Présentation projet de partenariat-Données Enrichies pour les Transactions Immobilières en France : un Outil clé d'Estimation de Prix Immobilier pour les Professionnels et les Particuliers

Tristan Ancel, tristan.ancel@ens-paris-saclay.fr

June 2024

1 Introduction

1.1 Présentation du projet

Le projet vise à fournir des données enrichies sur les ventes de maisons et d'appartements en France[1] au cours des cinq dernières années[2] et est réalisé par Ancel Tristan, élève normalien en master d'économie à l'ENS Paris-Saclay et stagiaire au CREST (Center for Research in Economics and Statistics). Ces données permettront d'apporter des informations supplémentaires pertinentes pour les particuliers et les professionnels de l'immobilier. Le but principal est d'améliorer la compréhension des caractéristiques des biens immobiliers en fournissant des données contextuelles sur les transactions survenues depuis 2019 telles que la présence en zone bruyante due aux aéroports, la proximité des infrastructures clés, les risques environnementaux, de cambriolage ou encore des données relatives à l'attractivité de la commune dans laquelle se situe le bien (voir section 2). Ces données utilisées avec des moyens statistiques avancées et des modèles d'analyse permettront la création d'une estimation juste et transparente de la valeur des biens, prenant en compte l'ensemble des caractéristiques géographiques et sociales de ce dernier.

[1] excepté les départements du Bas-Rhin, Haut-Rhin, Moselle et Mayotte pour des raisons de législation

[2] Les données vont actuellement de 2019 à 2023, la prochaine mise à jour aura lieu de octobre 2024 pour intégrer les données du premier semestre de l'année en cours. Ces nouvelles données pourront être ajoutées.

1.2 Utilités

L'utilisation du Big Data, ou grandes bases de données, permet à ceux qui parviennent à bien l'exploiter et le maîtriser de mieux comprendre leur écosystème, de développer des offres innovantes et d'optimiser leur stratégie de développement. Il est donc nécessaire pour l'ensemble des professionels de l'immobilier de tirer profit de ces données immobilières.

1.2.1 Mise en place d'un estimateur performant de la valeur des biens

Considérant la prolifération d'estimateurs en ligne de la valeur d'un bien immobilier, notre projet se propose d'enrichir ce type de modèle d'estimation en intégrant de nouvelles variables significatives.
L'utilisation de bases de données intégrant les transactions immobilières en France sur les cinq dernières années, enrichies de données contextuelles, permettent de disposer d'informations exhaustives et précises, essentielles pour le développement d'un estimateur économétrique fiable du prix des logements, pouvant être utilisé par les professionels de l'immobilier et mis à disposition des particuliers eux-mêmes. En intégrant ces multiples variables, il devient effectivement possible de modéliser de manière fine et précise les déterminants du prix immobilier, réduisant ainsi la marge d'erreur, garantissant ainsi une expertise certaine. De plus, cet outil interactif engage les utilisateurs, prolongeant leur visite et améliorant le SEO[3] du site internet du professionel de l'immobilier.

1.2.2 Intégration de données sur l'environnement des biens au service des particuliers

Les bases de données peuvent renseigner les vendeurs et les acheteurs sur la valeur approximative du bien qu'ils souhaitent vendre ou acheter en le comparant avec les transactions similaires effectuées dans le voisinage (au moyen de cartes interactives par exemple). Ainsi, les particuliers peuvent avoir une vision claire et réaliste du marché immobilier local, facilitant la fixation d'un prix de vente compétitif et justifié. Cela représente un gain de temps considérable pour les agents immobiliers. Cette approche permet donc aux professionnels de l'immobilier, aux investisseurs, et aux particuliers de bénéficier d'une expertise affinée et d'outils de prévision robustes pour la prise de décision, optimisant ainsi les transactions et les investissements dans le secteur immobilier.

1.2.3 Autres exemples concrets d'utilisation du Big Data dans le secteur immobilier à l'international

- LandVision (USA) est une plateforme d'analyse géospatiale développée par Digital Map Products (DMP). Elle est conçue pour aider les professionnels de l'immobilier, les promoteurs, les urbanistes et les investisseurs

[3] Référencement pour les moteurs de recherche

à prendre des décisions éclairées en matière d'acquisition et de gestion de terrains en identifiant des parcelles de terrain intéressantes avec des informations détaillées sur les propriétés, y compris les caractéristiques du terrain, les zonages, et les transactions récentes (au moyen notamment de cartes interactives).

- PlanetHome Investment AG permet aux investisseurs particuliers et institutionnels de financer des projets immobiliers en ligne. PlanetHome Investment AG utilise le Big Data pour analyser et évaluer les projets immobiliers proposés. Cela comprend l'analyse des tendances du marché, la performance passée des projets similaires, les risques associés et les prévisions de rentabilité.

- SmartZip (USA) utilise des technologies avancées basées sur l'analyse de données pour fournir des prédictions précises sur les comportements des vendeurs potentiels sur le marché immobilier. En utilisant des algorithmes de machine learning et des techniques de Big Data, SmartZip évalue une "propension à vendre" pour chaque propriétaire, indiquant la probabilité qu'ils mettent leur propriété en vente dans un délai donné.

- Zillow (USA) utilise le Big Data pour intégrer des données liés à l'efficacité énergétique, aux matériaux de construction durables, et à d'autres critères ESG pertinents dans son modèle d'estimation de la valeur des biens.

1.3 Processus de conception des données

Le projet utilise un fichier CSV initial provenant de la Base DVF (Demandes de Valeurs Foncières) contenant les détails des ventes de propriétés pour chaque année depuis 2019, tels que la valeur foncière, la surface de la maison, le nombre de pièces principales et de dépendances, la surface du terrain, etc. Avant d'être utilisé, ce fichier a nécessité un processus de nettoyage préalable en raison

de trois problèmes[4][5][6] qui rendent l'utilisation brute du fichier provenant de la Base DVF impossible. J'ai donc automatisé le traitement de ces doublons ou 'quasi-doublons' et rendu le fichier lisible et fonctionnelle (le code utilisé pour résoudre ces difficultés peut être fourni avec des explications détaillées).
À partir des fichiers reprenant les ventes d'appartements et de maisons de chaque année, plusieurs fichiers supplémentaires peuvent être créés, chacun ajoutant une colonne spécifique pour fournir des informations contextuelles supplémentaires sur les transactions en fonction de leur emplacement géographique (Section 2). Ces données supplémentaires sont collectées à partir de sources diverses telles que les bases de données publiques ou encore les cartes géographiques et sont appariées aux transactions immobilières par la méthode du 'spatial joint'.

[4]Pour des biens présentant des surfaces de terrain divisées en plusieurs catégories (par exemple, "sol" et "jardins" ou "landes"), chaque catégorie était initialement représentée par une ligne distincte dans les fichiers CSV, créant ainsi des lignes identiques en tout point excepté dans la colonne 'surface du terrain'. La solution a consisté à fusionner ces doublons en additionnant les surfaces de terrain correspondantes. Par exemple, pour un bien ayant initialement deux lignes séparées et presque identiques dans le fichier CSV excepté que l'une avait une surface terrain de "sol" de 100 m^2 et l'autre une surface terrain de "jardin" de 50 m^2, les deux lignes ont été fusionnées en une seule avec une surface terrain de 150 m^2.

[5]Un problème similaire au précédent concernait les transactions où l'achat de deux maisons distinctes sur un unique terrain était représenté par deux lignes distinctes dans le fichier CSV initial. Cependant, pour chacune des lignes, la valeur foncière indiquée correspondait à l'achat des deux maisons, et non d'une seule. La solution a consisté à fusionner ces lignes, en additionnant les surfaces habitables et le nombre de chambres, afin de refléter plus fidèlement la réalité. Par exemple, pour une transaction initialement représentée par deux lignes distinctes et presque identiques dans le fichier CSV excepté que l'une avait une surface habitable de 100 m^2 et 3 chambres, et l'autre 150 m^2 et 4 chambres pour la seconde, les deux lignes ont été fusionnées en une seule avec une surface habitable totale de 250 m^2 et 7 chambres, reflétant la valeur foncière commune des deux maisons.

[6]Une autre source de doublons, bien que moins fréquente, concernait l'achat de plusieurs biens à différentes adresses lors d'une même transaction. Ces transactions, souvent très coûteuses, étaient représentées par autant de lignes que de biens acquis, chacune indiquant la valeur foncière totale de la transaction, créant ainsi des prix au mètre carré faussés. Pour résoudre ce problème, les lignes ont été fusionnées en additionnant les surfaces habitables et le nombre de pièces principales, et une colonne a été ajoutée pour indiquer le nombre de biens concernés par la transaction. Par exemple, si 8 maisons ont été achetées lors de la transaction, '8' figure dans la colonne 'nombre de biens', et la surface totale et le nombre de pièces principales correspondent à la somme de celles des 8 maisons.

2 Données supplémentaires

2.1 Présence de transactions en zone bruyante due aux aéroports

Cette donnée indique si la propriété est située dans une zone exposée au bruit des aéroports. Ainsi, la colonne affichera un indicateur A (zone de bruit très fort: Lden>70 ou IP>96), B (zone de bruit fort: Lden entre 62 et 70 ou IP entre 89 et 96) , C (zone de bruit modéré: Lden entre 55 et 62 ou IP entre 72 et 89), D (zone de bruit faible: Lden entre 50 et 62) ou sera vide (absence de bruit).

2.2 Distance au centre-ville

Ces données indiquent la distance[7](en mètres) minimale entre la propriété et le centre-ville le plus proche, le nom de la commune dudit centre-ville, la taille de la population de la commune et celle vivant dans le centre ville.

2.3 Distance aux infrastructures clés

Ces données indiquent les distances (en mètres) entre la propriété et divers points d'intérêt, tels que :
L'hôpital le plus proche, en précisant le nom de l'hôpital.
La gare SNCF la plus proche, en précisant le nom de la gare.
Le point d'arrêt de transport en commun le plus proche (arrêt de bus, de tramway, etc.), en précisant le nom de l'arrêt.
Le restaurant le plus proche.
Le commerce le plus proche, avec une définition large (fleuriste, café, barbier, etc.) ou plus restrictive (seulement les supermarchés, par exemple).

2.4 Nombre de cambriolages survenus dans la zone

Cette colonne de données indique le taux annuels de cambriolages signalés dans la zone urbaine où se situe la propriété. Les données recensent le taux annuels de cambriolages et tentatives de cambriolages de logements, enregistrés par la police et la gendarmerie nationales, en lieu de commission, pour 1 000 logements dans les unités urbaines de plus de 100 000 habitants en France uniquement.

2.5 Caractéristiques d'éducation

Ces données indiquent la distance (en mètres) entre la propriété et l'école la plus proche, en précisant le type d'école (primaire, collège, lycée) ainsi que le nom de l'école. En outre, elles peuvent inclure une série de données décrivant l'offre éducative dans la commune, telles que : le nombre d'écoles maternelles,

[7]Les distances sont calculées à vol d'oiseau à partir des coordonnées GPS des points de transactions et des services ou infrastructures

élémentaires, de collèges, de lycées et d'enseignement supérieurs (général, technologique et/ou professionnel) en 2021-2022 ainsi que leurs effectifs, et enfin la part des 25-34 ans titulaires d'un diplôme de l'enseignement supérieur.

2.6 Carcatéristiques d'emploi au sein de la commune

Ces données de 2020 spécifient, à l'échelle de la commune dans laquelle se situe le bien, le taux de chômage des 15 ans et plus, la part des salariés de 15-64 ans en emploi précaire[8], la part des "agriculteurs exploitants", des "ouvriers", des "cadres et professions intellectuelles supérieures" et enfin celle des "retraités".

2.7 Couverture numérique

Ces données de 2022 spécifient, à l'échelle du département dans laquelle se situe le bien, la part de la surface couverte en 4G par a minima un opérateur et celle des locaux raccordables FttH (fibre optique).

2.8 Présence en zone inondable

Cette colonne indique si la propriété est située dans une zone inondable et le degré de risque associé. De fait, la colonne ajoutée quantifie le risque d'inondation associé à l'emplacement géographique du bien. Les valeurs pouvant être prises sont : 0 (absence de risque), 1 (risque faible), 2 (risque moyen dans une perspective à court terme, autrement dit un risque qui est aujourd'hui faible mais qui deviendra moyen sous 10 ans), 3 (risque moyen), 4 (risque fort). En plus de cela sont ajoutées les informations indiquant le coût moyen des sinistres totaux et par habitant[9] dans la commune où se situe le bien sur la période 1995-2019.

2.9 Zone au risque de retrait-gonflement des argiles

Ces données indiquent si la propriété est située dans une zone à risque de retrait-gonflement des argiles et le degré de risque associé. La colonne ajoutée quantifie le risque de retrait argileux associé à l'emplacement géographique du bien. Les valeurs pouvant être prises sont : 0 (absence de risque), 1 (risque faible), 2 (risque moyen), 3 (risque fort).

2.10 Niveau de pollution au sein de la commune

Ces données spécifient, à l'échelle de la commune dans laquelle se situe le bien, le total d'émissions de gaz à effet de serre hors puits [10] et l'émissions de gaz à

[8] Contrats en intérim, apprentissage, les emplois jeunes, contrats emploi solidarité, contrats de qualification ou autres emplois aidés, les stages rémunérés en entreprise, les CDD ou les postes saisonniers

[9] cet indicateur porte sur les coûts moyens des sinistres indemnisés par les assureurs au titre du régime des Catastrophes Naturelles pour le péril inondation au sens large (inondation et coulée de boue, inondation par remontée de nappes et inondation par submersion marine).

[10] Exprimé en milliers de tonnes équivalent CO2

effet de serre hors puits (PRG) par habitant.

2.11 Caractéristiques de Biodiversité et de patrimoine de la commune

Ces données spécifient, à l'échelle de la commune dans laquelle se situe le bien, la part des Zones Naturelles d'Intérêt Ecologique Faunistique et Floristique (ZNIEFF) de type 1 et la part des Zones Naturelles d'Intérêt Ecologique Faunistique et Floristique (ZNIEFF) de type 2 dans la superficie du territoire[11].

2.12 Caractéristiques énergétiques et année de construction du bien

Ces caractéristiques rassemblent l'étiquette DPE et GES du bien, la date à laquelle ont été faites ces évalutations, la consommation énergie en kWhEP/m^2/an et l'estimation GES en Kg eqCO2/m^2/an du bien ainsi que la période de construction du bien.
Cependant, bien que les diagnostics énergétiques soient obligatoires pour les transactions immobilières depuis 2007, le fait d'avoir un logement dans la base de données DPE/GES ne garantit pas de le retrouver dans la base de données DVF (des ventes de logements). Par conséquent, l'appariement des données diminue drastiquement le nombre de transactions identifiés (environ 20% des transactions concernant des maisons peuvent être appariées à leurs caractéristiques énergétiques).

2.13 Dynamique de la population

Ces données à l'échelle de la commune spécifient la densité de population, le taux d'évolution annuel de la population entre 2014 et 2020, le taux d'évolution annuel de la population dû au solde migratoire de 2014 à 2020, la part des immigrés dans la population et enfin celle des 65 ans et plus. D'autres données de ce type peuvent également être prises en compte dans l'élaboration des données.

2.14 Carcatéristiques du marché du logement au sein de la commune

Ces données spécifient, à l'échelle de la commune dans laquelle se situe le bien, le taux d'évolution annuel du nombre de logements entre 2014-2020, le prix moyen au m^2 en 2023 d'un loyer pour un bien type du parc privé locatif, pouvant notamment servir d'outil afin d'évaluer la rentabilité attendue d'un investissement dans un bien, et enfin le prix au m^2 moyen pour l'achat d'un bien type.

[11] L'inventaire des Zones Naturelles d'Intérêt Ecologique Faunistique et Floristique (ZNIEFF), identifie, localise et décrit des secteurs présentant de fortes capacités biologiques et un bon état de conservation. On distingue 2 types de ZNIEFF : les ZNIEFF de type I sont des secteurs de grand intérêt biologique ou écologique ; les ZNIEFF de type II sont des grands ensembles naturels riches et peu modifiés, offrant des potentialités biologiques importantes.

2.15 Dynamiques : Loisirs - culture - sports

Des données spécifiant les dynamiques au sein de la commune en terme de culture (nombre de cinéma en 2021, nombre d'entrées au cinéma) ou de sport (nombre de licenciés) peuvent être ajoutées.

2.16 Quartier prioritaire de la politique de la ville (QPV) en 2024

Ces données spécifient le nombre de QPV au sein de la ville où se situe le bien.

2.17 Finances locales

Ces données indiquent la situation financière de la commune, en particulier, sont spécifiés le montant 2023 des dotations de péréquation par habitant de l'Etat à la commune, le montant en 2019 des dépenses d'équipement par habitant de la commune en 2019[12] et enfin me montant 2022 d'encours de dette par habitant de la commune.

[12] L'effort d'équipement des communes correspond au financement des nouveaux investissements qu'elles réalisent et qui portent sur leur propre patrimoine. Il est significatif de leur engagement en matière d'équipements publics de proximité mis à disposition de leurs habitants et de ceux de leur aire d'influence dans le cadre de leurs compétences

3 Produits fournis

Ces données constituent des clés de compréhension majeures du marché immobilier d'un secteur. Ainsi, les produits fournis incluront les différentes bases de données souhaitées, et, sur demande, une explication détaillée des outils informatiques utilisés. Les données sont constitués d'environ 600 000 transactions de maisons par an et 450 000 appartements, soit un total de plus de 5 millions de transactions sur cinq ans.
Des cartes interactives seront également disponibles, affichant par des points les transactions immobilières et centrées sur des zones spécifiques. Elles incluront les caractéristiques désirées des biens, permettant ainsi aux particuliers d'évaluer eux-mêmes la valeur de leur propriété en la comparant à des biens similaires récemment vendus.

4 Budget et conditions de paiement

Le tarif pour ce projet sera déterminé en fonction du volume de données et de la complexité des ajouts. Je propose un système de tarification basé sur le nombre d'heures de travail nécessaires pour accomplir les tâches demandées. Le tarif de base est de 250 euros. Ce coût représente d'une part le processus de purification de la base de données DVF, d'autre part, la collecte et l'analyse de données contextuelles qui serviront à enrichir la base de données DVF en fournissant un contexte plus large et une compréhension approfondie des dynamiques socio-économiques qui influent sur le marché immobilier.
En plus du coût de base, il convient d'ajouter les frais associés à l'appariement des données avec les informations supplémentaires. La complexité de ce travail varie en fonction des ajouts spécifiques requis. Par conséquent, une tarification adaptée sera établie en fonction du nombre de fichiers souhaités et du temps nécessaire pour les élaborer (certains traitements de données nécessitent une quantité de travail importante, c'est par exemple le cas du 2.12). Les détails de cette tarification seront discutés en profondeur lors d'un échange direct afin de garantir une compréhension claire des besoins et des coûts associés.
Les modalités de paiement seront convenues entre les parties avant le début du projet. Les paiements peuvent être effectués par virement bancaire ou via une plateforme de paiement sécurisée.

5 Spécifications techniques

Les données seront stockées sous forme de fichiers Excel ou CSV reprenant l'ensemble des transactions sur une année ainsi que la/les donnée(s) supplémentaire(s) souhaitée(s). Les cartes sous HTML. Le partage des données se fera suite à un accord préalable encadré et signé concernant les données qui seront partagées, y compris leur nature, leur format, leur structure et leur contenu. Le partage de fichiers se fera via une plateforme sécurisée, l'utilisation de services de cloud computing, ou le transfert via un système de messagerie sécurisé.

6 Modalités de collaboration

Les services souhaitées doivent m'être communiqués par mail ou par téléphone (Section 7). Un entretien afin de bien saisir ce qui est attendu par les deux parties m'apparaît nécessaire avant de mettre en place un accord. La date limite de livraison du projet complet sera spécifiée une fois les détails finaux convenus. Les délais de livraison seront respectés. Un processus de validation des données et des livrables sera mis en place pour garantir la qualité et l'exactitude des données et la satisfaction du client quant à ces dernières.

7 Contact

tristan.ancel@ens-paris-saclay.fr
07 82 04 01 10

8 Références sur le Big Data dans le secteur immobilier

Comment le Big Data bouleverse l'immobilier, Haim Treistman La révolution Big Data, Viktor Mayer-Schönberger Les données immobilières : Des outils précieux pour la prise de décision, Dorian Zerroudi

Fait à Palaiseau, le 16/06/2024
Tristan Ancel