins de chances d'arbitrer entre des strates géographiques plus éloignées pour s'installer (région, province, France entière).

[36] Certaines strates peuvent être composées de logements appartenant à différents départements. Dans ce cas, le poids de l'indice de la strate au sein d'un département est calculé en utilisant uniquement les logements de la strate appartenant à ce département.

4.5 – Le chaînage

Les indices calculés jusqu'ici sont associés à un parc de référence et sont calculés par rapport à une date t_0 , date qui varie avec le changement de parc de référence tous les deux ans. Les séries d'indices sur l'ensemble de la période sont reconstituées par chaînage.

On introduit une base 100[37], qui sera la base de la série longue : cette base 100 est fixée en V4 à la moyenne annuelle des indices de 2015[38]. La valeur de l'indice en t en cette base 100 représentera l'évolution des prix entre t et la période choisie pour la base 100.

Il faut tout d'abord initialiser les séries. Si on calculait un indice pour la première fois, on pourrait par exemple utiliser la base t_0 du premier parc de référence comme base 100 et les séries calculées sur ce parc comme séries initiales[39]. Dans notre cas, les évolutions de prix, calculées en version 4 des indices à partir du T1 2018, sont chaînées aux séries existantes calculées à partir des versions précédentes des indices. On utilise donc lors du premier chaînage ces séries, référencées en moyenne annuelle 2015, jusqu'au T4 2017.
On note $I_{t/0}(z)$ l'indice en t de la zone z référencé sur la base 100.

Pour distinguer les différents parcs sur lesquels on travaille, on va noter p(i) le i-ème parc de référence utilisé en version 4 des indices et $t_0[p(i)]$ la période de base associée au parc p(i). On a par exemple p(1) le parc composé des transactions des années 2015 et 2016, utilisé pour calculer les indices de 2018 et 2019 avec $t_0[p(1)]=T\,42017$, puis p(2) le parc composé des transactions des années 2017 et 2018, utilisé pour calculer les indices de 2020 et 2021 avec $t_0[p(2)]=T\,42019$ …

Avec ces nouvelles notations, un indice non chaîné calculé sur le parc p(i) en base $t_0[p(i)]$, pour une zone z et une période t, est représenté par $I^{p(i)}_{t/t_0[p(i)]}(z)$.

L'indice chaîné en base 100 pour une zone z et une période t vaut alors :

$$I_{t/0}(z)=I^{p(i)}_{t/t_0[p(i)]}(z).\prod_{j=1}^{i-1} I^{p(j)}_{t_0[p(j-1)]/t_0[p(j)]}(z).I_{2017T4/0}(z)$$

---

37 Les évolutions de prix décrites par les indices sont les mêmes quelle que soit la base 100 choisie et il est facilement possible de changer de base 100 postérieurement aux calculs. Son nom vient du fait que les indices précédemment décrits sont multipliés par 100, ce uniquement pour en faciliter la lecture.

38 En d'autres termes, la base est choisie telle que la moyenne des quatre indices trimestriels bruts de 2015 vaille 100. C'est notamment en moyenne annuelle 2015 qu'Eurostat diffuse différents indices européens.

39 Les notaires calculent également des indices mensuels sur trimestre glissant (chaque indice mensuel est calculé avec trois mois de données). Pour une valeur du parc = 100 en t_0, l'indice en t_0' prendrait la valeur $100.\frac{\hat{W}_{s,t_0'}}{\hat{W}_{s,t_0}}$ et l'indice en t_0'' prendrait la valeur $100.\frac{\hat{W}_{s,t_0''}}{\hat{W}_{s,t_0}}$.

Les indices de janvier, avril, juillet et octobre sont chaînés sur les bases t_0' des différents parcs. De même, les indices de février, mai, août et novembre sont chaînés sur les bases t_0'' des parcs.

Exemple numérique pour une zone z :

L'indice au 2017T4 de la série préexistante en base 100 (moyenne annuelle 2015) est connu.
Supposons qu'il vaille 101 : $I_{2017\,T4/0}(z)=101$

Lorsqu'on se place en 2018T1 : l'indice non chaîné au 2018T1, en base 2017T4, est calculé d'après le parc p(1)=[2015-2016].
Supposons qu'il vaille 1,03 : $I^{[2015-2016]}_{2018\,T1/2017\,T4}(z)=1,03$
L'indice chaîné en base 100 au 2018T1 vaut alors :
$I_{2018\,T1/0}(z)=I^{[2015-2016]}_{2018\,T1/2017\,T4}(z).I_{2017\,T4/0}(z)=1,03*101=104,03$

De même, lorsqu'on se place en 2019T4 : l'indice non chaîné au 2019T4, en base 2017T4, est calculé d'après le parc p(1)=[2015-2016].
Supposons qu'il vaille 1,1 : $I^{[2015-2016]}_{2019\,T4/2017\,T4}(z)=1,1$
L'indice chaîné en base 100 au 2019T4 vaut alors :
$I_{2019\,T4/0}(z)=I^{[2015-2016]}_{2019\,T4/2017\,T4}(z).I_{2017\,T4/0}(z)=1,1*101=111,1$

Plaçons-nous cette fois au 2020T1 : l'indice non chaîné au 2020T1, en base 2019T4, est calculé d'après le parc p(2)=[2017-2018].
Supposons qu'il vaille 0,95 : $I^{[2017-2018]}_{2020\,T1/2019\,T4}(z)=0,95$
L'indice chaîné en base 100 au 2020T1 vaut alors :
$I_{2019\,T4/0}(z)=I^{[2017-2018]}_{2020\,T1/2019\,T4}(z).I^{[2015-2016]}_{2019\,T4/2017\,T4}(z).I_{2017\,T4/0}(z)=0,95*1,1*101=105,545$
équivalent à :
$I_{2019\,T4/0}(z)=I^{[2017-2018]}_{2020\,T1/2019\,T4}(z).I_{2019\,T4/0}(z)=0,95*111,1=105,545$

4.6 – Désaisonnalisation des indices

Les séries d'indices sont publiées en deux versions : une version brute, obtenue après l'étape de chaînage, et une version corrigée des variations saisonnières visant à présenter l'évolution des prix après élimination de l'effet des fluctuations saisonnières au cours de l'année. On observe en effet une saisonnalité sur les indices des prix des logements anciens, même si elle n'est pas très marquée : de manière générale, on observe une légère augmentation des prix au troisième trimestre de l'année, soit pendant la période des vacances plus propice aux déménagements en vue de la rentrée scolaire.

La correction des variations saisonnières (CVS) des séries s'effectue avec la méthode X12-ARIMA[40], *via* le logiciel Demetra, sur les séries d'indices trimestriels glissants diffusées

[40] Cette méthode repose sur un principe itératif d'estimation des différentes composantes, cette estimation étant faite à chaque étape grâce à des moyennes mobiles adéquates.

mensuellement par les notaires[41]. Cette méthode permet notamment de prolonger les séries à l'aide de prévisions par des modèles ARIMA, ce qui permet également de bonnes estimations des coefficients CVS sur les périodes postérieures à l'estimation. Il est également possible d'appliquer une correction pour jours ouvrables, mais les indices n'en font pas l'objet, la composition en jours ouvrables des trimestres n'ayant pas d'impact notable sur la mesure des prix.

Les techniques de désaisonnalisation reposent sur le principe de décomposition des séries en différentes parties : une composante de tendance traduisant l'évolution de fond de la série, une composante saisonnière représentant les fluctuations au sein de l'année et une composante irrégulière comprenant les fluctuations résiduelles. Pour le calcul des coefficients de désaisonnalisation, on peut choisir un modèle additif ou multiplicatif. Le modèle additif suppose que les composantes de la série sont indépendantes les unes des autres. Ainsi, le niveau des variations saisonnières est indépendant du niveau de la série. Le modèle multiplicatif suppose, quant à lui, que les composantes de la série sont dépendantes les unes des autres. Pour les prix des logements, on préfère utiliser le schéma multiplicatif sur l'ensemble des séries[42].

Pour les indices agrégés, il existe deux méthodes de désaisonnalisation. Dans la première, dite directe, on désaisonnalise chaque série indépendamment, quel que soit son niveau d'agrégation. La méthode indirecte, au contraire, consiste à désaisonnaliser d'abord les séries élémentaires, puis à les agréger entre elles pour obtenir les séries désaisonnalisées au niveau supérieur. D'un point de vue théorique, aucune méthode n'est meilleure que l'autre ; cependant, la méthode indirecte présente l'avantage d'assurer la cohérence entre les évolutions sur les différents niveaux d'agrégation. Les désaisonnalisations sont donc ici réalisées de manière indirecte. Pour l'agrégation, les indices élémentaires CVS sont pondérés de la même manière que les indices bruts (cf. partie 4.4). Les séries élémentaires utilisées pour le calcul des séries agrégées sont fournies dans la figure 4-3.

Figure 4-3 : Séries élémentaires utilisées pour le calcul des séries agrégées CVS

Séries agrégées	Séries élémentaires correspondantes
Province	
Appartements des agglomérations de plus de 10 000 habitants	- Appartements villes-centres - Appartements banlieue
Appartements	- Appartements villes-centres - Appartements banlieue - Appartements rural
Appartements et maisons	- Appartements villes-centres - Appartements banlieue

[41] Les modèles CVS sont déterminés sur les séries d'indices « trimestriels glissants » et les coefficients saisonniers sont donc mensuels. Pour les CVS trimestrielles, les coefficients saisonniers appliqués sont ceux du dernier mois de chaque trimestre.

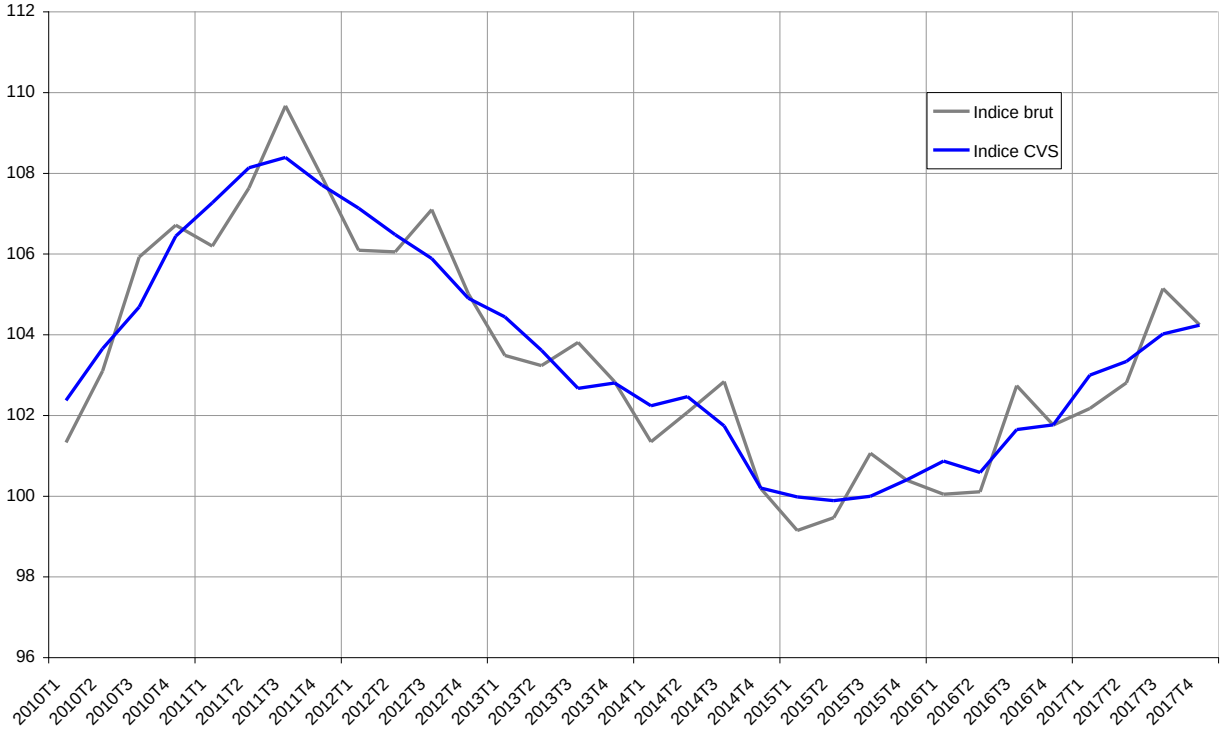
[42] Le modèle optimal n'est pas forcément le même sur les différentes zones de calcul. Pour certaines séries, le schéma additif peut s'avérer meilleur, mais les différences entre les séries CVS produites avec un schéma additif et celles produites avec un schéma multiplicatif sont tellement faibles qu'on préfère appliquer ce dernier sur l'ensemble des séries par souci d'homogénéité de traitement.

	- Appartements rural - Maisons
Appartements et maisons : Rhône-Alpes	- Appartements Rhône-Alpes - Maisons Rhône-Alpes
Appartements et maisons : Provence-Alpes-Côte d'Azur	- Appartements Provence-Alpes-Côte d'Azur - Maisons Provence-Alpes-Côte d'Azur
Appartements et maisons : Nord-Pas-de-Calais	- Appartements Nord-Pas-de-Calais - Maisons Nord-Pas-de-Calais
Appartements et maisons : Auvergne-Rhône-Alpes	- Appartements Provence-Alpes-Côte d'Azur - Maisons Provence-Alpes-Côte d'Azur - Appartements Auvergne - Maisons Auvergne
Appartements et maisons : Hauts-de-France	- Appartements Nord-Pas-de-Calais - Maisons Nord-Pas-de-Calais - Appartements Picardie - Maisons Picardie
Île-de-France	
Appartements	- Appartements par département
Appartements Île-de-France hors Paris	- Appartements par département
Appartements grande couronne	- Appartements par département (dép. 77, 78, 91 et 95)
Appartements petite couronne	- Appartements par département (dép. 92, 93 et 94)
Maisons	- Maisons par département
Maisons grande couronne	- Maisons par département (dép. 77, 78, 91 et 95)
Maisons petite couronne	- Maisons par département (dép. 92, 93 et 94)
Appartements et maisons	- Appartements par département - Maisons par département
France métropolitaine	
Appartements	- Appartements villes-centres - Appartements banlieue - Appartements rural - Appartements Île-de-France par département
Maisons	- Maisons province - Maisons Île-de-France par département
Appartements et maisons	- Appartements villes-centres - Appartements banlieue - Appartements rural - Maisons province - Appartements Île-de-France par département - Maisons Île-de-France par département

France hors Mayotte	
Appartements	- Appartements villes-centres - Appartements banlieue - Appartements rural - Appartements Drom hors Mayotte - Appartements Île-de-France par département
Maisons	- Maisons Province - Maisons Drom hors Mayotte - Maisons Île-de-France par département
Appartements et maisons	- Appartements villes-centres - Appartements banlieue - Appartements rural - Appartements Drom hors Mayotte - Maisons province - Maisons Drom hors Mayotte - Appartements Île-de-France par département - Maisons Île-de-France par département

Les coefficients saisonniers sont actualisés une fois par an pour prendre en compte les derniers indices dans les séries brutes utilisées pour le calcul. Leur mise à jour a lieu courant juillet une fois disponibles les données définitives du 4^{e} trimestre de la dernière année écoulée. Une mise à jour annuelle des CVS apparaît préférable à une mise jour trimestrielle ou mensuelle, dans la mesure où cela permet de limiter les révisions de coefficients, d'autant plus que ces derniers peuvent se révéler fragiles s'ils sont estimés à partir de données non définitives.
À l'occasion de la révision annuelle des coefficients CVS, on calcule également, *via* des projections, les coefficients CVS utilisés pour le calcul d'indices postérieurs à la période de désaisonnalisation (ces coefficients seront ensuite actualisés sur données réelles lors d'une prochaine mise à jour). Ces projections sont faites sur deux ans.
Un exemple de série brute / CVS est fourni en figure 4-4.

Figure 4-4 : **Indices bruts et CVS des maisons de province**



5 – Diffusion

5.1 – Séries publiées depuis septembre 2018

Les indices trimestriels publiés sous l'appellation « indices Notaires-Insee », résultant des conventions entre l'Insee et les organismes notariaux, sont calculés sur des zones d'intérêt permettant une bonne précision de calcul. Ils ont été choisis par le Conseil scientifique des indices Notaires-Insee selon plusieurs critères : les taux de couverture, l'ampleur des révisions, les volumes d'actes pris en compte, les délais d'intégration des actes dans les bases. Ces indices Notaires-Insee ont également été labellisés par l'autorité de la statistique publique (ASP) après examen des méthodes de collecte des données et de calcul : avis du 10 avril 2018 pour les indices de province (NOR : ECOO1810840V), avis du 27 avril 2018 pour les indices en Île-de-France (ECOO1812771V).

Les indices Notaires-Insee publiés conjointement par l'Insee et par les organismes notariaux (depuis septembre 2018) sont répertoriés dans la figure 5-1, chaque série étant disponible en version brute et en version CVS.

Figure 5-1 : Séries publiées conjointement par l'Insee et les organismes notariaux

Niveau géographique	Type de logements	Début de la série
France (hors Mayotte)	Ensemble	2010T1
France (hors Mayotte)	Appartements	2010T1
France (hors Mayotte)	Maisons	2010T1
France métropolitaine	Ensemble	1996T1
France métropolitaine	Appartements	1996T1
France métropolitaine	Maisons	1996T1
Province	Ensemble	1994T4
Province	Appartements	1994T4
Province	Maisons	1994T4
Agglomérations de province de plus de 10 000 habitants	Appartements	1994T4
Agglomérations de province de plus de 10 000 habitants – Villes-centres	Appartements	1994T4
Agglomérations de province de plus de 10 000 habitants - Banlieues	Appartements	1994T4
Agglomérations de province de moins de 10 000 habitants et zones rurales	Appartements	1994T4
Provence-Alpes-Côte d'Azur	Ensemble	1994T4
Provence-Alpes-Côte d'Azur	Appartements	1994T4
Provence-Alpes-Côte d'Azur	Maisons	1994T4
Commune de Marseille	Appartements	1994T4
Rhône-Alpes	Ensemble	1994T4
Rhône-Alpes	Appartements	1994T4
Rhône-Alpes	Maisons	1994T4
Commune de Lyon	Appartements	1994T4
Nord-Pas-de-Calais	Ensemble	2007T4
Nord-Pas-de-Calais	Appartements	2007T4
Nord-Pas-de-Calais	Maisons	2007T4
Agglomération de Lille	Maisons	2007T4
Hauts-de-France	Ensemble	2010T1
Hauts-de-France	Appartements	2010T1
Hauts-de-France	Maisons	2010T1
Auvergne-Rhône-Alpes	Ensemble	2010T1
Auvergne-Rhône-Alpes	Appartements	2010T1
Auvergne-Rhône-Alpes	Maisons	2010T1
Île-de-France	Ensemble	1996T1
Île-de-France	Appartements	1996T1
Île-de-France	Maisons	1996T1
Paris	Appartements	1991T1
Île-de-France hors Paris	Ensemble	1996T1

Niveau géographique	Type de logements	Début de la série
Île-de-France hors Paris	Maisons	1996T1
Île-de-France hors Paris	Appartements	1996T1
Île-de-France - petite couronne	Ensemble	1991T1
Île-de-France - petite couronne	Appartements	1991T1
Île-de-France - petite couronne	Maisons	1991T1
Île-de-France - grande couronne	Ensemble	1996T1
Île-de-France - grande couronne	Appartements	1996T1
Île-de-France - grande couronne	Maisons	1996T1
Seine-et-Marne	Ensemble	1996T1
Seine-et-Marne	Appartements	1996T1
Seine-et-Marne	Maisons	1996T1
Yvelines	Ensemble	1996T1
Yvelines	Appartements	1996T1
Yvelines	Maisons	1996T1
Essonne	Ensemble	1996T1
Essonne	Appartements	1996T1
Essonne	Maisons	1996T1
Hauts-de-Seine	Ensemble	1991T1
Hauts-de-Seine	Appartements	1991T1
Hauts-de-Seine	Maisons	1991T1
Seine-Saint-Denis	Ensemble	1991T1
Seine-Saint-Denis	Appartements	1991T1
Seine-Saint-Denis	Maisons	1991T1
Val-de-Marne	Ensemble	1991T1
Val-de-Marne	Appartements	1991T1
Val-de-Marne	Maisons	1991T1
Val-d'Oise	Ensemble	1996T1
Val-d'Oise	Appartements	1996T1
Val-d'Oise	Maisons	1996T1

Cette liste est susceptible de s'agrandir à l'avenir si de nouveaux indices respectent les critères de qualité surveillés par le Conseil scientifique (notamment si les taux de couverture se renforcent à la suite de l'intégration dans la mission de service public des notaires de la collecte d'informations statistiques).

5.2 – Modes de publication des indices

Les indices Notaires-Insee sont diffusés à travers différents supports.

L'Insee met les indices à disposition dans sa banque de données macro-économiques (BDM), permettant de visualiser et de télécharger les séries :
https://www.insee.fr/fr/statistiques/series/105071770?INDICATEUR=2878202.

Il publie tous les trimestres une analyse conjoncturelle du marché immobilier dans la collection « Informations Rapides » :
https://www.insee.fr/fr/statistiques?debut=0&theme=30&conjoncture=56.

L'indice de prix des logements anciens de France métropolitaine entre dans le calcul de l'indice des prix des logements (neufs et anciens), fourni tous les trimestres à Eurostat et publié dans la BDM : https://www.insee.fr/fr/statistiques/series/105071770?INDICATEUR=2770594, ainsi que dans la collection « Informations Rapides » :
https://www.insee.fr/fr/statistiques?debut=0&theme=30&conjoncture=55.

Les notaires publient tous les trimestres une note de conjoncture dans laquelle sont communiquées les dernières variations des indices globaux (ensemble, province, Île-de-France). Ce document de 4 pages présente une synthèse du marché immobilier en France. Les thèmes abordés sont principalement les volumes de ventes, les prix et leurs évolutions. Une conférence de presse est également organisée chaque fin d'année par le Conseil supérieur du notariat sur le marché immobilier français. À cette occasion, les dernières tendances sur les prix sont présentées, ainsi qu'une analyse du profil des acquéreurs et vendeurs. Des graphiques sur les séries des indices de prix sont notamment projetés. Les notes de conjoncture et conférences sont disponibles sur le lien suivant : https://www.immobilier.notaires.fr/fr/conseil-immobilier?idConseil=63.

La Chambre des notaires de Paris organise chaque trimestre une conférence de presse et diffuse sur son site les séries d'indices franciliennes : http://paris.notaires.fr/fr/liste-communiques-de-presse?debut=&fin=&type=1013.

Les notaires diffusent d'autres indicateurs sur l'immobilier, produits indépendamment des indices et du partenariat avec l'Insee. La mention « Notaires-Insee » vient différencier clairement les séries d'indices des autres statistiques proposées.

5.3 – Calendrier et révisions

Les séries sont actualisées tous les trimestres avec la diffusion des dernières valeurs des indices. Selon les conventions liant l'Insee et les services notariaux et conformément aux bonnes pratiques de la statistique européenne, la publication et la diffusion des indices par les différents acteurs sont soumises à embargo. La date et l'heure de levée de l'embargo, à compter desquelles les indices sont autorisés à être publiés, correspondent aux dates de publications de l'Insee et sont proposées par le Conseil scientifique des indices Notaires-Insee et validées par l'Insee au minimum un trimestre à l'avance.

Sur chaque publication des indices dans la collection « Informations Rapides » est précisée la date de la prochaine publication. Les dates des publications futures déjà arrêtées sont également notifiées dans l'agenda de publication de l'Insee[43] :
https://www.insee.fr/fr/information/1405540?debut=0

[43] La diffusion a lieu à la fin du deuxième mois suivant la fin du trimestre publié, hormis pour le deuxième trimestre de l'année, publié début septembre. L'heure de levée de l'embargo est toujours fixée à 8 h 45 (les informations sont communiquées aux agences de presse à 8 h 30).

La première diffusion d'un indice a lieu environ deux mois après la fin du trimestre sur lequel il porte, ce qui permet de suivre l'information conjoncturelle récente. Cependant, à ce stade, toutes les transactions immobilières du trimestre n'ont pas encore été enregistrées, étant donné les délais de réception et de traitement des actes. Par conséquent, l'indice est calculé sur un échantillon restreint[44], ce qui implique une précision plus faible. Les indices sont mis à jour ultérieurement, une fois la majorité des transactions prises en compte, afin d'obtenir un indice final plus robuste.

Pour l'Île-de-France, sur laquelle les taux de couverture sont relativement élevés, une seule mise à jour a lieu :

- les *indices provisoires* sont diffusés lors de la publication trimestrielle deux mois après la fin du trimestre auxquels ils se rapportent ;
- les *indices définitifs* remplacent les indices provisoires cinq mois après la fin du trimestre auxquels ils se rapportent, lors de la publication des indices provisoires du trimestre suivant.

Pour la province, sur laquelle les taux de couverture sont un peu plus faibles, on effectue quatre mises à jour (sous réserve de taux de couverture suffisants, voir figure 5-4) :

- les *indices provisoires avancés* sont diffusés lors de la publication trimestrielle deux mois après la fin du trimestre auxquels ils se rapportent, soit au même moment que le provisoire en Île-de-France. Seuls les indices sur l'ensemble de la province sont publiés ;
- les *indices provisoires* remplacent les indices provisoires avancés environ trois mois après la fin du trimestre auxquels ils se rapportent. Les indices à des niveaux géographiques plus fins que l'ensemble de la province sont diffusés à partir de cette étape ;
- les *indices semi-définitifs* remplacent les indices provisoires cinq mois après la fin du trimestre auxquels ils se rapportent, lors de la publication des indices provisoires avancés du trimestre suivant, soit au même moment que le définitif en Île-de-France ;
- les *indices définitifs* remplacent les indices semi-définitifs six mois après la fin du trimestre auxquels ils se rapportent, en même temps que le passage aux indices provisoires pour le trimestre suivant.

[44] Le développement de la télétransmission des actes et la hausse des taux de couverture permettent toutefois une amélioration de la qualité de l'indice.

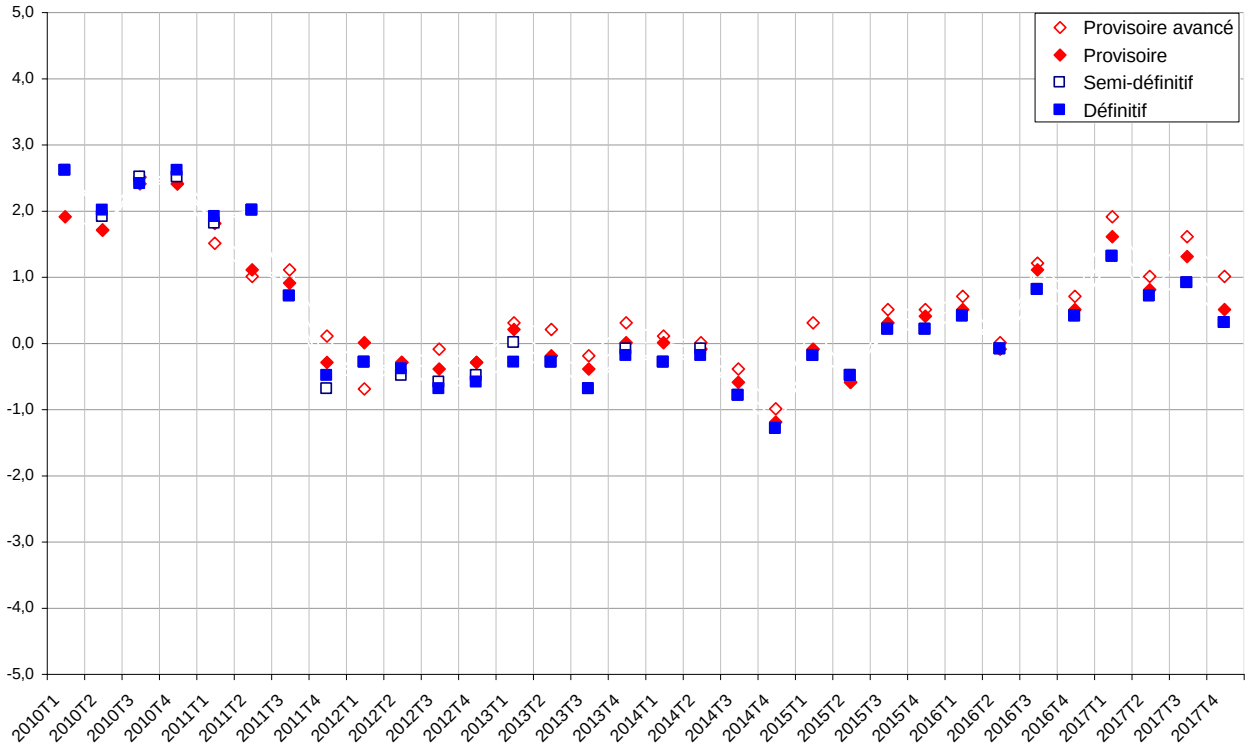
À titre d'exemple, les dates de publication du T3 2018 sont fournies dans la figure 5-2.

Figure 5-2 : Dates et statut de la publication de l'indice du T3 2018

	IdF	Province
29 novembre 2018 (Informations rapides du T3 2018)	Provisoire	Provisoire avancé
7 janvier 2019 (Actualisation des T2 2018 et T3 2018)	Provisoire	Provisoire
28 février 2019 (Informations rapides du T4 2018)	Définitif	Semi-définitif
1er avril 2019 (Actualisation des T3 2018 et T4 2018)	Définitif	Définitif

La figure 5-3 fournit les évolutions trimestrielles CVS de France métropolitaine observées en fonction de la publication. Ainsi, par exemple, l'évolution trimestrielle en France métropolitaine au T1 2018 était estimée à + 1,5 % au moment du provisoire avancé, 1,2 % au moment du provisoire et 1,1 % au semi-définitif et au définitif.

Figure 5-3 : Glissements trimestriels des prix des logements anciens CVS observés en France métropolitaine selon les publications



Enfin, des taux de couverture minimum pour la diffusion sont requis (figure 5-4). Ils ont été fixés pour chaque cran d'actualisation et par niveau géographique de façon à garantir que les révisions d'un indice soient restreintes[45]. Si ces seuils ne sont pas respectés, l'indice n'est pas diffusé.

[45] Le choix des seuils s'est fait lors de la mise en place de la version 3 des indices sur la base d'une analyse des liens entre les taux de couverture et l'ampleur des révisions sur les indices de province.

Figure 5-4 : Taux de couverture minimum imposés pour la publication de l'indice des prix des logements anciens selon la zone géographique

Niveau géographique	Provisoire avancé	Provisoire	Semi-définitif	Définitif
Métropole	20 %	20 %	20 %	20 %
Province	30 %	30 %	30 %	30 %
Autres indices suprarégionaux	Non diffusé	50 %	50 %	50 %
Région	Non diffusé	50 %	50 %	50 %

6 – Passage de la version 3 à la version 4

La version 4 des indices est entrée en application en 2018. Elle a été définie à l'occasion d'une expertise de la méthodologie existante. Ce travail est effectué par le Conseil scientifique des indices Notaires-Insee périodiquement, afin de faire évoluer la méthode en fonction de l'état des marchés. La version 4 ne provient donc pas d'une refonte complète du processus de production des indices mais d'une actualisation du dispositif utilisé en version 3[46], auquel ont été apportés un certain nombre d'ajustements. Ceux-ci sont décrits ci-dessous :

- Champ des indices :
    - le champ géographique comprend dorénavant la Corse et les Drom (Mayotte n'étant pas inclus) ;
    - les chalets ont été ajoutés dans le champ des indices des prix des maisons ;
    - les logements dont le nombre de pièces n'est pas renseigné sont maintenant supprimés. Ce traitement était déjà effectué en province en version 3, mais le nombre de pièces était estimé en Île-de-France, lorsque la surface était renseignée (l'observation était supprimée sinon), en fonction de la surface, de la strate et du type de bien. Il a été décidé d'homogénéiser les traitements en supprimant ces observations sur l'ensemble du territoire, étant donné le faible volume concerné (moins de 0,5 % des observations), plutôt que d'en faire une estimation incertaine ;
    - le choix des variables fiscales utilisées pour la définition des logements anciens a été homogénéisé entre l'Île-de-France et la province. L'ancien correspond toujours à la définition fiscale : logements non soumis en partie ou en totalité à la TVA. Les modifications dans la législation française en matière de TVA immobilière peuvent donc avoir un impact sur le champ de l'ancien utilisé pour les indices Notaires-Insee. Par exemple, les biens de moins de 5 ans revendus, auparavant soumis à la TVA et qui ne le sont plus depuis 2011, n'étaient pas comptés dans les logements anciens alors qu'ils le sont maintenant. Cela a pour conséquence d'augmenter les volumes de l'ordre de 2 % (calcul effectué sur le champ des logements anciens de province en 2017).

[46] Pour plus d'informations, la méthode de calcul des indices en version 3 est décrite dans l'*Insee Méthodes* n° 128 de juillet 2014 (https://www.insee.fr/fr/information/2569926).

- Définition du parc de référence : en version 3, le parc de référence correspondait au panier de biens des indices et était composé des transactions du champ sur la période de référence, auxquelles on retirait les 5 % d'observations avec les prix (au m² pour les appartements et total pour les maisons) les plus élevés et les 5 % d'observations avec les prix les plus faibles. L'équation hédonique était cependant estimée sur le « parc d'estimation » correspondant à l'échantillon utilisé pour l'estimation et composé de l'ensemble des transactions du champ auxquelles on retirait les valeurs atypiques détectées avec les résidus standardisés. Ces deux définitions avaient pour but d'exclure les valeurs extrêmes, avec des approches différentes. Dans la pratique, ces deux parcs étaient très proches[47], on a préféré conserver en version 4 un seul parc endossant ces deux rôles (appelé « parc de référence » mais dont la définition correspond à celle du « parc d'estimation » de la version 3).
- Mise à jour de la stratification en Île-de-France : de nouvelles strates ont été définies en Île-de-France ; la méthode de classification a été enrichie par rapport à la version 3 en ajoutant notamment des variables relatives aux communes, externes aux bases notariales (voir partie 3) et en modifiant les critères de volume afin d'obtenir des strates plus fines. Elle a de plus été éditée sur des données récentes.
- Le choix des variables candidates à l'inclusion dans les modèles : les variables incluses dans le modèle ont été homogénéisées entre Île-de-France et province. Des variables de distances communales (distance à la commune de plus de 50 000 habitants la plus proche pour la province, distance à Paris pour l'Île-de-France) et l'étiquette énergie ont été ajoutées aux variables candidates pour les appartements comme pour les maisons. Par ailleurs, le nombre de bâtiments, utilisé en version 3 uniquement pour les indices des maisons d'Île-de-France, a été retiré. Enfin, l'état du bien, pour les appartements et pour les maisons, ainsi que la présence d'une terrasse ou d'un balcon (ou d'une loggia depuis la version 4) pour les appartements, étaient utilisés seulement en province en version 3 et sont maintenant aussi utilisés en Île-de-France.
- Le traitement de la non-réponse des variables a également été homogénéisé : l'imputation des variables « époque de construction » et « présence d'un ascenseur » pour les appartements à partir de l'historique des ventes (voir partie 2.4) a été mis en place pour la province. En version 3, ce traitement n'était effectué que pour les appartements d'Île-de-France. Par ailleurs, le choix des modalités et le traitement de la non-réponse de certaines variables ont été actualisés. Ces changements sont présentés dans les figures 6-1 et 6-2.

[47] Des tests ont été effectués sur le parc des appartements anciens de province 2011-2012. Sur les indices 2014, le maximum des différences départementales est de 0,25 point, ce qui a très peu d'impact sur les évolutions d'indices.

Figure 6-1 : Synthèse des changements opérés sur le traitement de la non-réponse – Appartements

Appartements		
Modification	***Version 3***	***Version 4***
Le modèle de régression utilisé pour l'imputation de la **surface habitable** a été harmonisé entre l'Île-de-France et la province.	Province : les observations avec surface manquante étaient supprimées lors de la construction des parcs d'estimation ou de référence. En période courante, estimation en fonction du nombre de salles de bains, du nombre de pièces et de l'époque de construction. Île-de-France : estimation en fonction du nombre de salles de bains, du nombre de pièces, de l'époque de construction, de l'étage et du nombre de garages.	Province et Île-de-France : estimation de la surface (pour la construction du parc de référence ainsi qu'en période courante), en fonction du nombre de salles de bains, du nombre de pièces, de l'époque de construction, de l'étage et du nombre de garages.
Actualisation des modalités de l'**époque de construction**.	Modalités : Époque A : avant 1850 Époque B : de 1850 à 1913 Époque C : de 1914 à 1947 Époque D : de 1948 à 1969 Époque E : de 1970 à 1980 Époque F : de 1981 à 1991 Époque GHI : depuis 1992 Époque X : non renseigné	Modalités : Époque AB : avant 1913 inclus Époque C : de 1914 à 1947 Époque D : de 1948 à 1969 Époque E : de 1970 à 1980 Époque F : de 1981 à 1991 Époque G : de 1992 à 2000 Époque HI : depuis 2001 Époque X : non renseigné
Modification de l'imputation du **nombre de salles de bains**.	Modalités : (0, 1**,** 2 ou plus, non renseigné) Imputation déterministe à non renseigné	Modalités : (0, 1**,** 2 ou plus) Imputation déterministe à 1
L'imputation de la **présence d'une cave** est modifiée.	Modalités : (Pas de cave, Au moins une cave, Non renseigné) Imputation déterministe à "Non renseigné"	Modalités : (Pas de cave, Au moins une cave) Imputation déterministe à "Pas de cave"
La présence d'une terrasse ou d'un balcon est élargie à la **présence d'une terrasse ou d'un balcon ou d'une loggia.** L'imputation est de plus modifiée.	Modalités : (Sans terrasse ni balcon, Avec terrasse ou balcon) Imputation déterministe à "Sans terrasse ni balcon"	Modalités : (Sans terrasse ni balcon ni loggia, Avec terrasse ou balcon ou loggia, Non renseigné) Imputation déterministe à "Non renseigné"

Appartements		
Modification	***Version 3***	***Version 4***
Les modalités des classes définies par le croisement entre le **nombre de pièces et la surface moyenne par pièce** ont été modifiées.	Modalités : Studio - petit : <20 m² Studio - moyen : entre 20 et 30 m² Studio - grand : >30 m² 2 pièces - petit : <17 m² 2 pièces - moyen : entre 17 et 24 m² 2 pièces - grand : >24 m² 3 pièces - petit : <18 m² 3 pièces - moyen : entre 18 et 22 m² 3 pièces - grand : >22 m² 4 pièces ou + - petit : <17 m² 4 pièces ou + - moyen : entre 17 et 21 m² 4 pièces ou + - grand : >21 m²	Modalités : Studio - petit : <20 m² Studio - moyen : entre 20 et 30 m² Studio - grand : >30 m² 2 pièces - petit : <17 m² 2 pièces - moyen : entre 17 et 24 m² 2 pièces - grand : >24 m² 3 pièces - petit : <18 m² 3 pièces - moyen : entre 18 et 23 m² 3 pièces - grand : >23 m² 4 pièces - petit : <17 m² 4 pièces - moyen : entre 17 et 21 m² 4 pièces - grand : >21 m² 5 pièces ou + - petit : <16 m² 5 pièces ou + - moyen : entre 16 et 22 m² 5 pièces ou + - grand : >22 m²

Figure 6-2 : Synthèse des changements opérés sur le traitement de la non-réponse – Maisons

Maisons		
Modification	***Version 3***	***Version 4***
Le modèle de régression utilisé pour l'imputation de la **surface habitable** a été harmonisé entre l'Île-de-France et la province.	Province : les observations avec surface manquante étaient supprimées lors de la construction des parcs d'estimation ou de référence. En période courante, estimation en fonction du nombre de salles de bains, du nombre de pièces et de l'époque de construction. Île-de-France : estimation en fonction du nombre de salles de bains, du nombre de pièces, de l'époque de construction, du nombre de niveaux et du nombre de garages.	Province et Île-de-France : estimation de la surface (pour la construction du parc de référence ainsi qu'en période courante), en fonction du nombre de salles de bains, du nombre de pièces, de l'époque de construction et du nombre de niveaux.
Les modalités du **nombre de niveaux** ont été harmonisées entre la province et l'Île-de-France	Modalités province : (1, 2**,** 3 ou plus) Modalités IDF (1, 2 ou plus)	Modalités province et IDF : (1, 2 ou plus)
Les modalités et l'imputation du **nombre de garages** ont été modifiées.	Modalités : (0, 1, 2 ou plus) Imputation déterministe à 1 pour la province, à 0 pour l'IDF	Modalités : (0, 1, 2 ou plus, Non renseigné) Imputation déterministe à "Non renseigné" pour la province et l'IDF

Maisons		
Modification	***Version 3***	***Version 4***
Actualisation des modalités de l'**époque de construction**	Modalités : Époque AB : avant 1913 (inclus) Époque C : de 1914 à 1947 Époque D : de 1948 à 1969 Époque E : de 1970 à 1980 Époque FGHI : depuis 1981 Époque X : non renseigné	Modalités : Époque AB : avant 1913 (inclus) Époque C : de 1914 à 1947 Époque D : de 1948 à 1969 Époque E : de 1970 à 1980 Époque F : de 1981 à 1991 Époque GHI : depuis 1992 Époque X : non renseigné
La présence d'une cave et celle d'un sous-sol sont regroupées en **présence d'une cave ou d'un sous-sol** et harmonisées entre la province et l'Île-de-France. L'imputation est modifiée.	Modalités province : (pas de sous-sol, présence d'un sous-sol) Imputation déterministe à "pas de sous-sol" Modalités IDF : (pas de cave, présence d'une cave) Imputation déterministe à "pas de cave"	Modalités province et IDF : (pas de sous-sol ni de cave, présence d'un sous-sol ou d'une cave) Imputation déterministe à "pas de sous-sol ni de cave"
Modification de l'imputation du nombre de salles de bains.	Modalités : (0, 1**,** 2 ou plus) Imputation déterministe à 0	Modalités : (0, 1**,** 2 ou plus) Imputation déterministe à 1

- Correction des variations saisonnières : deux années de prévisions sont calculées avec les modèles ARIMA, contre une en version 3.

- Base 100 : les indices sont maintenant diffusés en base *moyenne annuelle 2015=100*, contre *2010 T1=100* en version 3.

Afin d'évaluer l'écart global entre les versions 3 et 4, un historique en version 4 a été recalculé. Les figures 6-3, 6-4 et 6-5 présentent la comparaison des indices entre les deux versions du premier trimestre 2010 au premier trimestre 2018.

Figure 6-3 : Indice des prix des logements anciens en France métropolitaine V3 / V4

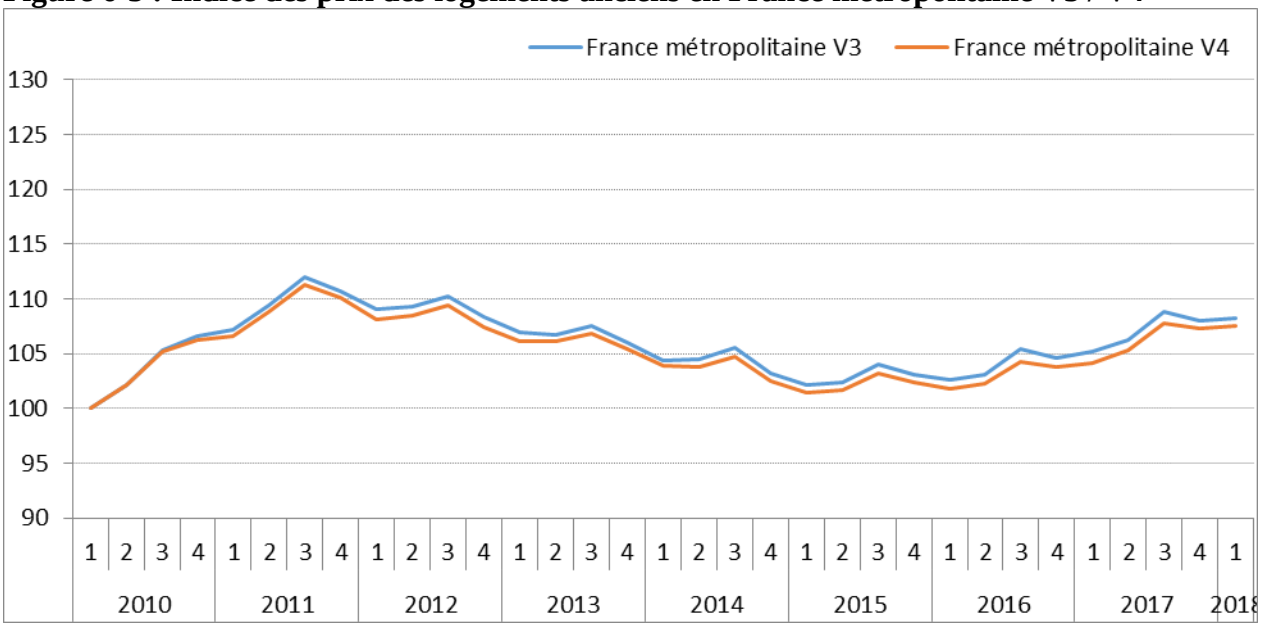


Figure 6-4 : **Indice des prix des logements anciens en Île-de-France V3 / V4**

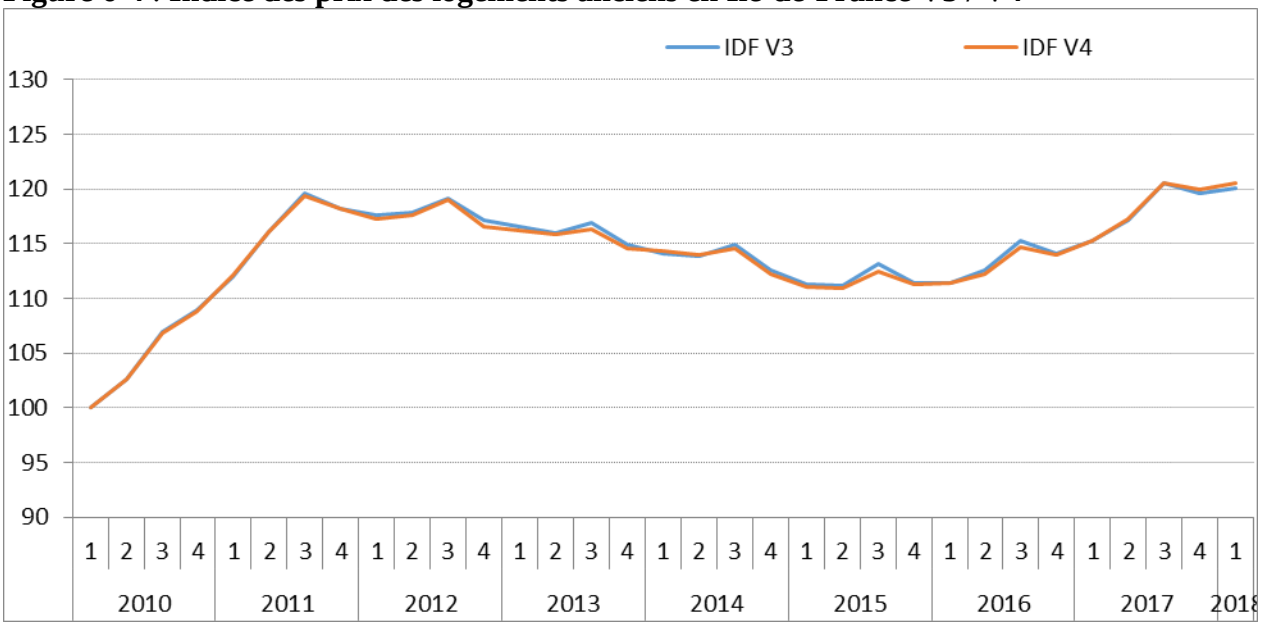


Figure 6-5 : Indice des prix des logements anciens en province V3 / V4