Mécanismes d'appariement et de formation des prix sur le marché immobilier : Trois études empiriques basées sur les données d'une plateforme numérique

Pierre Vidal

Thèse de doctorat

Pour l'obtention du titre de
Docteur en Sciences Economiques
délivré par

CY Cergy Paris Université

École doctorale n°405 : Économie, Management, Mathématiques, Physique et Sciences Informatiques (EM2PSI) Théorie Économique, Modélisation et Applications (THEMA)
CNRS UMR 8184

Mécanismes d'appariement et de formation des prix sur le marché immobilier

Trois études empiriques basées sur les données d'une plateforme numérique

Présentée et soutenue publiquement par
Pierre Vidal
le 26 février 2021

Directeurs de thèse : Michel BARONI, Doyen, ESSEC Business School
Frédéric CHERBONNIER, Professeur, Toulouse Schools of Economics

Jury

Gabriel DESGRANGES, Professeur, CY Cergy Université, Président du jury
François DES ROSIERS, Professeur, Université de Laval, Rapporteur
Stéphane GRÉGOIR, Doyen, Toulouse School of Economics, Rapporteur
Paloma TALTAVULL DE LA PAZ, Professeure, Universidad de Alicante, Examinatrice
Thomas LEFEBVRE, Directeur Scientifique, Meilleurs Agents, Examinateur

Avertissements

Cette thèse a été réalisée dans le cadre d'une Convention industrielle de formation par la recherche (CIFRE) entre CY Cergy Paris Université et la société Falguière Conseil, propriétaire de la marque Meilleurs Agents.

CY Cergy Paris Université et la société Falguière Conseil n'entendent donner aucune approbation ou improbation aux opinions émises dans les thèses. Ces opinions doivent être considérées comme propres à leurs auteurs.

Remerciements

Au moment de finaliser cette thèse, qui vient conclure un long parcours académique, il me semble naturel de remercier la longue chaine des professeurs et enseignants qui m'ont accompagné tout au long du chemin, depuis mes trois ans et ma première rentrée à l'école maternelle. Je remercie à la fois les excellents professeurs que j'ai eu la chance de croiser le long de ma carrière d'élève et d'étudiant, qui m'ont donné le goût d'apprendre en les écoutant. Je remercie aussi les autres, parfois moins passionnants, qui m'ont donné le goût d'apprendre seul, compétence ô combien utile dans un travail de recherche. Cette thèse leur est dédiée à tous.

Il est bien sûr deux ultimes maillons que je tiens à distinguer particulièrement dans ces remerciements : mes deux directeurs de thèse, Michel Baroni et Frédéric Cherbonnier. Je les remercie d'avoir accepté d'encadrer une thèse aux contours d'abord mal définis, puis tout au long de ces quatre années de m'avoir aidé à affiner et préciser mon travail. Merci pour votre exigence. Merci pour votre bienveillance. Je mesure la chance que j'ai eue de vous avoir eu à mes côtés pendant quatre ans.

Il n'est de travail académique valable sans jury et je me dois de remercier, encore une fois, quatre autres professeurs : Paloma Taltavull, François Des Rosiers, Stephane Grégoir et Gabriel Desgranges ; pour avoir accepté d'en faire partie. C'est beaucoup d'honneurs que vous faites à mon travail de prendre du temps pour le lire et l'évaluer. C'est d'autant plus vrai pour les professeurs Des Rosiers et Grégoir qui ont accepté d'être les rapporteurs de ma thèse, malgré un temps disponible que j'imagine précieux.

Privilège d'une thèse CIFRE, non content d'avoir deux directeurs de thèse, j'ai également eu la chance d'être soutenu tout au long de ce travail par un responsable scientifique en entreprise, en la personne du Directeur Scientifique de Meilleurs Agents, Thomas Lefebvre. Réduire l'impact que tu as eu sur ce projet à ce simple rôle serait d'une grande malhonnêteté de ma part. C'est toi qui as implanté cette graine dans mon esprit, lors de mon stage. C'est toi qui m'as mis en contact avec Michel et Frédéric. C'est toi qui m'as soutenu (supporté ?) quotidiennement. Tu sais l'amitié que je te porte, mais ça va parfois mieux en le disant. Merci pour tout.

En plus de Thomas, c'est à l'ensemble de l'équipe Data Science que j'adresse mes remerciements pour avoir partagé avec moi ces quatre années. D'une manière ou d'une autre,

vous m'avez tous aidé à un moment ou à un autre, par un coup de main sur une roquette SQL, un restore de base, un avis mathématique ou en gardant votre calme face à un « Je ne peux pas aujourd'hui, je bosse sur ma thèse ». Vous êtes malheureusement trop nombreux pour que je vous cite tous nommément, je n'en suis pas moins reconnaissant envers chacun d'entre vous. Privilège de l'ancienneté, je n'en mentionnerais que deux, qui auront été présents tout au long de ces quatre années. Rémi : merci d'avoir accepté que je vole si souvent du temps de ton équipe avec des demandes en dehors de toute roadmap, pour faire avancer cette thèse. Youcef : merci d'avoir été la cible privilégiée de ces demandes et pour l'inspiration du troisième chapitre.

Au-delà de la seule équipe Data Science, c'est à l'ensemble des membres présents et passés de l'entreprise Meilleurs Agents que j'adresse mes remerciements. Pour qu'un doctorant puisse écrire trois essais se basant sur les données d'une plateforme numérique, il faut que cette plateforme existe et fonctionne. Quelle que soit votre fonction, vous avez donc tous contribué à rendre possible cette thèse. Il me parait naturel de remercier tout particulièrement ceux qui ont fondé cette entreprise avec l'idée de placer un peu partout des « trappes » collectant les informations qui ont permis les travaux présentés ici. Spécifiquement, je remercie Sébastien de Lafond qui est également à la source de ce travail doctoral et je m'excuse de n'avoir pas « feed the beast » comme on en avait initialement convenu.

Au-delà de Meilleurs Agents, je remercie mes amis et les personnes qui ont été présentes dans ma vie pendant ces années. Merci de m'avoir écouté quand je dissertais sur d'obscurs concepts de microstructure du marché immobilier et de ne pas m'avoir trop souvent demandé comment avançait ma thèse (particulièrement quand elle n'avançait pas). Je remercie particulièrement Nicolas pour ses relectures de mon mauvais anglais et Pierre pour m'avoir fourni l'accès à un logiciel de correction orthographique. Sans vous, cette thèse aurait été illisible et aurait fait honte à mes institutrices de primaires.

Je remercie également ma famille pour leur soutien moral pendant ces quatre années. Je remercie mes parents, non seulement pour le soutien logistique pendant les deux mois de confinement qui m'ont permis de boucler ce travail, mais surtout, pour avoir rendu possible le fait de pousser des études jusqu'à un doctorat. Trop d'enfants n'ont pas cette chance. Mes parents me l'ont offerte. Entre mille autres choses, merci pour ça.

Pour finir, je remercie l'ensemble des services de l'université de Cergy et en particulier du laboratoire THEMA pour avoir répondu aux sollicitations de l'étudiant fantôme que j'ai été. Il y avait bien du travail derrière tous ces mails et vous l'avez rendu possible.

Table des matières

Liste des tables

Liste des illustrations

1. Introduction

I. Motivation et objet de la thèse

Cette thèse ne traite ni des épidémies ni du marché de l'emploi. Pourtant elle n'est pas sans rapport avec l'article de blog posté par Karin Kimborough le 20 mars 2020. Alors qu'à la suite de la Chine, l'Europe se confinait pour tenter d'enrayer la propagation de l'épidémie de Covid-19 et que gouvernement et observateurs se demandaient quel pourrait être l'impact de cet arrêt brutal de l'activité sur l'économie, l'économiste en chef du réseau LinkedIn publiait le graphique reproduit ci-dessous.

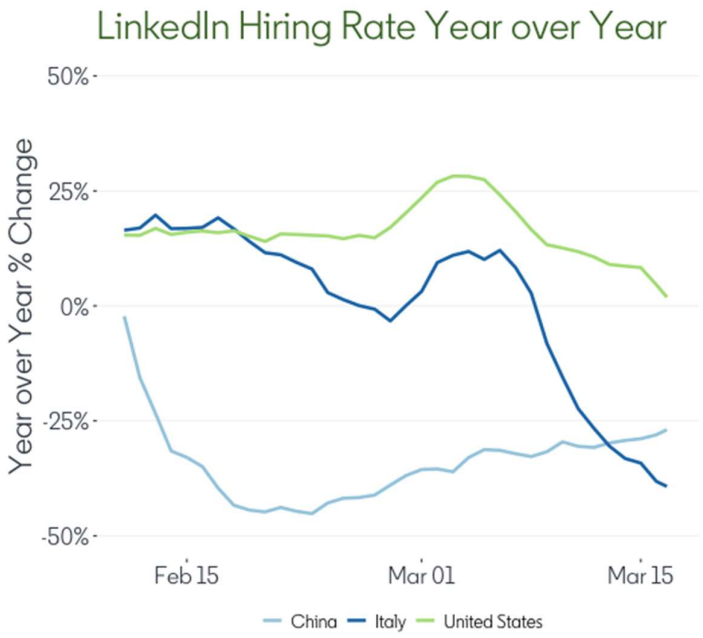

Figure 1-1 : Impact du confinement dû à l'épidémie de Covid-19 sur les embauches selon LinkedIn

En observant les embauches déclarées sur leur plateforme, les équipes du premier réseau social professionnel estimaient que la quarantaine généralisée mise en place en Chine puis en Italie, premier pays européen touché par la pandémie, avait réduit le nombre d'embauches de plus de 40% et prédisaient le même sort au marché de l'emploi américain alors que le nombre de cas commençait à s'y multiplier.

Face au besoin de comprendre les implications de cette crise inédite, notamment pour éclairer des décideurs devant prendre des décisions dans l'urgence, ils n'étaient pas seuls à utiliser des données non conventionnelles. L'Institut National de la Statistique et des Études Économiques (INSEE), dans sa note de conjoncture de mars 2020, faisait pour la première fois

l'usage de données dites de « haute fréquence » pour chiffrer la chute de l'activité. Ce passage de questionnaires à l'efficacité et l'exactitude éprouvée à des indices indirects et bruités comme la consommation électrique ou les statistiques de paiement par carte bancaire est dans la droite ligne du passage du sur-mesure (*customemade*) au prêt-à-porter (*readymade*) dans la recherche en sciences sociales que décrit Matthew Salganik dans son livre *Bit by Bit, Social Research in the Digital Age* (Salganik 2017).

La capacité de ces données non conventionnelles à capter ce qui échappe aux indicateurs classiques s'avère utile au-delà du suivi et de la prévision économique en période de crise. Elles rendent également possible l'étude empirique de problématiques structurelles dont les manifestations restaient invisibles à l'économétricien, dans le monde analogique. C'est une telle possibilité qui est explorée dans cette thèse. À travers l'utilisation de données issues de la plateforme spécialisée meilleursagents.com, elle vise à améliorer la compréhension des mécanismes de rencontre entre acheteurs et vendeurs, et de fixation des prix sur le marché de l'immobilier résidentiel.

Avec plus de 7 500 milliards d'euros, selon les comptes patrimoniaux 2019 de l'INSEE, les logements et les terrains bâtis représentent plus de la moitié du patrimoine brut des ménages français. Pour la majorité des ménages propriétaires, la résidence principale en est souvent de loin l'actif le plus important. L'intérêt d'une analyse fine du fonctionnement de ce marché et des processus qui aboutissent à la formation du prix apparait ainsi clairement. Au-delà des seuls ménages, les parties prenantes à ces questions sont nombreuses. En tant qu'objet d'imposition au moment des mutations ou, selon les pays, comme assiette de la taxe foncière, les pouvoirs publics ne peuvent qu'être intéressés par les questions qui entourent la découverte du prix des biens immobiliers. De plus, par son intrication avec le marché du travail, la compréhension du fonctionnement du marché du logement est critique pour tout gouvernement qui voudrait optimiser l'appariement entre employeurs et employés. D'autre part, le montant d'une transaction immobilière représente plusieurs années de revenus, elle donc fait souvent l'objet d'un crédit dont le financeur a tout intérêt à être en mesure d'appréhender les mécanismes en jeu. Le lien entre la crise financière globale de 2008 et le marché immobilier américain en est, en la matière, un exemple caricatural. Enfin, les différents acteurs de l'industrie qui s'est développée pour accompagner les transactions immobilières profiteraient bien évidemment d'une meilleure compréhension académique de l'objet même de leur travail. C'est en premier

lieu vrai pour les intermédiaires comme les agents immobiliers, mais aussi les entreprises, notamment numériques, qui vendent outils et services à ces derniers.

Ce n'est donc pas un manque d'intérêt, mais bien un manque de données, qui est la cause du développement limité de la littérature empirique traitant des mécanismes microéconomiques présidant à une vente immobilière. Jusqu'à présent, l'essentiel des études économétriques sur le marché du logement se basait sur les registres fiscaux ou notariaux, qui ne renseignent que sur les transactions finales et ne disent rien des mécanismes ayant abouti à ces transactions. Dans le meilleur des cas, les auteurs ayant accès aux historiques des annonces passées par les vendeurs pour promouvoir leur bien en vente obtenaient ainsi une vision partielle du processus, mais demeuraient aveugles à l'activité des acheteurs. Cette situation se trouve en partie abolie par l'arrivée et l'importance croissante de plateformes numériques intervenant au cours de ces transactions. Comme il est d'usage pour les entreprises du net, ces plateformes apportent un soin particulier à la collecte des traces que laissent derrière eux les utilisateurs de leurs services. Par ailleurs, en tant qu'intermédiaire sur un marché multifaces, reliant acheteurs, vendeurs et agents immobiliers, elles ont vocation à intervenir à chaque étape des processus amenant à la transaction. Ces deux facteurs en font des lieux privilégiés d'observation permettant d'accéder à une vision holistique des mécanismes à l'œuvre dans le marché immobilier résidentiel. Les apports empiriques qu'entend développer cette thèse sont essentiellement bâtis sur l'avantage qui découle de l'accès à de telles données inédites.

Une présentation détaillée des trois études, qui forment le corps du travail de cette thèse, est faite dans la cinquième et dernière section de cette introduction. Auparavant, et pour permettre de pleinement apprécier la démarche entreprise ici, nous nous proposons de présenter au lecteur une réflexion sur l'utilisation des données issues du monde numérique dans le cadre de recherche académique. Nous commencerons par identifier les avantages et inconvénients de ces données, du point de vue des sciences sociales. La section suivante analyse le cas particulier que représentent les plateformes numériques pour les économistes. Enfin, nous identifions les opportunités offertes par les données des plateformes spécialisées pour la recherche en immobilier.

II. Les données numériques dans la recherche en science sociale

La particularité de cette thèse réside donc dans les données qu'elle mobilise. Pour l'essentiel, la littérature empirique en économie et en finance immobilière s'appuie sur des

données de transactions ou d'annonces collectées par des institutions publiques, comme le fisc (base des Demandes des Valeurs Foncières), les notaires (Base d'Informations Économiques Notariales), ou des acteurs du marché comme les agents immobiliers (base *Multi Listing Services* aux États-Unis). Sont également utilisées des enquêtes menées auprès des particuliers comme l'Enquête Logement de l'INSEE ou les déclarations de patrimoine des ménages, particulièrement dans les pays où la fiscalité foncière est basée sur la valeur vénale des biens immobiliers. Les trois études présentées ici exploitent un autre type de données : les traces laissées par les utilisateurs d'une plateforme immobilière en ligne, le site meilleursagents.com. En cela ces travaux participent au mouvement commun à l'ensemble des sciences sociales d'utilisation des données numériques à des fins de recherches académiques.

Ce changement de paradigme dans la recherche en sciences sociales est le fruit d'un changement sociétal et anthropologique majeur lié au rôle important et toujours grandissant des technologies de l'information et de la communication depuis la fin du XXe siècle. Une des très nombreuses conséquences de ce passage vers une société de l'information (Webster 2002) est la croissance exponentielle de la quantité d'information produite et stockée à l'origine de ce que l'on a appelé les Big Data, ou données massives (voir Hilbert et Lopez, 2011, pour une mesure de cette croissance). Par leur omniprésence, les ordinateurs, smartphones et divers capteurs connectés documentent la vie des êtres humains dans des proportions inégalées. Même si on laisse de côté les traces numériques inconscientes que nous semons (positions GPS, informations de paiement par carte bancaire, données d'utilisations des sites et applications, entre autres), la production volontaire de contenu à travers le web participatif reste vertigineuse. On se restreindra ici à un seul exemple, le réseau social Twitter estime que 500 millions de tweets sont écrits et envoyés chaque jour sur sa plateforme. À raison de 33 caractères en moyenne[1], c'est plus de 1,6 milliard de signes qui sont écrits chaque jour, soit l'équivalent de 1 700 fois *À la Recherche du Temps Perdu* (qui compte 9 609 000 signes, ou 1,5 million de mots, répartis sur 7 tomes). Sans compter que le roman de Marcel Proust ne contient ni image ni vidéo.

[1] techcrunch.com/2018/10/30/twitters-doubling-of-character-count-from-140-to-280-had-little-impact-on-length-of-tweets/

Ce gisement de données décrivant le comportement de leur(s) sujet(s) d'étude constitue une opportunité pour les sciences sociales. Opportunité qui a déjà été saisie dans de nombreuses études. La figure 2 ci-dessous décrit l'évolution du nombre de publications en langues anglaise et française, utilisant le mot-clef « Big Data », accessibles depuis l'outil d'accès aux ressources documentaires du Centre National de Recherche Scientifique (bib.cnrs.fr), pour la section INSHS (Institut National des Sciences Humaines et Sociales). Si ce n'est une étude de 1974, qui apparait presque comme anachronique bien que traitant déjà d'une méthode de traitement automatisée de grande base de données (Howard et al. 1974), c'est au tournant des années 2000 qu'émergent les premières études faisant référence à ce concept alors récent et utilisé pour la première fois[2] dans sa signification actuelle dans une publication de l'*Association for Computing Machinery* en 1997 (Cox et Ellsworth 1997). Depuis 2013, coïncidemment l'année des révélations Edward Snowden sur l'utilisation de telles techniques à des fins de surveillance à grande échelle, le nombre de publications décolle et commence une inflation toujours d'actualité en 2020 (en moins de trois mois ce nombre atteint 32% du total de l'année précédente).

Cette inflation s'explique par un avantage décisif que présente ces données d'un nouveau genre : elles sont ubiquitaires. Comme déjà dit plus haut, l'omniprésence des outils digitaux et connectés dans nos sociétés a fait exploser la quantité d'information disponible. Loin de ne faire qu'enregistrer plus souvent, avec plus de détail ou auprès de plus d'individus les mêmes choses, le monde numérique garde des traces de phénomène et de comportements qui restaient jusque-là invisibles.

Cependant, comme le rappelle Matthew Salganik en conclusion de son ouvrage d'analyse du phénomène (Salganik 2017), la règle du « *no-free lunch* » s'applique aussi à l'utilisation de données numériques à des fins de recherche. La première caractéristique de ces bases numériques qui vient modérer l'enthousiasme soulevé par leur utilisation par les chercheurs est qu'elles n'ont pas été constituées dans cette finalité. Salganik distingue ainsi les données « sur-mesure » (*custommade*) habituellement utilisées par les chercheurs et les données « prêt-à-porter » (*readymade*) que l'on peut trouver dans les bases numériques

[2] www.forbes.com/sites/gilpress/2013/05/09/a-very-short-history-of-big-data/#1b43033465a1

constituées par des entreprises ou des institutions gouvernementales dans un tout autre but. Aux chercheurs de s'adapter à la nature des données disponibles et d'effectuer d'éventuels traitements et redressements pour en tirer une information exploitable. En la matière, les économistes ont sûrement une longueur d'avance sur les autres disciplines qui exploitent plus volontiers les résultats d'enquêtes ou d'expériences. En effet, bien avant la vague des « Big Data », les informations fiscales, établies en premier lieu pour lever l'impôt, constituent depuis longtemps une des matières premières les plus utilisées en économétries. Pour se cantonner au domaine immobilier, la mobilisation depuis 2002 des bases notariales par l'INSEE dans ses Indices des Prix des Logements (David et al., 2002) est un exemple de réutilisation de données non numériques.

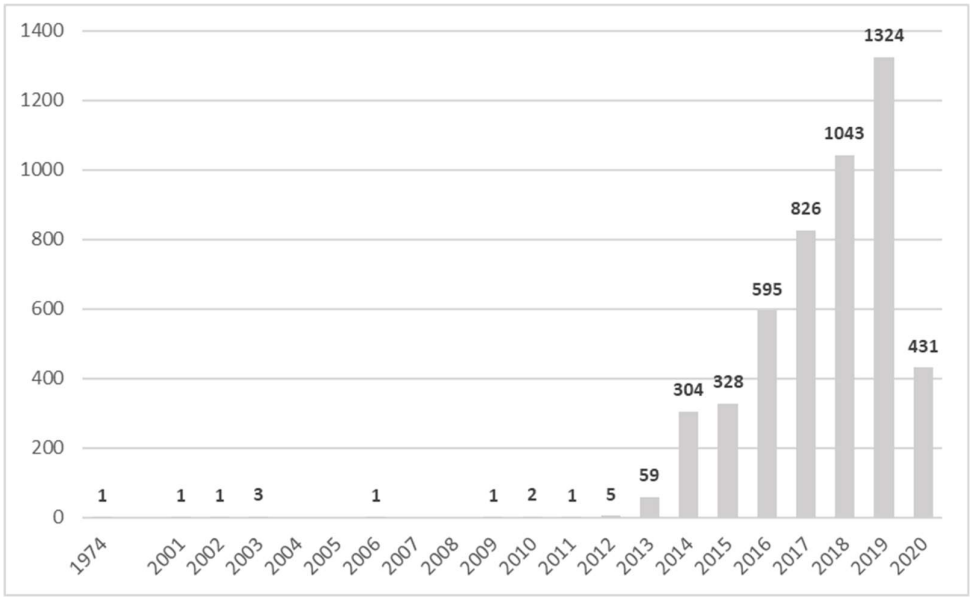

Figure 1-2 : Nombres de publications, en langues anglaise et française, le mot-clef "Big Data" accessibles depuis le portail bib.cnrs.fr (le 27/03/2020), section INSHS, classées par année de publication

Toujours dans le même ouvrage Salganik identifie dix caractéristiques des données numériques. Trois sont à ces yeux des avantages pour leur utilisation par des chercheurs et sept des inconvénients qui poussent à les manipuler avec précaution. Pour pleinement comprendre les enjeux liés à l'utilisation qui en est faite dans cette thèse, il est important que le lecteur ait cette caractérisation à l'esprit. Nous nous proposons donc de la résumer ici, en l'illustrant autant que possible par des exemples tirés de travaux présentés ici. Pour plus de détails et d'exemples, le lecteur voulant approfondir le sujet est renvoyé à l'ouvrage original. Ces données sont, selon lui :

- Massives
- Enregistrées en continu
- Inertes
- Inaccessibles
- Incomplètes
- Non représentatives
- Dérivantes
- Perturbées algorithmiquement
- Sales
- Sensibles

Le premier point peut sembler évident à première vue : les *Big Data* sont *big*. Pour autant, c'est la caractéristique sur laquelle il nous semble nécessaire de s'appesantir, car elle peut créer des attentes injustifiées. Toutes les données recueillies dans le monde numérique ne constituent pas nécessairement des bases de données volumineuses. Le petit exercice ayant permis la construction de la figure 2, plus haut, est incontestablement le fruit de l'utilisation de données numériques, pourtant on aura du mal à affirmer qu'un tableau de 21 lignes et 2 colonnes soit massifs. De la même façon, la taille des jeux de données utilisés dans les trois études présentées dans cette thèse ne sera pas constituée de millions d'observations. Même si dans les deux cas, il s'est agi d'isoler et agréger des éléments au sein de très larges bases de données, ils illustrent bien que toutes recherches mobilisant des données numériques ne sont pas nécessairement de l'ordre du « Big Data ».

Il n'en demeure pas moins que par le faible coût marginal et l'échelle globale, via Internet, dans laquelle elles peuvent être mises en œuvre, les technologies numériques permettent de constituer des bases sensiblement plus grandes que ce qui pouvait se faire dans le monde analogique. Cette taille représente un atout pour les chercheurs qui voudraient les exploiter. On mettra bien évidemment de côté la dimension de talisman dont pourraient se parer certaines études (« ces résultats sont inattaquables, car basés sur un grand nombre d'observations »). En revanche, grâce à leur profondeur, i.e. le nombre d'observations, et leur largeur, i.e. le nombre de variables décrivant chaque observation, de telles bases permettent d'aller au-delà de l'effet moyen et d'étudier l'hétérogénéité des phénomènes. Elles rendent également possible l'observation d'événements rares. En reprenant l'exemple de la figure 2, la publication d'un papier utilisant un mot-clef donné est quelque chose de rare au regard du nombre d'études publiées chaque année, mais la profondeur de la base documentaire du CNRS permet l'observation de l'évolution de ce phénomène. Enfin, la taille représente bien

évidemment un atout pour mesurer avec une certaine assurance statistique des différences faibles.

Le deuxième avantage est que les systèmes qui génèrent ces données sont pour la plupart en permanence en fonctionnement. Cet enregistrement de l'état du système en continu permet l'étude de phénomènes imprévisibles, mais surtout il permet au chercheur d'étudier les réactions des sujets sur le vif, tout en profitant d'information sur contexte pré et post-choc. L'exemple de l'utilisation faite par les économistes de LinkedIn de leurs données, présenté en figure 1, est un excellent exemple de comment il est possible de tirer avantage de cette particularité des données numériques. Premièrement, les équipes de Karin Kimborough sont capables de détecter l'impact de la crise sanitaire sur les embauches « en direct », chose impossible pour les enquêtes trimestrielles suivant les méthodes préconisées par le *Bureau International du Travail*. Deuxièmement, la figure 1 décrit la dynamique du phénomène, avec une chute brutale puis une recrue lente, bien au-delà de ce que l'on pourrait attendre d'une enquête menée a posteriori auprès d'un panel de DRH, qui auraient dû faire appel à une mémoire imparfaite et soumise à des biais cognitifs bien connus.

Cette comparaison nous permet également d'illustrer le troisième atout des données numériques. Elles constituent des données d'observation inertes, en ce que leur collecte est peu susceptible de modifier le comportement des personnes qui font l'objet de l'étude. Pour reprendre l'exemple des DRH, il semble possible qu'un certain nombre se refuse à avouer l'effet de sidération visible dans la chute abrupte des taux d'embauche sur la figure 1, de peur d'avoir à reconnaitre une certaine impréparation ou une surrection au choc. On pourra opposer à cette idée que les utilisateurs de ces différents services numériques savent très bien que leurs actions sont enregistrées et observées. C'est cette prise de conscience populaire qui a d'ailleurs poussé les gouvernements à prendre des mesures pour encadrer ces pratiques, comme le Règlement Général sur la Protection des Données (RGPD). Pour autant, on imagine mal un ménage qui allumerait moins (ou plus) ses appareils électriques, car il sait que sa consommation est suivie par un compteur intelligent, ou l'utilisateur quotidien d'une plateforme numérique se comporter autrement que selon ce qui lui semble être son intérêt, parce qu'il sait que ses actions seront compilées par les analystes travaillant pour ce site.

Tournons-nous maintenant vers l'autre versant, moins avantageux, de cette nouvelle matière première à disposition des chercheurs en sciences sociales. Le premier problème que posent ces données c'est que, précisément, elles ne sont pas disponibles. Les bases les contenant

ont été constituées par des organisations, publiques ou privées, qui ne sont pas nécessairement disposées à les partager. On peut comprendre qu'à l'heure où la « data est le nouvel or noir », communiquer à un tiers, fût-il universitaire, des informations constituant un avantage compétitif ne soit pas le premier réflexe des entreprises. Au-delà des considérations stratégiques, il existe une très bonne raison de ne pas partager ces données qui sera exposée plus bas : elles contiennent des informations sensibles. La loi, entre autres à travers le RGPD déjà mentionné, encadre ainsi très strictement les conditions d'un tel partage. Encore une fois, par leur utilisation des bases fiscales protégées, les économistes ont sans doute déjà sur cette question une maturité plus grande sur le reste des chercheurs en sciences sociales et ont fait la preuve qu'il existe des solutions pour accéder à de telles données. Pour ce qui est de celles détenues par des entreprises la meilleure stratégie consiste sans doute à nouer un partenariat duquel l'entreprise pourra retirer intérêt. C'est une des préconisations d'Andrey Fradkin, qui a collaboré avec le site Airbnb, dans son *Guide To Using Corporate Data for Academic Research*[3]. C'est aussi le moyen utilisé dans cette thèse, rendu possible par l'établissement d'une Convention Industrielle de Formation par la Recherche (CIFRE) entre l'entreprise qui opère la plateforme, le laboratoire THEMA et l'auteur, embauché comme un employé à part entière par la société. En revanche, une fois l'accès établi, de cette façon ou d'une autre, les mesures faites et l'étude écrite, le partage des données à des tiers reste la plupart du temps impossible. Cela empêche notamment les résultats d'être reproduits à moins que d'autres équipes n'entreprennent la même démarche auprès d'une organisation similaire.

Deuxième défaut de ce type de données, elles sont incomplètes. Pour rappel, les organisations qui collectent ces données ne le font pas en premier lieu à des fins de recherche. Manquera donc quasi systématiquement des éléments habituellement présents dans des données *custommade*, comme les profils sociodémographiques des participants ou une variable permettant de rendre opérant un concept théorique. Par ailleurs, le caractère « inaccessible » de ces données fait que le chercheur qui réussirait à observer le comportement des utilisateurs sur une plateforme resterait aveugle à leurs actions sur d'autres plateformes concurrentes ou complémentaires. Dans notre cas, les particuliers qui utilisent la plateforme que nous observons peuvent très bien s'informer sur les prix de l'immobilier auprès d'autres sources en ligne ou

[3] andreyfradkin.com/posts/2014/02/08/how-to-obtain-proprietary-datasets-for-research-part-1

hors-ligne (agents immobiliers, bouche-à-oreille, notaires). Impossible de s'assurer que l'influence de l'information diffusée sur le site n'est pas perturbée par l'action de ces autres sources. De plus, l'accès à une plateforme dédiée à l'estimation en ligne ne nous révèle qu'une partie de l'histoire des projets de ventes et d'achats. Le comportement des acheteurs, par exemple, sur les sites d'annonces immobilières permettrait certainement de compléter nos analyses.

Au-delà du problème de complétion, il existe également un problème de non-représentativité. En plus de ne pas avoir tous les éléments pour les personnes qui utilisent le service sur lequel sont faites les mesures, il faut faire face à l'absence totale de données sur celles qui ne l'utilisent pas. Si les usages numériques se généralisent, l'asymétrie de leur diffusion au sein des différentes strates de la société, on peut cependant affirmer que cet aveuglement ne concerne pas les différentes populations à parts égales[4]. Le manque de variables sociodémographiques déjà évoqué, compliquant un redressement statistique, ne fait qu'accentuer ce problème de non-représentativité. Nos mesures ne sont bien évidemment pas épargnées par cet écueil, malgré des indices nous laissant penser qu'il est limité (prépondérance d'Internet dans la préparation d'un projet immobilier[5] et notoriété de la plateforme dans sa catégorie[6] notamment). Pour garder la maitrise de ce risque planant sur nos résultats, et en l'absence de données sur les utilisateurs, une comparaison systématique entre les caractéristiques de biens estimés sur le site et ceux ayant fait l'objet d'une transaction (enregistrée dans les bases fiscales ou notariales) est réalisée pour les trois études présentées ici. Notons par ailleurs que si le manque de représentativité pose un problème quant à la généralisation « hors échantillon », pour faire des prédictions par exemple, elle n'empêche en rien la comparaison des effets au sein de l'échantillon, ce à quoi se cantonnent nos mesures économétriques.

[4] Pour la France, se référer au « Baromètre du Numérique 2019 » du CREDOC pour des éléments quantitatifs : www.credoc.fr/publications/barometre-du-numerique-2019

[5] 81% des acheteurs immobiliers ont utilisé un service en ligne, source Baromètre Digimmo de Pierre & Vacances Conseil Immobilier 2018

[6] 76% de part d'audience chez les internautes ayant visité un site d'estimation immobilière en mars 2019, source : Médiamétrie pour Meilleurs Agents

Quatrième inconvénient, comme les environnements sur lesquels elles sont réalisées, les mesures dans le monde numérique « dérivent », les rendant de ce fait peu adaptés à des études longitudinales. Salganik note trois changements qui peuvent impacter ce nouveau type de données d'observation. D'abord, les modifications que les équipes des sites web apportent en continu. Modifications d'autant plus nombreuses que les entreprises de la nouvelle économie sont friandes des méthodes de développements dites « agiles » (Beck et al., 2001) impliquant des mises à jour fréquentes. De ce point de vue, l'accès privilégié aux équipes assurant le fonctionnement de la plateforme dont l'auteur a bénéficié en sa qualité de salarié de l'entreprise est un atout. Cela a permis d'identifier des périodes où les perturbations des observations dues à de tels changements techniques étaient faibles, pour réaliser les mesures nécessaires aux travaux présentés ici. Deuxièmement, dans un environnement concurrentiel fort et au rythme des évolutions technologiques, les utilisateurs changent selon la popularité croissante ou décroissante des outils étudiés. En la matière, et malgré une maturité importante de l'entreprise à l'échelle du digital (création en 2008), nos mesures exclues des périodes où l'audience varie trop fortement (période initiale de croissance avant 2011, diffusion d'une publicité à la télévision à partir de 2019). Enfin, une dérive peut intervenir dans les usages, c'est-à-dire comment les utilisateurs se servent des outils et services numériques. Par exemple, en observant la liste des vidéos les plus visionnées sur la plateforme de vidéo YouTube, exclusivement constitué de clips musicaux, on pourrait conclure que c'est là le contenu que préfèrent *regarder* les internautes. Ce serait omettre que les utilisateurs ont détourné l'outil de son usage premier pour s'en servir comme d'un service streaming de musique pour *écouter* ces chansons. Encore une fois l'accès facilité aux équipes du site, notamment les équipes produits qui rencontrent régulièrement les utilisateurs pour étudier l'usage qu'ils font des outils fournis par l'entreprise, minimise ce risque dans notre cas.

La perturbation algorithmique, cinquième écueil de la liste, est quelque chose de nouveau, propre aux données du monde numérique. Les comportements des utilisateurs n'y sont pas parfaitement naturels, mais orientés par la technologie. Les concepteurs du système ont intérêt à encourager certains comportements et en prévenir certains autres, c'est particulièrement évident dans le cas de sites commerciaux détenus par des entreprises. Que les comportements des personnes soient modifiés par la finalité du système n'est pas nouveau. On pensera à la sous déclaration des revenus et du patrimoine dans le cadre des données fiscales. Ce qui est nouveau c'est que cette modification des comportements peut ici être encouragée par les développeurs des outils numériques, ce qui n'est pas le cas du fisc, on en conviendra

facilement. Ce défaut est particulièrement dangereux, du point de vue de la recherche, quand les concepteurs utilisent des résultats établis dans la littérature pour guider leur design, créant ainsi un effet de performativité. Dans leur étude des liens d'amitié à travers le réseau Facebook, Ugander et al. (2011) vérifiait un résultat connu, voire proverbial : « les amis de mes amis sont mes amis ». Ils n'étaient cependant pas dupes que ce résultat était fortement influencé par le moteur de suggestion de « personnes que vous connaissez peut-être » de Facebook, basé sur cette idée, qui rapproche volontairement des gens ayant de nombreux *amis* en communs. De tels effets n'ont pas été identifiés dans le cadre des travaux présentés ici.

Le problème suivant, lié à ce qu'on pourrait appeler la saleté des observations, n'est lui en revanche pas nouveau, mais peut être particulièrement aigu pour ce nouveau type de données. Le risque de voir des résultats perturbés par des observations aberrantes est exacerbé dans le monde numérique par la nature prêt-à-porter des données. Le chercheur ne maitrise pas les conditions dans lesquelles elles ont été produites et collectées. De plus, le caractère massif des bases ainsi constituées exclut toutes vérifications ou méthode de filtrages individuels. Beaucoup d'observations utilisées dans ce travail sont le fruit de la saisie manuelle par des utilisateurs du site, en dehors de toute supervision. Ces derniers peuvent faire des erreurs ou volontairement entrer des valeurs absurdes pour voir comment le système réagit. Ces valeurs anormales sont dans la majorité des cas facilement détectables par comparaison avec les caractéristiques standards des biens immobiliers (une maison de 1 m² ou 10 000 m² n'existe certainement pas). La méfiance de certains internautes vis-à-vis d'un site commercial peut également être source d'erreur. Par exemple, quand il leur est demandé de décrire la nature de leur projet. Cette méfiance entraine une sous-déclaration de certaines caractéristiques comme le fait d'être propriétaire d'un bien ou d'en envisager la mise en vente. Si cette sous-déclaration éventuelle n'a pas posé de problème dans nos mesures, elle nous a forcé à être d'autant plus vigilants sur une dérive d'usage la concernant. Enfin, comme tout site Internet, la plateforme siège de nos observation est exposé aux *spammeurs,* ces utilisateurs qui surutilisent les outils mis à leur disposition. Il a, par exemple, fallu systématiquement ne prendre en compte qu'une seule estimation faite par un utilisateur donné, à une adresse donnée et à une date donnée, pour éviter de voir les observations bruitées par une personne essayant « d'optimiser » la valeur de son bien ou une autre jouant avec les paramètres pour comprendre comment le modèle fonctionne. On peut rapprocher ce problème avec le cas bien connu de l'étude de Back et al. (2010), analysant les sentiments exprimés dans les messages envoyés par les Américains via leur *pager* le 11 septembre 2001, qui mesurait une montée de la colère le long de la journée.

Une année plus tard, Pury (2011) montrera qu'un très grand nombre de ces messages de « colère » étaient le fait d'un seul robot envoyant des messages d'erreurs « CRITIQUE ».

Pour finir, la dernière caractéristique des données numériques, liée à leur ubiquité et justifiant leur inaccessibilité, est qu'elles contiennent des informations sensibles. Les utilisateurs font confiance aux entreprises et organisations gouvernementales en leur déclarant certains aspects de leur vie ou même en utilisant des services qui révèlent des choses qu'ils préfèrent voir restées secrètes. Les bases auxquelles nous avons eu accès contiennent ainsi les noms, prénoms, adresses, numéros de téléphone, valeurs d'achats et estimées du patrimoine immobilier de millions de français. Plus sensible encore, certains en déclarant la raison de leur projet de vente révèlent des détails intimes de leur vie comme le décès d'un proche, la naissance d'un enfant, un divorce ou un surendettement. Il va sans dire qu'il est du devoir du chercheur de protéger les informations sensibles auxquelles il a accès. Deux règles ont ainsi été suivies dans cette optique pendant les recherches présentées ici. D'abord, ne pas extraire des bases des informations inutiles aux mesures effectuées, par exemple les noms et coordonnées des utilisateurs. Deuxièmement, travailler dans le respect strict des normes de sécurité informatique pour limiter les risques de fuites de données. Cela se traduit par le chiffrement du disque dur de l'ordinateur servant à la manipulation des données et par l'accès aux bases uniquement depuis les locaux de l'entreprise ou sur une connexion sécurisée et à travers un réseau privé virtuel. Bien sûr ces règles d'éthique de travail s'ajoutent à l'impossibilité déjà mentionnée de publier les données ayant servi aux mesures, même si cela réduit la portée des travaux. En la matière, le cas de Abdur Chowdhury chef de la recherche chez AOL a fait école. Publiant en 2006 une base anonymisée de requêtes à un moteur recherche, il fut contraint de démissionner après qu'il fut prouvé qu'il était possible de réidentifier les personnes à l'origine de ces requêtes (Barbaro and Zeller 2006).

En conclusion, l'enthousiasme des chercheurs en sciences sociales pour les « Big Data » (voir figure 2) n'est pas l'effet d'une mode, mais est bien motivé par des avantages inédits que présentent les données tirées du monde numérique en tant que données d'observations. L'utilisation de bases constituées par des entreprises ou des gouvernements nécessite cependant des précautions méthodologiques pour éviter les pièges tendus par ces données qui ne sont littéralement pas *faites* pour la recherche. Il n'en demeure pas moins que l'existence de cette masse d'information sur nos actions constitue un changement de paradigme majeur dans l'histoire de l'étude des comportements humains. Dans une paraphrase de George Box (Box

1976), des observateurs de l'industrie du web[7] allant même à penser que « tous les modèles sont faux, de plus en plus vous pouvez réussir sans eux » (Anderson 2008). Si nous n'avons pas encore vu, et ne verrons certainement jamais, cette « fin de la théorie », ce changement d'échelle dans le nombre et l'étendue des observations verra certainement la démarche inductive progresser, au détriment de la méthode hypothético-déductive, dans toutes les sciences sociales. Pour ce qui est de l'économie, il est un type particulier de données qui est à même de changer la façon de faire de la recherche. Ce sont les traces laissées par les utilisateurs des plateformes numériques.

III. Les plateformes en ligne, reflets numériques de l'économie

En chimie, un catalyseur est un élément qui déclenche ou accélère une réaction. C'est par ce terme que David Evans et Richard Schmalensee (2007) décrivent ces entreprises du web qui, implémentant un modèle économique de plateforme, ont pris une position dominante dans de nombreuses industries : publicité (Google), distribution (Amazon), tourisme (Airbnb), transport (Uber), média (Facebook), pour ne citer que les plus emblématiques. La notion de plateforme n'est pas nouvelle en économie. Elle a été pour la première fois formalisée dans un article fondateur par Tirole et Rochet (2003). Ils les définissent comme des organisations permettant à utilisateurs et fournisseurs d'interagir sur des marchés bi- ou multifaces présentant des externalités de réseau croisées positives. Grâce à sa capacité de mettre en relation un très grand nombre d'individus à un coût marginal négligeable, Internet a permis l'émergence de ce type d'entreprises. Les données générées par les acteurs de ces marchés jouent un rôle primordial dans l'organisation et la conception des plateformes web. Un grand soin est donc apporté à leur collecte. Dans une analyse néomarxiste de cette nouvelle économie, Nick Srnicek (2017) qualifie même ces dernières d' « appareils d'extraction de la donnée utilisateur ». Du point de vue du chercheur en économie, elles sont une occasion sans précédent d'observer *in situ* le comportement granulaire des agents économiques. Peter Coles, ancien professeur d'économie à Harvard et aujourd'hui chef économiste chez Airbnb, qualifie ces entreprises de « magasin de bonbon pour économiste »[8].

[7] www.wired.com/2008/06/pb-theory/

[8]www.nytimes.com/2016/09/04/technology/goodbye-ivory-tower-hello-silicon-valley-candy-store.html?_r=2

Les grandes entreprises technologiques, conscientes de l'apport que peuvent avoir sur leurs affaires des économistes munis de tels outils, ont d'ailleurs constitué de larges équipes parfois débauchées dans de prestigieuses universités américaines. Althey et Luca (2019) soulignent qu'Amazon emploie aujourd'hui largement plus d'économistes à plein temps que les plus grands départements d'économie universitaires. Si une partie de ces embauches vise à apporter des arguments aux lobbyistes combattant la régulation par les puissances publiques, une autre entend penser les règles régissant les marchés internes de ces plateformes. La notion d'économiste ingénieur (Roth 2002) prend ici tout son sens. Ces entreprises ne vendent pas un produit, mais un marché et elles entendent s'appuyer sur la littérature économique pour maximiser leurs profits par l'appariement entre acheteurs et vendeurs. C'est donc assez logiquement qu'une littérature de conception de marché s'est développée avec l'immense avantage de pouvoir s'appuyer sur la manne de données collectées sur ces plateformes. Andrey Fradkin (2017a) identifie trois problèmes qui se posent à ces entreprises qui relèvent du domaine de compétence des économistes : mécanisme d'appariement, fixation des prix et contrôle de la qualité.

Le premier relève d'un arbitrage entre les coûts de recherche, la qualité des *matchs* et leur nombre (Einav et al. 2016). Par exemple, un appariement algorithmique, qui est la norme pour les applications de VTC, sera préférable si les préférences du consommateur sont homogènes et qu'il est indifférent à l'identité du fournisseur à condition qu'un cahier des charges simple soit respecté. Dans le cas d'une course VTC : adresses de prise en charge et de destination et temps d'attente minimal. Un tel mécanisme centralisé permet de réduire les coûts de recherche et maximise le nombre de *matchs*. Il sera en revanche moins adapté pour un marché où la propension à payer dépend fortement de la qualité intrinsèque de l'appariement entre un acheteur et un vendeur donné, comme dans le cas d'un logement de vacance. C'est pourquoi Airbnb et les sites similaires laissent le choix à l'utilisateur tout en le guidant grâce à un moteur de recherche. Là encore des questions sur l'ordre (Ursu 2018), du filtrage (Fradkin 2017b) et de la quantité d'information mise à disposition du consommateur (Romanyuk 2017) dans les résultats de ce dispositif de recherche se posent.

Les options pour fixer les prix sont également multiples : enchères, prix choisis par les fournisseurs ou ajustement automatique par la plateforme. En la matière, la question de l'information sur le niveau de demande est clef. Les mécanismes d'enchères sont efficaces quand l'incertitude est forte et que le bénéfice lié à la découverte du prix dépasse le coût

important lié à ce processus (Einav et al. 2018). Dans les cas plus classiques, l'avantage informationnel des vendeurs justifie qu'ils fixent les prix. Cependant, la masse d'information recueillie dans le passé et la vision à l'instant présent de l'état de la demande permettent aux plateformes d'intervenir directement sur les prix sous la forme de prix conseillés (Airbnb) ou fixés (Uber). Hall et al. (2016) montre comment ce mécanisme appelé le « *surge* » sert à équilibrer demande et offre géographiquement et temporellement pour maximiser le nombre total de trajets parcourus grâce à la plateforme.

Enfin le dispositif d'avis client, permettant d'assurer l'intégrité d'une transaction entre de parfaits inconnus, a pris une importance prépondérante dans l'économie numérique. Cette fonctionnalité a si bien fait ses preuves (Luca 2016, Dellacoras 2003) qu'il parait aujourd'hui difficile d'imaginer acheter un service ou un produit sur un site qui ne la proposerait pas. Son implémentation, en revanche, pose de sérieux problèmes. Comment éviter les représailles (Bolton et al. 2013) ? Y a-t-il un biais de sélection parmi les gens qui prennent le temps d'écrire une critique (Masterov et al. 2015) ? Faut-il inciter les utilisateurs à laisser un commentaire et comment (Fradkin et al. 2017) ? Quel est l'impact de la fraude aux faux avis (Mayzlin et al. 2014) ?

Loin de ne faire que répondre aux questions que se posent les plateformes pour optimiser leur fonctionnement, les économistes peuvent en retour les utiliser pour investiguer des problématiques d'ordre général. Ces environnements peuvent en effet servir de terrain d'observation et d'expérimentation pour réaliser des mesures de quantités jusque-là invisibles et/ou à une échelle inédite. En revanche, ce « magasin de bonbons » présente un risque, celui de ne pas savoir où donner de la tête. Comme souligné dans Salganik (2017), ce n'est pas parce qu'un phénomène n'a jamais été mesuré que sa mesure est intéressante pour autant. Savoir quelles questions poser est l'un des trois enjeux majeurs de l'analyse économique à l'heure des données numériques pour Einav et Levin (2014).

Un tel exemple de réponse à une question intéressante et bien définie est Einav et al. (2014). En comparant les ventes inter et intra état, aux États-Unis, sur eBay, ils mesurent l'élasticité de la demande aux taxes. Elle s'avère positive, mais bien moins forte que ce à quoi on pourrait s'attendre pour un changement de prix. Cette étude est un modèle du genre. Elle traitre d'une question de recherche importante : l'impact des taxes sur la demande. Elle utilise au mieux une fonctionnalité du site : les niveaux de taxe ne sont calculés et affichés qu'après que l'acheteur ait indiqué sa volonté d'achat. Elle se base sur une situation inédite dans le

monde physique, ou même le site Internet d'un revendeur unique : la plateforme regroupe l'offre de plusieurs fournisseurs de produits similaires ou même identiques et génère des ventes sujettes à taxation (ventes intra état) et d'autres qui ne le sont pas (ventes inter état).

Si elles atteignent une masse d'utilisateurs suffisante, les plateformes peuvent se révéler utiles pour suivre l'activité en temps réel. L'exemple le plus connu de *nowcasting*[9] reste l'outil Google Flu Trends, décrit dans Ginsberg et al. (2009), sensé suivre l'évolution des épidémies de grippes au plus près en se basant sur le nombre de recherches en décrivant les symptômes. Par sa quasi-hégémonie parmi les moteurs de recherche Google est effectivement un excellent point de passage pour placer un capteur et construire des mesures indirectes de statistiques économiques : chômage, consommation, confiance des ménages, demande d'information par les investisseurs (Choi et Varian, 2012, Vosen et Schmidt, 2011, Drake et al. 2012). Mais les données de plateformes ne sont pas nécessairement la panacée dans ce genre d'étude. Cavallo et Rigobon (2016) préfèrent par exemple suivre les prix sur les sites web des grandes chaînes de distribution plutôt qu'Amazon, dans leur projet de suivi de l'inflation à grande échelle[10]. L'argument étant que les prix sur ces sites sont plus sûrement liés à ceux pratiqués en magasin, où se passe encore la majorité des achats, notamment alimentaires. Ils s'inquiètent également des perturbations sur leurs mesures que pourraient engendrer les filtres et autres mécanismes mis en place par ces plateformes et décrits plus haut. On en revient aux problèmes de perturbation algorithmique et de dérives propres aux données du monde numérique (Salganik 2017).

En revanche, parce qu'elles sont enregistrées en continu, les données générées par les plateformes numériques peuvent être le lieu de réelles expérimentations. Avec un tel dispositif de collecte, il n'y a qu'à attendre qu'un choc sépare certains utilisateurs des autres pour obtenir une expérience naturelle. Si ces derniers restent rares, d'autres chocs sont régulièrement organisés par les gestionnaires mêmes de ces plateformes au travers de l'*A/B testing* (Kohavi et al. 2009). Cette méthode consiste à proposer au même moment à différentes parties des utilisateurs plusieurs versions du site avec chacune des variantes sur une ou plusieurs

[9] De l'anglais *now*, maintenant, et *forcasting*, prévision.

[10] www.thebillionpricesproject.com

fonctionnalités. Les économistes ayant accès à ces données peuvent se servir des résultats des expériences passées pour établir des liens de causalités avec un contrôle proche de celui des expériences de laboratoire, mais avec le réalisme d'une expérience de terrain (Salganik 2017). Par exemple Fradkin (2017b) utilise des variations dans l'algorithme triant les résultats des recherches sur Airbnb pour estimer le coût des frictions sur le volume total de vente. Cependant, la nature de ces plateformes, soumises aux effets d'équilibre, peut rendre caduc le résultat de ces expériences (Blake et Coey 2014).

Avoir accès aux données internes générées par ces *A/B test* n'est pas le seul moyen d'utiliser ces plateformes à des fins d'expérimentation. Les chercheurs peuvent en effet se s'en servir pour réaliser des expériences de terrains à moindre coup. De nombreuses études utilisent les plateformes de travail « à la tâche » comme *Amazon Mechanical Turk* ou *UpWork* et une littérature sur les bonnes pratiques à suivre pour réaliser de telles expériences s'est développée (Paolacci et al. 2010, Masson et Suri 2012). Pour démontrer la pertinence de cette démarche Horton et al. (2011) la suivent pour répliquer les résultats de trois expériences classiques en théorie des jeux et économie comportementale : dilemme du prisonnier, effet d'amorçage, effet de cadrage. Les chercheurs ne sont pas les seuls à expérimenter sur ces plateformes. Einav et al. (2015) utilisent les variations dans les modalités de commercialisation pour un même produit que testent simultanément les vendeurs sur eBay pour mesurer leurs influences sur le prix et la probabilité de vente.

Jusqu'ici nous n'avons évoqué que des études basées sur de plateformes web complètes, c'est-à-dire où la totalité des interactions et notamment la contractualisation s'effectue à travers la plateforme. D'autres plateformes plus partielles n'interviennent que sur une partie du processus, le reste se passant hors ligne ou même ailleurs sur le web. Contrairement aux premières, elles ne sont pas des marchés à part entière, et ne présentent donc pas l'avantage du contrôle et de la visibilité totale sur les mécanismes de recherche et d'appariement. En revanche, elles opèrent bien souvent sur des marchés plus traditionnels aux poids économiques importants. C'est le cas par exemple des plateformes intervenant sur le marché de l'emploi (LinkedIn, Indeed) ou le marché immobilier (Zillow, Se Loger, Meilleurs Agents).

La recherche en immobilier ne s'est pour l'instant que peu saisie du potentiel de ces plateformes pour la recherche empirique. Avant de faire l'analyse de l'apport que l'on peut en espérer dans la section suivante, il est intéressant de noter comment la recherche sur le marché de l'emploi s'est emparée avec un peu d'avance de ces données. Outre la similitude déjà

soulignée quant au type de plateforme opérant sur ces deux marchés, la comparaison se justifie par le cadre théorique commun à la modélisation de ces deux marchés. La littérature immobilière a emprunté de nombreuses idées à celle du travail pour comprendre comment acheteurs et vendeurs se rencontrent et s'accordent sur les prix. On pensera bien évidemment, en premier lieu, au modèle Diamond-Mortensen-Pissarides d'appariement qui s'est avéré tout aussi bien adapté au marché du logement (Wheaton 1990, Krainer 2001, Novy-Marx 2009).

Pour Horton et Tambe (2015), les « Big Data » offrent un « microscope » aux chercheurs s'intéressant au marché de l'emploi (Faberman et Kudlyak 2016 offrent une analyse similaire du phénomène). Ils notent que jusqu'alors, l'économie du travail s'appuyait empiriquement sur des données sur-mesure, essentiellement des enquêtes menées auprès des particuliers par les instituts officiels de statistiques nationaux ou les services publics de l'emploi. Les avantages de ces sources, qualité et profondeur, allant de pair avec le fort coût de leur mise en œuvre, qui entraine une faible fréquence de collecte, une limitation de la portée des questions posées et l'impossibilité d'une utilisation à l'échelle granulaire. Ils voient par contraste dans les données collectées sur les plateformes les atouts des données prêt-à-porter permettant d'éclairer des questions en suspens.

Le premier apport notable est encore une fois la possibilité d'observer des comportements que les chercheurs ne pouvaient jusque-là mesurer. Les actions des demandeurs d'emploi pendant leur recherche font partie de ceux-là. La notion d'intensité de la recherche est déjà présente depuis longtemps dans la littérature tant théorique qu'empirique, à travers des mesures comme le nombre de candidatures envoyées ou le temps consacré à la recherche (Layard et al. 1991, Shimer 2004). Les sites d'annonce d'emploi en ligne qui enregistrent le nombre et l'identité de ceux parmi leurs utilisateurs ayant postulé aux différentes offres, permettent une description bien plus complète du phénomène. En suivant les candidatures déposées sur le snagajob.com, Faberman et Kudlyak (2019) enregistrent une décroissance de l'intensité de la recherche au cours du temps. En revanche, contrairement à la théorie, les personnes subissant de plus longues périodes de chômage postulent à plus d'offres d'emplois par semaine que ceux retrouvant vite du travail. Au-delà de la seule intensité, ces sites offrent aux chercheurs une meilleure vision de la façon dont les gens cherchent un emploi. Marinescu et Rathelot (2018) mesurent sur careerbuilder.com qu'un demandeur d'emploi à 35% de chance en moins de postuler à une offre si celle-ci est localisée à plus de 16 km (10 miles) de chez lui,

mais que le coût sociétal de l'inadéquation géographique reste faible, car la plupart habitent près des postes vacants.

Les personnes en recherche d'emploi ne sont pas seules à passer sous le microscope. Le comportement des entreprises est également disséqué. Azar et al. (2019) observent sur careerbuilder.com que dans les marchés, définis géographiquement et par secteur d'activité, où la demande est concentrée autour d'un ou de quelques employeurs les salaires proposés sont plus bas, toutes choses égales par ailleurs. L'analyse faite par Brencic (2012) des annonces publiées sur le site monster.com montre que les employeurs sont d'autant plus susceptibles d'indiquer un salaire explicitement qu'ils cherchent un travailleur peu qualifié. L'étude des indications explicites et implicites, à travers l'intitulé du poste ou les compétences requises, du niveau de salaire croisée au décompte des candidatures apporte des arguments empiriques aux modèles de recherches dirigées. Banfi et Villena-Roldan (2019) et Marinescu et Wolthoff (2020), utilisant respectivement les données de trabajando.com et de careerbuilder.com, confirment tous les deux que les annonces indiquant un salaire plus élevé attirent plus de candidats.

Les plateformes ne se limitent pas aux seules annonces, mais offrent également une vision des carrières de leurs membres dans le temps grâce à des *curriculum vitae* en ligne. En la matière le réseau professionnel LinkedIn est sûrement l'une des sources les plus riches. Li (2017) s'en sert pour construire une mesure de similarité entre les entreprises basée sur la similarité des parcours et des compétences de leurs employés. Mesure qui explique selon eux mieux la covariance des performances boursières de ces firmes que les regroupements standards. Le suivi des parcours à l'échelle individuelle que permet ce site est par ailleurs exploité dans Ge et al. (2015) qui suivent les passages d'une entreprise à l'autre des inventeurs. Ils confirment que ces derniers sont particulièrement visés par des employeurs intéressés par le capital humain qu'ils représentent. Dans la même ligne, les C.V. collectés sur careerbuilder.com par Tambe et Hitt (2013) révèlent que les investissements des firmes technologiques entrainent des gains de productivité pour les autres entreprises du secteur, via les employées passant d'un employeur à l'autre.

Enfin, ces plateformes offrent l'occasion d'appliquer à l'économie du travail les méthodes empiriques d'expérience aléatoire à grandes échelles et en conditions réelles. Au-delà des cas déjà cités plus hauts d'expériences sur des plateformes de travail à la tâche, comme dans Horton (2017), les sites généralistes permettent d'étendre la démarche à l'ensemble du marché.

En partenariat avec LinkedIn, Gee (2019) a testé l'impact d'une nouvelle fonctionnalité indiquant le nombre de personnes ayant déjà postulé à une offre d'emploi. Elle mesure que l'apport de cette information, en diminuant l'ambiguïté liée à la chance d'obtenir le poste, augmente de 1% les chances de postuler à une offre. Cet effet étant plus fort pour les femmes que pour les hommes.

Par le terrain d'observation et d'expérimentation qu'elles offrent aux économistes, les plateformes Internet apparaissent comme une opportunité sans précédent pour l'étude économétrique. Les besoins de ces plateformes de réguler leurs propres places de marché, pour optimiser les revenus qu'elles en tirent, sont une occasion pour des chercheurs de nouer des partenariats avec ces entreprises et de développer une littérature riche sur un secteur d'importance croissante. Ils peuvent également avec ou sans le concours des plateformes utiliser ce qui s'y passe pour traiter de questions empiriques laissées jusqu'ici en suspend faute de données. Les plateformes totales, où toutes les interactions entre acheteurs et vendeurs ont lieu, ont un intérêt évident, mais les plateformes partielles, n'intervenant que sur certains aspects du processus, peuvent également apporter à la recherche académique. En la matière, les utilisations faites en économie du travail peuvent servir d'exemple à d'autres champs de recherche et en particulier à la recherche en immobilier.

IV. Quel apport pour l'étude du marché immobilier résidentiel

Le marché immobilier résidentiel ne fait pas exception. Comme beaucoup d'autres secteurs traditionnels de l'économie, la montée en puissance d'Internet l'a restructuré autour d'acteurs numériques de type plateforme. Ce marché a la particularité de confronter essentiellement des amateurs. Acheteurs et vendeurs ne sont pour la plupart impliqués que dans une poignée de transactions immobilières, séparées de plusieurs années les unes des autres, au cours de leurs vies. Cette spécificité a permis l'émergence d'une industrie de l'intermédiation qui organise la rencontre entre ces agents granulaires, en tout premier lieu à travers les intermédiaires physiques que sont les agents immobiliers. Si l'importance des montants en jeu et certaines obligations réglementaires empêchent une digitalisation complète du processus de vente, des acteurs numériques se sont imposés comme intermédiaires à différents moments de la transaction, se positionnant entre acheteurs et vendeurs ou entre particuliers et agents immobiliers.

Les premières plateformes immobilières ont d'abord transposé en ligne le modèle de petites annonces agrégeant graduellement l'ensemble de l'offre disponible pour attirer toute la demande active. Certains de ces sites permettent uniquement aux professionnels de diffuser des annonces, d'autres prônant les ventes entre particuliers, les derniers ne faisant pas de distinction. En France, on pourra citer respectivement les sites Se Loger (fondée en 1992 sur Minitel), PAP (1998 pour la version en ligne, dérivé direct d'une publication papier) et Le Bon Coin (2006). Une deuxième vague s'est positionnée entre les différents acteurs du marché en agrégeant de l'information sur le marché. En répondant à des questions élémentaires, mais fondamentales sur les prix, leurs tendances, les informations sur les dernières ventes et sur la nature du stock de biens, elles sont venues compléter, voire concurrencer, les sites d'annonces. L'entreprise pionnière en la matière est l'américain Zillow (créé en 2006), vite imité au Royaume-Uni par Zoopla (2007) et en France par Meilleurs Agents (2008). Enfin, dans ces développements les plus récents, le marché de l'immobilier en ligne s'est attaché à redescendre la chaîne de valeur pour prendre une part plus active dans la transaction. L'exemple le plus frappant étant les « iBuyers », ces entreprises achetant les propriétés de particuliers après une estimation en ligne, pour ensuite les revendre soit via les réseaux classiques (cas d'Homeloop en France), soit par elles-mêmes avec un processus de visite également fortement digitalisé (Opendoor aux États-Unis).

Même si le sujet n'est pas sans intérêt et que les études sur le sujet se cantonnent à la première vague de digitalisation (Ford et al. 2005), il ne s'agira pas ici de faire l'étude de l'impact de ces transformations sur le marché. Cette numérisation croissante du processus de vente et d'achat nécessite et génère une production toujours plus importante de données le décrivant. C'est particulièrement vrai pour les entreprises de deuxième et de troisième génération décrite plus haut, dont les modèles d'affaires reposent sur la maitrise de ces informations. La suite de cette section propose donc une réflexion sur l'apport avéré et potentiel de ce nouveau gisement de donnée pour l'étude des mécanismes de rencontre et de fixation des prix sur le marché immobilier résidentiel. Notez qu'elle ne saurait non plus constituer une revue de la littérature traitant de la microstructure de ce marché. Pour cela, le lecteur est renvoyé vers Han et Strange (2015) qui en font une description riche et récente. L'enjeu est pour nous de tenter d'identifier quelles problématiques sont susceptibles d'être éclairées par ces informations. En nous appuyant sur la littérature existante exploitant des plateformes opérant dans d'autres secteurs, nous essaierons également de donner des pistes sur la démarche technique à mettre en place pour se faire.

Classiquement, les premières études réutilisant de telles données relèvent du domaine du *nowcasting*. L'utilité de telles informations à l'instant *t* est grande dans un secteur où les canaux classiques de production de statistiques souffrent d'un retard dû à la collecte d'information auprès d'acteurs de taille modeste disséminés dans l'espace. Par exemple, les indices Notaires-Insee (Cailly et al. 2019) qui font référence en France sont calculés à partir des actes authentiques enregistrés et remontés par les notaires sur l'ensemble du territoire. Le temps nécessaire à cette remontée et au traitement statistique engendre un retard d'un trimestre dans la publication des indices par rapport à la date des actes authentiques qui eux-mêmes n'interviennent en moyenne que trois mois après la conclusion des ventes[11]. Le rythme de publication trimestrielle fait que l'information « officielle » disponible aura entre 6 et 9 mois de retard. Le constat est le même avec les indices S&P CoreLogic Case-Shiller, aux États-Unis, qui gardent deux à trois mois de retard sur le marché[12]. En période de forte volatilité des prix, notamment lors des changements de cycle, cela rend ces informations, par ailleurs très adaptées au suivi longitudinal du coût du logement, peu pertinentes pour les agents économiques devant prendre des décisions dans l'instant (particuliers, organismes prêteurs, pouvoirs publics).

C'est d'ailleurs face à ce manque que sont construites les plateformes de deuxième génération qui diffusent des informations d'évolution de prix se voulant à jour (voir par exemple les Zillow Home Value Index[13] ou les Indices des Prix Immobiliers de Meilleurs Agents[14]). Elles profitent pour cela des informations des transactions remontées par les agents immobiliers clients, mais aussi des annonces de mise en vente publiées chaque jour sur leurs sites. Ce que ces approches perdent en représentativité, problème qui va par ailleurs en s'amenuisant avec le taux de pénétration grandissant de ces acteurs, elles le gagnent en réactivité, mais également en précision de description des biens, ce qui s'avère particulièrement utile dans l'approche hédonique (Loberto et al. 2020). Au-delà du suivi des prix de transactions, les données de ces sites recèlent des informations sur d'autres dimensions du marché. Chapelle et Eyméoud (2017) utilise les données d'annonces provenant de deux sites pour étudier l'offre de location dans les

[11] Par exemple, la mise à jour du 4 avril 2020 concerne les actes authentiques du quatrième trimestre 2019, soit des ventes conclues au troisième trimestre 2019.

[12] us.spindices.com/documents/methodologies/methodology-sp-corelogic-cs-home-price-indices.pdf

[13] www.zillow.com/research/data/

[14] www.meilleursagents.com/ipi/

plus grandes aires urbaines de France. On notera que les pouvoirs publics ont bien pris conscience de cette opportunité. Le projet de recherche « Un loyer de référence pour chaque commune française », lancé par le Ministère de la Cohésion des territoires et des Relations avec les Collectivités, qui s'appuie sur les données de plusieurs acteurs du numérique[15], en est la preuve. Nombreux sont les indicateurs qu'il est ainsi possible de produire et suivre dans le temps : délai de vente, décote entre prix de commercialisation et prix de mise en vente, stock de bien à vendre, indicateur de tension. En la matière le site américain Zillow propose l'offre la plus complète avec une échelle géographique qui va de l'ensemble du pays jusqu'au quartier. La mise à disposition aux chercheurs de l'historique de ces données laisserait entrevoir la possibilité d'une meilleure compréhension, par exemple, de la réaction des marchés immobiliers à des chocs exogènes (crise financière, épidémie, catastrophes naturelles, évolutions réglementaires).

Au-delà du *nowcasting*, l'analyse du comportement des utilisateurs qui préparent leur projet sur ces sites permet d'établir des prévisions sur les évolutions du marché. Dans Wu et Deng (2015), ce sont les recherches concernant l'immobilier faites sur Google qui sont utilisées pour prédire les prix et les volumes du trimestre suivant à l'échelle de chaque état américain. Prédictions qui s'avèrent meilleures que celles n'utilisant que les seules informations de marché passées ainsi que les prédictions d'experts de la *National Association of Realtors*. L'utilisation de données sur des sites dédiés, plutôt qu'un moteur de recherche généraliste, permet d'affiner ces prédictions. Van Dijk et Francke (2018) construisent un indicateur de tension du marché immobilier basé sur le taux de clic d'acheteurs potentiels sur les annonces du site funda.com, pour chaque municipalité aux Pays-Bas. Ils montrent que cet indicateur est prédictif des prix et des durées des ventes du trimestre suivant. Loberto et al. (2020) reprennent la même méthode grâce aux données du site immobiliare.it et affinent la granularité de la prédiction, en l'appliquant à l'échelle de quartiers italiens. Si ces articles n'améliorent pas par eux même notre compréhension du marché immobilier, ils apportent une preuve décisive quant à la capacité de mesurer sur ces plateformes une dimension jusque-là inobservable : le niveau de demande.

[15] www.cohesion-territoires.gouv.fr/connaitre-les-loyers-partout-en-france-le-ministere-lance-un-partenariat-inedit

Mesure décisive pour amener vers la vérification empirique la théorie dominante en matière de modélisation du marché immobilier : le *matching*.

Le modèle d'appariement DMP, du nom de ses précurseurs, Peter Diamond, Dale Mortensen et Christopher Pissarides, s'est développé dès les années 1980 pour expliquer l'existence simultanée de demandeurs d'emploi et de postes vacants sur un même marché (Diamond 1982a,b; Mortensen 1982a,b; and Pissarides 1984, 1985). Employeurs et employés ne peuvent se rencontrer instantanément, car chacun fait face dans sa recherche d'une contrepartie à des frictions : couts de recherche, problèmes de coordination, connaissance limitée des possibilités existantes, incertitude quant à l'utilité générée par l'échange. Une situation analogue à celles des acheteurs et des vendeurs sur le marché immobilier résidentiel. C'est donc naturellement que ce cadre théorique a été transporté à l'économie du logement, d'abord par Wheaton (1990) rapidement suivi par d'autres (Krainer 2001, Novy-Marx 2009, Albrecht et al. 2007). Cependant, à l'inverse du marché de l'emploi où les vérifications empiriques sont nombreuses (voir Petrongolo et Pissaridies 2001, pour une revue de littérature), ces modèles n'ont que peu été confrontés à des données. Particulièrement à cause du manque de mesure quantitative de la demande.

On peut cependant relever deux exceptions récentes. D'abord, Genesove et Han (2012) qui approximent un ratio entre les nombres d'acheteurs et de vendeurs à travers la différence des évolutions de leurs temps de recherches médians respectifs. La seconde, Piazzesi et al. (2020), emploie des données issues du site d'annonce américain trulia.com. Grâce aux nombres d'inscrits aux alertes email prévenant de la publication d'une nouvelle annonce, comme proxy du nombre d'acheteurs actifs, les auteurs tracent le premier équivalent à la courbe de *Beveridge* dans un contexte immobilier. Ce résultat ajouté à ceux de Van Dijk et Francke (2018) et Loberto et al. (2018) laissent penser que les données d'une plateforme immobilière peuvent permettre d'estimer ce qui fait le cœur des modèles d'appariement aléatoire : la fonction de *matching*.

Cette fonction, qui relie les nombres d'acheteurs et de vendeurs à celui des ventes, a pour but de synthétiser l'ensemble des frictions. Pour ce qui est du nombre d'embauches, les estimations permises dès les années 1980 par les statistiques du taux chômage ont permis d'établir un consensus autour d'une forme de Cobb-Douglass de degré 1 avec rendements d'échelle constants (Petrongolo et Pissaridies 2001). C'est donc assez naturellement que les auteurs transposant la théorie au contexte immobilier ont conservé cette spécification, sans qu'il existe de justification empirique pour l'appuyer. La mesure de l'offre et de la demande sur une

plateforme immobilière, si elle s'avère suffisamment représentative de l'ensemble du marché, peut combler ce manque.

À l'inverse des modèles précédents, qui agrègent la totalité des frictions pesant sur acheteurs et vendeurs dans une seule et même fonction, un large pan de la littérature se base sur les modèles de recherche introduits par Stigler (1961) pour analyser séparément les problèmes qui se posent à l'acheteur (Courant 1978, Turnbull et Sirmans 1993, entre autres) et au vendeur (Haurin 1988, Salant 1991 pour les premiers). Dans cette approche, le processus d'achat ou de vente consiste en une séquence de rencontres consécutives. Pour un vendeur, il doit à chaque fois accepter ou rejeter l'offre d'achat qui lui est faite, sans certitude quant aux offres issues des rencontres futures. Le problème se réduisant dans sa forme la plus simple à une condition d'arrêt portant sur le prix minimum de vente qu'il est prêt à accepter compte tenu des contraintes qui lui sont propres (coûts de recherche, préférences temporelles) et de ses croyances (distributions des offres futures, rythme espéré de rencontre avec les acheteurs). Le problème de l'acheteur étant symétrique, avec une suite de rencontres avec des vendeurs demandant des prix différents pour leurs biens.

Profitant des informations qu'apportent les annonces publiées par les vendeurs, les études empiriques se sont largement concentrées sur le processus de recherches de ces derniers et notamment sur le lien entre prix de vente et durée de commercialisation. Conformément au modèle théorique, les preuves d'une corrélation positive entre prix et temps de vente sont nombreuses (Cubbin 1974, Miller 1978, Yavas et Yang 1995, Anglin et al. 2003). Différentes raisons sont identifiées comme sources de cette hétérogénéité entre les préférences des vendeurs : biens atypiques (Haurin 1988), motivation (Zuehlke 1987, Glower et al. 1998), contrainte financière (Genesove and Mayer 1997). Cependant, le rôle stratégique du prix de mise en vente, identifié par Yavas et Yang (1995), Anglin et al. (2003) ou Han et Strange (2016), est mal prise en compte par ces premiers modèles théoriques. En effet, si *ex ante* un vendeur a intérêt à afficher un prix d'annonce bas, pour attirer un maximum d'acheteurs, il préfèrera *ex post* un prix de commercialisation élevé pour éviter les offres trop basses. Ce type de situation est mieux prise en compte dans les modèles de recherche dirigée comme ceux de Carrillo (2012) ou Merlo et al. (2013).

Le jeu de données documentant la vente de 1 000 maisons à Londres utilisé dans Merlo et Ortalo-Magné (2004) et Merlo et al. (2013) leur permet une modélisation bien plus complète des problèmes stratégiques se posant à un vendeur. Outre les prix de mis en vente, les prix

finaux et les durées de commercialisation, y sont consignés les différents changements de prix d'annonce, les dates des visites et les dates et montants des offres reçues. Une description aussi minutieuse du processus de vente fait exception aujourd'hui dans le monde analogique. L'émergence des plateformes immobilières de 3e génération, qui prennent part activement dans la vente, nous laisse envisager que tels jeux de données pourraient se multiplier. L'approche dite *data driven* étant si cher aux entreprises du numérique, on peut facilement imaginer que ces nouveaux acteurs collecteront systématiquement ce type de données pour piloter leurs ventes et noueront des partenariats avec des économistes pour les exploiter.

Suivant la même logique, ces vendeurs d'un nouveau genre semblent plus susceptibles que les vendeurs traditionnels d'exploiter la possibilité d'expérimentation à moindre coup offerte par une commercialisation via une plateforme numérique. C'est là une démarche plus en phase avec la culture de ces entreprises numériques. De plus, contrairement aux agents immobiliers ou aux particuliers, elles auraient les moyens d'analyser les résultats de ces expériences, à travers leurs systèmes de collecte de données, et d'en tirer réellement profit, compte tenu du flux important de biens qu'ils traitent. À l'instar des commerçants expérimentant différentes options de commercialisation sur eBay (Einav et al. 2015), ils pourraient aléatoirement exposer des annonces différentes pour un même bien à certains visiteurs de leur site. Si, pour des raisons juridiques, il semble compliqué d'inclure le prix dans de telles expérimentations, une modulation de la quantité et de la nature de l'information relative au bien peut être riche d'enseignements. Carrillo (2008) a déjà montré qu'une description visuelle abondante (photographie, visite virtuelle) a un impact positif sur la probabilité de vendre et raccourci la durée de commercialisation. Une expérience aléatoire permettrait d'aller plus loin en distinguant l'impact avant la visite, sur le taux de contact après visualisation, et après la visite, sur la probabilité d'une offre et son montant.

Cependant, c'est dans notre compréhension du processus d'achat que l'exploitation des données numériques a le plus grand potentiel immédiat. En effet, avant l'avènement d'Internet comme principal canal de diffusion des annonces, et donc d'outil de prospection pour les acheteurs, ces derniers ne laissaient aucune trace de leur recherche. Ainsi les études appliquant le cadre de la théorie de la recherche au problème de l'acheteur (Turnbull et Sirmans 1993, Lambson et al. 2004, et Ihlandfeldt et Mayok 2012) identifiaient principalement, dans leurs volets empiriques des facteurs extérieurs au processus d'achat comme la distance au lieu de résidence précédent, les effets d'ancrage ou le degré expérience en tant qu'acheteur. Anglin

(1997) et Wilhelmsson (2008) sont à notre connaissance les seules études basées sur des éléments concernant le déroulement de la recherche elle-même. Utilisant des enquêtes menées auprès d'acheteurs, le premier souligne l'importance de l'information pour expliquer l'intensité de la recherche et le second établit une corrélation négative entre le nombre de visites et le prix d'achat. L'usage massif par les acquéreurs des outils numériques peut combler ce vide de données d'observation. L'accès par des chercheurs à de tels éléments et leur exploitation restent à ce jour rares. Elles ont cependant prouvé leur intérêt pour décrire la façon dont les acheteurs organisent leur recherche. Rae et Sener (2016) et Piazzesi et al. (2020) se servent notamment de leur activité sur les sites d'annonces pour documenter la segmentation des marchés immobiliers de Londres et San Francisco.

Les plateformes immobilières de seconde génération ne font pas que mettre en rapport acheteurs et vendeurs. Certaines se sont également positionnées entre les particuliers ayant un projet et les agents immobiliers leur proposant leurs services. L'analyse de leurs données permettrait de comprendre sur quels éléments ces particuliers basent leurs choix dans un marché où la différenciation produit est limitée (Han et Strange 2015) et la compétition sur les tarifs faible (Schnare and Kulick 2009). En effet, conformément aux pratiques habituelles du web, ces sites incitent les agents à mettre en avant leurs performances passées pour rassurer les consommateurs à travers des avis clients ou le catalogue de leurs ventes récentes. Des acteurs comme getagent.co.uk ou homelight.com proposent même des métriques comme le temps de vente ou le taux de décote moyen, bien que les modèles de recherches précédemment discutés indiquent que ces quantités dépendent de nombreux autres facteurs sur lequel l'agent n'a pas de prise. La littérature sur la réputation des fournisseurs sur les plateformes (voir section III) laisse penser que le choix des particuliers est dirigé par de tels critères. L'identification des plus influents et la magnitude de leurs impacts restent un terrain de recherche vierge. Encore une fois, l'implémentation d'expériences aléatoires où les informations présentées à l'utilisateur seraient modulées est une piste prometteuse. D'autant plus que les plateformes, ayant un intérêt direct aux résultats de telles expériences pour établir un algorithme de recommandation par exemple, sont susceptibles d'accepter d'en être le terrain.

Par leur rôle d'intermédiaire entre particuliers vendeurs et agents immobiliers, elles sont également susceptibles de collecter des informations touchant au problème d'alignement d'intérêt entre principal et agent. Une large littérature documente cette situation. On pensera par exemple aux travaux de Rutherford et al. (2005) ou Levitt et Syverson (2008) qui comparent

les performances d'agents vendant leur propre bien avec leurs résultats lorsqu'il s'agit de ceux des clients. Selon le degré de contrôle qu'ont les plateformes sur les échanges entre les deux parties et les informations collectées avant leur rencontre (prix espéré par le propriétaire, motivation de la vente) et à la fin de leur coopération (aboutissement de la vente, prix de mise en vente, prix final, durée), elles peuvent apporter de nouveaux éclairages sur la question. Par exemple, Cherbonnier et Lévêque (2019) exploitent les données recueillies par Meilleurs Agents pour étudier l'impact de la concurrence et des opportunités de collusion entre agents immobiliers sur leur tendance à abuser de leur avantage informationnel. Ils mesurent que les situations où la compétition entre agents est plus forte, correspondent à une augmentation de leurs estimations et ainsi que des prix de commercialisation et de vente des biens de leurs clients.

Ceci nous emmène naturellement vers le dernier aspect de l'étude du fonctionnement du marché immobilier qui peut tirer parti de l'existence de ces plateformes : l'assimilation de l'information par les agents économiques. Loin de n'être que des instruments de mesure de l'utilisation de la rente informationnelle par les intermédiaires, ces sites, en particulier ceux de la deuxième vague, se sont construits avec l'objectif de l'abolir. L'inefficience du marché immobilier résidentiel est bien documentée (Case et Shiller 1988, Gatzlaff et Tirtiroglu 1995) et la découverte du prix a longtemps était dominé par les estimations d'experts (Geltner et al. 2003). L'ouverture et la diffusion à grande échelle d'information de prix ainsi que la généralisation des outils d'estimation en ligne a très certainement modifié son fonctionnement. Ben-Shahar et Golan (2019) mesurent que la mise en open data des bases fiscales de transactions immobilières a entrainé une diminution de la variance des prix, ajustés de la qualité, en Israël. À notre connaissance, aucune étude ne s'est penchée à ce jour sur l'impact de la création des outils de traitement de données et de production d'information par les plateformes immobilières. Pourtant leurs déploiements au cours du temps et de façon asymétrique sur des territoires comparables ont généré des situations de quasi-expériences naturelles probablement exploitables pour un travail de recherche.

Par ailleurs comme sources d'informations qui enregistrent des comportements des acteurs du marché, ces sites permettent d'en mesurer l'impact et d'analyser les mécanismes d'assimilation par ces derniers. Une courte littérature sur la diffusion de l'information immobilière à travers les plateformes généralistes commence à se constituer. Au niveau agrégé, Wu et Deng (2015) suivent cette diffusion à travers les flux de requêtes liées au marché du

logement sur le moteur de recherche de Google. Ils montrent que la cascade d'information descendant des plus grandes villes du pays ou des capitales régionales vers le reste des territoires entraine les prix de ces derniers dans le sillage des évolutions des marchés des premières. À une échelle individuelle, Bailey et al. (2018) mesure que les personnes dont les « amis » Facebook vivant loin connaissent un boom immobilier, ont un comportement enthousiaste vis-à-vis de leur propre marché immobilier local. Ils sont plus susceptibles de devenir propriétaires, d'acheter des biens plus grands et de payer plus cher pour un bien donné. Ces exemples illustrent parfaitement l'importance de l'information reçue (ou perçue) par les particuliers dans les mécanismes de fixation des prix. C'est d'autant plus vrai dans un marché où les propriétaires se trompent largement et systématiquement sur la valeur de leur actif. D'après Goodman et Ittner (1992), leurs prédictions du montant d'une vente future accusent une erreur absolue moyenne de 14% et un biais de surestimation de +6%. En identifiant précisément les informations de prix qu'un acheteur ou un vendeur a lues sur une des plateformes, on peut espérer en mesurer les conséquences sur ses croyances et donc sur ses actions.

Les champs de recherche ouverts par les données issues des plateformes immobilières sur Internet sont nombreux. Au-delà des premières études de *nowcasting*, l'apparition de traces mesurables des acheteurs devrait permettre des avancées empiriques tant sur les modèles d'appariements que sur ceux de recherche séquentielle. La description minutieuse des ventes et les possibilités d'expériences aléatoires sont de nature à améliorer notre compréhension du problème du vendeur. En tant qu'intermédiaires entre particuliers et agents immobiliers, elles apportent des éléments sur le choix d'un agent plutôt qu'un autre et sur le comportement de ces derniers. Enfin, les mesures des impacts des informations de marché qui y sont diffusées amélioreraient nos connaissances du mécanisme de découverte des prix. Toutes ces plateformes ne sont pas identiques et chacune a ses propres avantages et inconvénients du point de vue du chercheur voulant exploiter les données qu'elles génèrent. Nous avons tenté d'établir différentes pistes qui ne sont pas bien sûr pas toutes suivies dans le travail présenté ici. La section qui suit présente les spécificités du site à travers lequel nous avons pu observer le comportement des acteurs du marché immobilier français et les trois études que cette démarche a permises.

V. Présentation de la thèse

Cette thèse entend donc utiliser les données issues d'une plateforme immobilière française de deuxième génération pour étudier comment acheteurs et vendeurs se rencontrent et s'accordent sur les prix dans ce marché. Les trois études présentées ici s'intéressent chacune à un aspect différent du problème : mécanisme d'appariement pour la première, problème de l'acheteur pour la seconde, assimilation de l'information de prix par les particuliers dans la dernière. Avant de présenter la démarche suivie dans chacune d'entre elles, une rapide présentation du terrain d'observation est nécessaire. Comme expliqué plus haut, les données tirées de chaque plateforme présentent des avantages et des défauts qui leur sont propres quant à leur utilisation à des fins de recherche, et chaque site représente des opportunités d'observations ou d'expérimentation différentes. Le lecteur ne saurait donc pleinement apprécier la démarche de ce travail sans cette description préalable[16].

Présentation de la Plateforme

Meilleurs Agents est une plateforme immobilière fondée en 2008 qui se décrit elle-même comme le « n°1 de l'estimation immobilière en ligne ». Elle est un exemple type de ce que nous avons appelé dans la section précédente les plateformes immobilières de deuxième génération. Son modèle d'affaires est fondé sur sa capacité à capter une large audience de particuliers ayant un projet immobilier par la diffusion gratuite d'informations sur le marché immobilier, et de monétiser cette audience en la redirigeant vers des agents immobiliers. Bien que ses outils soient utilisés par des particuliers ayant des projets divers, c'est la mise en relation avec des vendeurs qui est promue et facturée aux agents. Elle se distingue ainsi des plateformes de première génération, les sites d'annonces, qui leur offrent une visibilité auprès des acheteurs. Selon son degré d'implication dans cette mise en relation, l'entreprise se rémunère soit en tant qu'apporteur d'affaires, soit comme un média offrant une simple visibilité à l'agent.

Les informations disponibles sur les sites sont de quatre ordres. Premièrement, dans un exercice qui relève du *nowcasting*, une bibliothèque de 800 indices suivant l'évolution temporelle des prix à travers la France est mise à jour chaque mois. Ces indices sont diffusés

[16] Il est par ailleurs invité à examiner le site par lui-même, accessible à l'adresse : www.meilleursagents.com

directement sur le site, par mail aux particuliers abonnés et par voie de presse au travers de nombreux partenariats de relations publiques entretenus par Meilleurs Agents. Le plus emblématique étant la publication mensuelle des *Indices des Prix Immobiliers Meilleurs Agents – Les Échos*, avec le quotidien économique de référence en France. Compte tenu des différences dans les données effectivement disponibles, l'évolution de chaque marché n'est pas suivie avec la même précision. Si en Île-de-France 95% des appartements et 45% des maisons sont couverts par un indice propre à la ville dans laquelle ils sont situés[17], dans le reste des régions métropolitaines seuls les prix des logements collectifs des 50 plus grandes villes sont suivis à l'échelle communale, le reste l'étant au niveau des départements. Le second niveau d'information de prix est de nature géographique. Meilleurs Agents calcule et affiche sur son site des cartes de prix pour un bien standardisé sur l'ensemble de l'hexagone. La granularité de ses prix est encore une fois variable. Pour la quasi-totalité des communes françaises (hors DOM-TOM), un prix standardisé est calculé au moins au niveau municipal. Dans l'ensemble de l'Île-de-France et dans 60 des plus grandes villes françaises et leurs proches banlieues, un prix différent est produit pour chaque adresse. Une centaine de communes de tailles intermédiaires bénéficient de carte au niveau des IRIS[18]. Le troisième outil mis à disposition des particuliers est le produit phare de l'entreprise. Il s'agit d'un estimateur permettant d'évaluer la valeur d'une propriété située en France métropolitaine. C'est à travers l'utilisation de cet outil que nous avons observé le comportement des particuliers. Une description plus détaillée en est donc faite plus bas. Notons simplement ici que, si l'outil est en apparence le même partout, ses performances sont variables selon la localisation du bien, dans la mesure où il est en partie basé sur le modèle de prix géographique produisant les cartes de prix. Le seul chiffre communiqué par l'entreprise sur la précision du moteur d'estimation est une erreur médiane absolue de 6% dans l'estimation des appartements parisiens. Enfin le dernier type d'information présenté sur le site concerne les agents immobiliers. Ces derniers, en s'inscrivant sur le site, pour certains gratuitement, sont incités à renseigner les ventes qu'ils ont réalisées dans les vingt-quatre derniers mois, à maintenir à jour la liste des annonces des biens qu'ils ont à vendre, ainsi que de solliciter leurs clients pour qu'ils donnent leur avis sur la qualité des services fournis. Ces informations sont présentées aux particuliers à différents endroits du site

[17] Source : Meilleurs Agents

[18] Ilots Regroupés pour l'Information Statistique, soit les microquartiers définis par l'INSEE

comme des indicateurs de performances censés les aider dans le choix du professionnel avec qui faire affaire.

Le choix d'observer le comportement des particuliers sur le marché immobilier à travers l'outil d'estimation est motivé par deux éléments. D'abord, parce que l'estimation en ligne est l'expertise de l'entreprise. Elle est d'ailleurs notoirement reconnue comme leader du secteur en France, comme tant à le montrer la figure 3 ci-dessous présentant les parts d'audiences relatives à l'estimation en ligne d'après Médiamétrie en mars 2019. Un particulier cherchant à se renseigner sur les prix sur Internet a donc de grandes chances d'utiliser ce service. Or comme l'indiquait un sondage Opinion Way de 2014 cité par Lefebvre (2015), il s'agit là d'une pratique répandue chez les Français : 79% d'entre eux estiment qu'Internet est utile dans la collecte d'information relative à l'immobilier et 87% de ceux s'étant rendu sur un site du secteur l'on fait pour se renseigner sur les prix. Si ces éléments sont certes rassurants quant à notre démarche générale et la part non négligeable d'utilisateurs observables relativement au nombre de particuliers actifs sur le marché immobilier, ils n'assurent en rien la représentativité des données collectées. Ce point sera traité spécifiquement dans chacune des études.

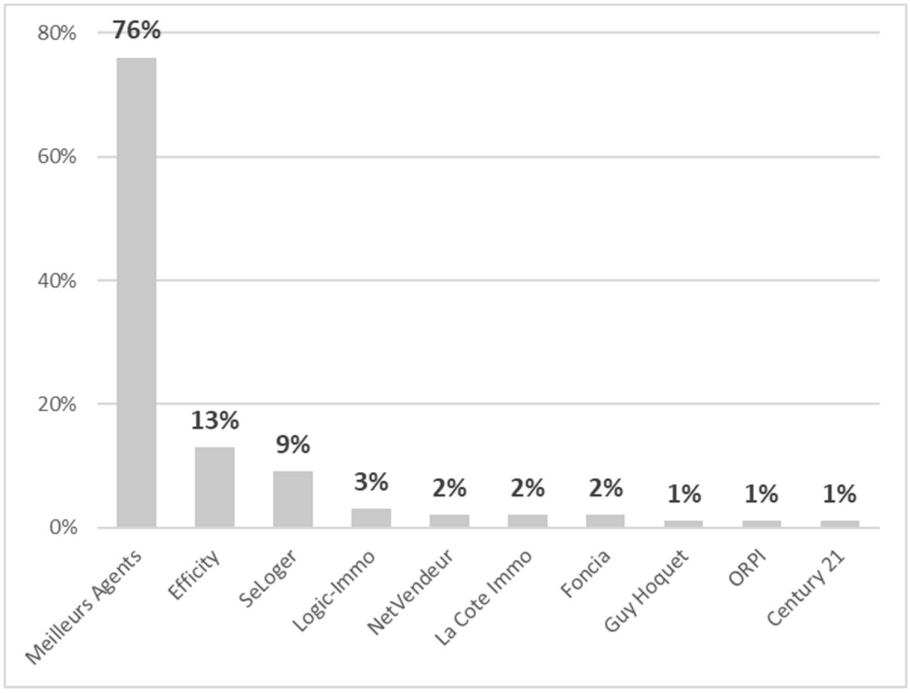

Figure 1-3 : Part d'audience en mars 2019 parmi les internautes ayant visité un site d'estimation immobilière ou de prix - Source : Médiamétrie pour Meilleurs Agents

Deuxièmement, l'outil d'estimation présente un avantage technique par rapport aux autres portions du site qui facilite le suivi de l'activité des utilisateurs. Pour réaliser une

estimation, les utilisateurs doivent s'identifier avec un email et un mot de passe ce qui limite par exemple le risque de compter plusieurs fois la même personne et permet de suivre les actions d'une même personne au cours du temps. Par ailleurs, une fois connecté, le parcours pour obtenir le résultat est plus long que la consultation d'un prix à une adresse ou dans une ville donnée. Tout au long de ce parcours, l'utilisateur va renseigner plusieurs informations permettant de mieux caractériser son projet immobilier. Les éléments qu'il doit indiquer sont :

- L'adresse du bien
- Une description de ses caractéristiques hédoniques
- S'il est propriétaire du bien
    - Si oui :
        - S'il s'agit de sa résidence principale, secondaire ou d'un investissement
        - S'il envisage de le vendre et, le cas échéant à quelle échéance
        - Le mode commercialisation dans le cas d'une vente en cours
        - S'il accepte d'être recontacté par une agence ou par Meilleurs Agents
        - (Optionnel) La date et le montant d'achat
    - Si non :
        - S'il cherche à acheter un bien ou simplement s'informer
        - Si acheteur : s'il est intéressé par un logement neuf ou ancien
        - Si acheteur : l'avancement de ses recherches
        - Si acheteur : s'il vend un bien pour financer cet achat
        - Si acheteur : s'il souhaite faire un emprunt pour financer cet achat
- Une fois le résultat affiché, il a la possibilité de donner son avis sur l'estimation :
    - La trouve-t-il : beaucoup trop basse, trop basse, juste, trop haute ou beaucoup trop haute ?
    - Son estimation personnelle du bien

La description précise, à grande échelle et au cours du temps de millions de projets immobiliers représente une opportunité certaine du point de vue du chercheur. D'autant plus qu'elle est accompagnée d'une estimation, calculée par Meilleurs Agents, de la valeur du bien concerné. En revanche, elle n'évite pas les écueils habituels des données prêt-à-porter. Les utilisateurs remplissant le formulaire sans aucune supervision, on peut redouter des erreurs de saisie, des incompréhensions face à certaines questions ou une volonté de cacher certains éléments à un site commercial. La plateforme évoluant au cours du temps, le visuel du questionnaire, les questions et les réponses possibles ont changé, forçant à restreindre les études sur des périodes permettant des mesures stables. Enfin, l'incertitude quant à la qualité de l'estimation et la variance temporelle (évolution du moteur d'estimation) et spatiale (contexte informationnel différent selon les zones) de cette qualité pousse à la prudence quant à son utilisation et surtout peut engendrer des comportements différents parmi les utilisateurs. Une

grande partie du travail réalisé au cours de cette thèse a consisté à tirer partie de ces avantages tout en maitrisant du mieux possible les risques liés à ces inconvénients.

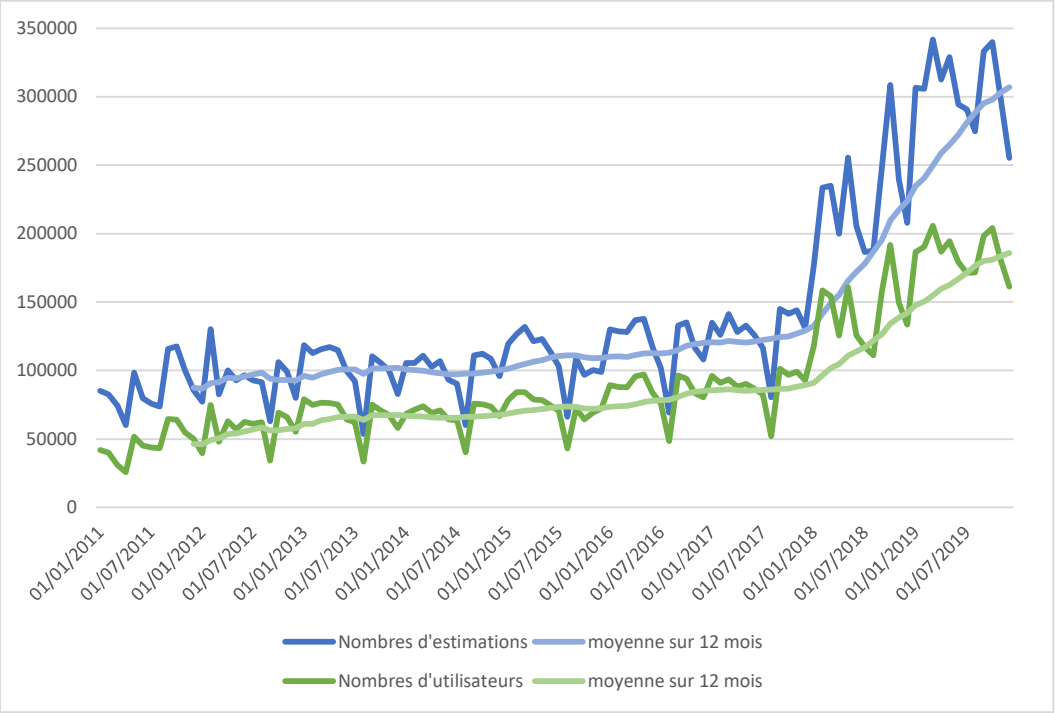

Figure 1-4 : : Nombres d'utilisateurs et d'estimations réalisées par mois sur Meilleurs Agents de 2011 à 2019

La figure 4, ci-dessus, présente l'évolution de l'audience de l'outil d'estimation de Meilleurs Agents entre début 2011 et fin 2019. Elle est symptomatique du caractère dérivant des données issues du monde numérique. Après la fin d'une phase de croissance plus chaotique, les nombres d'utilisateurs de l'outil et d'estimations réalisées par ces derniers connaissent une phase de stabilité de 2013 à 2017. Pendant cette période, les variations mensuelles suivent une régularité s'expliquant par la saisonnalité du marché immobilier : chutes au mois d'aout et décembre, pics pendant les premiers et deuxièmes trimestres. L'année 2018 rompt avec ce régime de variations régulières, avec le début d'une forte croissance dans l'utilisation de l'outil d'estimation qui double entre fin 2017 et fin 2019. Cette augmentation est liée au lancement d'une campagne de publicité à la télévision par Meilleurs Agents à partir du premier trimestre 2018. Remarquez que les évolutions des courbes se font en parallèle. C'est bien l'accroissement du nombre d'utilisateurs, dû à une notoriété croissante du site, qui entraine le doublement du nombre d'estimations et pas un changement brutal dans l'usage qu'ils en font. Le nombre moyen estimations par utilisateur reste environ à 1,5 avant comme après cette phase d'expansion.

Au-delà de leur nombre d'estimations, le profil des utilisateurs change également au cours du temps. Comme le montre le tableau 1 ci-dessous, dès 2012, la proportion d'estimation faite par des propriétaires se stabilise autour de 70%, ce qui s'explique par le modèle d'affaire du site. La part des vendeurs déclarés, elle, augmente de 2011 à 2014 puis diminue jusqu'en 2019 et celle des acheteurs chute brutalement de 25% à 10% dans les 4 premières années puis se stabilise autour de 10% des estimations. Ces changements sont en partie dus à l'élargissement de l'audience, avec de plus en plus de gens n'ayant pas de projet immobilier défini utilisant le service.

	2011	2012	2013	2014	2015	2016	2017	2018	2019
Propriétaire	62%	68%	71%	72%	69%	71%	70%	72%	71%
Vendeur	24%	33%	37%	38%	36%	35%	33%	32%	29%
Acheteur	25%	16%	13%	10%	11%	10%	10%	10%	9%
Île-de-France	72%	59%	52%	48%	44%	43%	44%	39%	38%
Paris	32%	24%	20%	18%	16%	16%	16%	13%	12%
Top 10 province	4%	6%	7%	8%	8%	8%	9%	9%	10%
Appartement	71%	63%	60%	57%	56%	56%	56%	53%	52%
Maison	29%	37%	40%	43%	44%	44%	44%	47%	48%
Prix médian	313 000 €	290 500 €	278 300 €	267 100 €	255 100 €	256 800 €	271 200 €	270 800 €	277 700 €
Prix au m² médian	4 770 €	3 687 €	3 398 €	3 153 €	2 948 €	2 949 €	3 117 €	3 060 €	3 172 €

Table 1-1 : Caractéristiques principales des estimations réalisées sur Meilleurs Agents de 2011 à 2019

Plus critique en ce qui concerne la représentativité des données à des fins de recherches, la répartition spatiale des biens estimés affiche une surreprésentation des logements situés en Île-de-France et plus particulièrement à Paris. Ce biais, s'expliquant par le déploiement du service de Meilleurs Agents uniquement dans cette région dans un premier temps, diminue en deux temps de 2011 à 2014 puis à partir de 2018, mais reste cependant fort. En effet, l'immobilier parisien, qui représente 4% du total des biens de France métropolitaine selon l'INSEE, concentre toujours 12% des estimations. Pour la région capitale dans son ensemble, c'est 38% des estimations faites sur Meilleurs Agents pour 17% des logements totaux. Ce fort tropisme parisien a des effets sur les autres caractéristiques des biens estimés, qui sont plus souvent des appartements que des maisons, alors que c'est l'inverse dans le parc Français (56% d'habitat individuel en 2019 selon l'INSEE) et qui sont également plus cher (prix médian en 2020 : 2 100€/m² selon les notaires de France). Nous détaillerons dans chacun des chapitres comment nous prenons en compte ces biais, par des redressements statistiques ou par le choix de la période étudiés.

Présentation des Études

La première étude consiste en une application empirique des modèles de *matching* sur le marché immobilier. Nous transposons la démarche déployée dans de nombreuses études traitant du marché de l'emploi, notamment rassemblées dans Petrongolo et Pissaridies (2001), pour confronter à des données les éléments théoriques avancés dans Wheaton (1990), Krainer (2001), Novy-Marx (2009) ou Albrecht et al. (2007). Plus précisément, grâce aux estimations faites sur Meilleurs Agents, nous construisons des proxys des nombres d'acheteurs et de vendeurs actifs sur les marchés de 44 grandes aires urbaines de France entre 2013 et 2017. Croisant ces indicateurs avec le nombre de ventes enregistrées par le fisc dans la base des Demandes de Valeurs Foncières sur ces marchés, nous réalisons à notre connaissance la première estimation d'une fonction d'appariement sur le marché du logement. L'estimation des paramètres effectifs de cette fonction permet, en premier lieu, d'ajouter à Genesove et Han (2012) et Piazzesi et al. (2020), une nouvelle confirmation empirique du bien-fondé de l'adaptation du modèle de *matching* DMP au marché immobilier. Mais surtout, elle permet d'entrevoir pour la première fois l'intérieur de « la boîte noire » (Petrongolo et Pissaridies, 2001) qui se trouve au cœur des modèles théoriques.

Les questions auxquelles nous apportons des premiers éléments de réponses sont essentielles. La forme fonctionnelle de Cobb-Douglas, adoptée par l'ensemble des modèles, permet-elle d'expliquer les variations observées dans les données ? Les rendements d'échelles sont-ils constants comme dans le marché de l'emploi ? Ou bien croissants ? Une augmentation de la taille du marché, en augmentant les chances de trouver une contrepartie adaptée, accroissant d'autant plus le nombre de ventes. Ou au contraire décroissants ? Les effets de congestions s'empirant avec le nombre de participants dans un marché fortement hétérogène et dépourvu de systèmes de coordination. Il s'agit également de savoir si le mécanisme d'appariement ne réagit qu'aux nombres d'acheteurs et de vendeurs, à la manière d'un processus purement aléatoire, ou si certains comportements et caractéristiques de ces derniers entrent en jeu. Des réponses à ces interrogations dépendent la crédibilité des modèles théoriques, mais surtout leur caractère actionnable en dehors du seul champ académique. Comme nous le verrons, nos mesures confirment une grande partie des éléments qu'ils présupposent, mais pas tous. En particulier, contrairement au postulat de la littérature théorique, nous montrons que les rendements d'échelles de cette fonction ne sont pas constants, mais décroissants.

Le chapitre suivant analyse le processus d'achat d'un bien immobilier à travers les parcours d'acheteurs parisiens, entre 2015 et 2018. Comme nous le soulignons dans la section IV de la présente introduction, notre connaissance des mécanismes en jeu dans la vente d'un bien s'est construite depuis longtemps en se basant sur les annonces immobilières. Cela permet d'avoir une bonne compréhension des contraintes (Genesove and Mayer 1997), des frictions (Miller 1978, Haurin 1988) et des stratégies (Yavas et Yang 1995, Anglin et al. 2003) à l'œuvre de ce côté-ci de la transaction, mais nous laisse aveugle quant à l'autre versant. Pourtant, autant que celle du vendeur, la recherche de l'acheteur est soumise à de grandes incertitudes et se fait au prix de beaucoup de temps et d'effort. Ces frictions, qui pèsent sur les mécanismes de découverte et de fixations des prix, ont essentiellement été étudiées à travers des variations dans les conditions initiales en termes d'expertise et d'origine de l'acheteur (Turnbull et Sirmans 1993, Lambson et al. 2004, et Ihlandfeldt et Mayok 2012). L'influence des éléments propres au parcours d'achat n'a quant à elle jamais été mesurée, quand bien même une adaptation continue aux informations recueillies au fur et à mesure, comme dans Kohn et Shavell (1974) ou Rothchild (1974), semble pleinement justifiée compte tenu du peu d'information à disposition de l'acheteur au début de ses recherches.

Pour combler ce manque, nous utilisons les estimations successives faites par un utilisateur se déclarant d'abord comme un acheteur en recherche puis comme étant devenu propriétaire d'un appartement. Le jeu de données ainsi constitué permet une étude empirique du problème de l'acheteur qui prend en compte le déroulement de la recherche elle-même, et non plus seulement les contraintes liées aux conditions initiales. Grâce à cet aperçu de la chronologie des visites, des biens qu'elles concernent et des estimations de la valeur de ces biens, nous analysons comment l'histoire de l'acquisition influence le prix payé pour un appartement donné. Il apparait que confronté à l'incertitude liée à la valeur des appartements, les acheteurs ajustent leur prix de référence interne en fonction de leur expérience récente. En effet nous mesurons qu'un acheteur qui visite des appartements plus (resp. moins) chers que celui finalement acquis le paie plus (resp. moins) cher, toutes choses égales par ailleurs.

Pour finir, la plateforme siège de nos observations étant en premier lieu un site diffusant de l'information à propos des prix immobiliers, elle représente une opportunité d'observer comment acheteurs et vendeurs intègrent ces informations. Si l'existence de telles sources contredit la vision d'un marché où l'opacité règne, les particuliers n'en demeurent pas moins des amateurs opérant avec une quantité d'information disponible limitée. L'économie

comportementale décrit comment les agents économiques prennent des décisions dans de tels contextes et les limites de ces mécanismes. Les heuristiques introduites par Tversky et Kahneman (1974) ont été appliquées au contexte immobilier dans des expériences contrôlées (Nothcraft and Neale 1987, Black and Diaz 1996, Corina and Chenavaz 2011) et les effets d'ancrage ou d'aversion à la perte sont invoqués pour expliquer des comportements effectivement observés sur le marché (Genesove and Mayer 2001, Einio et al., 2007, Ihlandfelt and Mayock 2012). C'est également dans le cadre offert par ces théories qu'est analysée la capacité des propriétaires à évaluer la valeur de leur bien (Kish and Lansing 1954, Goodman and Ittner, 1995, Van der Cruijsen et al., 2018).

Le travail présenté dans la dernière partie de cette thèse s'intéresse également aux capacités des particuliers à en prédire le prix, mais se distingue des études existantes sur deux points. D'abord, elle compare les facultés de personnes à la fois en position de vendeur ou d'acheteur. Deuxièmement, les personnes interrogées le sont alors qu'elles mènent effectivement leur projet immobilier. Nous observons donc la formation des croyances en matière de prix au moment où elles sont le plus importantes, car pesant alors sur le montant de la transaction. Pour ce faire, nous comparons les prix des ventes enregistrées dans les bases notariales avec les estimations qu'en ont faites les utilisateurs eux-mêmes, après avoir découvert le résultat du calcul de Meilleurs Agents, dans l'année précédant la date de la transaction. Contrairement aux vendeurs qui, conformément au résultat établi dans la littérature, surestiment la valeur de leurs biens, les acheteurs ne démontrent aucun biais positif ou négatif. L'opinion qu'ils se font du juste prix est également moins influencée par le résultat de l'estimation de Meilleurs Agents que celle des vendeurs, même si l'influence qu'elle peut avoir sur ces derniers tend à diminuer alors qu'ils avancent dans le processus de vente.

On le voit, les trois études qui constituent cette thèse se distinguent tant par les sujets qu'elles traitent (mécanismes de rencontres, parcours d'achats, croyances sur les prix) que par les cadres théoriques dans lesquelles elles s'inscrivent (modèles d'appariement, économie de l'information, décision face à l'incertitude). Pourtant, toutes trois partagent la même ambition. Celle de faire progresser, grâce à l'apport de données inédites issues d'une plateforme Internet, la compréhension des processus à l'œuvre dans une transaction immobilière jusqu'à la rencontre entre l'acheteur et le vendeur, et la fixation par eux de son prix.

2. Estimating the housing market matching function through Internet traffic analysis

Abstract:

In the absence of any empirical estimation, standard theoretical models of the housing market follow the framework developed for the job market that assume a Cobb-Douglas form with constant returns to scale for the matching function. The present study aims to fill the gap and provides a first estimate of a matching function for the housing market. To that end, we assemble a unique dataset, through Internet traffic analysis, with measurements of the numbers of home buyers and home sellers active in the forty-four largest urban areas in France, from October 2013 to June 2017. We confirm that a Cobb-Douglas form fits the data well. However, our results indicate decreasing rather than constant returns to scale. Two extensions of the baseline model random matching are explored. The first accounts for the intensity of the buyers' search, the second, for some characteristics of the market participants. Both bring evidence that these elements matter in the way buyers and sellers meet and transact and that the matching on the housing market is not purely random.

Résumé :

En l'absence de mesure empirique, les modélisations classiques d'appariement sur le marché immobilier imitent celles du marché de l'emploi et postule une fonction de matching de forme Cobb-Douglas à rendements d'échelle constants. L'étude présentée ici vise à combler ce manque et propose une première estimation d'une fonction d'appariement du marché immobilier. Dans ce but, nous construisons un jeu de données unique, basé sur une analyse du trafic d'une plateforme immobilière en ligne. Ces données contiennent des mesures des nombres d'acheteurs et de vendeurs actifs sur les marchés immobiliers des quarante-quatre plus grandes aires urbaines françaises, d'octobre 2013 à juin 2017. Nous confirmons qu'une fonction de Cobb-Douglas correspond bien aux données, mais nos résultats indiquent pour cette fonction des rendements d'échelles décroissants plutôt que constants. Deux extensions du modèle d'appariement sont explorées. Le premier rend compte de l'intensité de la recherche des acheteurs, le second de certaines caractéristiques des acteurs du marché. Tous deux apportent la preuve que ces éléments importent dans la manière dont les acheteurs et les vendeurs se rencontrent, et que l'appariement sur le marché du logement n'est pas purement aléatoire.

I. Introduction

In the lecture he gave on December 8, 2010, when he received the Bank of Sweden Prize in Economic Sciences in Memory of Alfred Nobel, Dale Mortensen told how he purchased an apartment "with a view of Lake Michigan." According to him: "All of the time and effort spent by both sides of such a transaction represent search and matching frictions." Indeed, if first developed to understand how there can be simultaneously unemployed job seekers and open job vacancies, the DMP matching model has proven to be suited for the study of the housing market as well.

Wheaton (1990) first drawn in the labor matching literature to build a theoretical model that explains the structural vacancy of the housing market and the relationship between these vacancies with the expected prices and time-on-market. He was followed by Krainer (2001), Albrecht et al. (2007) and Novy-Marx (2009). They each add some specifications to the general framework, respectively: difference in the cost of mismatches in booms and busts, participants ' increasing state of despair along their search, and endogenous entry rate proportional to the participants expected participation values.

On the empirical side, the literature appears scarce due to the difficulty to measure the number of active buyers, as they do not declare their search through listings. A first exception is Genesove and Han (2012), who proxy market tightness, i.e. the buyer-to-seller ratio, with the log difference between the buyer and the seller time-on-market. They show how search behaviors and market outcomes tend to overshoot a demand shock. Another way to overcome this measurement issue is explored in Franke and van Dijk (2018) and Piazzesi et al. (2020) who both use Internet data to measure the demand. Franke and van Dijk (2018) show that the click rate on the listings posted on line Ganger causes price fluctuations at city level, in the Netherlands. Even though they do not estimate a matching model, their work demonstrates that the housing demand can be measured through the analysis of users 'activities on a listing website. Piazzesi et al. (2020) use email alert subscriptions to study the segmentation of the market in the San Francisco Bay Area. Similarly, to the job market, offer and demand appear to be negatively correlated when aggregated across areas that do not share many common searchers. In contrast, across segments searched by many common prospective buyers, their data exhibit a reverse Beveridge curve, with searchers and vacancies positively correlated.

Following these two examples, we compute proxies for the numbers of sellers and buyers active in the housing markets of the fifty largest urban areas in France from October 2013 to June 2017. We base these measurements on the analysis of the traffic on MeilleursAgents.com (thereafter MA) a French real estate Internet platform. This dataset allows us to measure both supply and demand in the housing market, as the labor market empirical literature is used to measure the numbers of job seekers and job vacancies. This research focus is to use these measures and transactions figures recorded in the tax databases to estimate a matching function of the real estate market.

To date, the matching functions used in the housing literature (Novy-Marx 2009,Genesove and Han 2012) mimic the homogenous Cobb-Douglas form with constant returns to scale of the job market. However, if such specification for job matching has been proven to fit the data numerously (see Petrongolo and Pissaridies 2000 for a survey), there is no empirical justification to support this hypothesis for housing. The present study aims to fill this gap. Our results confirm that a Cobb-Douglas function is suited for housing matching. However, they lead to reject the constant returns to scale hypothesis. Characteristics of the residential real estate market advocate for the matching efficiency to be sensitive to the number of market participants. On the one hand, thick market effects (Ngai and Tenreyro 2014) ease the search for a suitable match in larger markets. On the other hand, coordination failures among isolated participants and the uncertainty caused by the properties heterogeneity should worsen as the market expand. Our results indicate that the latter dominate. Indeed, we find returns to scale to be decreasing in all specifications and robustness checks.

While this first specification proves that random matching offers a fair modeling of the way we meet and transact in the housing market, we decided to go beyond and investigate how market participants' behavior and characteristics impact this process. In our first extension, we account for a proxy of the average number of visits per buyer. This parameter, which can be seen as a measure of the intensity of the search, had a rather negative effect on the number of sales and did not increase the matching efficiency. This result advocates for the existence of two coordination failures that operate on both sides of the transaction or can be seen as evidence of a two-period stochastics matching scheme. In our second extension, we investigate the impacts of the characteristics of the sellers and buyers on the numbers of sales. It appears that the time the participants already spent on the market matters, with opposite effects on the offer and demand sides. The share of incoming sellers happens to be positively correlated with the

sales figures, as if they were more likely to find a suitable match than incumbents, as in "stock-flow" mathcing. On the other hand, the more buyers at the very beginning of their search are on the market, the fewer matches are formed, as if they were missing motivation or confident at first. Overall, both extensions demonstrate that more nuanced mechanisms operate in addition to the general random matching process.

The rest of the paper is organized as follows. Section II presents a short literature survey on the matching function, with a focus on empirical studies. Section III describes how data are collected on the MA website and outlines these data. Section IV analyzes empirical results of a first estimate of a matching function. Results of section IV are tested for robustness in section V. Sections VI and VII are extensions of this model, that account respectively for the intensity of the buyers' search and for the composition of the pools of market participants. Section VIII concludes.

II. Literature

The role of the matching function is to summarize the whole process of search with frictions for a counterpart, participants on both sides of a decentralized market go through to meet and contract. For a given number of participants on each side, the function output is the number of productive matches in the market. Matching functions can describe any market with frictions such as real estate (Wheaton 1990) or the marital market (Burdett and Coles 1997). However, it is in the labor market that this theoretical tool has been mostly used, starting with the seminal works of Peter Diamond, Dale Mortensen and Christopher Pissarides on wages and employment level in the 1980s (Diamond 1982a,b; Mortensen 1982a,b; and Pissarides 1984, 1985). Therefore, the literature presented hereunder mostly focuses on the job market. We also adopt notations used in labor economics so as to present how results from the housing literature can be interpreted within the matching framework.

In its simplest form, a matching function can be described as in (1)

$$M = m(U, V) \qquad (1)$$

With U the number of job seekers, V the number of job vacancies and M the number of matches or hires. Commonly accepted restriction are:

$$m(0, V) = m(U, 0) = 0$$

And :

$$m(U,V) \leq min\ (U,V)$$

The equality case corresponds to a frictionless market.

The elasticity with respect to the numbers of job seekers is:

$$\eta_U = \frac{\partial m(U,V)}{\partial U}$$

And the elasticity with respect to the job vacancies is :

$$\eta_V = \frac{\partial m(U,V)}{\partial V}$$

While several functional forms of the matching functions have been proposed, the most commonly supposed in empirical studies is the Cobb-Douglas production function (Petrongolo and Pissaridies 2000):

$$M = AU^{\alpha}V^{\beta} \qquad (2)$$

With $\alpha = \eta_U$ and $\beta = \eta_V$ that can easily be estimated with a log-linear specification.

The availability of long-term national statistics on the job market allowed to validate of the matching theory, early on. The first empirical studies used aggregated data. Pissarides (1986) estimates a matching function for Great Britain through the 1967-1983 period, followed by Layard, Nickell, and Jackman (1991, ch. 5) for the 1968-1988 period. Later on, authors switch to more disaggregated data and provide cross-section estimates of the labor markets matching function (Coles and Smith 1996, Bennet and Pinto 1994). Finally, Anderson and Burgess (2000) use panel data to estimate matching over time and across regions and industries.

In the Petrongolo and Pissaridies (2000) survey, the authors underline two results inherited from twenty years of estimation of the job market matching function: constant returns to scale and a larger elasticity with respect to the number of job seekers. They report an elasticity with respect to job seekers η_U between 0.5 and 0.7 for most studies. Following these results, Novy-Marx (2009) andGenesove and Han (2012) built theoretical matching frameworks for the

housing market on the Cobb-Douglass form and the constant returns to scale hypothesis, even though no empirical result supports these specifications.

Beyond the first formulation of equation (1), more complex matching functions that account for the heterogeneity among market participants explain the difference in individual hazard rates. One can consider the disparity in the search intensity among participants as in (Pissarides 2000, ch. 5). Each job seeker i supplies a different amount of search units s_i at a marginally increasing cost. The greater s_i the faster i finds a job. At the aggregate level, with s the average search units supplied by a job seeker, the matching function becomes:

$$M = m(sU, V) \qquad (3)$$

With $\frac{\partial m(sU,V)}{\partial s} > 0$.

Measuring how active a job seeker is raises an observability issue. Different proxies have been proposed: the number of applications, time spent, the amount of money spent or the number of search channels (Layard et al. 1991, Shimer 2004, Krueger and Muller 2010). Most of these papers bring evidence that search intensity is counter-cyclical: the effort job seekers put in their search seems to be an attempt to compensate poor economic conditions. At the micro-level, Faberman and Kudlyak (2019) measures the longer-duration job seekers to be the ones that send the most applications. Hence, an attempt to incorporate a direct intensity measure within the estimation of the matching function may probably lead to a negative effect, in contradiction with the theory. To the authors' knowledge, such an estimation has never been performed. However, authors (Burgess 1993, Bell 1997, Coles and Smith 1996) have used indirect measures based on segmentation of job seekers by unemployment duration, reason of unemployment, labor market status or demographic criteria.

Housing market participants also exhibit heterogeneity in the effort they put in the search for a counterpart. Even though the literature on homebuyers is thin, because of a lack of data, empirical researchers have used the number of homes visited, the time-on-market or how often they consult listings to characterize their search intensity. Anglin (1997) measures that information-related variables are the most important to explain the differences among buyers. In line with search frameworks developed in Turnbull and Sirmans (1993) or Lambson et al. (2004), Elder et al. (1999) argue search costs are central to explain buyers ’efforts. They measure individuals with high within-period search costs (high-income) search less intensively

but longer, whereas the ones with high across-period search costs (out-of-town buyers) search more intensively.

A priori, sellers 'efforts are harder to characterize since they seem to passively wait for buyers. However, results on the sellers 'time-on-market variance demonstrates heterogeneity in their willingness to close a deal. Time (Glower et al. 1998) or financial constraints (Genesove and Mayer 1997) have been put forward to explain differences in the sellers 'search efforts. The most prolific line of research on the matter is the relationship between the marketing time and the asking price. Yavas and Yang (1995) or Anglin et al. (2003), show that, for a given property, a higher asking price leads to longer time-on-market. According to directed search models (Carrillo 2012, Merlo et al. 2015), the meeting process is not purely random and sellers can signal their motivation with low asking prices to attract buyers. The asking price may be viewed as a way for the seller to signal his/her reservation price (Albrecht et al. 2016), another channel of heterogeneity that has been extensively investigated in the job market matching literature.

Indeed, job seekers and firms also differ in the wage they are eager to accept and to pay respectively (Jovanovic 1979, Burdett and Mortensen 1998, Pissarides 2000, ch. 6). In such a setting, every match does not lead to a hiring. For each pair of (worker; firm) that meets through the matching technology, a wage w equal to the match productivity is drawn from a distribution G. If this wage is below the worker and firm reservation values, there is no job creation. This two-step matching procedure is called stochastic matching. With R the average reservation wage, the matching function becomes:

$$M = (1 - G(R))m(U, V) \qquad (4)$$

Novy-Marx (2009) andGenesove and Han (2012) both consider stochastic matching in their models of the housing market. Indeed, despite the description of the property available through listing, the realization of the match value occurs only after the buyer's visit and the negotiation process. The same mechanism is also central to the direct search models of Carrillo (2012).

Third, according to Blanchard and Diamond (1994), firms may have preferences between short-term job seekers, who just began their search, over long-term job seekers. Such "ranking" of the workers would explain the decrease in the exit rate of a given job seeker over

time, documented in Jones (1989) or Machin and Manning (1999). The global matching function becomes:

$$M = m(U^S + U^L, V) \quad (5)$$

And we have:

$$\frac{m^s(U^s, V)}{U^s} > \frac{m(U^S + U^L, V)}{U^L} - \frac{m^s(U^s, V)}{U^L}$$

With U^S the short-term job seekers, U^L the long-term job seekers and $m^s(U^s, V)$ the matching function for short-term job seekers. The left-hand side of the inequality represents the hazard rate of short-term job seekers, and the right-hand side, the one of long-term job seekers.

The results of Burgess (1993), Mumford and Smith (1999) or Bell (1997) on the negative impact of long-term unemployment on the matching rate have equivalent in real estate. Taylor (1999) develops the idea that a long time-on-market is a negative signal in the context of housing. Homebuyers interpret long marketing time as a sign that other buyers have already inspected the property and have judged it is of poor quality. This negative herding phenomenon creates a stigma, which decreases the sale likelihood of slow-moving homes. The opportunist matching of Albrecht et al. (2007) also considers the difference in the exit rate between entrants and incumbents with the exact opposite consequence. In their model, buyers and sellers become desperate and looser on their matching criteria after a while, which boosts their chance of matching.

Finally, Coles and Smith (1998) introduce the so-called stock-flow approach as another explanation for the greater hazard rate of the recently unemployed workers and the recently opened vacancies. They propose a matching mechanism with perfect information and no coordination failure. In such setting, no acceptable match is unconsumed and all job-seekers apply to all acceptable job offers at the beginning of their search. The workers and vacancies that remain unmatched after this first round will not try to match with each other afterward. Thus this "stock" of participants has to wait for newcomers, the "flow", to unlock them. Hence, stock from one side tries to match with the flow of the other side, whereas the flows of both sides can also match together. Gregg and Petrongolo (1997) confirm the empirical results of Coles and Smith (1998). The stock-flow approach could be applied to housing in line with the

idea of homebuyers first scanning the entire stock of for-sale homes and then only focusing on new listings.

In the 2010 decade, the rise of the online platforms has opened new application opportunities for the matching theory. These centralized two sided markets have flourished in multiple segments of the so-called "sharing economy", in sectors including transport (Uber, Lyft), tourism (Airbnb, Home Away) or outsourcing (Amazon Mechanical Turk, Task Rabbit). The seminal paper of Rochet and Tirole (2003) formalizes how these companies create value for their customer by enabling mutually beneficial interactions. Extensions of the theory accounting for the specific of these platforms are being built (Arnosti et al., 2018) but most of the academic perspectives they open are empirical. The matching mechanism being critical to them, they massively collect data on their users' behaviors and have partnered with academics to document and optimize it (Fradkin 2017a). Such partnerships have produced numerous studies that lead to more diverse results than the ones surveyed in Petrongolo and Pissaridies (2000). Returns to scale are measured to be decreasing for a peer-to-peer holiday property-rental platform in Australia (Li and Netessine 2020), constant for an outsourcing task website in the US (Cullen Farronato 2016) and increasing in demand ride-hailing in Singapore (Kabra et al. 2016). Li and Netessine (2020) deserve particular attention. They not only apply search and matching to a new context, but also take advantage of this new context to solve the identification issue that lies in the application of the framework in traditional markets. They use the acquisition of two competing platforms as a quasi-experiment and show the doubling in size that results from this merger, keeping the matching technology constant, leads to the loss of 5.6% of the potential matches. This context also allows investigations on previously overlooked issues like the efficiency of the matching technology, in this context the search engine (Fradkin 2017b), or the buyer uncertainty about seller availability (Horton 2019).

Recent developments in the labor market literature, surveyed by Faberman and Kudlyak (2016), also leverage datasets drawn from online job search websites. Brenčič (2012) studies the incidence of wage posting. Faberman and Kudlyak (2016), Marinescu (2017) and Baker and Fradkin (2017) identify causes for the search efforts to vary. Text recognition is performed on job titles and descriptions in Marinescu and Wolthoff (2020), Banfi and Villena-Roldán (2019) and Kuhn and Shen (2013). Websites also offer the possibility to run experimental research as in Kroft et al. (2013), Pallais and Sands (2016) or Belot et al. (2019). To our knowledge, the literature does not present any study linking the number of total offline matches

with online measurement of the demand and offer. The present paper is the first to estimate a matching function that way.

III. Dataset

The primary innovation of our research comes from the nature of data. Through the analysis of Internet users 'activities, we are able to build measures of the numbers of buyers and sellers active in the housing markets of the forty-eight largest French metro areas, from 2014 to 2017. These measurements are performed on a quarterly basis allowing a dynamic analysis over time. These measures are performed on *MeilleursAgents.com*, a website that makes various real estate information available online: price maps, indexes and an automated valuation model (AVM). With 2 million unique visitors and 200,000 property estimates every month, it claims to be the French national leader for "Real estate online estimation."[19]

To measure the number of active buyers and sellers in a market at a given moment, we use a feature of the AVM. This valuation tool gives an estimate of any housing unit based on its address and a description of the property. The online form asks for basic characteristics: livable surface, number of rooms, floor, and number of bathrooms; and advanced ones: the presence of an elevator, of a cellar, recent facade renovation, etc. A hedonic model combines those characteristics with spatial econometrics models to compute a predicted market value.

The user specifies the reason why she or he performs an estimate of the property. If she/he is the owner, she/he is asked the property usage, if the property is currently on sale or if she/he is planning on selling in the near future. She/he may optionally give the date and the price of its property when she/he acquired it. If she/he is not the owner, she/he is asked if she/he is currently contemplating the acquisition of a property and, if so, at what stage she/he currently is in her/his search. In any case she/he needs to have an account on MA which allows to uniquely identify her/him. Through the response of these questions we are able to perform a measure of the relative size of the two populations. More precisely:

[19] Source : www.meilleursagents.com

- A user is considered a seller if :
    - he/she indicates that he/she owns the property
    - he/she answers the question: “Do you consider selling this property?” by :
        - “Yes, I have already started to sell it”
        - “Yes, as soon as possible”
        - “Yes, within three months”
- A user is considered a buyer if :
    - he/she indicates she does not own the property
    - he/she answers the question: “Why do you estimate this property?” by “ I wish to buy a home”

We quarterly measure the number of buyers and sellers active in a metro area by counting the number of distinct users, identified by their MA accounts, that fold into one of these aforementioned two categories. We compute this measure for the fifty largest urban areas in France, as defined by INSEE[20], for every quarter from Q4-2013 to Q2-2017.

The use of digital traces in academic research implies some risks and calls for caution. The principal issue Salganik (2017) points out is that, in contrast with “custommade” datasets traditionally used in social sciences, such “readymade” data drawn from the Internet have not been collected to be used for research. Indeed MA has recorded the activities of users on the AVM primarily for strategical and operational reasons rather than building measures of market tightness. We use purely declarative information, like many survey-based studies. However, unlike usual surveys, users fill out the form without any supervision and can declare erroneous information regarding their project. Yet, two reasons plead for the good faith of users. First, they have no incentive to hide the truth, except perhaps being reluctant to disclose information to a commercial website, which would only lead to under-declaration. Second, we get a confirmation - at least for the sellers - that their information is accurate in the existence of MA as a company, for now more than 10 years. Its primary business model is to connect prospective

[20] www.insee.fr/fr/metadonnees/definition/c2070 and www.insee.fr/fr/information/2115011

sellers with real estate agents. If the projects declared on MA were not accurate then realtors would have canceled their subscriptions to the MA service.

	DVF						
	Numbers of observations : 1,507,105						
	mean	std	min	25%	50%	75%	max
area	79	63	1	50	72	97	9,677
rooms	3.4	1.5	1	2	3	4	112
price	238,762	3,143,078	0	120,000	181,900	275,000	3,300,000,000
pm2	3,268	27,914	0	1,832	2,579	3,747	2,5781,250

core city	33%	2013	5%
Paris	34%	2014	22%
top 11	34%	2015	26%
apartment	44%	2016	29%
house	56%	2017	17%

	Seller						
	Numbers of observations : 450,312						
	mean	std	min	25%	50%	75%	max
area	94	60	1	55	80	119	950
rooms	4.0	1.9	1	2	3	5	10
price	327,556	356,963	0	163,900	244,500	383,600	83,584,350
pm2	3,829	2,864	0	2,148	2,923	4,568	522,402

core city	33%	2013	6%
paris	52%	2014	27%
top 11	24%	2015	27%
apartment	61%	2016	26%
house	39%	2017	13%

	Buyer						
	Numbers of observations : 302,387						
	mean	std	min	25%	50%	75%	max
area	82	51	1	47	71	103	998
rooms	3.5	1.7	1	2	3	5	10
price	367,172	2,864,003	0	175,550	267,300	426,100	1,487,427,700
pm2	4,851	10,977	0	2,510	3,838	6,860	4,958,092

core city	43%	2013	6%
Paris	61%	2014	23%
top 11	23%	2015	27%
apartment	68%	2016	28%
house	32%	2017	16%

Table 2-1 : Statistics comparison of properties estimated by MA users and the ones sold and recorded in DVF

Another well-know pitfall of custom-made datasets collected by Internet companies is that there is no good reason to consider them representative of the whole population. Hence, we should assess how buyers and sellers using AVM are different from the rest of the market.

MA does not ask for personal information other than names and genders, thus its users cannot be compared with standard buyers and sellers on a socio-demographic standpoint. However, the aforementioned procedure enables to gather information about the properties they are selling or touring. A comparison between these properties and all the properties sold during the period, recorded in the tax database *Demande de Valeur Foncière* (thereafter DVF), is proposed in Table 1. Transactions in Alsace and Moselle department are not recorded in DVF, thus the scope of the study is reduced from forty-eight to forty-six urban areas.

The DVF dataset release by French tax administration records all real estate transactions and not only the ones about individual residential properties. Moreover, it tracks fiscal allotments property transfer, therefore a single transaction can be presented over several rows. For instance, the sale of a single-family house that is built over two cadastral parcels is presented over two rows. In order to fit with our empirical objective, we apply the method from *Groupe National DVF*[21] that enables to identify individual sale. Then, we filter out all transactions that do not concern a single residential property.

First, the volume of the three datasets illustrates that if MA is not exhaustive, it captures a significant part of the market. For ten actual sales, three properties are estimated by sellers on MA and two by buyers, during the period in the forty-six urban areas considered. As mentioned above, the MA business model is based on sellers. Thus, one cannot be surprised to see an over-representation of sellers compared to buyers as the company has an incentive to optimize its traffic acquisition funnel to attract sellers.

As the datasets composition is concerned, the key difference stems from the geographical distribution of the properties. A third of the properties sold during the period is located in the Paris urban area, another third in the next ten biggest urban areas and the last third in the thirty-five others. The proportions are strongly different among the properties estimated by sellers and buyers on MA. The Paris metro area is over-represented: 51% for sellers and 61% for buyers. This may be due to historical reasons: MA started in 2008 with a price map only for the city of Paris before expanding first to the Paris suburban region in 2009 and then gradually to the rest of France starting in 2010. The even stronger over-representation

[21] www.groupe-dvf.fr/vademecum-fiche-n3-precautions-techniques-et-qualite-des-donnees-dvf/

of the Parisian market among buyers can be due to the market being tighter in the French capital region. We can make the same argument for the over-representation of homes located in the main cities of the metro areas (43% among the ones estimated by buyers against 33% for both DVF and sellers).

From these differences in the geographical reparation of the estimated and sold properties, differences in prices and types arise. Estimated properties are more expensive both in absolute value and in price per meter square: 238,762 € or 3,268 €/m² for the sold properties against 327,556 € or 3,829 €/m² for the sellers 'and 367,172 € or 4,851 €/m² for buyers'. Paris area is more expensive which explains these differences. The stronger divergence in the estimated price for the buyers 'properties is explained by the fact that they are often located in the main cities than in the suburbs. The Paris area and major cities are more urbanized. Their housing stock comprise mostly apartments rather than individual houses. That is why the proportion of apartments jump from 45% for the sold properties to 61% and 68% for the sellers ' and buyers 'properties respectively.

The temporal evolution in the number of properties sold illustrated an acceleration of the market turnover rate during the period (for 2017 only the first semester is considered). Such an observation is consistent with the French housing market dynamic. According to INSEE, French prices dropped by 2.4% in 2014 then 0.5% in 2015 and increased by 1.5% in 2016 and 2.9% in 2017. The number of estimated properties by buyers follows almost the exact same dynamics, whereas sellers have estimated almost the same number of properties every year of our samples. Implicitly, it means that an increase in the market tightness happens as the market goes from "cold" to "hot", in accordance with the literature (Novy-Marx 2009, Genesove and Han 2012).

Altogether, this comparison reinforces our conviction that using the traffic on MA can be meaningful to describe the housing market evolution. However, the differences between properties estimated on MA and properties sold recorded in DVF impose some specifications of the models that are tested in part IV. We use a log-linear specification with metro areas fixed effects to account for the difference in both the penetration rates among buyers and sellers and the geographical repartition of the MA audience (see Appendix A to see how we can get an unbiased estimation of the matching function from biased measures). We also run robustness tests on subsamples after splits of the dataset according to the location, the type of housing and the prices.

Let us now finally move on to the dataset that we used to estimate the matching function of the housing market in part IV. Here we aggregate the estimation by user (seller or buyer) and not by property as above. For each of the forty-six metro areas, the number of distinct buyers and sellers that performed an estimation is computed for every quarter from Q1-2014 to Q2-2017. For representativeness concern, we have discarded 2 metro areas that did not count more than ten buyers and ten sellers on every contemplated quarter. This reduces the subset from forty-six to forty-four metro areas. In order to estimate the matching function, the number of market participants have to be compared with the output of the matching function, here the number of transactions recorded in DVF. To avoid any risk of reverse causality, we match the number of sellers and buyers in a given period with the number of sales in the following period. Table 2 below gives a statistical description of the dataset.

	count	mean	std	min	25%	50%	75%	max
Number of sales		2,282	5,101	272	658	1,009	1,824	43,954
Number of buyers		336	1,232	11	50	85	179	9,728
Number of estimations		493	1,881	13	70	118	239	14,630
Estimations per buyer		1.38	0.14	1.00	1.29	1.37	1.45	2.29
Estimations per address		1.03	0.03	1.00	1.01	1.03	1.04	1.16
Wondering ratio		10%	6%	0%	8%	11%	13%	45%
Starting ratio		26%	7%	3%	22%	26%	30%	59%
Active search ratio		41%	11%	6%	36%	43%	48%	82%
Offer ratio		16%	5%	0%	13%	16%	19%	42%
Number of sellers	660	625	2,107	35	129	200	362	16,896
Main residence ratio		76%	5%	53%	73%	77%	80%	89%
Vacation home ratio		8%	5%	1%	5%	7%	9%	33%
Rental investment ratio		16%	4%	2%	13%	16%	18%	31%
Already selling ratio		35%	5%	19%	31%	34%	38%	52%
As soon as possible ratio		47%	5%	30%	44%	47%	50%	72%
Within three months ratio		19%	4%	7%	16%	19%	22%	32%
With realtor ratio		45%	9%	10%	39%	45%	51%	76%
FSBO ratio		28%	9%	7%	21%	27%	34%	63%
Realtor and FSBO ratio		27%	7%	8%	23%	27%	31%	70%

Table 2-2 : Summary statistics of the matching panel dataset

On top of computing the numbers of sellers and buyers, we use other questions of the estimation forms to describe more precisely the search market participants. We assess how far the average market participant is in her/his search, both for buyers and sellers. The average number of estimations per buyer, taken as a proxy for the number of visits, carries information

about the buyers' search. Sellers also specify the nature of property put on the market (i.e. primary, vacation home or rental investment) and, for properties already on sale, the kind of commercialization used (i.e. For Sale By Owner, realtor or both).

The strongly skew distribution of the numbers of sales, sellers and buyers underlines how different in size the urban areas are from Paris metro, which gathers 12.6 million inhabitants, to Quimper's, with 127,000 persons. This heterogeneity suggests a metro area-specific fixed effect control in the econometrics of parts IV to VI. Taken individually, these figures confirm what table 1 suggested. First, MA traffic is not exhaustive but seems to capture a significant share of market participants. On average, it detects one thousand buyers and sellers taken together for a little bit more than two thousand sales on average. Second, the number of sellers active on MA is significantly higher than the number of buyers. As previously discussed, this bias is more probably due to MA seller-oriented business model rather than any market-related reason. The business model of MA being seller-oriented since its creation and the evolution of the number of buyers over time exhibited in table 1 in line with sales count, we make the assumption that this bias has been stable over time. Thus, it should not interfere with the results of the log-linear model estimated in part IV.

Other variables are more evenly distributed. Across all metro areas/quarters, buyers estimated between 1 and 2.39 properties on average. The vast majority of properties, identified by their address, appears only once among the buyers 'estimations, yet some are estimated by different buyers. Over all quarter/metro area pairs, 1 to 1.16 distinct buyers visit a given address on average. Not all buyers used MA and the ones that do do not necessarily estimate every property they visit. Yet, we argue that these proxies can help us quantify the effects of the coordination failures, in section V. Buyers use MA's valuation tool mostly during the "active search" phase of the purchase process (41%). Yet, 10% on average use it from the very beginning of the search as they are "wondering" whether to buy or not. Sellers use the AVM just before they put their property on the market, as they are about to choose a listing price. 47% on average, across all markets and periods, request for an estimation wanted to start the sale "as soon as possible." Main residences, occupied by the owners, make the majority of the market in all urban areas across all studied periods. Other properties are composed of two third of rental investment and one third of vacation home on average. On average, a little bit less than half of the sellers delegate the sale entirely to a realtor but more than a quarter sells their property by themselves. These figures a consistent with the share of transactions in which a

realtor is involved in, ranging between 60% and 65% from 2005 to 2015 according to Friggit (2018). A large portion of the sellers do both, give a sale mandate to a realtor and search for a buyer by themselves, as exclusive mandates are a minority in France.

IV. Simple Matching Function

Our empirical strategy to estimate the real estate matching function over panel data is derived from the established methodology of the labor market (Anderson and Burgess, 2000). At first, we consider a matching function of the simplest form, a two-factor Cobb-Douglas function, where the number of sales is the weighted product of the numbers of buyers and sellers. We estimate this matching function through a log-linear specification in which sales, sellers and buyers are aggregated geographically at metro area level and temporally over quarters.

These levels of aggregation are standard for housing statistics. In particular, INSEE computes the reference housing price indices for the French market on a quarterly basis. However, this temporal aggregation imposes a three-month horizon on every real estate project, while cross-periods and shorter searches are possible. This raises a possible reverse causality issue. For example, a spike in matches at the beginning of the period could reduce the number of market participants. To prevent such issue in our method, we use lagged measures of the numbers of active buyers and sellers. Our estimated model is the following.

$$ln(sales_{i,t}) = \alpha\, ln(buyers_{i,t-1}) + \beta\, ln(sellers_{i,t-1}) + FixedEffects_{i,t} + \varepsilon_{i,t} \qquad (1)$$

With $sales_{i,t}$,$buyers_{i,t-1}$, and $sellers_{i,t-1}$ respectively the number of sales recorded in the tax database *DVF*, the number of buyers and the number of sellers detected on *MA* (see part III) in metro area i in quarter t for the sales and quarter $t-1$ for sellers and buyers. The coefficient α is the elasticity with respect to the number of buyers and β the elasticity with respect to the number of sellers. The sum of these elasticities, the returns to scale of the housing markets, is tested against the constant returns to scale hypothesis: $\alpha + \beta = 1$.

Section III shows MA users are not necessarily representative of the real market participants. However, our specification allows us to get an unbiased estimation of the elasticities from biased measures the numbers of market participants. Let us consider that our measures of the numbers of buyers and sellers can be linked with errors to the actual total numbers trough the following equations:

$$total\ buyers_{i,t} = MA\ correction_i^{buyer} * buyers_{i,t} * e^{\varepsilon_{i,t}^{buyer}} \qquad (2)$$

And

$$total\ sellers_{i,t} = MA\ correction_i^{seller} * sellers_{i,t} * e^{\varepsilon_{i,t}^{seller}} \qquad (3)$$

Where $MA\ correction_i^{buyer}$ and $MA\ correction_i^{seller}$ are the inverses of the MA average penetration rates in metro area *i,* among buyers and sellers respectively, that we consider as constant over the study period. The multiplicative error terms represent the random variation around these average penetration rates over times.

We postulate a Cobb-Douglass form for the matching function. Formally:

$$sales_{i,t} = A * total\ buyers_{i,t-1}{}^{\alpha} * total\ sellers_{i,t-1}{}^{\beta} * e^{\varepsilon_{i,t}} \qquad (4)$$

And in the log-linear form:

$$\ln(sales_{i,t}) = \ln(A) + \alpha \ln(total\ buyers_{i,t-1}) + \beta \ln(total\ sellers_{i,t-1}) + \varepsilon_{i,t} \qquad (5)$$

From (2) and (3); (5) becomes:

$$\ln(sales_{i,t}) = \ln(A) + \alpha \ln(MA\ correction_i^{buyer}) + \beta \ln(MA\ correction_i^{seller}) + \alpha \ln(buyers_{i,t-1}) + \ln(sellers_{i,t-1}) + \varepsilon_{i,t}^{buyer} + \varepsilon_{i,t}^{seller} + \varepsilon_{i,t} \qquad (6)$$

Hence, an OLS regression of the log-linear form, based on our biased measures of $total\ buyers_{i,t}$ and $total\ sellers_{i,t}$, enables to reach an unbiased estimation of α and β. Under the hypothesis of a constant penetration rate of MA overall urban areas, the first three terms are gathered in the intercept. Otherwise, we only need to introduce specific fixed effects for the metro area to capture $\alpha \ln(MA\ correction_i^{buyer})$ and $\beta \ln(MA\ correction_i^{seller})$. The validity of the above relies on the assumption we made that the penetration rate of the platform in each metro area is constant over time. Figure (1) in Appendix A shows that we perform our measure in a period when the global popularity of the platform was stable. We challenge this hypothesis in the robustness tests of section V.

We estimate model (1) through an OLS regression. To account for the temporal serial correlation of the housing market activity of a metro area, we clustered the errors at the metro

level. All statistical tests are performed against a normal distribution. Table (3) bellow presents the results.

As expected, estimation of both buyers' and sellers' elasticities are as positive and significant at the 0.1% level across the four regressions. Moreover, the explanatory power of the model is strong. The R^2 range from 94% to 98%. Hence, our estimation of the housing matching function, the first to the authors' knowledge, confirms that the application to real estate of the matching framework is well funded. Also, it proves the usefulness of using data from Internet platforms to overcome matching function measurement issues, in general, and the representativeness of MA data for the French market, in particular. Indeed, in column (1), these data alone explain 94% of the variance in the number of sales recorded in 44 urban areas, over 15 quarters.

Dependent variable	ln(sales)			
	(1)	(2)	(3)	(4)
intercept	3.1136*** (0.1595)	3.2121*** (0.1586)	2.8858*** (0.3685)	3.9335*** (0.5414)
ln(buyers)	0.4712*** (0.0663)	0.4909*** (0.0657)	0.4222*** (0.0435)	0.4518*** (0.0399)
ln(sellers)	0.3295*** (0.0699)	0.3062*** (0.0695)	0.3897*** (0.0517)	0.2579*** (0.0642)
Metro FE	No	No	Yes	Yes
Seasonality	No	Yes	No	Yes
RTS	0.8007***	0.7971***	0.8119***	0.7097***
Observations	660	660	660	660
R^2	0.945	0.949	0.979	0.983
Adj-R^2	0.944	0.949	0.977	0.982
Wald-statistic	123942.9	129663.5	215.1	6396.3

Standard errors between parenthesis
*** p<0.001, ** p<0.01, * p<0.05, ° p<0.1
RTS test is with H_0=1
Errors are clustered at the metro area level and normal distribution is used for inference

Table 2-3: Model (1) OLS estimation over 44 metro areas from 2013-Q4 to 2017-Q2

The strongest result of table 1 is that, in contrast with the hypothesis made so far in real estate matching models (Novy-Marx 2009,Genesove and Han 2012, Diaz and Jerez 2013), the assumption of constant returns to scale is rejected in all our regressions, with a strong statistical significance (p-value below 0.1%). We find returns to scale to be decreasing ranging between 0.81 and 0.71 depending on the fixed effects accounted for. It differs from empirical measures

from the job market (Petrongolo and Pissarides 2000) the constant returns to scale hypothesis was based on. In the absence of empirical results, relying on the labor market literature the matching model emerged from made sense. Our first estimation of its matching function suggests the housing market behaves differently.

Specific features of the residential real estate market can explain how strong congestion effects may arise with the number of participants. First, it is dominated by amateurs that exchange highly heterogenous assets. Homebuyers need time and effort to integrate the information they need to make a decision. A larger market means more information to integrate and therefore more time and effort. Second, these individuals behave without any kind of coordination. A given seller is not aware of marketing strategies of its peers nor on the searches of the buyers - the opposite also being true. Hence, sellers can use the listing price as a signal to attract potentially multiple buyers (Han and Strange 2016, Yavas and Yang 1995, or Anglin et al. 2003). On the other hand, buyers may tour several properties simultaneously to increase their chance to find a suitable match (Anglin 1997, Elder et al. 1999). The fragmentation of the industries in France limits the ability of the realtors to mitigate this and to act as market coordinators . Because of low entry barriers (Han and Strange 2015), this might be especially true in thicker markets. Finally, the sequential nature of the search for a counterpart can explain these decreasing returns to scale. Theoretical search models for both sides of a real estate transaction that are derived from Stigler (1961) (Haurin 1988, Salant 1991, Turnbull and Sirmans 1993, Lambson et al. 2004) predict that the decision maker tightens his/her stopping rule as the expected meeting rate increases. In thick market, a "good deal" can always seem to be "just around the corner". On the contrary, buyers and sellers are more likely to make concessions and to reach an agreement in tighter situations, as they are willing to avoid waiting for the distant next encounter. In the next section, we enhance models (1) to identify these congestion effects.

One may object that the thick-market effect identified by Ngai and Tenreyro (2014) contradicts our results. They measure that the matches that occur during the "hot" housing season, from April to September, are of higher quality (longer afterward tenure, fewer repairs and alterations within the first two years). They model how this effect explains the seasonality in the sales volume over a year. However, they emphasize that their mechanism is not a comparison between two different steady states. It is based on the existence of a deterministic fluctuation of economic conditions within a given market and a given year that agents are aware

of. Such predictability allows them to coordinate their decision to move. Even though sellers have the choice over when to list their property and buyers over which market to enter, the strong constraints related to the need for a suitable home make such coordination unlikely to happen over a longer period or across markets. Moreover, our results also exhibit a seasonal pattern coherent with Ngai and Tenreyro (2014) observations. From model (3) to model (4) controlling for the seasonality decreases the return to scale from 0.8119 to 0.7097, indicating a thick-market effect has an influence over the year, within a given market.

Table 3 results underline another difference with the job market. The elasticity with respect to the number of buyers is greater than the one to the number of sellers. Hence, the number of matches is more sensitive to the intensity of the demand than to the offer. The situation is somehow the opposite of the one from the job market which creates jobs with a larger elasticity to the number of job seekers, the offer side, than to the number of job vacancies, the demand side. However, in both cases, the ones that explicitly advertise they are looking for a counterpart are also the ones that suffer the most from congestion. Directed search models may offer an explanation to this similarity. In these frameworks, information on price or wage is disclosed to signal quality or motivation in an attempt to attract buyers for a property and applicants for a job. In the housing market, Han and Strange (2016) show how the number of bids is sensitive to the listing price. The competition among sellers and firms through listings creates a negative externality on each other's matching rate as it generates an urn-ball friction. The competition is less explicit on the other end of the transaction, which might explain weaker congestion.

From this first estimation, we can drawn two preliminary conclusions. First, our results bring proof that the random matching framework is well suited for the housing market. The elasticities with respect to the measures of the numbers of buyers and sellers are positive and significant. Moreover, model (1) explains most of the variance in our panel of sales records of 44 urban areas, over 16 quarters. Second, the matching function we estimate thanks to our Internet-based measures of the numbers of market participants exhibits decreasing returns to scale in contrast with the constant return to scale hypothesis adopted in the theoretical literature. The following section assesses how robust these two conclusions are.

V. Robustness Tests

The static analysis of section III raises some possible shortcomings of the dataset we used in our estimation of the housing market matching function. The following presents some variations in the empirical specification that aim at assessing to which extent these limits may affect our result.

The first concern that is usually brought forward when datasets stemming from Internet. are used in an academic perspective regards their representativeness. Section III shows that the properties estimated by sellers and buyers differ significantly from the ones in sales records, across several dimensions. The Table 4 displays the estimations of model (1) based on datasets that result from the division of the housing markets considered in section IV along these dimensions. Columns (1) and (2) present respectively the estimations of the function within the sole core city of the metro area and the suburbs. We consider apartments and individual houses separately in columns (3) and (4). Columns (5), (6) and (7) each represent the transaction that occurred in the three price tiers of each metro area. Note that sellers and buyers are distributed within each price tiers in consideration with the estimated value of the AVM but that the tiers are computed on the transaction prices observed in the fiscal records. Finally, in column (8), as the Parisian urban area is structurally different in size and in MA penetration rate, we regress model (1) over the same dataset used for table 3 but without the Parisian area. As before, the markets - here market segments - with fewer than ten buyers and ten sellers in every contemplated quarter.

Our first results are confirmed. Elasticities with respect to buyers and sellers are positive and significant for all but one regression (the upper price tier for which the elasticity with respect to sellers is not significantly different from 0). The elasticity with respect to the number of buyers is larger than the elasticity with respect to sellers, for all splits but the lower tier segments where they are not significative different. Finally, we reject the constant returns to scale hypothesis in all regressions. Note that the returns to scale measured for these sub-markets are smaller than in table 3, except for column (8) which is the only one which does not focus on a specific market segment. That is primarily due to the elasticities with respect to the number of buyers, which are systematically lower. Because of the demand spillover between market segments, our measure of this number is biased upward after segmentation. As an example, a buyer who is visiting both apartments and houses is counted twice in regressions (1) and (2) but only once in the estimations of table 3. A weighted measure as proposed by Piazzesi et al.

(2020) would be better suited to account for this sub-markets spillover effect but is not in the scope of the present study. Overall, table 4 results consolidate section IV conclusions and rule representativeness issues out.

Dependent variable	ln(sales)							
	(1) Core City	(2) Suburbs	(3) Apartments	(4) Houses	(5) Low tier	(6) Middle tier	(7) Upper tier	(8) WO Paris
intercept	4.6704*** (0.4905)	4.7288*** (0.5747)	4.3253*** (0.6232)	5.2601*** (0.4535)	5.8704*** (0.4096)	6.0605*** (0.5101)	6.6238*** (1.0972)	3.8315*** (0.3961)
ln(buyers)	0.3456*** (0.0453)	0.3657*** (0.0390)	0.3581*** (0.0519)	0.3417*** (0.0394)	0.2184*** (0.0424)	0.3026*** (0.0392)	0.4118*** (0.0540)	0.4461*** (0.0396)
ln(sellers)	0.2342** (0.0716)	0.2455*** (0.0657)	0.2808*** (0.0818)	0.1819*** (0.0512)	0.2299*** (0.0596)	0.1083* (0.0477)	-0.0947 (0.1127)	0.2767*** (0.0620)
Metro FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Seasonality	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
RTS	0.5798***	0.6112***	0.6389***	0.5237***	0.4482***	0.4110***	0.3171***	0.7229***
Observations	540	540	525	630	300	540	555	645
R^2	0.982	0.977	0.982	0.973	0.984	0.980	0.960	0.976
Adj-R^2	0.980	0.975	0.981	0.971	0.982	0.979	0.957	0.974
Wald-stat	3487.1	4178.7	3449.5	4627.0	4442.8	4064.1	2218.8	3106.1

Standard errors between parenthesis
*** p<0.001, ** p<0.01, * p<0.05, ° p<0.1
RTS test is with H0=1
Errors are clustered at the metro area level and normal distribution is used for inference

Table 2-4: Model (1) OLS estimations over segments of the biggest French metro areas 2013-Q4 to 2017-Q2

As pointed in section IV, to hold table 3 elasticities estimates as unbiased, one has to make the hypothesis that the MA popularity is constant among users across metro areas. In the absence of any reference figure for the numbers of buyers and sellers outside of our own estimates, it is impossible to back this hypothesis with a definitive quantitative argument. Yet, to prove our results remains robust despite this,we change our empirical strategy and to move from a panel regression to a cross-section approach which does not rely on any intertemporal stability hypothesis. Table 5 presents the fifteen cross-metro area estimations of the housing market matching function, one for each quarter between 2013-Q4 to 2017-Q2.

Our two main conclusions from section IV are confirmed. The elasticities of the number of sales in a metro area with respect to buyers and sellers are both positive and statistically significant in all but one quarter and the returns to scale remain statistically inferior to one across all regressions. If the exact estimates of the elasticities vary from one quarter to the other, they remain close to the ones of table 3: 0.4-0.5 for buyers and 0.3-0.4 for sellers. Moreover,

they do not exhibit any particular pattern over time, which is reassuring with respect to our "constant popularity hypothesis".

Dependent variable	ln(sales)							
	2013-Q4	2014-Q1	2014-Q2	2014-Q3	2014-Q4	2015-Q1	2015-Q2	2015-Q3
intercept	2.8151*** (0.2274)	3.2855*** (0.2242)	3.0688*** (0.2602)	3.0302*** (0.2482)	2.7781*** (0.2536)	3.1566*** (0.2550)	3.1987*** (0.2587)	2.9955*** (0.2307)
ln(buyers)	0.3129* (0.1250)	0.5458*** (0.1093)	0.4615** (0.1474)	0.3613** (0.1083)	0.3929** (0.1345)	0.4204*** (0.1093)	0.6561*** (0.1388)	0.4902*** (0.1205)
ln(sellers)	0.5034*** (0.1405)	0.2233° (0.1258)	0.3411* (0.1628)	0.4198** (0.1241)	0.4305** (0.1493)	0.3646** (0.1287)	0.1531 (0.1553)	0.3199* (0.1373)
Metro FE	No	No	No	No	No	No	No	No
Seasonality	No	No	No	No	No	No	No	No
RTS	0.8162***	0.7691***	0.8026***	0.7811***	0.8233***	0.7850***	0.8092***	0.8101***
Observations	44	44	44	44	44	44	44	44
R^2	0.955	0.963	0.952	0.941	0.948	0.955	0.958	0.962
Adj-R^2	0.953	0.961	0.949	0.938	0.946	0.953	0.955	0.960
F-statistic	439.9	535.7	403.6	328.4	376.0	435.9	462.0	521.4
	2015-Q4	2016-Q1	2016-Q2	2016-Q3	2016-Q4	2017-Q1	2017-Q2	
intercept	2.8975*** (0.2347)	2.8187*** (0.2003)	3.3053*** (0.2309)	3.1337*** (0.2156)	3.0822*** (0.1979)	3.2887*** (0.2202)	3.3475*** (0.2243)	
ln(buyers)	0.4143** (0.1369)	0.2348** (0.0861)	0.5197*** (0.1054)	0.4558*** (0.1019)	0.4229*** (0.1049)	0.4327*** (0.0992)	0.4735*** (0.0982)	
ln(sellers)	0.3985* (0.1538)	0.6051*** (0.1014)	0.2723* (0.1219)	0.3324** (0.1171)	0.3945** (0.1185)	0.3638** (0.1179)	0.3120* (0.1170)	
Metro FE	No	No	No	No	No	No	No	
Seasonality	No	No	No	No	No	No	No	
RTS	0.8129***	0.8398***	0.7920***	0.7882***	0.8174***	0.7966***	0.7855***	
Observations	44	44	44	44	44	44	44	
R^2	0.955	0.964	0.958	0.961	0.962	0.962	0.956	
Adj-R^2	0.953	0.963	0.956	0.959	0.960	0.960	0.954	
F-statistic	434.8	553.5	469.2	502.7	520.6	521.0	448.5	

Standard errors between parenthesis
*** p<0.001, ** p<0.01, * p<0.05, ° p<0.1
RTS test is with H0=1
The student distribution is used for inference

Table 2-5: Model(1) cross-metro areas OLS estimates, each column corresponds to a quarter

Another temporal issue that may worry the reader. Our temporal aggregation over quarters with a one-period lag between the dependent variable and regressors imposes a fix three-month horizon for searches of buyers and sellers. Table 6 present alternative specifications with a monthly aggregation for column (1) and two- and three-period lags in

columns (2) and (3). Note that "the ten buyers and sellers" rule over every period reduces the number of available metro areas to thirty, and that each additional each lag withdraw a quarter. Table 6 results lead to the same conclusions as section IV: housing matching mechanism happens according to a Cobb-Douglass function with decreasing returns to scale. Yet, the fact that the elasticities of the number of sales in period *t* with respect to the sellers of periods *t-2* and *t-3*, than to the number of sellers in *t-1* period raises a question. A specification that takes into account the degree of maturity of sales project is proposed in section VII and help to explain this phenomenon.

Dependent variable	ln(sales)		
	(13) Monthly aggregation	(14) Two lags	(15) Three lags
intercept	5.4377*** (0.3863)	2.8116*** (0.5842)	4.5402*** (0.6293)
ln(buyers)	0.3912*** (0.0426)	0.3100*** (0.0400)	0.2167*** (0.0391)
ln(sellers)	0.0870* (0.0436)	0.5100*** (0.0657)	0.4102*** (0.0754)
Metro FE	Yes	Yes	Yes
Seasonality	Yes	Yes	Yes
RTS	0.4782***	0.8200**	0.6269***
Observations	1350	616	572
R^2	0.960	0.983	0.983
Adj-R^2	0.958	0.981	0.981
F-statistic	8144.6	2859.4	6151.9

Standard errors between parenthesis
*** p<0.001, ** p<0.01, * p<0.05, ° p<0.1
RTS test is with H0=1
Errors are clustered at the metro area level and normal distribution is used for inference

Table 2-6: Model(1) OLS estimations with alternative temporal specifications

The final dimension along which we want to check that our results are robust to variations is in the specifications of our measures of the numbers of active buyers and sellers in a market. Indeed, the way we categorize a user of the AVM as a seller or a buyer, if not totally arbitrary, could yet be challenged. The following table presents the results we obtain for model (1) estimation with slightly different rules in the way we label users. For both sellers and buyers we used three different definition based on the response of the users in AVM form. Buyer (1) follows the same definition as presented in section 3. Buyer (2) is a stricter definition that excludes the buyers that responded “I am wondering” or “I have just made an offer to the question “How far are you in your buying process?”, keeping only the ones that reply “I am

starting" or "I am actively searching". Buyer (3) is an even stricter definition in which we only consider the one that said they search actively. On the other hand, Seller (1) corresponds to the same definition as above, Seller (2) to a stricter definition ("As soon as possible" and "Within three months" sellers only) and Seller (3) to a wider definition (we add sellers replying "Within six months" to the main definition).

Dependent variable	ln(sales)							
	Buyer 2 Seller 1	Buyer 3 Seller 1	Buyer 1 Seller 2	Buyer 1 Seller 3	Buyer 2 Seller 2	Buyer 3 Seller 2	Buyer 2 Seller 3	Buyer 3 Seller 3
intercept	5.6237*** (0.5955)	5.2404*** (0.6899)	3.7217*** (0.4524)	2.5189*** (0.4987)	5.3965*** (0.5349)	5.0264*** (0.5194)	3.8012*** (0.5419)	3.2100*** (0.5510)
ln(buyers)	0.3201*** (0.0245)	0.2919*** (0.0271)	0.4444*** (0.0380)	0.3925*** (0.0361)	0.3123*** (0.0229)	0.2841*** (0.0248)	0.2714*** (0.0267)	0.2385*** (0.0275)
ln(sellers)	0.2176** (0.0726)	0.2986*** (0.0834)	0.2976*** (0.0565)	0.4477*** (0.0545)	0.2575*** (0.0648)	0.3397*** (0.0624)	0.4385*** (0.0676)	0.5394*** (0.0634)
Metro FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Seasonality	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
RTS	0.5376***	0.5905***	0.7420***	0.8402**	0.5698***	0.6238***	0.7099***	0.7779***
Observations	585	405	660	660	585	405	585	405
R^2	0.982	0.983	0.984	0.985	0.983	0.983	0.984	0.984
Adj-R^2	0.981	0.982	0.983	0.983	0.981	0.982	0.982	0.983
Wald-stat	4159.4	3556.2	8620.8	10639.0	6394.2	4976.1	2807.7	2339.5

Standard errors between parenthesis
*** p<0.001, ** p<0.01, * p<0.05, ° p<0.1
RTS test is with H0=1
Errors are clustered at the metro area level and normal distribution is used for inference

Table 2-7: Model (1) OLS estimations with alternative definitions for "Seller" and "Buyer"

Table 7 presents the results of the estimations of model (1) for the 8 possible combinations, with the exception of Buyer (1) and Seller (1), already in Table 3. Our two main conclusions from section IV prove to be robust to these changes in the way we compute our measure of the numbers of market participants. The variance in the magnitude of the estimated elasticities enhances the necessity of a specification of the matching function that accounts for the maturity of the searches of the participants, as in section VII below.

VI. Search Intensity of the Buyers

In the previous sections, we have brought a first empirical evidence that a Cobb-Douglass matching function with decreasing returns to scale is a good modeling for the searches of buyers and sellers in the French residential housing market. In this section, we offer to test an extension of this first model that accounts for the intensity of the searches of the buyers as

in Pissarides (2000, chap. 5). To do so, we add to model (1) $estimates\ per\ buyer_{i,t-1}$, the average number of estimates the buyers of metro area *i* performed on MA during quarter *t-1*. This quantity is a good measure of the average intensity of their search. Indeed the number of estimates a buyer does on MA can be seen as a proxy of the number of visits he or she does. Let us remember, users must describe the property and indicate its address to get an estimate. A buyer that do so is likely to have a visited it, or at least shows a particular interest in it. Hence, we estimate model (2) below, with the same specificity as what was presented for model (1) in section IV, and present the results in table 8.

$$\ln(sales_{i,t}) = \alpha \ln(buyers_{i,t-1}) + \beta \ln(sellers_{i,t-1}) + \gamma \ln(estimates\ per\ buyer_{i,t-1}) + Fixed\ Effects_{i,t} + \varepsilon_{i,t} \quad (2)$$

Dependent variable	ln(sales)			
	(1) No FE	(2) Seasonality FE	(3) Geo FE	(4) Full
intercept	3.1824*** (0.1679)	3.2794*** (0.1674)	2.9252*** (0.3746)	3.9291*** (0.5357)
ln(buyers)	0.5034*** (0.0626)	0.5211*** (0.0611)	0.4249*** (0.0414)	0.4528*** (0.0393)
ln(sellers)	0.3092*** (0.0676)	0.2880*** (0.0663)	0.3865*** (0.0492)	0.2627*** (0.0627)
ln(estimates per buyer)	-0.3613** (0.1114)	-0.3658** (0.1141)	-0.0801 (0.0762)	-0.1196° (0.0707)
Metro FE	No	No	Yes	Yes
Seasonality	No	Yes	No	Yes
RTS	0.8126***	0.8091***	0.8115***	0.7155***
Observations	660	660	660	660
R^2	0.948	0.952	0.980	0.984
Adj-R^2	0.948	0.952	0.978	0.982
Wald-statistic	142026.8	149521.7	240.0	6316.2

Standard errors between parenthesis
*** p<0.001, ** p<0.01, * p<0.05, ° p<0.1
RTS test is with H0=1
Errors are clustered at the metro area level and normal distribution is used for inference

Table 2-8: Model (2) OLS estimations over 44 metro areas from 2013-Q4 to 2017-Q2

First, let us note that results of section IV are left unchanged by the addition of our measure of intensity: elasticities with respect to buyers and sellers are both positive and significant, the former with greater magnitude than the latter, and returns to scale are decreasing. On the other hand, table 8 estimates of γ are all negative and it remains statistically significant at the 10% level, in column (4), as we control for the market specific fixed effect

and the seasonality. The results of robustness tests, similar to the one of section V, are gathered in Appendix B. If it fails to be statistically significant in most of them, the effect of the number of estimates per buyer remains negative it the vast majority of the variations we put model (2) through.

The fact our search intensity measure does not have a positive impact on the number of sales but a rather negative one may seem counterintuitive. A priori, it would seem fair to believe that the harder buyers search the more sales are agreed on. Search and matching mechanisms such as the ones introduced in Albrecht et al. (2003) and Albrecht et al. (2006), which accounts for two coordination failures that operate simultaneously, offer an explanation for this. In their model for a directed job search with multiple applications, they account first for the well-known urn ball friction, but also for the problem induced by job seekers applying to several positions simultaneously. The first implies that the most appealing vacancies receive more than one application while others do not receive any, the second that some open positions that have received applications fail to hire any worker in the end, as related applicants get hired elsewhere. Such a framework is perfectly suitable for housing search: buyers often consider several properties simultaneously and sellers organize visits for multiple potential buyers. The comparative static result of such a similar matching scheme presented in Albrecht, at al.(2003) shows how, past a certain point, an increase in the average number of applications, or visits in a real estate context, may reduce the total number of matches. Hence, the absence of any kind of coordination, that generates those two kinds of friction, can explain why buyers efforts fail to pay off.

It could also be that an increase in the buyers' efforts is a sign of a difficult market rather than a cause for a drop in the matching efficiency. It would be consistent with most empirical results from the job search literature, which finds that equivalent intensity measures, such as the number of job applications, to be counter-cyclical with the overall market activity. As matches get more difficult to form buyers, just as job seekers do, may search harder in an attempt to compensate for this. See Faberman and Kudlyak (2019) for a recent example using data from an online job board.

Finally, the stochastic matching modeling of the housing market proposed by Novy-Marx (2009) andGenesove and Han (2012), following Mortensen (2000 chap. 5), offers an alternative interpretation of the negative effect of $ln(estimatesperbuyer)$. In such frameworks, every match does not lead to a sale. A transaction happens only on the condition

of an ex-post match value, specific to each buyer/seller pair and unknown ex-ante, to exceed the sum of the seller's and buyer's reservation values. Hence, if an increase in the average number of visits per buyer is not a sign of a more intense search but a result of a decrease in this sale hazard rate conditional to a meeting, then a negative correlation with the number of sales make perfect sense. Note that such a decrease is likely to happen as markets grow, the acceleration in the expected meeting rate increasing the values of external options for both buyers and sellers. In that perspective, our results are in line withGenesove and Han (2012). They measure that positive demand shocks, proxied by income and population variations, have a negative impact on the buyer contract hazard rate.

Overall, the introduction of the number of estimates per buyers in our model, as a proxy of the number of visits, shows that housing matching is not totally random but also depends on the behaviors of market participants. Here we have been able to incorporate this proxy of the average intensity of the search on one side of the transaction. A natural and probably fruitful extension would be to include a measure of the degree of motivation their counterparts. Degree of over (or under) pricing as in Anglin et al. (2003) may represent such a proxy. Our dataset does not contain asking price information and MA is not well-known for its listings of for-sale properties. This extension is therefore left for future research.

VII. Market Participants' Characteristics

Beyond the matching at random, that are modeled by a simple function as in section IV, most of the theoretical descriptions of the housing search attach importance to the individual behaviors of the market participant. To have a glance on what factors influence the matching mechanism of the real estate market, outside of the sole numbers of buyers and sellers, we extend our first model to account for their characteristics. MA, as a company, being more focus on sellers, it is mostly the impact of their specifics on matching that we can study but not only. First, following the "stock-flow" matching approach developed and tested in Coles and Smith (1998) for the job market, we consider the impact of the share of buyers and sellers who are at an early stage of their respective project. Second, we account for the impact of the share of rental investment and secondary homes among the on-sale properties (buyers do not disclose the motivation of their purchased to MA). Finally, for the sellers whose properties are already on the market, we consider the proportion that sells by themselves without the help of a real estate agent (assisted buyers are almost non-existent in France because of regulatory issues).

The following model (3) is estimated. The results are presented in Table 9 and variations similar to the ones of section V are gathered in Appendix C.

$$\ln(sales_{i,t}) = \alpha \ln(buyers_{i,t-1}) + \beta \ln(sellers_{i,t-1}) + \omega_1\, early\ stage\ buyers + \omega_2\, early\ stage\ sellers\ + \omega_3\, rental\ investments + \omega_4\, vacation\ homes + \omega_5\, for\ sale\ by\ owners\ + Fixed\ Effects_{i,t} + \varepsilon_{i,t} \quad (3)$$

Where $earlystagebuyers$ is the ratio of buyers who answer "I am wondering" to the question: "How far are you in your buying process?"; $earlystagesellers$ is the share of sellers who answer "Within three months" to the question: "When do you plan to sell this property?"; $rentalinvestments$ and $vacationhomes$ are the share of sellers who reply "A rental investment" and "A secondary home" to the question: "What kind of property it is?"; $forsalebyowners$ is the share of sellers who reply "By myself" to the question: "How do you sell it?" among the ones who reply "I already started the sales" to the question: "When do you plan to sell this property?".

The main results of the section IV are confirmed. Table 9 brings evidence of the impact of participants' heterogeneity on the matching process efficiency, especially regarding sellers. In models (1), (2), (5) and (7) the share of sellers in an early stage of their project impacts positively the number of sales, whereas the share of buyers in an early stage impacts it negatively. This effect is consistently significant for sellers, at the 0.1% level, and remains so in the robustness tests of Appendix C.

This result regarding sellers is in line with the so-called stock-flow approach introduced in Coles and Smith (1998), for the job market matching. They argue that exit rates are higher for participants who just enter the market because job seekers and firms with vacancies scan the entire market as they enter. Thus, participants who do not match in their first round would have to wait for a newcomer to get them out of the market. A similar argument can be made for property that just entered the market, by contrast with a property that has been for sale for months and most potential buyers already toured. Of course, this represents only first evidence of stock-flow matching that remains to be confirmed through a custom specification as in Coles and Petrongolo (2008).

Dependent variable	ln(sales)						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
intercept	5.0068*** (0.5637)	4.6909*** (0.5318)	4.6499*** (0.5391)	4.8742*** (0.5362)	4.7582*** (0.5227)	4.3327*** (0.4520)	4.4001*** (0.4497)
ln(buyers)	0.4376*** (0.0369)	0.4157*** (0.0369)	0.4377*** (0.0388)	0.4128*** (0.0355)	0.4137*** (0.0366)	0.3894*** (0.0365)	0.3866*** (0.0360)
ln(sellers)	0.1662** (0.0610)	0.1908** (0.0617)	0.1878*** (0.0555)	0.1971*** (0.0574)	0.1876** (0.0602)	0.2365*** (0.0491)	0.2340*** (0.0480)
early stage buyers	-0.1311 (0.1192)				-0.2001° (0.1101)		-0.2099* (0.1004)
early stage sellers		1.1113*** (0.2110)			1.1352*** (0.2116)	1.0821*** (0.1917)	1.1080*** (0.1898)
rental investment			0.9146*** (0.2065)			0.8386*** (0.2082)	0.8324*** (0.2022)
vacation home			0.2215 (0.2954)			0.1800 (0.2929)	0.1945 (0.2893)
FSBO				0.1826** (0.0562)		0.1846*** (0.0512)	0.1896*** (0.0483)
Metro FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Seasonality	Yes	Yes	Yes	Yes	Yes	Yes	Yes
RTS	0.6038***	0.6065***	0.6256***	0.6099***	0.6014***	0.6259***	0.6206***
Observations	572	572	572	572	572	572	572
R^2	0.985	0.987	0.986	0.986	0.987	0.988	0.988
Adj-R^2	0.984	0.985	0.985	0.984	0.986	0.987	0.987
Wald-statistic	6721.9	5362.5	5251.1	5262.9	5687.4	4229.9	4185.7

Standard errors between parenthesis
*** p<0.001, ** p<0.01, * p<0.05, ° p<0.1
RTS test is with H_0=1
Errors are clustered at the metro area level and normal distribution is used for inference

Table 2-9: Model (3) estimation over 44 metro areas from 2014-Q2 to 2017-Q2

This is also in line with the idea that a long time-on-market creates a stigma that drives potential buyers away. Applying the works of Banerjee (1992) and Bikhchandani et al. (1992) on herding to real estate, Taylor (1999) argues that a property that stays too long on the market is interpreted as a signal of bad quality, which decreases the sale probability. These predictions find an empirical confirmation in the effect of a policy that ban relisting implanted in Massachusetts studied in Tucker et al. (2013).

On the other hand, early-stage buyers would be less likely to match according to the results of table 9. A reason could be that the implied perfect information hypothesis of the stock-flow approach does not hold here. Contrary to job seekers who usually know which job to apply, homebuyers do not know, at first, which homes will suit them. On the contrary, they need time to learn and adjust their preferences and to reduce the incertitude of a deeply committing

financial decision. Another convincing explanation is the "opportunist matching" mechanism introduces in Albrecht et al. (2007): over time, following a Poisson process, participants go from relaxed to desperate, and strongly decrease their flow value of being unmatched, which lead them to accept matches they would have declined previously. The "wondering" buyers as they just enter the market seem way more likely to be in a "relax" state than others. Note that if this effect remains negative and somehow significant in most segments (robustness test of Appendix C), its sign changes and grows in magnitude as we increase the number of lags between the dependant variable and regressor. Far from discounting our results, this finding make sense as a *early stage buyer* in a period becomes *active* in the next, while the ones with more mature project may have left the market.

The dataset collected on MA enables a further investigation of the impact of the heterogeneity among sellers. Models (3), (6) and (7) (as of most regressions in Appendix C) show that the share of sellers who put a *rental investment* property on the market, rather than a *permanent residence*, has a positive and significant impact on the number of sales. The magnitude of the coefficient implies that rental investors are twice more efficient than standard sellers. Such sellers focus only on maximizing the return on their investment and attach less importance to the non-financial utility. They may also be less subject to the endowment effect and either have an informational advantage in comparison to the rest of the market or present a preference for liquidity over price.

We measure that the larger the ratio of FSBO among the properties on the market the higher the sales number. This result, which appears to be robust to variation presented in Appendix C, was unexpected. The role of real estate agents is supposed to be facilitators, a role they appear not to perfectly fulfill according to this result. A reasonable argument made in Han and Strange (2015), for this negative effect on the number of matches, is that the commission charged by realtors make some otherwise mutually beneficial transactions impossible. Without an intermediary involved, for the transaction to happen, the buyer's valuation must exceed the seller's. With an intermediary, it must also be over the fees in addition to the seller's valuation. Empirically, Dachis et al. (2011) exploits the land transfer tax increase to estimate the impact of transfer costs. They measure that a 1.1% increase in the tax causes a drop of 15% in the sales volumes. Of course, unlike transfer taxes, realtor fees come with informational and organizational benefits that should compensate their costs, making the comparison at most partial.

On the benefits of using an agent, Hendel et al. (2009) show they decrease TOM and increase the probability of sale compared to an FSBO platform, in the city of Madison, Wisconsin from 1998 to 2005, which seems counterfactual with our results. However, they measure that about a quarter of the home first listed as FSBO moves latter to MLS, whereas the opposite almost never happen. They interpret this phenomenon as evidence of a stock-flow kind of matching. Sellers first prefer to match with the impatient and motivated buyers that are looking for home on FSBO. If they fail, they migrate to MLS and its larger stock of buyers. A low degree of confidence in agents 'ability to find a good match, or a deficiency generated by the misalignment of their respective interests, are reasons for a seller to prefer selling by himself. According to Larceneux et al. (2015), the French market exhibits an extreme illustration of this phenomenon with 70% of the sellers in their studies first listing their property on an FSBO portal (against 25% in Hendel et al. 2009) but with 67% of them finally selling with an agent. This opens up a way for an alternative explanation for the positive correlation of the higher share of FSBO with the number of matches. If sellers massively prefer to sell their home on their own first and hire a realtor only if they fail to do so, then a higher proportion of sales involving a realtor is rather a consequence of a market with complicated matching than a cause.

Overall, section VII confirms that mechanism overlooked by the random matching modeling, which gives a good macro understanding of how housing market function, actually play a role. We obviously cannot aspire to exhaust the subject here. The empirical study of the effects of the many characteristics of participants through data from web platforms is a promising field of research. In particular, extending our investigations by taking into account the specificities of buyers (first time versus experienced, out-of-town versus local, just to name a few) seems a first natural extension of our work.

VIII. Conclusion

This article introduces a new dataset extracted from an analysis of the traffic of a French real estate web platform. This dataset allows us to measure the numbers of buyers and sellers active in the housing market of forty-four of the largest urban areas in France, from 2014 to mid-2017, on a quarterly basis. With these measures and an exact count of the real estate transactions that happens in these areas, we compute the first estimation of the housing market matching function.

The first and most obvious conclusion we can draw from this estimation is that it brings evidence that the matching framework is well suited to depict how the housing market works. The most simple Cobb-Douglas using our measure of the numbers of sellers and buyers fits the data well and exhibits meaningful and statistically significant estimates for their respective elasticity. On top of that, this first estimation brings insightful new results. Foremost, in contrast to the standard assumption of constant returns to scale, that stems from empirical evidences specific to the job market, we find that the housing matching function has decreasing returns to scale. The congestion dominates the enhancement of the matching opportunities from a larger pool of possible counterparts. Not only is this result stable and robust in a basic specification, it is also effective in richer models that account for search intensity or heterogeneity among participants.

Beyond the specific effects that have been analyzed and discussed in section VI and VIII, these two extensions show that the behavior of the market participants and their characteristics matter. As a first approximation, random matching enabled a good understanding of the way the real estate residential market behaves at a macro level. Yet, considering such second-order factors shows that the way buyers and sellers meet and exchange properties is not fully random and unravel elements of the housing market microstructure. In our case, the negative effect on the matching rate of our proxy of the number of visits per buyer advocates for the existence of two coordination failures happening simultaneously on both side of the transaction. Also, in line with the stock-flow matching approach, we find that a greater proportion of sellers in an early stage of their search increases the number of matches, while, on the contrary, the share of early-stage buyers impacts it negatively. It can be explained by the time they might need to learn and adjust their preferences or caused by a transition from a "relax" to a "desperate" state, as in Albrecht et al. (2007). We also find that a greater proportion of rental investors among sellers produces more transactions and that the share of sellers with no realtor is negatively correlated with the number of sales.

The contribution of the paper is twofold. First, the estimation of the matching function of the housing market fills a gap in the empirical housing market microstructure literature. As a fundamental element in the matching framework, such an estimation called for confirmation. With the Internet taking a predominant position in the housing selling and buying process, a similar methodology appears replicable on other platforms. A comparison between countries

with different market organizations and institutions should also be fruitful. Such a comparison may allow researchers to identify features to improve the liquidity of the housing market.

Second, it participates in the growing econometrics literature that leverages Internet datasets to tackle open research questions. After *nowcasting* researches that used activity on generic Internet tools such as Google to follow influenza outbreaks (Preis and Moat 2014) to predict employment figures (Ettredge et al. 2005 ; Askistas and Zimmerman 2009) or private consumption (Kholodin et al. 2010), recent studies use specific web platform to address open empirical questions. The real estate literature like Piazzesi et al. (2020) or the present paper follows the path laid out by labor market economists, such as Iona Marinescu and her coauthors (Marinescu and Rathelot 2018, Azar et al. 2020, Marinescu and Wolthoff 2020), who use data from online job boards to enhance our understanding of the job market. The portion of the human activities that happens on Internet platforms has been growing in many segments in recent years. These platforms, as digital representation of the economy with every interaction and action of market participants logged, represent an opportunity for econometrics researches in the years to come.

Appendix

Appendix A: Evolution of the numbers of estimates and users of the AVM from 2011 to 2019

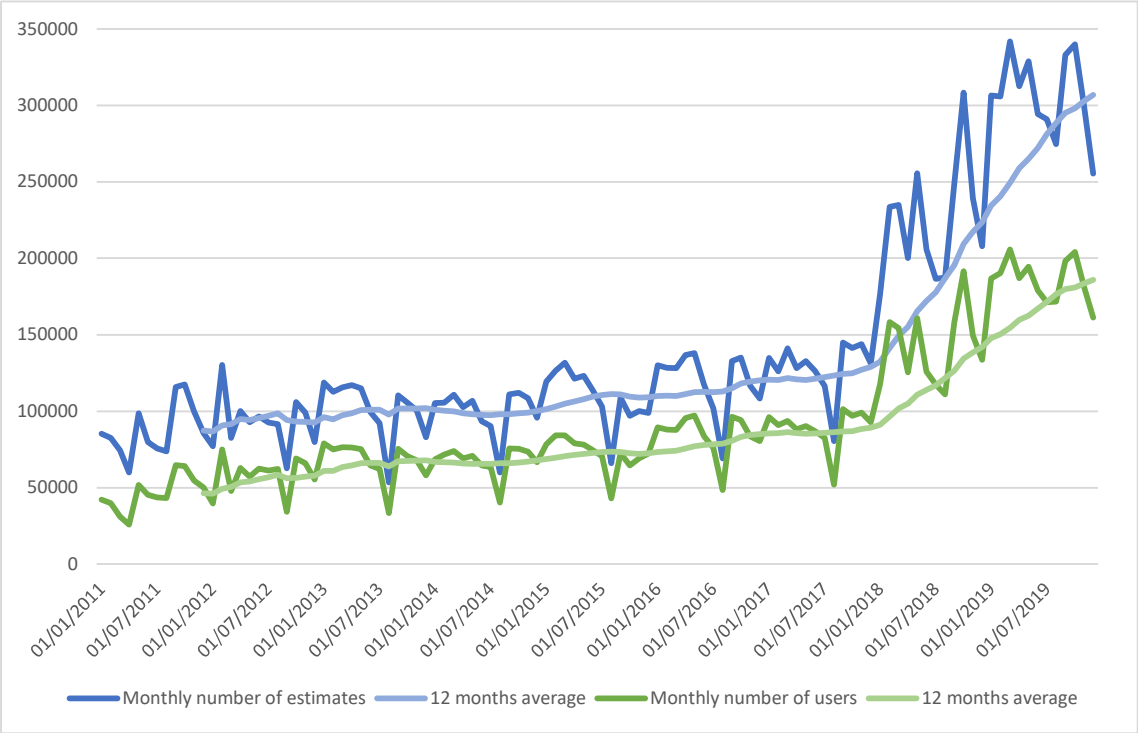

Figure 2-1: Evolution of the monthly numbers of estimates and users of the Meilleurs Agents AVM from january 2011 to december 2019

The measures of the numbers of sellers and buyers used in the present study are within the mid-2013 to late 2017 period of the platform stable popularity. One can notice than during this period both the average level and the periodicity of the numbers of estimates and users are steady. We interpret the moderate increase in these numbers from the spring 2015 to 2017 as an effect of the transition from a bust to a boom market at that time. The more turbulent periods that precede and follow the moment we perform our measure correspond, on the one hand, to the initial growth of the company and the extension of its activities over the whole country and, on the other hand, to the beginning of a television ads campaign.

Appendix B: Robustness tests Model (2)

Dependent variable	ln(sales)							
	(1) Core City	(2) Suburbs	(3) Apartments	(4) Houses	(5) Low tier	(6) Middle tier	(7) Upper tier	(8) WO Paris
intercept	4.7025*** (0.4868)	4.7009*** (0.5844)	4.3144*** (0.6201)	5.2443*** (0.4559)	5.8798*** (0.4135)	6.0599*** (0.5117)	6.6330*** (1.0935)	3.8309*** (0.3926)
ln(buyers)	0.3447*** (0.0450)	0.3660*** (0.0394)	0.3613*** (0.0501)	0.3422*** (0.0392)	0.2174*** (0.0423)	0.3026*** (0.0392)	0.4136*** (0.0536)	0.4471*** (0.0389)
ln(sellers)	0.2351*** (0.0702)	0.2502*** (0.0674)	0.2815*** (0.0810)	0.1854*** (0.0524)	0.2287*** (0.0599)	0.1078* (0.0478)	-0.0944 (0.1122)	0.2809*** (0.0606)
ln(estimates per buyer)	-0.0777 (0.0573)	-0.0542 (0.0946)	-0.0560 (0.0820)	-0.0628 (0.0709)	0.0258 (0.0720)	0.0166 (0.0645)	-0.0697 (0.1080)	-0.1134 (0.0705)
Metro FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Seasonality	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
RTS	0.5799***	0.6162***	0.6429***	0.5276***	0.4460***	0.4104***	0.3192***	0.7280***
Observations	540	540	525	630	300	540	555	645
R^2	0.982	0.977	0.982	0.973	0.984	0.980	0.960	0.976
Adj-R^2	0.980	0.975	0.981	0.971	0.982	0.979	0.957	0.974
Wald-statistic	3986.6	4246.6	3612.4	4693.1	4270.9	4259.6	2197.9	3185.8

Standard errors between parenthesis
*** p<0.001, ** p<0.01, * p<0.05, ° p<0.1
RTS test is with H0=1
Errors are clustered at the metro area level and normal distribution is used for inference

Table 2-10: Model(2) OLS estimations over segments of the biggest French metro areas 2013-Q4 to 2017-Q2

Dependent variable	ln(sales)		
	(1) Month	(2) Two lags	(3) Three lags
intercept	5.4372*** (0.3862)	2.8229*** (0.5769)	4.5426*** (0.6271)
ln(buyers)	0.3914*** (0.0426)	0.3114*** (0.0403)	0.2172*** (0.0391)
ln(sellers)	0.0873* (0.0434)	0.5105*** (0.0649)	0.4105*** (0.0754)
ln(estimates per buyer)	-0.0125 (0.0707)	-0.0715 (0.0707)	-0.0224 (0.0668)
Metro FE	Yes	Yes	Yes
Seasonality	Yes	Yes	Yes
RTS	0.4787***	0.8219**	0.6277***
Observations	1350	616	572
R^2	0.960	0.983	0.983
Adj-R^2	0.958	0.981	0.981
Wald-statistic	8057.5	2856.6	6017.0

Standard errors between parenthesis
*** p<0.001, ** p<0.01, * p<0.05, ° p<0.1
RTS test is with H0=1
Errors are clustered at the metro area level and normal distribution is used for inference

Table 2-11: Model(2) OLS estimations with alternative temporal specifications

Dependent variable	ln(sales)							
	2013-Q4	2014-Q1	2014-Q2	2014-Q3	2014-Q4	2015-Q1	2015-Q2	2015-Q3
intercept	2.9472*** (0.2344)	3.3129*** (0.2043)	3.2690*** (0.2842)	3.1446*** (0.2470)	2.9165*** (0.2626)	3.0177*** (0.2932)	3.2428*** (0.2571)	3.0041*** (0.2356)
ln(buyers)	0.3487** (0.1237)	0.5036*** (0.1005)	0.4959** (0.1463)	0.3867*** (0.1055)	0.4229** (0.1331)	0.3907** (0.1137)	0.6318*** (0.1380)	0.4830*** (0.1249)
ln(sellers)	0.4816** (0.1377)	0.2907* (0.1166)	0.3115° (0.1608)	0.4130** (0.1201)	0.4097** (0.1469)	0.4010** (0.1342)	0.1945 (0.1559)	0.3297* (0.1438)
ln(estimates per buyer)	-0.5418° (0.3097)	-0.6143** (0.1998)	-0.6273 (0.3911)	-0.6184° (0.3150)	-0.5785 (0.3540)	0.2398 (0.2489)	-0.4900 (0.3383)	-0.0812 (0.3105)
Metro FE	No	No	No	No	No	No	No	No
Seasonality	No	No	No	No	No	No	No	No
RTS	0.8303***	0.7943***	0.8073***	0.7997***	0.8327***	0.7916***	0.8263***	0.8127***
Observations	44	44	44	44	44	44	44	44
R^2	0.959	0.970	0.955	0.946	0.952	0.956	0.960	0.962
Adj-R^2	0.956	0.968	0.951	0.942	0.948	0.953	0.957	0.959
F-statistic	309.0	434.0	280.2	235.5	261.8	290.4	316.9	339.7
	2015-Q4	2016-Q1	2016-Q2	2016-Q3	2016-Q4	2017-Q1	2017-Q2	
intercept	2.9827*** (0.2377)	2.8328*** (0.2061)	3.5052*** (0.2412)	3.2504*** (0.2244)	3.0961*** (0.2030)	3.3106*** (0.1925)	3.3639*** (0.2257)	
ln(buyers)	0.4394** (0.1357)	0.2339* (0.0871)	0.5635*** (0.1034)	0.4520*** (0.1002)	0.4251*** (0.1061)	0.4200*** (0.0867)	0.4665*** (0.0988)	
ln(sellers)	0.3901* (0.1515)	0.6092*** (0.1031)	0.2294° (0.1189)	0.3392** (0.1151)	0.3953** (0.1198)	0.4195*** (0.1041)	0.3336** (0.1200)	
ln(estimates per buyer)	-0.5328 (0.3497)	-0.1027 (0.2813)	-0.4918* (0.2333)	-0.4080 (0.2593)	-0.1015 (0.2538)	-0.9088*** (0.2455)	-0.3384 (0.3868)	
Metro FE	No	No	No	No	No	No	No	
Seasonality	No	No	No	No	No	No	No	
RTS	0.8294***	0.8431***	0.7929***	0.7913***	0.8204***	0.8394***	0.8002***	
Observations	44	44	44	44	44	44	44	
R^2	0.957	0.964	0.962	0.963	0.962	0.972	0.957	
Adj-R^2	0.954	0.962	0.960	0.960	0.959	0.970	0.954	
F-statistic	300.0	361.3	340.6	348.0	340.0	459.6	297.5	

Standard errors between parenthesis
*** p<0.001, ** p<0.01, * p<0.05, ° p<0.1
RTS test is with H0=1
Student distribution is used for inference

Table 2-12: Model(2) cross-metro areas OLS estimates, each column corresponds to a a given quarter

Dependent variable	ln(sales)							
	Buyer 2 Seller 1	Buyer 3 Seller 1	Buyer 1 Seller 2	Buyer 1 Seller 3	Buyer 2 Seller 2	Buyer 3 Seller 2	Buyer 2 Seller 3	Buyer 3 Seller 3
intercept	5.6128*** (0.5960)	5.2903*** (0.6785)	3.7170*** (0.4564)	2.5270*** (0.4932)	5.3768*** (0.5340)	5.0744*** (0.4907)	3.7625*** (0.5506)	3.2537*** (0.5411)
ln(buyers)	0.3202*** (0.0245)	0.2898*** (0.0268)	0.4454*** (0.0372)	0.3936*** (0.0352)	0.3124*** (0.0230)	0.2825*** (0.0248)	0.2713*** (0.0269)	0.2370*** (0.0273)
ln(sellers)	0.2191** (0.0731)	0.2922*** (0.0823)	0.3029*** (0.0558)	0.4510*** (0.0526)	0.2605*** (0.0651)	0.3333*** (0.0592)	0.4438*** (0.0693)	0.5338*** (0.0627)
ln(estimates per buyer)	-0.0096 (0.0502)	0.0597 (0.0608)	-0.1249° (0.0703)	-0.1201° (0.0661)	-0.0188 (0.0491)	0.0505 (0.0574)	-0.0299 (0.0495)	0.0497 (0.0592)
Metro FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Seasonality	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
RTS	0.5392***	0.5820***	0.7483***	0.8446**	0.5729***	0.6158***	0.7151***	0.7708***
Observations	585	405	660	660	585	405	585	405
R^2	0.982	0.983	0.984	0.985	0.983	0.983	0.984	0.984
Adj-R^2	0.981	0.982	0.983	0.984	0.981	0.982	0.982	0.983
Wald-statistic	4470.3	3609.1	8827.0	10660.3	6576.0	4981.9	3141.6	2364.2

Standard errors between parenthesis
*** p<0.001, ** p<0.01, * p<0.05, ° p<0.1
RTS test is with H0=1
Errors are clustered at the metro area level and normal distribution is used for inference

Table 2-13: Model(2) OLS estimations with alternative definitions for "Seller" and "Buyer"

Appendix C: Robustness tests Model (3)

Dependent variable	ln(sales)							
	(1) Core City	(2) Suburbs	(3) Apartments	(4) Houses	(5) Low tier	(6) Middle tier	(7) Upper tier	(8) WO Paris
intercept	4.7501*** (0.5629)	5.4130*** (0.5452)	4.4008*** (0.5405)	5.8792*** (0.4059)	5.9847*** (0.4101)	6.5524*** (0.5261)	6.5389*** (0.9865)	4.7501*** (0.5629)
ln(buyers)	0.3190*** (0.0396)	0.3263*** (0.0423)	0.3132*** (0.0460)	0.2878*** (0.0377)	0.1912*** (0.0432)	0.2806*** (0.0396)	0.3734*** (0.0490)	0.3190*** (0.0396)
ln(sellers)	0.2232** (0.0696)	0.1755** (0.0591)	0.2958*** (0.0705)	0.1371** (0.0469)	0.2474*** (0.0551)	0.0660 (0.0525)	-0.0801 (0.1083)	0.2232** (0.0696)
early stage buyers	-0.0243 (0.1103)	-0.0949 (0.1273)	0.0373 (0.1437)	-0.2744* (0.1083)	-0.0724 (0.1142)	-0.2814* (0.1382)	-0.0816 (0.2070)	-0.0243 (0.1103)
early stage sellers	0.8186*** (0.2028)	0.9696*** (0.1689)	0.5887** (0.1821)	0.9039*** (0.1423)	0.0843 (0.1349)	0.5239*** (0.1280)	1.1668*** (0.2301)	0.8186*** (0.2028)
rental investment	0.3655* (0.1769)	0.7999** (0.2899)	0.3699* (0.1653)	0.6874* (0.2793)	-0.0394 (0.1369)	0.0514 (0.1589)	0.1833 (0.3207)	0.3655* (0.1769)
vacation home	0.1314 (0.2617)	-0.0465 (0.3820)	-0.0706 (0.2489)	0.1721 (0.3222)	-0.1013 (0.2256)	-0.1621 (0.2652)	0.0038 (0.4035)	0.1314 (0.2617)
FSBO	0.1045 (0.0640)	0.1783** (0.0543)	0.1661** (0.0537)	0.1338*** (0.0305)	0.0094 (0.0582)	0.0670 (0.0424)	0.2062** (0.0798)	0.1045 (0.0640)
Metro FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Seasonality	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
RTS	0.5422***	0.5018***	0.6090***	0.4249***	0.4386***	0.3466***	0.2932***	0.5422***
Nb Observation	468	468	455	546	260	468	481	468
R^2	0.986	0.983	0.986	0.980	0.986	0.984	0.969	0.986
Adj-R^2	0.984	0.981	0.984	0.978	0.984	0.982	0.966	0.984
Wald-statistic	2658.8	3270.2	3467.7	3763.7	4094.1	3507.3	3681.8	2561.7

Standard errors between parenthesis
*** p<0.001, ** p<0.01, * p<0.05, ° p<0.1
RTS test is with H0=1
Errors are clustered at the metro area level and normal distribution is used for inference

Table 2-14: Model(3) OLS estimations over segments of the biggest French metro areas 2014-Q2 to 2017-Q2

Dependent variable	ln(sales)		
	(1) Month	(2) Two lags	(3) Three lags
intercept	5.9942*** (0.3476)	3.7295*** (0.5308)	5.2937*** (0.5372)
ln(buyers)	0.3167*** (0.0348)	0.2586*** (0.0342)	0.1846*** (0.0334)
ln(sellers)	0.0626 (0.0402)	0.4409*** (0.0579)	0.3456*** (0.0598)
early stage buyers	-0.0004 (0.0643)	0.2702** (0.0827)	0.6484*** (0.0879)
early stage sellers	0.7919*** (0.0965)	0.6264*** (0.1593)	0.4196* (0.1985)
rental investment	0.4791*** (0.0948)	0.5516* (0.2560)	0.2308 (0.2282)
vacation home	-0.0963 (0.2161)	-0.5638* (0.2344)	-1.0696** (0.3367)
FSBO	0.1471*** (0.0332)	0.2542*** (0.0586)	0.2381** (0.0785)
Metro FE	Yes	Yes	Yes
Seasonality	Yes	Yes	Yes
RTS	0.3793***	0.6995***	0.5301***
Observations	1170	572	572
R^2	0.976	0.987	0.986
Adj-R^2	0.975	0.986	0.985
Wald-statistic	11596.5	5436.1	8096.1

Standard errors between parenthesis
*** p<0.001, ** p<0.01, * p<0.05, ° p<0.1
RTS test is with H0=1
Errors are clustered at the metro area level and normal distribution is used for inference

Table 2-15: Model(3) OLS estimations with alternative temporal specifications

3. The home buying problem: evidence from the Internet

Abstract:

Search models offer a framework to explain why different buyers pay different prices for similar homes. Previous studies focus on the premium paid by distant and first-time buyers. We introduce a new dataset based on users´ behavior on MeilleursAgents.com, a French real estate website. It contains the purchases of 2,372 apartments in the Parisian Region. Besides the apartment price and characteristics, this dataset contains information about the searches that led to those purchases. The main focus of this paper is to explore how the search for a home influences the outcome of the negotiation. According to search-with-learning theories, we find that if the buyers encounter higher (resp. lower) prices during their search then they will value and thus pay a higher (resp. lower) price for a given apartment. We also test the effects on the transaction price of the buyer's time-on-market and the size of the market segment scanned.

Résumé :

Les modèles de recherche offrent en cadre théorique pour expliquer pourquoi des acheteurs paient des prix différents pour des biens immobiliers similaires. Les études antérieures se concentrent sur la prime que paient les acheteurs venant de loin et les primo-accédants. Nous présentons ici un nouveau jeu de données basé sur le comportement d'utilisateurs sur le site MeilleursAgents.com, une plateforme immobilière française. Ces données documentent l'achat de 2 372 appartements de région parisienne. Au-delà des prix et des caractéristiques des biens, sont renseignées des indications sur les parcours de recherche qui aboutissent à ces acquisitions. L'objectif principal de ce papier est d'explorer comment la recherche d'un bien influence l'issue de la négociation lors de l'achat. En accord avec la théorie de la recherche avec apprentissage, nous montrons que les acheteurs ayant été confrontés à des prix plus (resp. moins) élevés pendant leur recherche paient plus (resp. moins) cher pour un appartement donné. Nous testons également les effets du temps passé sur le marché et de la taille du segment envisagé sur le montant de la transaction.

I. Introduction

While real estate assets usually represent a large share of a standard household, the process of buying a house or an apartment is crucial decision. As consumers only make that decision a few times over their life at most, most sellers and buyers are “amateurs” in the housing.

Starting with Cubbin (1974) and Miller (1978) who bring evidence of the price and liquidity relationship, the literature addressing the seller side of the transaction is rich. Salant (1991) is the first to use a theoretical sequential search framework for the home selling problem. Yavas and Yang (1995) stretch the strategic importance of the listing price. Anglin et al. (2003) show how the asking price “degree of overpricing” compared to a hedonic sale price impacts the seller time-on-market and Carillo (2012) and Merlo et al. (2015) develop rich directed search models of the problem.

In comparison, the home buying problem is overlooked in the literature, at least empirically. Some noteworthy exceptions are Turnbull and Sirmans (1993), Lambson et al. (2004) and Ihlandfeldt and Mayok (2012) who match transaction data and information about buyers within a search model framework. Their focus is to understand how the geographical origin and previous experience of buyers affect the price they end up paying. Turnbull and Sirmans (1993) conclude that the housing market is efficient as they do not measure first-time buyers nor out-of-town buyers to pay a premium. On the contrary, Lambson et al. (2004) find that out-of-state buyers pay a higher price for a given apartment but fail to identify whether it is caused by anchoring or information asymmetry. Finally, Ihlandfeldt and Mayok (2012) measure that both anchoring and search cost, estimated by the distance from the previous residence, impact the transaction price.

Anglin (1997) directly studies the determinants of the buyer’s search through survey data. He investigates how some precise features of the search for a home, such as the market segments considered or the sources of information used, affect its intensity. He measures that information-related variables have significant effects on the time-on-market and the count of visited houses whereas buyers’ personal characteristics do not. To our knowledge, Wilhelmsson (2008) is the only research that empirically investigates the impact of the buyers ’searches on the price they pay. He finds that the buyers who tour more homes prior to the purchase pay less *ceteris paribus*.

The reason for the apparent lack of interest for home buyers in the literature is the absence of data on their search. Sellers signal themselves through listing, buyers do not. For a long time, they have not left any trace of their searching process such as how long they have been looking for a home or which other properties they have visited. The prominence of the Internet in the buyers 'search has changed that. Piazzesi et al. (2020) is an example of how academic research can leverage Internet data. They use alert subscriptions on the listing website trulia.com to study the segmentation in the San Francisco housing market.

The present study aims at better understanding the impact of the buyer search on the price paid at the bottom line. Theoretically, we develop a search model in line with the ones introduces by Turnbull and Sirmans (1993), Lambson et al. (2004). Besides search costs and beliefs about the price, our model takes into account the hazard rate of meeting a seller and the size of the market segment searched. It states that the segment's size impacts the transaction price through two competing channels: a larger segment increases the encounter rate but may increase the search cost. Empirically, we use data from the French real estate platform MeilleursAgents.com (thereafter MA) to trace back the story behind 2,372 apartment purchases in the Parisian region. We collect information from buyers' use of MA on-line estimation tool. It allows to know if home buyers already owned a property prior to their purchase and, if so its location and type, and which other properties they have visited during their search. Being able to observe the home buyers 'search process is the key novelty of this study.

Our principal finding is that buyers who have visited on average more expensive apartments than the one they purchase pay a premium - the opposite being true to a lesser extent. Considering the price per square meter, this finding is in line with search-with-learning theory developed in Kohn and Shavell (1974). We measure that buyers subject to budget constraints are likely to pay less for the same home and confirm the anchoring effect from higher prices market, already observed in Lambson et al. (2004) and Ihlandfeldt and Mayok (2012). Although not entirely robust, we find first evidence of a negative correlation between the buyers 'searches duration and the transaction prices. Finally, the impact of the segment size on price is tested.

Following this introduction, this paper is organized as follows. Section II presents a theoretical framework for the home buying problem. From this model we derive predictions about the effects on transaction price of several search characteristics. As we use unusual data, section III presents how we collect it and build the dataset used in this study. Section IV deepens

the variables that describe key characteristics of the buyer search. Section V exhibits our empirical results and section VI tests its robustness. Section VII concludes.

II. Theoretical Model

We consider a sequential search model with infinite horizon, similar to the ones developed in Turnbull and Sirmans (1993) or Lambson et al. (2004). As in previous models, the buyers differ from one another in their search costs, their belief about the price distribution and their discount rate of the future. In addition, our model takes into account differences in the size of the segments they are scanning and the market tightness of these segments.

The setup is as follows, a buyer enters a housing market with sellers ready to sell their homes for different prices. These reservation prices follow a distribution $F(p)$ with $p \in]0; +\infty[$. The buyer does not know each seller's price until she meets them, but she has a belief about the price distribution: $F(p, b)$. The parameter b represents a shift in the perceived probability compared to the true distribution $F(p) = F(p, 0)$. A positive (resp. negative) b implies that she believes the prices are higher (resp. lower) than they actually are. Thus, $F(p, b)$ is the probability expected by a buyer with beliefs b that any seller will sell his home for less than a price p and $f(p, b)$ is the corresponding probability density.

A search for a seller is costly and uncertain: a buyer meets sellers according to Poisson arrival rate of parameter $\lambda(s)$ and each meeting generates to her a fixed cost $c(s)$, where s is the size of the inventory of the market segment scanned. The fixed costs c account for information acquisition costs and the time and resource-consuming costs of physically inspecting a property. It depends both on the buyer herself and, for a given buyer, in the size of the market segment scanned by the buyer. On the buyer, because she can live far from the market or very close. Also, she can be knowledgeable or have no clue about this real estate market and thus needs more time and effort to gather the information necessary to make a decision. On the size of the segment, because a geographically extended market segment implies more time-consuming visits or because a broad diversity within the segment increases the information acquisition cost. The encounter rate λ depends only on the market segment, the larger it is and the less it is searched by other buyers, the shorter the buyer has to wait to meet a new seller. We note τ the time until the next visit.

The buyer expects to derive the same value v from every property on the segment she searches and she discounts the future at a rate r. Her problem is to find an optimal stopping rule p^*. This rule implies she will continue to search until she meets a seller willing to give his home away for a price $p \leq p^*$. If the beliefs of the buyer about the price distribution are stable along her search, so does her stopping rule. Then p^* is the solution of:

$$(v - p^*) = E[e^{-r\tau}]((1 - F(p^*, b))(v - p^*) + E[v - p | p \leq p^*] - c) \quad (1)$$

The left-hand side of equation (1) is the expected surplus of the transaction for the buyer. The right-hand side is composed of three terms, all discounted at a rate r for the expected time τ: the cost c she has to pay to meet a new seller, the continuation value she gets for resuming the search with the probability $1 - F(p^*, b)$ if the seller asks for too much, and the expected surplus she gets from meeting a seller that agrees to sell for less than p^*.

With:

$$E[e^{-r\tau}] = \int_0^{\infty} e^{-rt}\, \lambda\, e^{-\lambda t} dt = \frac{\lambda}{r + \lambda}$$

And:

$$E[v - p | p \leq p^*] = \int_0^{p^*} (v - p)\, f(p, b)\, dp$$

Rearranging the terms, we get:

$$c + \frac{r}{\lambda}(v - p^*) = \int_0^{p^*} (p^* - p)\, f(p, b)\, dp \qquad (2)$$

Equation (2) has a straightforward interpretation: a buyer continues her search unless the continuation cost, which is the sum of the fixed cost c of new sampling and cost of waiting to encounter the next seller $\frac{r}{\lambda}(v - p^*)$, is greater than the marginal benefit of searching for a price lower than p^*.

As in Lambson et al. (2004), we differentiate this optimal condition for the stopping rule with respect to all the parameters that differ from one buyer to another: c, s, r, λ and b.

Differentiating (2) with respect to c gives:

$$1 - \frac{r}{\lambda}\frac{\partial p^*}{\partial c} = \frac{\partial p^*}{\partial c}\frac{\partial}{\partial p^*}\left(\int_0^{p^*} (p^* - p)\, f(p,b)\, dp\right)$$

With the Leibniz integral rule:

$$\frac{\partial}{\partial p^*}\left(\int_0^{p^*} (p^* - p)\, f(p,b)\, dp\right)$$

$$= 1 * (p^* - p^*)\, f(p^*, b) \; - \; 0 \; * p^* f(0,b) + \int_0^{p^*} \frac{\partial}{\partial p^*}\big((p^* - p) f(p,b)\big) dp$$

$$= \int_0^{p^*} f(p,b)\,) dp + \int_0^{p^*} (p^* - p) * 0 \, dp$$

$$= \; F(p^*, b)$$

We get:

$$\frac{\partial p^*}{\partial c} = \frac{1}{r/\lambda + F(p^*,b)} > 0 \qquad (3)$$

Buyers with higher search costs set higher reservation prices to reduce the risk of many unsuccessful visits. The impact on the price paid because of such a difference in search cost has been investigated. As the distant buyers face greater fix costs, because visits imply longer trips, equation (3) is the theoretical justification of the "distant buyer hypothesis" made in Ihlandfeldt and Mayock (2012). They find a positive, statistically significant, impact of the distance between the newly purchased homes and their former addresses on the prices paid by buyers. For Turnbull and Sirmans (1993) and Lambson et al. (2004), search costs can also increase through an information acquisition channel. They test if first-time buyers pay a premium. Even though it is at most weakly significant, they both find a positive impact of the lake of experience on transaction prices. A possible interpretation of the premium paid by buyers who tour only a few homes Wilhelmsson (2008) measures is as a consequence of higher search costs.

Differentiating with respect to r:

$$\frac{\partial p^*}{\partial r} = \frac{1/\lambda\,(v - p^*)}{r/\lambda + F(p^*,b)} > 0 \qquad (4)$$

Buyers in a hurry, who discount the future more, set higher reservation prices to decrease the risk of having to meet a seller that ask for more than $p^*$and having to search for another period. Evidence of the effect of time preference on the transaction price exists for the selling side, see for example Anglin et al. (2003). On the other hand, the effect of the search duration on the transaction price has never been measured, to our knowledge, for the buying side. Indeed, buyer time-on-market is harder to track than the one of sellers. This is one of the gaps the present paper aims to fill in section 5.

Differentiating with respect to λ:

$$\frac{\partial p^*}{\partial \lambda} = \frac{-r/\lambda^2\,(v-p^*)}{r/\lambda + F(p^*,b)} < 0 \qquad (5)$$

The easier it is to meet a seller, the lower a buyer sets her reservation price. Thus, in a tight market where for-sale properties are scarce, a buyer would be ready to pay a premium to outpace the other buyers.

Differentiating with respect to b:

$$\frac{r}{\lambda}\frac{\partial p^*}{\partial b} = \frac{\partial}{\partial b}\left(\int_0^{p^*} (p^* - p)\, f(p,b)\, dp\right)$$

With the Leibniz integral rule:

$$\frac{\partial}{\partial b}\left(\int_0^{p^*} (p^* - p)\, f(p,b)\, dp\right) = \int_0^{p^*} \frac{\partial}{\partial b}\left((p^* - p)\, f(p,b)\right) dp$$

$$= \int_0^{p^*} \frac{\partial p^*}{\partial b}\, f(p,b)\, dp + \int_0^{p^*} (p^* - p)\frac{\partial f(p,b)}{\partial b}\, dp$$

$$= \frac{\partial p^*}{\partial b} F(p^*,b) + \int_0^{p^*} (p^* - p)\frac{\partial f(p,b)}{\partial b}\, dp$$

Thus:

$$\frac{\partial p^*}{\partial b} = \frac{-\int_0^{p^*} (p^* - p)\frac{\partial f(p,b)}{\partial b}\, dp}{r/\lambda + F(p^*,b)} > 0 \quad if \quad \int_0^{p^*} (p^* - p)\, \frac{\partial f(p,b)}{\partial b}\, dp < 0 \qquad (6)$$

Intuitively the sign of $\int_0^{p^*} (p^* - p)\, \frac{\partial f(p,b)}{\partial b}\, dp$ is negative (resp. positive) if an increase in the parameter b shifts the perceived distribution to the right (resp. left) of the true distribution.

Thus, a buyer who has a positive bias would set a higher reservation price, which is what we expect a priori.

This prediction draws the attention of previous studies. Both Lambson et al. (2004) and Ihlandfeldt and Mayock (2012) test for an anchoring bias of homebuyers. They find that the price difference between the buyer's former residence housing market and the market in which the transaction occurs has a positive effect on price. However, only Ihlandfeldt and Mayock (2012) measure it as statistically significant. Turnbull and Sirmans (1993) argue that the cause for such a shift in the perceived distribution may also be inexperience. As they fail to find a significant difference in the price paid by first-time buyers and experienced buyers, they conclude housing markets are informally efficient enough to eliminate such an effect.

In equation (1), we hold the beliefs of the buyer to be stable over time. So far, the literature has considered only factors impacting beliefs prior to the homebuyer's search. Most home buyers are amateurs with regard to the real estate market and even the experienced buyers may not know its latest trends. A plausible conjecture would be that buyers learn about market prices as they search. Kohn and Shavell (1974) have been the first to introduce a search framework where the decision-maker updates her stopping rule at each sample. Rothschild (1974), Rosenfield and Shapiro (1981) or Bikhchandi and Sharma (1996) discuss the conditions for an optimal stopping rule to hold in a search with learing. Morgan (1985) applies such a model to job seekers. He derives from the general Kohn and Shavell's model a search model in which they learn from the job offers they have received. He shows the reservation wage of a job seeker is an increasing function of sampled job offers. Transposing to the real estate market, one can expect a buyer who had encountered sellers who ask for high prices in her previous visits to increase her own reservation price.

Differentiating with respect to s, we yield:

$$\frac{\partial p^*}{\partial s} = \frac{\frac{\partial c}{\partial s} - r/\lambda^2\,(v-p^*)\frac{\partial \lambda}{\partial s}}{r/\lambda + F(p^*,b)} > 0 \;\; if \;\; \frac{\partial c}{\partial s} > r/\lambda^2\,(v-p^*)\,\frac{\partial \lambda}{\partial s} \qquad (7)$$

To scan a larger segment leads to set a higher reservation price if the marginal search cost of increasing the segment size is greater than the value that the buyer gets from waiting less between each visit. Intuitively, $\frac{\partial \lambda}{\partial s} > 0$ as scanning a larger segment increase the chance to meet a seller. Anglin (1997) shows empirically that buyers who look in more neighborhoods

visit more houses. On the other hand, $\frac{\partial c}{\partial s} \geq 0$ as scanning a larger segment may increase fixed search costs. The buyer has two options to increase the size of the inventory she scanned: either she extends the geographic area of her search or she loosens some constraints on the apartment characteristics. In the first instance, $\frac{\partial c}{\partial s}$ is strictly positive. Indeed, to expand the search geographically implies longer travel time and a greater quantity of information to process about the different neighborhoods considered. In the latter $\frac{\partial c}{\partial s}$ can be null. Loosening some of her criteria about the apartment characteristics is not more time consuming nor it increases the information to process. Thus, buyers are facing a trade-off, between meeting more sellers and facing a higher search cost.

Before getting further in the analysis, let us note that, so far, all our theoretical results are related to the optimal stopping rule and not the price paid by the buyer in the end. It may create a problem as only the transaction price is trackable. A high reservation price buyer is equally likely to draw a seller ready to sell for a low price out of the distribution than as one with a low reservation. However, she remains more likely to accept worse deal if the seller has high price expectation. Thus, on average, she will pay a higher price as she is more likely to accept a price on the right end of this distribution. Mathematically speaking, as $f(p)\,dp$ is positive over $[0; +\infty]$:

$$\int_0^{p_1^*} p\, f(p)\, dp \geq \int_0^{p_2^*} p\, f(p)\, dp \quad \text{if} \quad p_1^* > p_2^*$$

Thus:

$$\frac{\partial E[p_{transaction}]}{\partial p^*} \geq 0$$

Hence, all results of equations (3) to (7) hold for the transaction price, in expected value. We test this prediction in section IV through a regression of the transaction price. The explanatory variables of this regression are proxies for the search costs, the time preference, the arrival rate, the buyers' beliefs about the price distribution and the size of the segment scanned.

III. Data Gathering

If there is no empirical evidence of the effect of the buyer's search on the transaction price in the literature, it is because of the lack of data describing it. To fill the gap, we use the

data of the French real estate Internet platform MA. MA produces information about the French housing market and makes it available on its website for free. One of the tools it offers is an automated valuation model (or AVM). This AVM computes for the users an estimated value for any housing unit in France based on its address and a description of the property. Here, we consider the estimates of apartments in Île-de-France, the Paris Region, made between 2015 and 2018.

To get an estimate, the user needs to give an address and to fill out a form that asks for basic characteristics such as the livable surface, the number of rooms, floor, and the number of bathrooms. He may also provide more advanced details: the presence of an elevator, of a cellar, recent façade renovation, etc. All those characteristics are used by a hedonic model to compute an estimate of the property value. The model also relies on a geographical price per meter square, based on a spatial econometric model, specific to each address.

The user also has to specify the reason he wants an estimate of the value of the property. If he reports owning the property, he is asked the property usage, if he wishes to sell it, and the date and the price of its acquisition (optional). If he does not, he is asked if he is currently looking to buy a home and at what stage he is in his search. The users that we are interested in in this study are the latter as they are buyers.

We make a hypothesis: if a buyer asks for an estimate of an apartment, it is because she has toured the property and she might consider buying it. What supports this hypothesis is the fact that in France most listings do not indicate the address of the property because exclusive selling mandates are a rare thing (Larceneux et al. 2015) and real estate agents do not want to openly disclose the address to potential competitors. A buyer able to know the address of an apartment for sale should have, at least, called to request a visit.

We then identify who among these buyers are also part of another group of users: the ones that report owning a property they have bought between 2015 and 2018 and indicate the date and the net price of the transaction. For all those users, identified by their account on MA, we gather the estimates they have made as buyer (as defined above) in the 12 months prior to the purchase. Thus, we get the sequence of the apartments visited by a given buyer together with the information regarding the transaction of the apartment actually bought. Among these users, we also identify the ones that previously owned another property. We look for valuations

asked for a property, they have reported owning, at another address, anywhere in France, prior to the purchase date of the transaction considered.

Dataset	Variable	count	mean	std	min	25%	50%	75%	max
Purchased	surface area	3,086	60	31	10	39	55.02	74	450
	nb of rooms	3,086	2.8	1.2	1	2	3	3	10
	nb of bathrooms	3,086	1.2	0.4	1	1	1	1	4
	floor	3,086	2.8	2.4	0	1	2	4	29
	elevator	3,086	58%	49%	0	0	1	1	1
	nb of floors	3,086	5.4	2.9	0	4	5	6	49
	parking	3,086	34%	47%	0%	0%	0%	100%	100%
	exterior	3,086	47%	50%	0%	0%	0%	100%	100%
	construction date	1,496	1,942	56	1,070	1,905	1,952.5	1,974	2,019
	estimated price	3,086	436,671	359,812	0	216,400	333,000	525,300	4,135,600
	estimated pm²	3,086	7,062	2,799	0	4,859	7,044	8,953	2,3920
	transaction price	3,086	406,403	515,678	0	199,000	305,000	475,000	14,000,000
	transaction pm²	3,086	6,692	5,945	0	4,384	6,395	8,184	117,763
Searched	surface area	15,500	62	216	4	39	56	75	26,035
	nb of rooms	15,500	2.8	1.2	1	2	3	3	10
	nb of bathrooms	15,360	1.1	0.4	1	1	1	1	10
	floor	15,500	2.9	2.4	0	1	2	4	30
	elevator	15,500	57%	50%	0	0	1	1	1
	nb of floors	15,500	5.5	2.7	0	4	5	6	47
	parking	15,500	31%	46%	0	0	0	1	1
	exterior	15,500	44%	50%	0	0	0	1	1
	construction date	12,486	19,41	50	1,070	19,00	1,950	1,975	2,018
	estimated price	15,500	427679	1819529	0	208775	318600	493450	219688800
	estimated pm²	15,500	6,635	2,600	0	4,631	6,546	8,365	27,558
Owned	surface area	1,940	77	157	5	40	60	90	6,540
	nb of rooms	1,940	3.2	1.7	1	2	3	4	10
	nb of bathrooms	1,931	1.3	0.6	1	1	1	1	5
	floor	1,630	3	2.7	0	1	3	4	25
	elevator	1,940	0.6	0.5	0	0	1	1	1
	nb of floors	1,630	5.7	3.6	0	4	5	6	38
	parking	1,940	41%	49%	0	0	0	1	1
	exterior	1,940	35%	48%	0	0	0	1	1
	construction date	1,668	1,947	58	1,250	1,914	1,960	1,987	2,019
	is an apartment	1,940	84%	37%	0	1	1	1	1
	estimated price	1,939	454,956	990,674	29,900	196,700	326,200	533,900	40,489,000
	estimated pm²	1,939	6,302	3,042	748	3,750	6,028	8,417	23,119

Table 3-1 : Purchased, searched and owned apartments statistical description

Our dataset contains 3,086 transactions between 2015 and 2018. Table 1 contains a statistical description of the apartments. It is divided into three sections: Purchased, Search and Owned; they respectively regroup the apartments finally purchased by MA users, the

apartments those users have estimated during their search prior to the transaction and finally the properties they owned before the purchase.

The average purchased apartment in the dataset is a 60m² flat of two to three rooms and one bathroom. They are equally distributed vertically, across the typical six floors Parisian building, with an elevator for more than half of them. A half enjoys an exterior: a balcony, a terrace or a private ground garden; and 34% a private parking spot. It was sold on average for 389k€ at a €6,353 per square meter rate and estimated for +6% more by MA. The average visited apartment is similar to the one purchased. The only noticeable difference is a –6% estimated price per meter square. The previously owned properties differ from the apartments of the two other datasets by their size. They are larger by +16m² or +26%. Indeed, a minority of people, 16%, previously owned a detached house, generally larger than a standard apartment. Regarding the differences in volume of the three datasets, they are only natural. On the one hand, a given buyer visits several apartments before buying one. On the other hand, some homebuyers are first-time buyers and do not have any owned property to estimate the value of. Moreover, there is no reason to believe that every experience homebuyer in our dataset has used the MA valuation tool to get an estimated value of her property.

The fact that we use raw data from the Internet is visible in table 1. The dataset contains some serious outliers: a visited apartment of 4m² (below the 8m² legal limit in France), an apartment purchased for €0 or an estimate price of €0/m². Various reasons can explain those outliers from misreported value by the user, when filling out the form, to a bug in the valuation model. These erroneous data points may distort the results of the econometric analysis. The pre-processing that aims at minimizing the potential perturbations are presented in part IV.

Datasets drawn from the Internet are known to have representativeness issue (Salganik 2017). Because of privacy issues, the dataset does not contain any socio-demographic information about the MA users. Thus, we cannot check if the subset of buyers in our dataset differs significantly from all the buyers active, in the Paris area, from 2015 to 2018. However, we can compare the apartments those buyers bought with all the ones that were subjects of a transaction in 2015 and 2016, thanks to the public fiscal record of all real estate sales, *Demandes de Valeurs Foncières* (thereafter DVF). According to the French fiscal administration, DVF is exhaustive but the number of variables that describes transactions is limited. Hence, the comparison between the two datasets only concerns the location, year of the sale, size of the apartment and price. We aggregate the location within three segments: Paris proper, “Petite-

Couronne" which is the French nickname for the three departments directly surrounding the city of Paris, and "Grande-Couronne" which is the four departments further away from Paris that surround the "Petite-Couronne". Statistics are presented in Table 2. Note that the open DVF dataset need to be aggregated in order to represent each transaction on a single line, to do so we follow the *Groupe National DVF* methodology[22]. We filter out only the transactions that do not concern single residential properties, thus outliers on surface areas, the numbers of rooms and prices are present.

	DVF (count: 184,872)							MA-Purchased (count: 3,086)						
	mean	std	min	25%	50%	75%	max	mean	std	min	25%	50%	75%	max
Paris	33%	47%	0	0	0	1	1	46%	50%	0	0	0	1	1
Petite-Couronne	39%	49%	0	0	0	1	1	41%	49%	0	0	0	1	1
Grande-Couronne	28%	45%	0	0	0	1	1	13%	34%	0	0	0	0	1
2015	48%	50%	0	0	0	1	1	46%	50%	0	0	0	1	1
2016	52%	50%	0	0	1	1	1	31%	46%	0	0	0	1	1
2017	-	-	-	-	-	-	-	21%	41%	0	0	0	0	1
2018	-	-	-	-	-	-	-	1%	11%	0	0	0	0	1
price	307,583	331,474	2	149,000	217,650	356,500	28,500,000	406,403	515,678	0	199,000	305,000	475,000	14,000,000
pm2	5,457	4,794	0	3,097	4,804	7,490	10,76,667	6,692	5,945	0	4,384	6,395	8,184	117,763
surface area	58	73	1	36	54	71	9,590	60	31	10	39	55.02	74	450
rooms	2,7	1,2	1	2	3	3	47	2.8	1.2	1	2	3	3	10

Table 3-2 : Comparison between sales collected on MA and all sales recorded in fiscal bases

The two datasets are different in almost every dimension. Only in terms of surface and number of rooms do they are comparable, over the whole distributions. Along this dimension, the transactions recorded in the MA dataset seems to be representative of the whole Paris region market.

By contrast, the geographical distributions of the sales differ drastically. Indeed, the Paris market is overrepresented in the MA dataset (46% vs 33%), whereas the "Grande-Couronne" is underrepresented (13% vs 28%). The history of the commercial extension of MA explains this imbalance. The company operated solely in Paris, when it was first launched in 2008, then it expanded its activities to the rest of the Paris region in 2011. Moreover, the housing

[22] www.groupe-dvf.fr/vademecum-fiche-n3-precautions-techniques-et-qualite-des-donnees-dvf/

prices in Île-de-France follow a center/periphery decreasing gradient[23]. It is economically meaningful for MA to optimize its users' acquisition funnel for Paris instead of the great suburb. This over-representation of the more expensive Parisian market explains the difference in the transaction price and price per meter square, respectively +32% and +24% higher on average in the MA dataset.

Finally, the transactions are evenly distributed between 2015 and 2016 in DVF, whereas it decreases significantly over the years in the MA dataset to reach a residual share of 1% in 2018. What operates here is the "drifting" characteristic of digital datasets (Salganik 2017). Indeed, because of completion issues and business reasons, the product team of MA decides to bring less to the fore the questions regarding acquisition conditions over time. They end up removing them entirely early in 2018, which explains the drop.

Altogether, these differences do not seem to be likely to threaten the significance of the results. However, we take them into account in the robustness tests presented in part VI.

IV. Characteristics of the Search

Once aggregated, the data collected on MA creates a novel dataset that describes the buyers' searches and the details of the transactions for apartment sales, in the Paris region, from 2015 to 2018. In order to test for the predictions of the model presented in section II, we create variables that describe characteristics of the buying process. They are listed below. For each of them, we state how it is computed, which characteristic of the search it represents and the interpretation we make of it.

- *tom*, or time-on-market, is the number of days between the first estimate the buyer has made and the date of the purchase. By construction, it cannot exceed 365 days as we only consider the estimate made during the 12 months that precede the purchase. There is no reason to believe buyers estimate every apartment they visit. Moreover, they may have started their search before their first estimate. Thus, our time-on-market constitutes a lower bound for the actual time the buyer has spent looking for an apartment. In

[23] See the Paris region price map on www.meilleursagents.com/prix-immobilier/ile-de-france/

section V, we use *normalized tom* instead, which is the time-on-market divided by the mean time-on-market for the observed purchases.

- *number of estimates* represents how many different apartments, uniquely identified by their address, the buyer has estimated on MA during her search.
- *anchor* is the log difference between the average estimated prices of the properties that the user has reported owning before the purchase and the estimated price of the purchased apartment. We consider here price per meter square, and we use the MA estimates. Indeed, we only have the transaction price of the latest purchase and not for the previously own properties. Moreover, those are the estimates the buyer has received for her properties. If the buyer is subject to anchoring, they should have participated in the creation of this reference point. The anchoring bias has no reason to have a symmetrical effect, thus we create two variables $anchor^{+}$, which is equal to anchor if it is positive and zero otherwise, and symmetrically $anchor^{-}$. For users who had never declared owning other property before, *anchor* as well as $anchor^{+}$ and $anchor^{-}$ are equal to zero.
- *learning* is the log difference between the average estimated prices of all the apartments estimated before the purchase, during the search, and the estimated price of the purchased apartment. We use per square meter prices and estimates from MA again. It might have been more into line with the search-with-learning theory to use the listing prices here, but unfortunately they are unavailable to us. Nevertheless, the estimated prices of the visited properties are figures that the buyer has seen during her search and most likely participate in her learning process. Moreover, as they are parts of the same search process, the buyer may consider that she could have derived a similar utility from owning any visited apartment or the one she finally purchases. In order to determine how symmetrical the learning effect is, we create two variables: $learning^{+}$ and $learning^{-}$; similarly to $anchor^{+}$ and $anchor^{-}$.
- *budget* is the log difference between the average estimated prices of the apartments estimated before the purchase and the estimated price of the purchased apartment. Unlike the anchor and learning variables, we use here the full estimated price, not the price per meter square. This variable intends to represent how much off-budget (if positive) or under-budget (if negative) the acquisition is. We also create two variables: $budget^{+}$ and $budget^{-}$. Note that computing the *budget* variable based on the estimated

prices of the property the buyer already owns might have been meaningful. However, it would have reduced our dataset to only a third of the buyers.

- *number of boroughs* is the number of distinct boroughs, as defined by the French national bureau of statistics INSEE, the buyer tours apartments in.
- *size flexibility* is a simple measure of diversity among the visited apartments, which focuses only on their difference in terms of size. It is equal to the difference between the surface areas of the largest and the smallest apartment visited, divided by the average of these two values.

As mentioned above, our dataset, drawn from the Internet and partly based on user filling form unsupervised, contains some obvious outliers. We choose to filter transactions regarding apartment: smaller than 8m², bigger than 300m², sold for a price below €10,000 or above €20,000,000 or for a price per meter square below €100/m² or above €50,000/m². On top of that, 0.5% of the extreme values on both ends of the distributions for these three variables are filtered. Last, we get rid of the apartments that the same user inconsistently reports having bought for different prices. Altogether we reduce the dataset from 3,086 to 2,714 apartment purchases.

On top of outliers regarding the purchased itself, we check for abnormal search patterns. First, one of the key variables describing a search for an apartment is its length. As noted before, the way we compute the buyer time-on-market is actually a lower bond of this length. To minimize the approximation, we choose to filter the purchased for which the buyer selects "I already made an offer" out of the four possible stages[24] for her first estimate. Second, the distribution of the number of estimates computed by the same user is strongly skewed. It reaches a maximum of 357 estimations when the median is 3 and the third quarter 8. Figure 1 plots this distribution.

One can easily see how the right end of the curve is related to strongly abnormal behavior. To prevent these extreme data points to distort our results, we choose to filter the top 2 percent of users with the highest number of estimates (i.e. the buyers who have used the

[24] The three other stages are "I am wondering", "I am starting", "I am actively searching"

valuation tool more than 48 times). These two filters reduce the dataset to 2,372 transactions. Summary statistics of the final dataset are presented in table 3.

A little more than a third of the buyers owned a real estate property prior to the purchase, almost exclusively in the same department. On average they compute a little less than seven estimates during their search, but more than 25% of them only perform one. Most concentrate their search in only one city[25], but they look in different places within the cities: on average more than two boroughs. If the majority of the transactions concerns primary residences, 14% are about rental investments and 4% vacation homes.

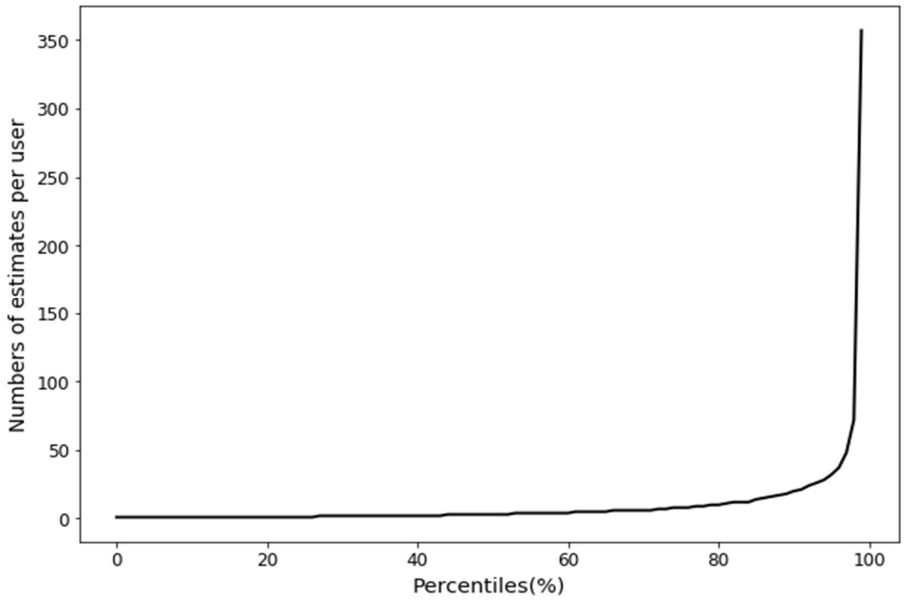

Figure 3-1 : Percentile distribution of the number of estimates per user..
From the minimum to the 75th percentiles the number of estimates grows slowly from 1 to 8. This growth accelerate up to 48 estimates for the 98th percentiles. From ther, it is multiplied by more than seven for the user at the very end of this strongly skew distribution.

On average, buyers who owned a property before owned a less expensive one as price per meter square is concerned. A large part of the buyers does not report owning a property prior to the purchase, thus *anchor* is null for most of our sample (64%). Despite this, *anchor* as a large standard deviation, of 0.32 compared to 0.15 for *learning*. The distribution of *learning* indicates that, on average, buyers have visited apartments of the same price per meter square

[25] Parisian arrondissements are considered as individual cities

than the one they finally buy. However, the distribution of *budget* indicates than most of them visited less expensive apartments in absolute value during their search. They stay on average 5.5 months on the market, which seems reasonable, but the duration of the search has a strong variance: more than 25% searched for less than 3 months and a week, whereas 25% more than 7 months. The average spread in the surface areas of the visited apartments is of 24%. It represents 15 m², almost an entire room, for the average purchased apartment of 60 m².

	count	mean	std	min	25%	50%	75%	max
transaction price	2,372	380,713	251,007	67,000	206,750	310,000	475,000	1,750,000
transaction pm2	2,372	6,442	2,362	1,030	4,550	6,482	8,182	13,191
estimated price	2,372	406,062	288,430	57,710	216,432	322,268	500,505	2,257,300
estimated pm2	2,372	6,772	2,575	1,659	4,762	6,716	8,550	18,058
nb of properties	2,372	0.7	1.4	0	0	0	1	24
already owner	2,372	34%	47%	0	0	0	1	1
already owner same city	2,372	19%	39%	0	0	0	0	1
already owner same depart.	2,372	31%	46%	0	0	0	1	1
mean estimated owned pm2	812	6,282	2,867	748	4,046	6,119	8,124	23,119
mean estimated owned price	812	420,020	332,022	37,400	215,450	332,883	506,700	3,121,000
nb of estimates	2,372	6.7	8.2	1	2	4	8	48
nb of distinct boroughs	2,372	2.3	2.1	1	1	2	3	19
nb of distinct cities	2,372	1.5	1.1	1	1	1	2	17
mean estimated searched pm2	2,372	6,526	2,356	1,513	4,686	6,556	8,181	17,475
mean estimat. searched price	2,372	528,474	3,904,325	52,100	212,540	314,436	476,950	10,9958,250
primary residence	2,372	81%	39%	0	1	1	1	0
vacation home	2,372	15%	35%	0	0	0	0	1
rental investment	2,372	4%	20%	0	0	0	0	1
tom	2,372	164	92	1	96	142	227	365
norm. tom	2,372	1.07	0.60	0.01	0.63	0.93	1.49	2.39
anchor	2,372	-0.03	0.27	-1.95	0.00	0.00	0.00	1.16
learning	2,372	0.00	0.17	-1.24	-0.07	0.00	0.06	1.20
budget	2,372	-0.02	0.35	-2.40	-0.14	-0.04	0.05	6.11
size flexibilty	2,372	0.24	0.34	0.00	0.00	0.10	0.33	2.00

Table 3-3 : Regression dataset statistics

V. Empirical Analysis and Results

To test for the prediction made in section II, we follow an empirical specification close to the ones of Turnbull and Sirmans (1993), Lambson et al. (2004), Ihlandfeldt and Mayock (2012) or Wilhelmsson (2008). A single hedonic model of the prices of the apartments, detailed in equation (8), is estimated through an OLS regression. In line with the literature, the variables related to the features of the apartments and the ones related to the buyers are accounted for all at once. Another method would have been to perform first a classic hedonic regression of the prices, based only on the characteristics of the apartments, then a second regression of the

residuals of this first regression by the buyers' characteristics. Such an approach is tested on the robustness tests, in part VI.

A minor difference between equation (8) and the regressions estimated in the literature is that the dependent variable is the log of the price per meter square, not the log of the absolute price. The reasons for this choice are twofold. First, it is the most commonly used specification for the study of French apartment markets. In particular it is the one used by the French Bureau of Statistics (INSEE) in its indexes (*Insee Methodes*-132, 2019). Second, it allows the use of the address-specific price per meter square computed by MA as a regressor (see detail below).

$$\begin{aligned}\ln(pm^2) = {} & \beta_1 + \beta_2 \ln(tomNorm) + \beta_3 \ln(nbEstimations) + \beta_4 learning^+ + \\ & \beta_5 learning^- + \beta_6 anchor^+ + \beta_7 anchor^- + \beta_8 budget^+ + \beta_9 budget^- + \\ & \beta_{10} nbBoroughs + \beta_{11} nbBoroughs^2 + \beta_{12} areaElasticity + \\ & \beta_{13} vacationHome + \beta_{14} rentalInvestment + \beta_{15} alreadyOwner + \\ & \beta_{16} ln(MApm^2) + \beta_{17} ln\left(\frac{area}{40}\right) + \sum_{i\in[1,6]\setminus 2} \beta_{18,i} \mathbb{1}_{nbRooms=i} + \\ & \sum_{i\in[2,3]} \beta_{19,i} \mathbb{1}_{nbBathrooms=i} + \sum_{i\in[0,6]\setminus 2} \beta_{20,i} \mathbb{1}_{floo} + \beta_{21} noElevator + \\ & \beta_{22} outdoor + \beta_{23} noCellar + \beta_{24} smallBuilding + \beta_{25} bigBuilding + \\ & \beta_{26} view + \beta_{27} supRelativePerception + \beta_{28} infRelativePerception + \\ & \beta_{29} renovationNeeded + \sum_{i\in[1,9]} \beta_{30,i} \mathbb{1}_{builtPeriod=i} + \\ & \sum_{i\in[2016,2018]} \beta_{31,i} \mathbb{1}_{saleYear=i} + \varepsilon \qquad (8)\end{aligned}$$

The two groups of variables that explain the transaction price are: the apartment's attributes, regressors 16 to 30, as in a classic hedonic model; the characteristics of the buyer's search, regressors 2 to 15. For the hedonic part of the model, we choose a reference apartment of 40m², two rooms, one bathroom, one cellar, on the second floor of five or six-story building equipped with an elevator, sold in 2015. As for geographical control, we use a price per meter square computed by MA, thanks to its proprietary spatial econometrics model. This price is specific to the exact address of the apartment and the month of the transaction for our standard apartment (hence it does not take into account the characteristics of the purchased apartment).

Such a granularity in the geographical control allows for greater confidence in the results as it mitigates the risk of unobserved qualities. As pointed out by Lambson et al. (2004), any analysis of the impact of buyers' and sellers' characteristics can be subject to selection bias. One cannot rule out the possibility that if, for example, distant buyers pay more, it is because

they buy specific properties, which look similar to the others from the econometrician perspective. To avoid this pitfall Ling et al. (2016) regress the price of 100,000 transactions made by real estate investors with an extended vector of property characteristics and characteristics regarding the buyers and sellers. They assume symmetry for the demand for unobserved characteristics between buyers and sellers, and an equivalent symmetry hypothesis on the effects of search cost and anchoring. Hence, they can distinguish the real anchoring or cost effects from the selection bias, thanks to a linear combination of the dummies characterizing buyers and sellers. Here such a technique is impossible. First, because there is no such thing as a "distant home seller" in residential real estate. Second, more generally, because some characteristics of a buyer's search have simply no equivalents for sellers, e.g. the size of the scanned inventory.

The results of the estimation of equation 8 are presented in Table 4. For readability issues, only the coefficients of the variables relative to the search are presented in the main text. The complete table is available in the appendix.

Hedonic features

The hedonic features alone explain most of the observed variance in the apartment prices. R^2 of model (1) is 84% against 85% for model (6) with all the search characteristics variables on top of the hedonic ones. Paraphrasing the stochastic matching terminology (Novy-Marx 2009,Genesove and Han 2012), the idiosyncratic component of the sale price, proper to the buyer/seller match, is small compared to the common component value.

The MA's estimate of price per meter square is almost taken into account one for one and extremely statistically significant. This justifies to use it as a control for the quality of the location, a posteriori. Overall larger apartments are cheaper per meter square as the log of the surface area has a negative impact, as expected under the law of diminishing marginal utility. However, the largest apartments of six rooms and more, which are also the most likely to have more than one bathroom, seem to violate this law. Their scarcity in the dense urban market of the Paris region may explain this phenomenon. Ground floor apartments are sold for a discount of -4% whereas top floors apartments enjoy a premium of +4% to +6%. Buildings built between the end of World War II and the '90s are considered as inferior when compared to historical buildings and new constructions.

Dependent variable	ln(pm²)					
	Hedonic (1)	Cost (2)	Time (3)	Info (4)	Size (5)	Full (6)
intercept	0.6630*** (0.095)	0.6675*** (0.096)	0.6412*** (0.096)	0.2802** (0.099)	0.6703*** (0.095)	0.2580** (0.099)
rental investment				-0.0227* (0.011)		-0.0240* (0.011)
other						-0.0314 (0.038)
vacation home		0.0367* (0.018)		0.0296° (0.018)		0.0269 (0.018)
ln(TOM norm)			-0.0106* (0.005)			-0.0108* (0.005)
ln(nb estimation)		-0.0006 (0.003)				-0.0037 (0.005)
learningEffectMinus				0.1372*** (0.039)		0.1375*** (0.039)
learningEffectPlus				0.2416*** (0.035)		0.2497*** (0.035)
anchorEffectMinus				0.0028 (0.017)		0.0037 (0.017)
anchorEffectPlus				0.0800* (0.032)		0.0753* (0.032)
budgetEffectMinus				0.1303*** (0.024)		0.1230*** (0.024)
budgetEffectPlus				-0.0213 (0.013)		-0.0100 (0.014)
nb bouroughs					0.0110* (0.004)	0.0109* (0.005)
nb bouroughs²					-0.0007* (0.000)	-0.0007* (0.000)
already owner				-0.0030 (0.009)		0.0029 (0.010)
size flexibility					-0.0391** (0.013)	-0.0340* (0.014)
ln(MA pm²)	0.9273*** (0.011)	0.9267*** (0.011)	0.9296*** (0.011)	0.9713*** (0.011)	0.9254*** (0.011)	0.9727*** (0.011)
Hedonic controls	Yes	Yes	Yes	Yes	Yes	Yes
Nb Observations	2372	2372	2372	2372	2372	2372
R^2	0.841	0.842	0.853	0.842	0.853	0.854
Adj-R^2	0.839	0.839	0.850	0.839	0.850	0.850
F-statistic	343.7	326.0	298.8	318.5	298.8	265.3

Standard errors between parenthesis
*** p <0.001, ** p <0.01, * p <0.05, ° p <0.1

Table 3-4 : Equation (8) OLS regression results

One can see that apartments have been described by their owners. The laudatory subjective characteristic "superior relative perception" fails to have a significant effect, whereas the deprecatory one "inferior relative perception" has a negative and statistically significant effect on the price of -10%. However, homeowners still exhibit some objectivity as a declarative exceptional view does increase the price by +2.5% and a need for renovation decreases it by -5.5%.

Search Costs

According to equation (3), buyers with higher search costs fixed higher reservation prices and thus should pay more for a given apartment. Our dataset does not contain information about such cost or a proxy. Lambson et al. (2004) and Ihlandfeldt and Mayok (2012) use the distance to the former residence, for example.

However, we record the number of estimates, which can be seen as a good proxy for the number of visits. These numbers should be related to the buyer search costs: the higher they are, the fewer visits she is likely to do. Indeed, the coefficient related to *ln (nb. estimations)* is negative but fails to be significant, both in model (2) and in model (6). This proxy measure of the number of visits is maybe too much of an approximation. Moreover, the fixed costs are not its only determinant of the number of visits. The size of the segment scanned or the encounter rate must affect it also.

The dummy *vacation home* may indicate greater search costs. Indeed, a buyer who purchases a vacation home is a "distant buyer". The positive, significant at the 5% percent level in model (2), impact on the purchased price can be interpreted as an effect of higher search costs. However, its statistical significance decrease as other characteristics of the search are accounted for. It remains significant, but at the 10% level, in model (4), and fails to be significant in model (6). The premium we measure could also be related to information disadvantage such buyers face or to hidden characteristics secondary homes share: accessibility from transport hubs, proximity to local attractions, views on iconic monuments, specific design, etc.

Time Preference

Equation (4) predicts buyers with higher discount rates optimally set higher reservation prices and should pay a premium to shorten the search. The buyer's time preferences are not

directly trackable. However, the realized time-on-market is directly impacted by these preferences and should represent a good proxy. Indeed, we observed that the coefficient related to the *ln (norm. tom)* variable is negative and statistically significant at the 5% level in both model (3) and (6). To our knowledge, such a relationship between time and price has never been evaluated for the demand side of a real estate transaction, in the literature. The magnitude of the coefficient indicates that a search twice longer results in a -0.75% discount, or €-2,800 for the average transaction observed in the dataset. As pointed out in section IV, our measure of the buyer's search duration is imperfect. One can conjecture the effect to be stronger with a more accurate value of the buyers' time-on-market. Note that, as underlined in Anglin et al. (2003) for the selling side, the price and the time on market are determined simultaneously. A paper whose main focus would be to study this relationship may prefer a two-stage least square regression approach to avoid endogeneity issues.

Size of the Segment

According to equation (7), the size of the segment scanned by a buyer does not have a straightforward effect on her reservation price as it affects it through two concurrent channels. On the one hand, a larger segment increases the inventory scanned and thus the encounter rate. On the other hand, it may increase the search cost. Our empirical results illustrate this trade-off.

The most striking result in that matter is the effects of the geographic spread of the search on the transaction price. In both model (5) and (6), the estimated coefficients of the number of distinct boroughs the buyer has toured apartments in are positive, whereas its square has a negative effect. The positive first-order effect of a geographically wider search on the transaction price shows that the rise in the search cost overcomes at first the increase in the encounter rate. However, the marginal effect of scanning an additional borough is decreasing. According to table 4 results, it turns negative for the 10th borough. Overall, the buyer is better off searching in seventeen or more boroughs rather than concentrating her search in only one. These are high numbers when 75% of the buyers in the dataset search in only three or less different boroughs. These results suggest a geographical focus search is a better strategy for most buyers. This result is somehow puzzling. One might be surprised that a rational buyer would choose to search in a geographically extended area at the expense of a greater transaction price. Especially as we control for the length of the search. A possible explanation may be that,

ex ante, buyers underestimate the increase in the search costs induced by scanning a geographically broad segment.

The segment size does not only vary with its geographical spread. On a given zone, a buyer with looser (resp. stricter) criteria regarding apartments' characteristics has higher (resp. lower) chances to meet a seller, for the same search cost. The significant negative effect of the *size flexibility* in model (5) and (6) indicates that being flexible about the size of the apartment allows the buyer to pay less at the bottom line.

Price Distribution Beliefs

Following prediction already made by Turnbull and Sirmans (1993) and Lambson et al. (2004), equation (6) states that buyers with a distorted perception of the price distribution to the right (resp. to the left) set higher (resp. lower) reservation prices.

This is the theoretical argument behind the anchoring bias observed by Inhafelt and Mayock (2012), that our results confirm. The coefficient of $anchor^{+}$ is positive and significant at the 5% level in both models (4) and (6). On the other hand, $anchor^{-}$ seems not to have any effect on the transaction price. One can easily imagine how a buyer whose price beliefs are anchored on higher values may end up paying more. Because of her inability to differentiate good deals from poor offers, everything looks like a good deal from her perspective. In contrast, someone used to lower prices either gives up, and thus do not appear in our dataset, or learns the prices during her search and pays a regular price. According to our measure, an increase of one percent in the price of the previous property corresponds to a 0.08 percent higher transaction price. Inhalfelt and Mayock (2012) measure a similar magnitude of 0.05 for 1, in Florida

Our second result related to the buyer's beliefs about the price distribution is the main contribution of the present paper. We interpret the positive effect, statistically significant at the 0.1% level, of $learning^{+}$ and $learning^{-}$ as evidence of a search-with-learning mechanism. More precisely our results show that buyers, who visit apartments on average more (resp. less) expensive, in terms of the price per square meter, pay on average more (resp. less) for the same apartment. One must consider the magnitude of the coefficient together with the low value of the variable itself. Leaning ranges between -0.07 and +0.06 only for half of the transactions. Buyers visit properties in a homogenous price range during their search, yet they seem to incorporate this small difference in their beliefs.

A search-with-learning mechanism perfectly makes sense in the context of the home buying problem. As mentioned in section II, there is no reason to hold the buyers' beliefs as fix over time. Most buyers are "amateurs", in the sense they will probably engage in less than a handful of real estate transactions in their life. Therefore, they know little about housing prices prior to their search. Going through listings and visiting apartments is for them a "crash course" during which they get an idea of housing prices. For a detailed specification of the update process of the beliefs theoretically predicted in Kohn and Shavell (1974), the reader is referred to Rothchild (1974). A straightforward way to look at it is as a Bayesian update of the belief's distribution $F(p, b)$. Every time the buyer draws a price from the distribution, on every encounter with a seller, the prior, i.e. the buyer's belief, is updated.

The estimated price per meter square of the apartment by MA takes into account its quality both in terms of hedonic attributes and of its location. It represents the values assigned by market forces to these qualities. Thus, the *learning* variable represents the difference in the qualities of the purchased apartment and the ones the buyer has visited. Therefore, the learning process we document here may be caused by buyers being myopic regarding the qualities of the qualities. As if, during the learning process, they only accumulate information and update their beliefs about the price per meter square, but do not relate the differences in prices to differences in the characteristics or in the locations. They might also consciously deviate from the market rate because they consider as perfectly substitutable properties of different qualities. In both cases, they end up assessing similar values to apartments of unequal qualities.

To the author's knowledge, few empirical results on such a search-with-learning process exist in the literature and none in the context of real estate. It is worth noting that one the few exceptions, De Los Santos et al. (2012) who investigates search patterns of bookstores customers, also uses Internet browsing data. In real estate, a different, yet comparable measure can be found in the behavioral literature with the results of Northcraft and Neale (1985) on anchoring-and-adjustment decision heuristic. They find a variation in the listing price of a home impact the appraisal value of amateurs with a magnitude of 0.2 for 1, comparable with the 0.25 for 1 and 0.14 for 1 we measure for $learning^{+}$ and $learning^{-}$, respectively.

The asymmetry of the learning effect, which is almost 80% larger for the buyers who visit more expensive apartments, makes sense. The exact same argument has for anchoring can be repeated here. The fact that $learning^{-}$ has any effect might actually be more puzzling. Why would a buyer who holds the price to be lower than they really are get a discount at all? It might

be that the beliefs formed during the search are more persistent than if they were related to some anchoring phenomena. If so, the buyer would fight harder during the final bargaining to get what she believes is a fair price. Our model predicts for the median –0.07 $learning^-$ a -1% discount, or €3,800 for the average transaction. Such an effect stays within the range of the standard negotiation discount a buyer can expect to get after bargaining (measured around 4% by Merlo and Ortalo-Magné 2004).

Alternative explanations can be proposed for this learning phenomenon. One may think we face here an example of selection bias. A buyer who visits more expensive, better quality, apartments would just buy a more expensive property because they see some hidden qualities in it. Such an explanation appears unlikely as the prices we use for the visited apartments as well as for the purchased ones are estimates of the MA valuation model. There is no reason to believe this model would systematically fail to encounter for some hidden characteristics of the purchased apartments but not for the ones the buyers visit.

A more convincing alternative explanation could be that what we measure is a consequence of budget constraints. Buyers with looser (resp. stricter) budget constraints visit more (resp. less) expensive apartments and end up negotiating less (resp. more) if they purchase a less (resp. more) expensive apartment. This explanation can be discarded as it is already accounted for through the $budget^+$ and $budget^-$ variables.

From the significant positive effects of $budget^-$ in model (4) and (6), budget constraints appear to play a role in the price formation process. Buyers who buy more expensive property in absolute value than the ones they have considered during their search, get a discount proportional to the budget effort they seem to consent. Buyers who go off-budget either only agree to transact if they can get a good deal out of the negotiation, or the ones subject to stricter budget constraints get better at identifying a good deal. They may even use the cheaper properties they have visited as arguments to drive sellers' prices down. The magnitude of this effect for the median $budget^-$ value of -0.14 (or -13%) corresponds to a discount of -1.7%.

The fact we do not measure any significant effect of the $budget^+$ variable on the transaction price strengthens the confidence we have in our interpretation of the effect $learning^+$. As explained in part IV above, the budget variables account for the difference in absolute price, whereas the learning variables track the difference in price per meter square. Buyers who visit more expensive apartments do not pay more because they can afford it. It is

rather their beliefs, about what should be a fair price per square meter, that lead them to pay a premium.

The buyers' beliefs about the price distribution may be the cause of other results in table 4. We fail to measure a significant impact of an informational advantage of experience buyers over first-time buyers. However, we find that buyers who buy an apartment as a rental investment get a discount of -2.3%, significant both in model (4) and (6). Such a discount can be attributed to the better knowledge of these specific buyers about the housing market, but once again one cannot rule out a selection bias here. Finally, as it is mentioned above, the premium paid by vacation home buyers can also be interpreted as a consequence of their lack of expertise in the local market.

VI. Robustness Checks

The first concern about the results of table 4 is an interaction between possibly collinear *learning*, *anchor* and *budget* variables. All three are built as a log difference with an MA price estimation in the denominator. Table 5 presents the results of regression where each of the three variables has been introduced without the two others. Once again, only the first half of the coefficient are reported here (the full table is available in the appendix). For models (7) to (9), coefficients related to these variables remain positive and significant as in model (4) and (6), which comforts the results of part V. However, their magnitudes, and significance in the case of anchoring, increase in model (7) to (9) compared to model (6). This variance in the magnitude coefficient casts doubt on a possible interaction of the three variables with *ln (MA pm²)*.

The coefficient of the geographical control seems sensitive to the introduction of these three variables at it increases from 0.93 in models (1), (2), (3) and (5) to 0.97 in models (4) and (6). To take into account for this, models (6) to (9) are estimated following a two-step approach. Instead of estimating the model of equation (8) as a whole, a first hedonic regression is performed (equivalent to model 1), then its residuals are regressed against the second group of variables. The results are presented in Table 6. The coefficients for models (6 bis) to (9 bis) are similar to the ones obtained for models (6) to (9) in sign, magnitude, and significance. The only two exceptions are the premium related to *vacation home***,** which gets significant at the 10% level in all four models, and the $anchor^{+}$ effect, which decreases but remains positive and significant at the 10% level.

Dependent variable	ln(pm²)			
	Learning (7)	Anchor (8)	Budget (9)	Full (6)
intercept	0.2902** (0.099)	0.5851*** (0.098)	0.5717*** (0.095)	0.2580** (0.099)
rental investment	-0.0252* (0.011)	-0.0254* (0.012)	-0.0167 (0.011)	-0.0240* (0.011)
other	-0.0221 (0.038)	-0.0506 (0.039)	-0.0323 (0.039)	-0.0314 (0.038)
vacation home	0.0216 (0.018)	0.0289 (0.018)	0.0327° (0.018)	0.0269 (0.018)
ln(TOM norm)	-0.0120* (0.005)	-0.0107* (0.005)	-0.0097* (0.005)	-0.0108* (0.005)
ln(nb estimation)	-0.0010 (0.005)	-0.0005 (0.005)	-0.0053 (0.005)	-0.0037 (0.005)
learningEffectMinus	0.2270*** (0.035)			0.1375*** (0.039)
learningEffectPlus	0.2739*** (0.034)			0.2497*** (0.035)
anchorEffectMinus		0.0046 (0.018)		0.0037 (0.017)
anchorEffectPlus		0.1289*** (0.032)		0.0753* (0.032)
budgetEffectMinus			0.1854*** (0.022)	0.1230*** (0.024)
budgetEffectPlus			0.0131 (0.014)	-0.0100 (0.014)
nb bouroughs	0.0115* (0.005)	0.0112* (0.005)	0.0132** (0.005)	0.0109* (0.005)
nb bouroughs²	-0.0007* (0.000)	-0.0006° (0.000)	-0.0007* (0.000)	-0.0007* (0.000)
already owner	0.0108 (0.007)	-0.0038 (0.010)	0.0109 (0.008)	0.0029 (0.010)
size flexibility	-0.0439** (0.014)	-0.0362** (0.014)	-0.0331* (0.015)	-0.0340* (0.014)
ln(MA pm²)	0.9682*** (0.011)	0.9349*** (0.011)	0.9376*** (0.011)	0.9727*** (0.011)
Hedonic controls	Yes	Yes	Yes	Yes
Nb Observations	2372	2372	2372	2372
R^2	0.852	0.844	0.848	0.854
Adj-R^2	0.849	0.841	0.845	0.850
F-statistic	283.7	267.9	267.9	265.3

Standard errors between parenthesis
*** p <0.001, ** p <0.01, * p <0.05, ° p <0.1

Table 3-5 : Equation (8) with information variables accounted for individually OLS regression results

The dataset collected on the MA website is not perfectly representative of the Paris region housing market from 2015 to 2018. The comparison with fiscal records shows that Paris is overrepresented compared to the suburbs. Likewise, more transactions in our dataset happen in 2015 than for the three other years. To assess how general the results of the part V are, we regress model (6) on some subsets of the dataset. The result over four different splits are presented in table 6: Paris / suburbs, small apartments (studio or 2 rooms apartments) / large apartments (3 rooms or more), sold in 2015 / sold in 2016 or after, experienced buyers (who report owning another property prior to the purchase) / other buyers. We show only the results regarding the search variables (complete results in the appendix).

Coefficients of the statistically significant effects discussed in part V remain of the same sign as in model (6) but only two are statistically significant across all regressions: $learning^+$ and $budget^-$. Interestingly, $learning^-$ fails to have any measurable effect in models (10) and (15). Those two datasets present sales that occurred respectively a very tight market (Paris proper)

and a boom period[26]. As if, in this specific market dynamics, buyers whose price beliefs are based on lower reference points have a hard time imposing their viewpoints during the negotiation phase. In that perspective, a study using an instrument that measures the bargaining power on both sides of the transaction, like the one proposed in Carillo (2013), would be a good extension of the present work.

Dependent variable	ln(pm²)			
	Learning (7 bis)	Anchor (8 bis)	Budget (9 bis)	Full (6 bis)
intercept	-0.0123 (0.008)	-0.0107 (0.008)	0.0058 (0.008)	-0.0024 (0.008)
rental investment	-0.0184° (0.010)	-0.0184° (0.010)	-0.0191° (0.010)	-0.0201* (0.010)
other	-0.0084 (0.037)	-0.0084 (0.037)	-0.0368 (0.038)	-0.0204 (0.037)
vacation home	0.0303° (0.017)	0.0303° (0.017)	0.0341* (0.017)	0.0321° (0.017)
ln(TOM norm)	-0.0111* (0.005)	-0.0106* (0.005)	-0.0087° (0.005)	-0.0094* (0.005)
ln(nb estimation)	-0.0009 (0.005)	-0.0010 (0.005)	-0.0050 (0.005)	-0.0032 (0.005)
learningEffectMinus	0.1956*** (0.034)			0.1137** (0.038)
learningEffectPlus	0.2302*** (0.032)			0.2057*** (0.034)
anchorEffectMinus		0.0008 (0.018)		-0.0056 (0.017)
anchorEffectPlus		0.1208*** (0.031)		0.0588° (0.031)
budgetEffectMinus			0.1690*** (0.021)	0.1139*** (0.023)
budgetEffectPlus			0.0132 (0.013)	-0.0049 (0.013)
nb bouroughs	0.0127* (0.005)	0.0115* (0.005)	0.0133** (0.005)	0.0125* (0.005)
nb bouroughs²	-0.0008* (0.000)	-0.0007° (0.000)	-0.0007* (0.000)	-0.0008* (0.000)
already owner	0.0107 (0.007)	-0.0038 (0.010)	0.0124° (0.007)	0.0038 (0.009)
size flexibility	-0.0402** (0.013)	-0.0334* (0.014)	-0.0310* (0.014)	-0.0325* (0.014)
Nb Observation	2372	2372	2372	2372
R^2	0.057	0.018	0.018	0.068
Adj-R^2	0.052	0.013	0.013	0.062
F-statistic	12.86	3.824	3.824	11.40

Standard errors between parenthesis
*** p <0.001, ** p <0.01, * p <0.05, ° p <0.1

Table 3-6 : OLS regression of model (1) residuals with search variables and information variables accounted for individually

The anchoring effect of buyers who have previously owned more expensive properties is not as robust as the learning and budget effect. It is only weakly significant for three out of seven models (p-value of 0.13 to 0.16). Yet, this lack of robustness does not cast serious doubt on a result already established in the literature (see Ihlandfelt and Mayock 2012). Indeed, only

[26] According to the French national statistics bureau INSEE, apartments prices in the Paris region increase by +8.8% from Q4-2015 to Q4-2017, whereas they decrease by -1.1% over the year 2015.

a third of the buyers declare to MA to own a home prior to the purchase, leaving the anchor variables at 0 for most of the dataset.

We cannot be so confident for the other two main results presented in part V related to the buyer's time preference and the size of the segment scanned. The effect of the time-on-market remains negative for all the models but is significant only for half of them. In model (11), the fact we do not control for the variance in search duration due to market conditions heterogeneity across very different cities, from the heavily urbanized direct surroundings of Paris to the rural areas of the "Grande-Couronne", may be an explanation. There is no such argument for its weak significance in models (13) and (15), with p-values of 0.13 and 0.19 respectively, and non-significance in model (16). The proxy of the search duration based on the activity of buyers on the MA website might reach its limit here.

Finally, as appealing as they might be and conformed to the predictions of equation (7), the results related to the trade-off in the choice of the size of the segment is not robust. The effect of the size flexibility is significant only in four out of eight models. The joint negative effect of *nb. boroughs* and positive effect of $nb.\ boroughs^2$ remain statistically significant only in models (15) and (16).

VII. Conclusion

This paper presents an extension of the standard theoretical search model for the homebuyer, developed in Turnbull and Sirmans (1993) and Lamnson et al. (2004). We test the model's predictions on 2,372 apartment purchases, in the Parisian region, between 2015 and 2018. This dataset is based on users 'activity on the French Internet real estate platform MeilleursAgents.com. For each apartment, it records information about the transaction per se (price, date, apartment 'features) and the search that have led to this purchase.

From a theoretical perspective, the model accounts for the buyer's optimal response to both the meeting hazard rate and the impact of the size of the segment she searches in. We derive the same results as Turnbull and Sirmans (1993) and Lamnson et al. (2004) about the effect on the buyer's reservation price of her search costs, time preference and bias in her beliefs about price distribution. The model predicts that a buyer facing a tighter market increases her stopping rules and an ambiguous impact of the market segment size. A segment size increase

reduces the reservation price and thus the expected value of the final purchase price if the increase in meeting rate overcomes the increase in the search costs.

Our main focus is empirical. This research introduces new results to the home buying problem literature. Overall, we find that the searches that led to these purchases partly explain the price difference for similar apartments. First, buyers that visit apartments more (resp. less) expensive per square meter than the one they purchase at the bottom line pay a premium (resp. get a discount). This result, robust to all the tests of section VI, is evidence that the homebuyer search follows a search-with-learning process as described in Kohn and Shavell (1974). Second, buyers 'budget constraint impacts the final transaction price: buyers who visit cheaper apartments, in absolute value, pay less for the same apartment. Third, we measure a negative, statistically significant, yet not entirely robust, relationship between the buyer's time-on-market and the price she ends up. Finally, we test the ambiguous effect of the segment size on the reservation price. We find the first evidence of a positive effect of its geographical spread and a negative one of looser characteristics constraints on the transaction price. However, these results fail the robustness tests of section VI. Altogether, the novelty of the results and the use of an unconventional data source call both for the study to be replicated.

Besides participating in the empirical literature on the home buying problem, the present paper also contributes to the real estate literature that relies on Internet datasets. Websites have become central to the housing market, both due to the share of users, the range of products offered and services provided. Internet datasets could help answer questions so far overlooked because of the absence of data. We believe that they should be the basis of numerous studies in the years to come - not only in the real estate literature. As the share of online economic activities grows, opportunities to confront theories with facts will grow as well.

Dependent variable	ln(pm²)							
	Paris (10)	Suburbs (11)	Small Apartment (12)	Large Apartment (13)	Sold in 2015 (14)	Sold in 2016+ (15)	Exp. buyers (16)	Non exp. buyers (17)
intercept	2.5096*** (0.275)	0.1556 (0.140)	0.7209*** (0.150)	0.0043 (0.133)	0.5086*** (0.138)	0.0345 (0.144)	0.3499° (0.187)	0.2271° (0.119)
rental investment	-0.0094 (0.016)	-0.0187 (0.016)	-0.0136 (0.013)	-0.0725** (0.023)	-0.0243 (0.016)	-0.0157 (0.016)	-0.0267 (0.019)	-0.0183 (0.015)
other	0.0060 (0.045)	0.0025 (0.066)	-0.0611 (0.048)	0.0376 (0.063)	-0.0136 (0.055)	-0.0605 (0.053)	-0.0699 (0.050)	0.0666 (0.066)
vacation home	0.0734** (0.022)	0.0085 (0.029)	0.0305 (0.021)	0.0369 (0.032)	0.0291 (0.025)	0.0329 (0.025)	0.0613* (0.029)	0.0085 (0.023)
ln(TOM norm)	-0.0206** (0.007)	-0.0016 (0.006)	-0.0130* (0.007)	-0.0100 (0.007)	-0.0121° (0.007)	-0.0086 (0.007)	-0.0017 (0.010)	-0.0146** (0.005)
ln(nb estimation)	-0.0027 (0.007)	-0.0073 (0.006)	-0.0127° (0.007)	0.0012 (0.006)	-0.0004 (0.007)	-0.0063 (0.007)	-0.0175* (0.009)	0.0023 (0.006)
learningEffectMinus	-0.0128 (0.049)	0.2066** (0.063)	0.0922° (0.055)	0.2164*** (0.056)	0.3043*** (0.056)	0.0058 (0.057)	0.1365* (0.063)	0.1436** (0.051)
learningEffectPlus	0.2774*** (0.079)	0.2125*** (0.044)	0.1866*** (0.052)	0.2543*** (0.056)	0.2133*** (0.047)	0.2603*** (0.056)	0.2322*** (0.062)	0.2597*** (0.045)
anchorEffectMinus	0.0018 (0.020)	-0.0116 (0.034)	0.0027 (0.024)	0.0177 (0.025)	0.0360 (0.023)	-0.0341 (0.026)	0.0064 (0.019)	
anchorEffectPlus	0.0506 (0.057)	0.0865* (0.039)	0.0436 (0.046)	0.1016* (0.044)	0.0519 (0.044)	0.0942* (0.046)	0.0802* (0.036)	
budgetEffectMinus	0.0848** (0.028)	0.2291*** (0.044)	0.2240*** (0.042)	0.0599° (0.031)	0.0598* (0.030)	0.2038*** (0.039)	0.0858* (0.041)	0.1388*** (0.030)
budgetEffectPlus	-0.0054 (0.015)	-0.0175 (0.034)	-0.0086 (0.015)	0.0026 (0.046)	-0.0098 (0.015)	0.0036 (0.031)	8.565e-05 (0.019)	-0.0191 (0.021)
nb bouroughs	0.0086 (0.007)	0.0040 (0.007)	0.0165* (0.008)	0.0061 (0.007)	0.0052 (0.007)	0.0140° (0.008)	0.0130 (0.008)	0.0109° (0.006)
nb bouroughs²	-0.0004 (0.000)	-0.0005 (0.000)	-0.0007 (0.001)	-0.0006 (0.000)	-0.0003 (0.000)	-0.0009° (0.001)	-0.0006 (0.001)	-0.0009* (0.000)
already owner	0.0043 (0.013)	-0.0053 (0.014)	0.0068 (0.014)	0.0040 (0.013)	0.0099 (0.013)	-0.0043 (0.014)		
size flexibility	-0.0217 (0.019)	-0.0093 (0.022)	-0.0438* (0.019)	-0.0210 (0.022)	-0.0380° (0.020)	-0.0354° (0.021)	-0.0383° (0.022)	-0.0335 (0.020)
ln(MA pm²)	0.7200*** (0.031)	0.9869*** (0.016)	0.9192*** (0.017)	1.0044*** (0.015)	0.9446*** (0.015)	0.9968*** (0.016)	0.9670*** (0.021)	0.9745*** (0.013)
Hedonic Control	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Nb Observation	1112	1260	1046	1326	1106	1266	813	1559
R^2	0.491	0.816	0.831	0.874	0.868	0.850	0.843	0.863
Adj-R^2	0.467	0.809	0.823	0.869	0.862	0.844	0.833	0.859
F-statistic	20.05	105.4	106.9	180.7	145.1	137.8	81.95	198.4

Standard errors between parenthesis
*** p <0.001, ** p <0.01, * p <0.05, ° p <0.1

Table 3-7 : Equation (8) OLS regression on data subsets result

Appendix

Appendix A: Table 4 complete

Dependent variable	ln(pm²)					
	Hedonic (1)	Cost (2)	Time (3)	Info (4)	Size (5)	Full (6)
intercept	0.6630*** (0.095)	0.6675*** (0.096)	0.6412*** (0.096)	0.2802** (0.099)	0.6703*** (0.095)	0.2580** (0.099)
rental investment				-0.0227* (0.011)		-0.0240* (0.011)
other						-0.0314 (0.038)
vacation home		0.0367* (0.018)		0.0296° (0.018)		0.0269 (0.018)
ln(TOM norm)			-0.0106* (0.005)			-0.0108* (0.005)
ln(nb estimation)		-0.0006 (0.003)				-0.0037 (0.005)
learningEffectMinus				0.1372*** (0.039)		0.1375*** (0.039)
learningEffectPlus				0.2416*** (0.035)		0.2497*** (0.035)
anchorEffectMinus				0.0028 (0.017)		0.0037 (0.017)
anchorEffectPlus				0.0800* (0.032)		0.0753* (0.032)
budgetEffectMinus				0.1303*** (0.024)		0.1230*** (0.024)
budgetEffectPlus				-0.0213 (0.013)		-0.0100 (0.014)
nb bouroughs					0.0110* (0.004)	0.0109* (0.005)
nb bouroughs²					-0.0007* (0.000)	-0.0007* (0.000)
already owner				-0.0030 (0.009)		0.0029 (0.010)
size flexibility					-0.0391** (0.013)	-0.0340* (0.014)
ln(area/40)	-0.1014*** (0.018)	-0.1000*** (0.018)	-0.0996*** (0.018)	-0.1171*** (0.018)	-0.1000*** (0.018)	-0.1145*** (0.018)
ln(MA pm²)	0.9273*** (0.011)	0.9267*** (0.011)	0.9296*** (0.011)	0.9713*** (0.011)	0.9254*** (0.011)	0.9727*** (0.011)
1 room	-0.0086 (0.015)	-0.0095 (0.015)	-0.0076 (0.015)	-0.0112 (0.015)	-0.0046 (0.015)	-0.0086 (0.015)
3 rooms	0.0120 (0.011)	0.0122 (0.011)	0.0115 (0.011)	0.0190° (0.011)	0.0101 (0.011)	0.0172 (0.011)
4 rooms	0.0117 (0.016)	0.0123 (0.016)	0.0117 (0.016)	0.0219 (0.016)	0.0105 (0.016)	0.0208 (0.016)
5 rooms	0.0218 (0.023)	0.0227 (0.023)	0.0213 (0.023)	0.0345 (0.022)	0.0192 (0.023)	0.0320 (0.022)

6 rooms	0.0646° (0.038)	0.0659° (0.038)	0.0653° (0.038)	0.1032** (0.037)	0.0668° (0.038)	0.1054** (0.037)
2 bathrooms	0.0424** (0.013)	0.0418** (0.013)	0.0431** (0.013)	0.0484*** (0.013)	0.0435** (0.013)	0.0496*** (0.013)
3 bathrooms	0.0453 (0.044)	0.0421 (0.044)	0.0444 (0.044)	0.0352 (0.043)	0.0440 (0.044)	0.0302 (0.043)
ground floor	-0.0437** (0.013)	-0.0443** (0.013)	-0.0451** (0.013)	-0.0392** (0.013)	-0.0459*** (0.013)	-0.0429** (0.013)
1st floor	-0.0052 (0.011)	-0.0044 (0.011)	-0.0055 (0.011)	-0.0005 (0.011)	-0.0065 (0.011)	-0.0015 (0.011)
3rd floor	0.0206° (0.011)	0.0207° (0.011)	0.0210° (0.011)	0.0196° (0.011)	0.0202° (0.011)	0.0194° (0.011)
4th floor	0.0128 (0.013)	0.0135 (0.013)	0.0131 (0.013)	0.0090 (0.012)	0.0132 (0.013)	0.0101 (0.012)
5th floor	0.0390** (0.015)	0.0392** (0.015)	0.0387** (0.015)	0.0387** (0.014)	0.0403** (0.015)	0.0397** (0.014)
6th floor and above	0.0632*** (0.015)	0.0636*** (0.015)	0.0634*** (0.015)	0.0499** (0.014)	0.0624*** (0.015)	0.0496** (0.014)
no elevator	-0.0186* (0.009)	-0.0184* (0.009)	-0.0183* (0.009)	-0.0194* (0.009)	-0.0192* (0.009)	-0.0197* (0.009)
has outdoor	0.0059 (0.008)	0.0056 (0.008)	0.0061 (0.008)	0.0059 (0.008)	0.0060 (0.008)	0.0064 (0.008)
no cellar	0.0229** (0.009)	0.0223* (0.009)	0.0222* (0.009)	0.0232** (0.009)	0.0245** (0.009)	0.0236** (0.009)
exceptional view	0.0251* (0.011)	0.0253* (0.011)	0.0254* (0.011)	0.0291** (0.010)	0.0245* (0.011)	0.0288** (0.010)
Inf. relative perception	-0.1086** (0.036)	-0.1108** (0.036)	-0.1048** (0.036)	-0.1060** (0.035)	-0.0949** (0.037)	-0.0841* (0.036)
Sup. relative perception	0.0043 (0.007)	0.0045 (0.007)	0.0049 (0.007)	0.0048 (0.007)	0.0046 (0.007)	0.0054 (0.007)
renovation needed	-0.0566** (0.017)	-0.0566** (0.017)	-0.0582** (0.017)	-0.0548** (0.017)	-0.0577** (0.017)	-0.0580** (0.017)
construction date <1850	0.0222 (0.025)	0.0166 (0.025)	0.0174 (0.025)	0.0085 (0.024)	0.0220 (0.025)	0.0032 (0.024)
construction date [1850; 1913]	0.0033 (0.011)	0.0030 (0.011)	0.0025 (0.011)	0.0009 (0.011)	0.0029 (0.011)	-0.0009 (0.011)
construction date [1914; 1947]	-0.0032 (0.012)	-0.0032 (0.012)	-0.0025 (0.012)	-0.0006 (0.012)	-0.0036 (0.012)	-0.0004 (0.012)
construction date [1948; 1969]	-0.0354** (0.012)	-0.0356** (0.012)	-0.0358** (0.012)	-0.0340** (0.012)	-0.0367** (0.012)	-0.0357** (0.012)
construction date [1970; 1980]	-0.0725*** (0.013)	-0.0727*** (0.013)	-0.0726*** (0.013)	-0.0615*** (0.012)	-0.0719*** (0.013)	-0.0624*** (0.012)
construction date [1981; 1990]	-0.0693** (0.022)	-0.0693** (0.022)	-0.0691** (0.022)	-0.0612** (0.021)	-0.0701** (0.022)	-0.0620** (0.021)
construction date [1991; 2000]	-0.0732** (0.021)	-0.0718** (0.021)	-0.0713** (0.021)	-0.0591** (0.020)	-0.0731** (0.021)	-0.0561** (0.020)
construction date [2001; 2010]	-0.0007 (0.019)	-0.0004 (0.019)	5.141e-05 (0.019)	0.0078 (0.018)	-0.0041 (0.019)	0.0063 (0.018)

construction date > 2011	0.1245*** (0.019)	0.1247*** (0.019)	0.1232*** (0.019)	0.1195*** (0.019)	0.1226*** (0.019)	0.1135*** (0.019)
sold in 2016	-0.0112 (0.008)	-0.0117 (0.008)	-0.0117 (0.008)	-0.0076 (0.008)	-0.0102 (0.008)	-0.0070 (0.008)
sold in 2017	-0.0261** (0.009)	-0.0260** (0.009)	-0.0265** (0.009)	-0.0150° (0.009)	-0.0260** (0.009)	-0.0159° (0.009)
sold in 2018	0.0334 (0.030)	0.0351 (0.030)	0.0255 (0.030)	0.0459 (0.029)	0.0352 (0.030)	0.0394 (0.029)
bulding levels <5	-0.0063 (0.009)	-0.0058 (0.009)	-0.0057 (0.009)	0.0037 (0.009)	-0.0055 (0.009)	0.0045 (0.009)
building levels > 7	-0.0442*** (0.013)	-0.0437** (0.013)	-0.0441*** (0.013)	-0.0414** (0.012)	-0.0441*** (0.013)	-0.0415** (0.012)
Nb Observations	2372	2372	2372	2372	2372	2372
R^2	0.841	0.842	0.853	0.842	0.853	0.854
Adj-R^2	0.839	0.839	0.850	0.839	0.850	0.850
F-statistic	343.7	326.0	298.8	318.5	298.8	265.3

Standard errors between parenthesis
*** p <0.001, ** p <0.01, * p <0.05, ° p <0.1

Table 3-8 : Table 3.4 complete

Appendix B: Table 5 complete

Dependent variable	ln(pm²)			
	Learning (7)	Anchor (8)	Budget (9)	Full (6)
intercept	0.2902** (0.099)	0.5851*** (0.098)	0.5717*** (0.095)	0.2580** (0.099)
rental investment	-0.0252* (0.011)	-0.0254* (0.012)	-0.0167 (0.011)	-0.0240* (0.011)
other	-0.0221 (0.038)	-0.0506 (0.039)	-0.0323 (0.039)	-0.0314 (0.038)
vacation home	0.0216 (0.018)	0.0289 (0.018)	0.0327° (0.018)	0.0269 (0.018)
ln(TOM norm)	-0.0120* (0.005)	-0.0107* (0.005)	-0.0097* (0.005)	-0.0108* (0.005)
ln(nb estimation)	-0.0010 (0.005)	-0.0005 (0.005)	-0.0053 (0.005)	-0.0037 (0.005)
learningEffectMinus	0.2270*** (0.035)			0.1375*** (0.039)
learningEffectPlus	0.2739*** (0.034)			0.2497*** (0.035)
anchorEffectMinus		0.0046 (0.018)		0.0037 (0.017)
anchorEffectPlus		0.1289*** (0.032)		0.0753* (0.032)
budgetEffectMinus			0.1854*** (0.022)	0.1230*** (0.024)
budgetEffectPlus			0.0131 (0.014)	-0.0100 (0.014)
nb bouroughs	0.0115* (0.005)	0.0112* (0.005)	0.0132** (0.005)	0.0109* (0.005)
nb bouroughs²	-0.0007* (0.000)	-0.0006° (0.000)	-0.0007* (0.000)	-0.0007* (0.000)
already owner	0.0108 (0.007)	-0.0038 (0.010)	0.0109 (0.008)	0.0029 (0.010)
size flexibility	-0.0439** (0.014)	-0.0362** (0.014)	-0.0331* (0.015)	-0.0340* (0.014)
ln(area/40)	-0.1212*** (0.018)	-0.1087*** (0.018)	-0.0946*** (0.018)	-0.1145*** (0.018)
ln(MA pm²)	0.9682*** (0.011)	0.9349*** (0.011)	0.9376*** (0.011)	0.9727*** (0.011)
1 room	-0.0102 (0.015)	-0.0010 (0.016)	-0.0043 (0.015)	-0.0086 (0.015)
3 rooms	0.0185° (0.011)	0.0129 (0.011)	0.0084 (0.011)	0.0172 (0.011)
4 rooms	0.0225 (0.016)	0.0149 (0.016)	0.0091 (0.016)	0.0208 (0.016)
5 rooms	0.0343 (0.022)	0.0250 (0.023)	0.0185 (0.023)	0.0320 (0.022)
6 rooms	0.0942* (0.037)	0.0751* (0.038)	0.0936* (0.037)	0.1054** (0.037)
2 bathrooms	0.0442** (0.013)	0.0443** (0.013)	0.0541*** (0.013)	0.0496*** (0.013)
3 bathrooms	0.0287 (0.043)	0.0463 (0.044)	0.0358 (0.043)	0.0302 (0.043)
ground floor	-0.0444*** (0.013)	-0.0462*** (0.013)	-0.0462*** (0.013)	-0.0429** (0.013)
1st floor	-0.0022 (0.011)	-0.0054 (0.011)	-0.0037 (0.011)	-0.0015 (0.011)
3rd floor	0.0184° (0.011)	0.0183 (0.011)	0.0217° (0.011)	0.0194° (0.011)
4th floor	0.0115 (0.012)	0.0127 (0.013)	0.0127 (0.012)	0.0101 (0.012)
5th floor	0.0386** (0.014)	0.0397** (0.015)	0.0414** (0.014)	0.0397** (0.014)
6th floor and above	0.0487** (0.014)	0.0590*** (0.015)	0.0612*** (0.015)	0.0496** (0.014)
no elevator	-0.0208* (0.009)	-0.0191* (0.009)	-0.0178* (0.009)	-0.0197* (0.009)
has outdoor	0.0076 (0.008)	0.0042 (0.008)	0.0038 (0.008)	0.0064 (0.008)
no cellar	0.0226** (0.009)	0.0241** (0.009)	0.0246** (0.009)	0.0236** (0.009)
exceptional view	0.0277** (0.010)	0.0251* (0.011)	0.0278** (0.011)	0.0288** (0.010)
Inf. relative perception	-0.0736* (0.036)	-0.0888* (0.037)	-0.0981** (0.036)	-0.0841* (0.036)
Sup. relative perception	0.0040 (0.007)	0.0043 (0.007)	0.0077 (0.007)	0.0054 (0.007)
renovation needed	-0.0558** (0.017)	-0.0561** (0.017)	-0.0614*** (0.017)	-0.0580** (0.017)
construction date <1850	0.0076 (0.024)	0.0090 (0.025)	0.0062 (0.025)	0.0032 (0.024)
construction date [1850; 1913]	-0.0009 (0.011)	0.0012 (0.011)	0.0007 (0.011)	-0.0009 (0.011)
construction date [1914; 1947]	-0.0003 (0.012)	-0.0038 (0.012)	-0.0026 (0.012)	-0.0004 (0.012)
construction date [1948; 1969]	-0.0347** (0.012)	-0.0380** (0.012)	-0.0367** (0.012)	-0.0357** (0.012)
construction date [1970; 1980]	-0.0648*** (0.012)	-0.0696*** (0.012)	-0.0694*** (0.012)	-0.0624*** (0.012)

construction date [1981; 1990]	-0.0622** (0.021)	-0.0655** (0.022)	-0.0730** (0.022)	-0.0620** (0.021)
construction date [1991; 2000]	-0.0536** (0.021)	-0.0701** (0.021)	-0.0645** (0.021)	-0.0561** (0.020)
construction date [2001; 2010]	0.0066 (0.018)	-0.0014 (0.019)	-0.0015 (0.019)	0.0063 (0.018)
construction date > 2011	0.1111*** (0.019)	0.1215*** (0.019)	0.1259*** (0.019)	0.1135*** (0.019)
sold in 2016	-0.0067 (0.008)	-0.0100 (0.008)	-0.0106 (0.008)	-0.0070 (0.008)
sold in 2017	-0.0188* (0.009)	-0.0255** (0.009)	-0.0194* (0.009)	-0.0159° (0.009)
sold in 2018	0.0464 (0.030)	0.0281 (0.030)	0.0278 (0.030)	0.0394 (0.029)
bulding levels <5	0.0044 (0.009)	-0.0034 (0.009)	-0.0021 (0.009)	0.0045 (0.009)
building levels > 7	-0.0436*** (0.012)	-0.0428** (0.013)	-0.0407** (0.012)	-0.0415** (0.012)
Nb Observations	2372	2372	2372	2372
R^2	0.852	0.844	0.848	0.854
Adj-R^2	0.849	0.841	0.845	0.850
F-statistic	283.7	267.9	267.9	265.3

Standard errors between parenthesis
*** p <0.001, ** p <0.01, * p <0.05, ° p <0.1

Table 3-9 : Table 3.5 complete

Appendix C: Table 7 complete

Dependent variable	ln(pm²)							
	Paris (10)	Suburbs (11)	Small Apart. (12)	Large Apart. (13)	Sold in 2015 (14)	Sold in 2016+ (15)	Exp. buyers (16)	Non exp. buyers (17)
intercept	2.5096*** (0.275)	0.1556 (0.140)	0.7209*** (0.150)	0.0043 (0.133)	0.5086*** (0.138)	0.0345 (0.144)	0.3499° (0.187)	0.2271° (0.119)
rental investment	-0.0094 (0.016)	-0.0187 (0.016)	-0.0136 (0.013)	-0.0725** (0.023)	-0.0243 (0.016)	-0.0157 (0.016)	-0.0267 (0.019)	-0.0183 (0.015)
other	0.0060 (0.045)	0.0025 (0.066)	-0.0611 (0.048)	0.0376 (0.063)	-0.0136 (0.055)	-0.0605 (0.053)	-0.0699 (0.050)	0.0666 (0.066)
vacation home	0.0734** (0.022)	0.0085 (0.029)	0.0305 (0.021)	0.0369 (0.032)	0.0291 (0.025)	0.0329 (0.025)	0.0613* (0.029)	0.0085 (0.023)
ln(TOM norm)	-0.0206** (0.007)	-0.0016 (0.006)	-0.0130* (0.007)	-0.0100 (0.007)	-0.0121° (0.007)	-0.0086 (0.007)	-0.0017 (0.010)	-0.0146** (0.005)
ln(nb estimation)	-0.0027 (0.007)	-0.0073 (0.006)	-0.0127° (0.007)	0.0012 (0.006)	-0.0004 (0.007)	-0.0063 (0.007)	-0.0175* (0.009)	0.0023 (0.006)
learningEffectMinus	-0.0128 (0.049)	0.2066** (0.063)	0.0922° (0.055)	0.2164*** (0.056)	0.3043*** (0.056)	0.0058 (0.057)	0.1365* (0.063)	0.1436** (0.051)
learningEffectPlus	0.2774*** (0.079)	0.2125*** (0.044)	0.1866*** (0.052)	0.2543*** (0.056)	0.2133*** (0.047)	0.2603*** (0.056)	0.2322*** (0.062)	0.2597*** (0.045)
anchorEffectMinus	0.0018 (0.020)	-0.0116 (0.034)	0.0027 (0.024)	0.0177 (0.025)	0.0360 (0.023)	-0.0341 (0.026)	0.0064 (0.019)	
anchorEffectPlus	0.0506 (0.057)	0.0865* (0.039)	0.0436 (0.046)	0.1016* (0.044)	0.0519 (0.044)	0.0942* (0.046)	0.0802* (0.036)	
budgetEffectMinus	0.0848** (0.028)	0.2291*** (0.044)	0.2240*** (0.042)	0.0599° (0.031)	0.0598* (0.030)	0.2038*** (0.039)	0.0858* (0.041)	0.1388*** (0.030)
budgetEffectPlus	-0.0054 (0.015)	-0.0175 (0.034)	-0.0086 (0.015)	0.0026 (0.046)	-0.0098 (0.015)	0.0036 (0.031)	8.565e-05 (0.019)	-0.0191 (0.021)
nb bouroughs	0.0086 (0.007)	0.0040 (0.007)	0.0165* (0.008)	0.0061 (0.007)	0.0052 (0.007)	0.0140° (0.008)	0.0130 (0.008)	0.0109° (0.006)
nb bouroughs²	-0.0004 (0.000)	-0.0005 (0.000)	-0.0007 (0.001)	-0.0006 (0.000)	-0.0003 (0.000)	-0.0009° (0.001)	-0.0006 (0.001)	-0.0009* (0.000)
already owner	0.0043 (0.013)	-0.0053 (0.014)	0.0068 (0.014)	0.0040 (0.013)	0.0099 (0.013)	-0.0043 (0.014)		
size flexibility	-0.0217 (0.019)	-0.0093 (0.022)	-0.0438* (0.019)	-0.0210 (0.022)	-0.0380° (0.020)	-0.0354° (0.021)	-0.0383° (0.022)	-0.0335 (0.020)
ln(area/40)	-0.0477* (0.023)	-0.1621*** (0.028)	-0.1146*** (0.023)	-0.0959** (0.029)	-0.1149*** (0.024)	-0.0976*** (0.026)	-0.1270*** (0.031)	-0.1045*** (0.022)
ln(MA pm²)	0.7200*** (0.031)	0.9869*** (0.016)	0.9192*** (0.017)	1.0044*** (0.015)	0.9446*** (0.015)	0.9968*** (0.016)	0.9670*** (0.021)	0.9745*** (0.013)
1 room	-0.0194 (0.020)	0.0016 (0.022)	-0.0160 (0.017)		0.0041 (0.021)	-0.0135 (0.022)	-0.0106 (0.027)	-0.0104 (0.019)
3 rooms	0.0072 (0.015)	0.0287° (0.015)		-0.0094 (0.021)	0.0060 (0.015)	0.0204 (0.016)	0.0172 (0.020)	0.0149 (0.013)
4 rooms	0.0215 (0.023)	0.0368° (0.022)		-0.0099 (0.017)	0.0008 (0.022)	0.0279 (0.023)	0.0266 (0.028)	0.0209 (0.020)
5 rooms	0.0378 (0.031)	0.0301 (0.031)			0.0461 (0.032)	0.0130 (0.031)	0.0531 (0.038)	0.0186 (0.028)

6 rooms	0.0806°	0.0798		0.0516	0.0770	0.1081*	0.1233°	0.0966*
	(0.046)	(0.061)		(0.034)	(0.052)	(0.053)	(0.063)	(0.046)
2 bathrooms	0.0183	0.0720***	0.0883	0.0320*	0.0681***	0.0336°	0.0545*	0.0402*
	(0.019)	(0.018)	(0.055)	(0.014)	(0.019)	(0.018)	(0.022)	(0.017)
3 bathrooms	0.0228	0.0555		0.0071	-0.0541	0.1154°	0.0064	0.0486
	(0.059)	(0.060)		(0.044)	(0.056)	(0.064)	(0.065)	(0.059)
ground floor	-0.1276***	-0.0001	-0.0315°	-0.0542**	-0.0166	-0.0603**	-0.0501*	-0.0437**
	(0.021)	(0.016)	(0.018)	(0.018)	(0.018)	(0.018)	(0.023)	(0.015)
1st floor	-0.0138	0.0133	0.0184	-0.0129	0.0017	-0.0057	-0.0129	0.0045
	(0.015)	(0.014)	(0.016)	(0.014)	(0.015)	(0.015)	(0.020)	(0.013)
3rd floor	0.0071	0.0319*	0.0355*	0.0116	0.0328*	0.0070	0.0086	0.0213
	(0.016)	(0.015)	(0.017)	(0.015)	(0.016)	(0.016)	(0.020)	(0.014)
4th floor	0.0252	0.0026	0.0468**	-0.0157	0.0347*	-0.0072	-0.0146	0.0218
	(0.017)	(0.017)	(0.018)	(0.017)	(0.017)	(0.018)	(0.022)	(0.015)
5th floor	0.0144	0.0532*	0.0234	0.0622**	0.0284	0.0502*	0.0296	0.0408*
	(0.018)	(0.022)	(0.021)	(0.019)	(0.019)	(0.021)	(0.028)	(0.017)
6th floor and above	0.0478*	0.0424°	0.0437*	0.0687***	0.0593**	0.0425*	0.0347	0.0527**
	(0.019)	(0.022)	(0.021)	(0.020)	(0.020)	(0.021)	(0.026)	(0.017)
no elevator	-0.0128	-0.0256*	-0.0108	-0.0144	-0.0154	-0.0155	-0.0513**	-0.0037
	(0.012)	(0.012)	(0.013)	(0.012)	(0.012)	(0.013)	(0.016)	(0.010)
has outdoor	0.0302*	0.0006	0.0105	0.0133	0.0058	0.0077	0.0016	0.0061
	(0.012)	(0.011)	(0.012)	(0.011)	(0.011)	(0.012)	(0.015)	(0.010)
no cellar	0.0160	0.0294*	0.0323**	0.0136	0.0189	0.0318*	0.0119	0.0290**
	(0.012)	(0.012)	(0.011)	(0.013)	(0.012)	(0.012)	(0.016)	(0.010)
exceptional view	0.0207	0.0455**	0.0275°	0.0208	0.0279*	0.0336*	0.0341°	0.0270*
	(0.015)	(0.014)	(0.016)	(0.013)	(0.014)	(0.015)	(0.018)	(0.013)
Inf. relative perception	-0.1145**	0.0052	-0.0384	-0.1391**	-0.0869°	-0.0988°	-0.0355	-0.1397**
	(0.043)	(0.062)	(0.051)	(0.051)	(0.046)	(0.056)	(0.062)	(0.045)
Sup. relative perception	0.0030	0.0030	-0.0131	0.0154°	0.0119	0.0012	0.0041	0.0039
	(0.010)	(0.010)	(0.010)	(0.009)	(0.009)	(0.010)	(0.013)	(0.008)
renovation needed	-0.0604**	-0.0706**	-0.0932***	-0.0266	-0.0715**	-0.0462°	-0.0387	-0.0657**
	(0.023)	(0.024)	(0.024)	(0.023)	(0.023)	(0.024)	(0.030)	(0.020)
construction date <1850	0.0006	0.1104°	0.0206	-0.0122	0.0276	-0.0329	0.0480	-0.0211
	(0.026)	(0.066)	(0.032)	(0.036)	(0.031)	(0.038)	(0.041)	(0.031)
construction date [1850; 1913]	-0.0079	-0.0043	0.0009	-0.0067	0.0122	-0.0154	-0.0176	0.0045
	(0.012)	(0.024)	(0.015)	(0.015)	(0.015)	(0.016)	(0.019)	(0.013)
construction date [1914; 1947]	0.0034	-0.0034	-0.0120	0.0032	-0.0093	0.0038	-0.0346	0.0134
	(0.018)	(0.016)	(0.016)	(0.017)	(0.016)	(0.018)	(0.023)	(0.014)
construction date [1948; 1969]	-0.0105	-0.0434**	-0.0271	-0.0441**	-0.0348*	-0.0377*	-0.0037	-0.0460**
	(0.019)	(0.015)	(0.019)	(0.015)	(0.015)	(0.019)	(0.023)	(0.014)
construction date [1970; 1980]	-0.0558**	-0.0672***	-0.0609**	-0.0641***	-0.0627***	-0.0571**	-0.0656**	-0.0634***
	(0.018)	(0.016)	(0.018)	(0.016)	(0.017)	(0.018)	(0.021)	(0.015)
construction date [1981; 1990]	-0.0185	-0.0565*	-0.0148	-0.0856**	-0.1311***	-0.0160	-0.1258**	-0.0318
	(0.042)	(0.024)	(0.035)	(0.027)	(0.032)	(0.029)	(0.038)	(0.026)
construction date [1991; 2000]	-0.0153	-0.0575*	-0.0825**	-0.0468°	-0.0374	-0.0741*	-0.1225**	-0.0364
	(0.053)	(0.022)	(0.029)	(0.028)	(0.028)	(0.030)	(0.044)	(0.023)
construction date [2001; 2010]	0.0033	0.0054	-0.0420	0.0255	0.0185	-0.0004	-0.0280	0.0230
	(0.056)	(0.019)	(0.031)	(0.022)	(0.025)	(0.026)	(0.034)	(0.022)

construction date > 2011	0.1007* (0.049)	0.1243** * (0.020)	0.1265** * (0.032)	0.1110** * (0.023)	0.0885** (0.027)	0.1331** * (0.026)	0.0922** (0.032)	0.1197** * (0.023)
sold in 2016	-0.0097 (0.011)	-0.0084 (0.011)	-0.0253* (0.012)	0.0068 (0.010)			-0.0075 (0.014)	-0.0068 (0.009)
sold in 2017	0.0211 (0.013)	-0.0321** (0.012)	-0.0109 (0.013)	-0.0161 (0.012)		-0.0087 (0.010)	-0.0066 (0.016)	-0.0182° (0.011)
sold in 2018	0.0865* (0.038)	-0.0480 (0.046)	0.0169 (0.047)	0.0688° (0.039)		0.0364 (0.032)	0.1002° (0.060)	0.0311 (0.034)
bulding levels <5	0.0249° (0.015)	0.0161 (0.012)	0.0146 (0.013)	-0.0074 (0.013)	-0.0028 (0.012)	0.0121 (0.013)	0.0198 (0.016)	-0.0046 (0.011)
building levels > 7	-0.0721** * (0.016)	-0.0277 (0.018)	-0.0089 (0.018)	-0.0779** * (0.016)	-0.0580** (0.017)	-0.0335° (0.017)	-0.0577* (0.022)	-0.0313* (0.015)
Nb Observation	1112	1260	1046	1326	1106	1266	813	1559
R^2	0.491	0.816	0.831	0.874	0.868	0.850	0.843	0.863
Adj-R^2	0.467	0.809	0.823	0.869	0.862	0.844	0.833	0.859
F-statistic	20.05	105.4	106.9	180.7	145.1	137.8	81.95	198.4

Standard errors between parenthesis
*** p <0.001, ** p <0.01, * p <0.05, ° p <0.1

Table 3-10 : Table 3.7 complete

4. Homesellers and homebuyers self-reported estimations

Abstract:

We use self-reported estimations collected on an Internet platform to study the formation of housing price beliefs of sellers and buyers of the Paris region. Comparing the users' estimates with real prices of transactions, we are able to evaluate how accurate they are in predicting the selling price and to identify factors that influence their opinion. We confirm the already studied upward bias of homeowners but bring a first evidence that buyers are unbiased. Our results confirm that market participants follow an anchoring and adjustment scheme to set their opinion on housing price. Users' beliefs are influenced by the results of the platform Automated Valuation Model, which acts as an anchor. Buyers are less influence by this reference point, and the more the sellers progress in their sale project, the less their opinions rely on the model estimations. Finally, the estimates of owners whose property has lost value since they purchase it are anchored to its purchase price and less correlated to the AVM results. We interpret these results as proof of an updating mechanism of the buyers' and sellers' beliefs along their search for a counterpart.

Résumé :

Grâce à des estimations déclaratives collectées sur une plateforme Internet, nous étudions la formation des croyances des acheteurs et vendeurs de régions parisiennes sur les prix de l'immobilier. En comparant les estimations des utilisateurs avec les montants des transactions, nous sommes en mesure d'évaluer le degré de précision de leurs prédictions et d'identifier les facteurs qui les influencent. Nous confirmons le biais à la surestimation déjà établi des propriétaires, mais nous apportons une première preuve que les acheteurs sont eux non biaisés. Nos résultats confirment que les particuliers actifs sur le marché immobilier suivent une heuristique d'ancrage et ajustement pour définir leur opinion sur les prix de l'immobilier. Le résultat de l'outil d'estimation de la plateforme agit comme un point d'ancrage et influence les croyances des utilisateurs. Les acheteurs apparaissent moins influencés par ce point de référence, et plus les vendeurs progressent dans leur projet, plus ils en affranchissent également. Enfin, les estimations des propriétaires dont les biens ont perdu de la valeur depuis leur acquisition sont en partie basées sur leur prix d'achat et moins sur le résultat du moteur d'estimations. Nous interprétons ces résultats comme une preuve de l'existence d'un mécanisme de mise à jour des croyances des acheteurs et des vendeurs immobiliers tout au long de leur recherche.

I. Introduction

The housing market is decentralized, illiquid and dominated by amateurs, who for most of them engage only in a handful of transactions over their whole life. Yet, those amateurs in such a complicated setting have to agree on a price for what is often their largest asset (INSEE 2018) and for which there is no perfect comparable. Facing so much uncertainty, buyers and sellers in the real estate market do what humans usually do when they have to assess an unknown numerical quantity: they rely on heuristics. From the three heuristics introduced by Tversky and Kahneman (1974), the more likely they are to follow to form their beliefs about housing prices is anchoring and adjustment. Nothcraft and Neale (1987) is, to our knowledge, the first research to have brought anchoring and adjustment out of the laboratory to a real-life setting, through a real estate related field experiment.

Such an economical and usually useful heuristic (Tversky and Kahneman, 1974) is not perfect and tend to bias the decision maker's opinion. Designed experiments have revealed the role played by listing prices as anchors (Nothcraft and Neale 1987, Black and Diaz 1996) and that the internal reference prices on both sides of the transaction are adjusting depending on market evolution and available information (Corina and Chenavaz, 2011). Outside the lab, anchoring and loss aversion have measurable impacts on how the housing market work. Genesove and Mayer (2001) show that Boston's sellers are eager to jeopardize their chance to sell by setting higher asking price to avoid a nominal loss (see also Einio et al., 2007, and Anenberg, 2011). Anchoring disturbs the buyers' behavior as well. In different settings, Lambson et al. (2004) and Ihlandfelt and Mayock (2012) measure that homebuyers who come from more expensive areas pay more for a given house.

The study of the homeowners' opinion about their property values offers a great opportunity to directly test for the effects of the judgment under uncertainty, in real life. For tax or census purposes, owners are asked to evaluate the value of their home for a long time and across countries. From Kish and Lansing (1954) in the US to Van der Cruijsen et al. (2018) in the Netherland, researchers have analyzed their responses to feed a rich literature. The main focus of this line of research is to detect and measure the positive bias of the homeowners' predictions of the value of their homes. Using appraised values (Kish and Lansing 1954, Kain and Quigley 1971, Van der Cruijsen et al., 2018), hedonic estimates (Ihlandfelt and Martinez-Vazquez, 1980, Windsor et al. 2014) or subsequent transaction prices (Goodman and Ittner, 1995, Benitez-Silva et al. 2015) for comparison, most papers observe an overestimation above

+5%. Van der Cruijsen et al. (2018) identify two explanatory factors for this upward bias: endowment effect and loss aversion.

However, because respondents are not engaged in a sale process, these studies unveil only a part of how the sellers form their opinions. They do not tell us anything about the buyers' price discovery process, either. This is the gap the present study aims to fill. We collect estimates of apartment values that actively searching buyers and sellers made in reaction to the result of an online Automated Valuation Model (thereafter AVM), on a French real estate website, MeilleursAgents.com (thereafter MA). We were able to match 10,337 of these estimates, with sales which occurred within a year after the estimates, in the Paris region, between March 2012 and December 2019.

Our analysis of the sale price prediction errors brings another evidence of the owners' positive bias, as we measure sellers overestimate their apartment by +5.7% on average, within the range of previous studies. Conversely, buyers appear to be unbiased, in what is to our knowledge the first measure of the quality of their predictions. The fact that they are confronted with the market reality right from the beginning of their search, through ads and visits, seems to prevent them from an overly optimistic attitude. The sellers' "dream" can last longer because they face a contradiction only in the last steps of their process. As in Goodman and Ittner (1995), we measure a positive correlation of the sellers' overestimation with the time that separate the estimate and the sale. The search for a counterpart somehow mitigates some of their initial, unrealistic price expectations. We also replicate their measure of the negative impact of the price changes on the homeowners' estimates. The endowment effect, through the positive impact of a longer tenure on overestimation, already measure in Van der Cruijsen et al. (2018), is confirmed by our result.

But the main contribution enabled by this novel dataset concern the way housing market participants form and update their price beliefs. In line with anchoring and adjustment we measure that they use the result of AVM as a reference point to make their estimate. Yet, depending on the context, the weight of this reference is not equal. We measure the magnitude of the impact of the AVM result on the sellers' opinions to be two times larger than on the buyers. In our opinion the availability of the listing price, as another reference point, to buyers can explain this difference. Indeed, unlike most of the sellers in our study, who have not yet started the commercialization, the minority who have already listed their apartment and chosen an asking price also exhibits a reduced sensitivity to the AVM results. More generally, we

measure that the more the sellers get further into the sale process, the less he or she anchors is opinion to the AVM valuation. Finally, loss aversion also plays a role in the process of weighing different reference points. Sellers whose purchasing price is greater in nominal value than the subsequent sales price anchor their appraisal to this previous transaction price. Their estimates are also less impacted by the AVM results than the ones who did not suffer from such a loss.

The rest of the paper is organized at it follows. The second section presents more precisely the way the estimates were collected on the website and matched with the notarial sales records. It also brings first statistical descriptions and tackles the selection bias issue specific to datasets drawn from the Internet. Then we present our first econometric results on the different driver of the users' sale price prediction error. The fourth part focus more specifically on the impact of loss aversion. In the fifth part, we bring a dynamic perspective to the analysis, by studying how our results evolve as we take the evolution within the search process into account. Finally, the sixth section concludes.

II. Data

This section describes how we have collected the data used in this study on the MA website. MA is an Internet company that makes information about the French housing market available on line, for free. This includes indices, price maps and an automated valuation model (thereafter AVM) which enable users to estimate the value of any property in France based on the location and description of the property. Two million unique visitors, who perform 50,000 estimations, visit the MA website every month (source MA).

In this article, we focus on the data collected through a feature of the AVM called by the company, "the users' feedback". At the end of the estimation funnel, after the user has filled in the online form, once the model estimation is completed, and the estimated value is displayed, the user is invited to optionally leave feedback about the estimated value. Two questions are asked: "What do you think about the estimated value?" and "How much do you think this property worth?" Five answers are proposed to answer at the first questions: "Way too low", "Too low", "Fair", "Too high" and "Way too high". To answer the second question, the user has to type in what he or she would consider to be a fair value for the property. These users' self-estimations of the properties' values are the variable of interest of the present work.

According to MA, 5% of the users of the AVM leave a numerical estimate, after they get the model result. The company gave us access to all feedback concerning apartments located

in the Paris region. MA started collect "feedback" in 2011, which accounts for a vast number of users' estimations that spread over a bust (2011-2015) and a boom (2016-2019). Yet, the datasets drawn from the numerical world are drifting (Salganik 2017). The users' classification by MA (the owner, seller and buyer categories) changed until early 2012. Thus, our sample contains feedback collected between March 2012 and December 2019. Moreover, only a fraction is effectively usable in our study. Indeed, to understand what factors impact the user's opinion, it must be compared to a fair assessment of the market value of this apartment. Following Goodman and Ittner (1992), we consider that the value of property at a given moment is best defined by its sale price. Hence, we only use feedback about properties that were sold in the year following the user estimate and that we were able to match with a transaction recorded in the notarial database.

MA have access to an almost exhaustive record of the Paris region since the early 1990's. For each transaction in the database the amount, the date of the authentic act of the transfer of ownership, a description of the apartment, the exact location and, for most of them, the amount and date of the previous transaction are recorded. As we want to use this previous purchase price in the analysis, we only consider the records with this information. To compare the users' estimations with transaction prices, we match the feedback listing and the notarial database. In detail we link a user estimate and a sale if:

- They concern the same apartment:
    - Same address
    - Same floor
    - Same living area (with a 10% margin)
    - Same number of rooms (with a margin of 1 rooms)
- The apartment has been sold within a year of the estimate:
    - There are at most one year and three months between the estimate date and the day of authentic act
    - There is at least three months between these two dates
- If more than one estimates, from the same user, match with a given transaction we keep only the more recent.

We offset the transaction date by three months because, once an agreement is reached between a buyer and seller, the transaction cannot occur right away. First, the notary launches a legal procedure which lasts two to three months and that leads to the actual transfer of

ownership at the date recorded in the database. Hence, we compute an approximation of the time elapsed between the estimate and the moment the sale is concluded. Note that, it lets us think that the user does not know the actual price of the transaction yet. We also take some margins regarding the living area and the number of rooms, as non-professionals may be confused with the difference between the common and the legal definitions of these characteristics.

We perform cleaning and filtering of the dataset obtained through the matching. Outliers, abnormal transactions are filtered out: living area below 8m² or above 300m², sale price, MA estimate price and user estimates under 1,000€ or over 100,000,000€. Finally, to prevent any confusion due to our matching or the misuse of the AVM, we discard top and bottom 0.5% along three dimensions: the transaction price, the error of the user estimate and the error of the AVM. These errors are computed as log differences with the transaction price. Altogether, we connect 10,337 user estimates and transactions. Table 1 presents statistics about the prices and the estimates.

	count	mean	std	min	25%	50%	75%	max
Transaction price	10,337	372,225	247,325	78,000	204,000	300,000	463,000	1843,900
Transaction price per m²	10,337	6,564	2,642	1,173	4,429	6,498	8,493	20,000
AVM estimate price	10,337	374,665	270,984	65,200	200,300	293,700	456,000	2546,300
AVM estimate price per m²	10,337	6,442	2,610	1,081	4,363	6,303	8,212	20,029
User estimate price	10,337	387,103	256,060	70,000	215,000	310,000	480,000	2360,000
User estimate price per m²	10,337	6,765	2,630	1,210	4,648	6,698	8,654	20,000
Log relative error user	10,337	4.4%	9.8%	-35.7%	-1.1%	3.7%	9.3%	48.2%
Log relative error AVM	10,337	-1.1%	15.4%	-45.8%	-11.5%	-2.0%	9.0%	45.0%
Log absolute error user	10,337	7.9%	7.3%	0.0%	2.7%	5.9%	10.9%	48.2%
Log absolute error AVM	10,337	12.3%	9.2%	0.0%	5.0%	10.5%	17.8%	45.8%
Elapse time	10,337	127	100	1	43	100	196	364
Price var. during elapse time	10,337	0.4%	1.6%	-3.9%	-0.6%	0.0%	1.2%	8.3%
Price var. of last year	10,337	1.0%	3.1%	-3.9%	-1.5%	0.1%	3.6%	8.7%
Purchase price	10,337	255,328	197,168	1	130,000	208,500	322,000	3750,000
Log purchase price diff.	10,337	46%	57%	-239%	11%	31%	67%	1,223%
Holding time	10,337	3,307	2,771	28	1,522	2,557	4,171	23,070
Nominal value lost	10,337	10%	31%	0	0	0	0	1
Dispointing estimate	10,337	15%	36%	0	0	0	0	1

Table 4-1 : Statistic of the matching procedure of estimates and transactions

With more than 370k€ for the average transaction, our sample is clearly more expensive than the standard French real-estate transaction[27]. The fact that we focus only on apartments sold in the Paris region, which housing market is tightened by the proximity of the capital, explains these higher prices. The distributions of both estimates are similar to the one of the actual transaction prices, both in absolute price and in price per meter square, with a slight positive difference for the users' estimates. Indeed, the average log relative error made by the users in their predictions of the sale prices is +4.4%. For owners, such a tendency to overestimate their property value has been repeatedly measured: +6% in Goodman and Ittner (1992), +8% in Benitez-Silva et al. (2015) or +8% to +11% in van Cruijsen et al. (2018). Our measure is below the ones of the literature but MA users are not only homeowners evaluating their property. For an analysis per category of users see table 2. The AVM, on the other hand, appears almost unbiased on average (-1.1% statistically different from 0). However, it is less accurate than users in terms of the absolute log error the predictions: 12.3% on average and a median of 10.5%, against 7.8% and 5.9% for users. Note that these absolute errors are not necessarily perfect assessments of the accuracy of the predictions of users and the AVM. On the one hand, users with a better knowledge of the housing market may be more likely to give their opinions. It would also explain why they are more accurate on average than the random sampling from Goodman and Ittner (1992). On the other hand, they may have a greater tendency to react to the AVM result, when it is further from reality. These are just examples of the selection biases that might be at work in our collection method. We address the representativeness issue later in this section.

The sale records database contains information about the previous transaction involving the apartment. On average, owners have held their apartment for more than 9 years on average and a median of 7 years. The turnover rate of housing in France being below 3%[28], such short holding times indicate the segment we study is more active than the average. Compared to the long-term trend, prices in the Paris region are high during the 2012-2019 period (see appendix

[27] The average amount for the transactions recorded in the fiscal databases *Demandes de Valeurs Foncières* is of 220k in 2019 (author's calculation).

[28] In 2019, the number of sales exceed on million for the first time (source French Ministry of housing www.cgedd.developpement-durable.gouv.fr/) for 35.4 millions housing units (source INSEE www.insee.fr/fr/statistiques/2533533).

A). As a consequence, on average, the owner has purchased the apartment for much less than he or she sells it: 255k€ against 372k€. But if most owners make a profit (46% in nominal on average), a small fraction of 10% sells for less, from the 15% that were predicted a nominal lost by the MA AVM. Genesove and Mayer (2001) results on the reluctance of the sellers to face such loss can explain the difference between the percentage of sales for lost predicted by AVM and the actual ratio.

We collected estimates that occurred before the transaction. The *elapse time* measure the number of days between the feedback and the sales. They spread out all over the one-year window we have chosen, with a higher concentration in the last three months before the transactions. For sellers, the closer they get to a potential agreement, the more the price issue is important and the more likely they are to use an AVM and to react to its result. Moreover, buyers estimate apartments that are on the market. In a tense market, like the Paris region, their marketing time rarely lasts longer than a few weeks. Price variations over this elapse time might disturb the accuracy of the estimates. To control for it, we used the departmental housing price indices produce by INSEE, the French Bureau of Statistics (see appendix A), to track *the price variation during the elapse time*. During the period covered here, the housing market went through a bust (2012-2015) and a boom (2016-2019). Hence, the price variations are centered around 0%. Goodman and Ittner (1992) show the price dynamics influences the perception of the value of American homeowners. To test for it we use the same indices to compute the price changes over the year before the estimate. On average for the Paris region, a sharp price increase of +25% over two years preceded the 2012-2015 bust of -5% which was lower in magnitude than the following boom of +20% (Notaire-INSEE indices are reproduced in the appendix). Hence, estimates happened equally following price increases and decreases (median of +0.1%) but the formers being more intense, the average of the variable *price variation of last year* is positive (+1%).

Because they are regularly and massively surveyed about the value of their property, the literature has focused almost exclusively on the homeowners' self-evaluations. Our empirical strategy to rely on data collected on an Internet platform make us able to broaden the scope of the present study. Homeowners use MA services to get estimate of their apartment our house value. But they are not the only ones for whom an automated appraisal represents an important information for. Buyers, that have to decide either make an offer or not and how much is a fair price for a given property, are also interested in the results of an AVM. Before they get the

estimation result, MA ask for its users to state the reason they use the estimation tools. They may indicate they own the property or not, if they plan to sell it and by when, or if they are buyers and how far in their search. The progress of the buyers' searches has been surveyed only from 2014. Hence, the difference between the total amount of buyer and the sum of four categories. Table 2 presents the relative and absolute error in the sale price prediction per category of user.

As expected, not all users own the apartment, they estimate on MA. Yet, they represent the large majority in our sample with 72%, when only 38% of the users indicate not to be the owner of the estimated property. MA business model relying on connecting homeowners to realtors, this imbalance is not a surprise. Among them four out of five (58% of the total) indicate they considered selling it, but only 13% has already put his or her apartment for sale. Note that the other fifth of the owners actually ended up selling their flat. This under-declaration may be due to a fear of disclosing intentions to a commercial website. The same kinds of suspicion make us think that some actual homeowner may have declared themselves as *non-owners*. In contrast, we do not see a good reason for a user to lie about its intention to sell or to buy an apartment.

		Share of	Log relative error			Log absolute error		
	count	users	mean	std	median	mean	std	median
Owners	7,463	72%	5.6%	9.9%	4.7%	8.5%	7.6%	6.5%
Sellers	6,042	58%	5.7%	9.6%	4.8%	8.4%	7.5%	6.5%
Sellers more than 6 months	889	9%	5.4%	11.0%	4.1%	9.2%	8.2%	7.0%
Sellers within 6 months	964	9%	4.1%	10.0%	3.0%	8.0%	7.3%	6.1%
Sellers within 3 months	1,012	10%	5.0%	9.2%	4.4%	7.8%	7.0%	6.1%
Sellers A.S.A.P.	1,872	18%	6.3%	9.7%	5.1%	8.6%	7.8%	6.5%
Sellers on sale	1,305	13%	7.1%	8.2%	6.2%	8.3%	7.0%	6.9%
Non owners	2,874	28%	1.1%	8.6%	1.4%	6.3%	5.9%	4.7%
Buyers	1,899	18%	1.0%	8.4%	1.3%	6.2%	5.7%	4.7%
Buyers wondering	89	1%	0.0%	9.2%	0.2%	6.9%	6.0%	5.6%
Buyer starting	268	3%	-0.7%	9.2%	0.0%	6.9%	6.1%	5.0%
Buyer active	692	7%	0.5%	8.1%	0.6%	5.9%	5.6%	4.4%
Buyer offer	310	3%	2.3%	7.5%	2.1%	5.5%	5.5%	4.0%

Table 4-2 : Relative and absolute error in the sale price prediction per category of user

Table 2 reveal important difference in the ability to predict transaction price among the categories of users. First of all, in line with the literature, sellers overestimate the value of their property by 5.7% on average. On the other hand, buyers, whose ability to predict price has never been measured before, exhibit a much lower bias of 1% (yet statistically significant). One might

have expected to see buyers show a symmetrical behavioral bias and underestimate the value of for-sale apartments, in line with their own interest. They do not, and the matching and negotiation process has more to do with lowering sellers' expectations than conciliating antagonistic positions. However, along the sale process the average level of the sellers' overestimation increase quasi-constantly, as if, far from learning the price overtime, sellers build up more and more unreal expectation. In terms of accuracy, buyers also do better than sellers predicting the sale with only a 6.2% log absolute error margin versus 8.4% for sellers. It is way better than what has been measured in the literature so far (14% mean absolute error of homeowners in Goodman and Ittner, 1992). The argument already made of better-informed users to be more likely to give feedback can be repeated here. Yet, the difference might be due to the increase in housing information available from the early 1990s to the 2010s because to the real estate market boom in Internet usage, as well.

Continuous Variables								
	count	mean	std	min	25%	50%	75%	max
Area	10,337	58	26	9	39	55	72	234
Room count	10,337	2,8	1,1	1	2	3	3	6
Bathroom count	10,337	1,1	0,3	1	1	1	1	3
Floor	10,337	2,7	1,8	0	1	2	4	6
Parking count	10,337	0,5	0,6	0	0	0	1	2
Build date	9,775	1735	604	-2	1900	1949	1972	2016
Sale year	10,337	2015	2	2012	2013	2015	2017	2019

Dummy Variables					
Secondary room	2%	Renovated shared parts	34%	In Seine-et-Marnes	6%
Outdoor	46%	To renovate	12%	In Essone	3%
Cellar	77%	Inf. relative perception	4%	In Hauts-de-Seine	24%
Elevator	61%	Sup. relative perception	28%	In Seine-St-Denis	6%
Exceptional view	14%	In Paris	43%	In Val-de-Marne	11%
Renovated facade	26%	In Yvelines	3%	In Val-d'Oise	3%

Table 4-3 : Hedonic description of the apartments

To conclude our description of our dataset, table 3 presents the hedonic characteristics of the apartment in our dataset. As it is more detailed than the notarial records, we kept the description made by the user through the AVM form. The average transaction in our sample concerns a three rooms, one bathroom, 58 m² apartment which is fairly standard in the Paris region. It is situated on the third floor with an elevator in most cases. One can tell that the majority of the users owns the apartment they describe. Indeed, 28% of the users have a superior relative perception of the apartment compared to the other apartment of the building, against only 4% who have an inferior relative perception. From table 3 the most important thing to

notice from the authors' point of view is the strong imbalance in the geographical distribution. Almost half of the sales in our dataset occurred in the city of Paris while only 2 of the 12 million inhabitants of the region lived there. The wealthy west suburb Hauts-de-Seine follows with 24% of our transactions for 1.6 million inhabitants, whereas the four more distant suburban departments of Yvelines, Seine-et-Marnes, Essone, and Val-d'Oise only represents 23% of our dataset for 44% of the total population. A first reason for that is that the latter departments are less urbanized, thus individual houses are more frequent than apartments. It may also be due to the commercial development of MA, that have a clear financial incentive to focus more on the most expensive area.

The very nature of the datasets based on digital traces collection prevents them from being representative as the random samples used in classic surveys can be (Salagnik 2017). Indeed, selection bias may occur all along the funnel that leads to the user giving feedback (going to the MA website, filling the AVM form and finally reacting to the results with a personal estimate). Thus, we should control for a potential lack of representativeness as much as possible . We have no information about the users of MA. Thus, we are unable to verify how similar or different they are from the rest of the buyers and sellers of the French housing market. However, having matched users' estimations with recorded sales on the notarial databases, we can easily compare the apartments estimated with all the ones that have been sold. According to MA estimates, the sales record database we use is not exhaustive but contains three sales out of four. Yet, it seems fair to us to hold it as representative of the whole market. To assess how likely any given sale, which occurred in the Paris region, from 2012 to 2019, is to be present in our dataset, we run a probit analysis of the following equation (1).

$$
\begin{aligned}
&Prob(Presence = 1 \mid X_i) \\
&\quad = \phi\Big(\sum_{i\in[1,6]\setminus 2} \beta_{1,i} \mathbb{1}_{nbRooms=i} + \sum_{i\in[0,3]\setminus 1} \beta_{2,i} \mathbb{1}_{nbBathrooms=i} \\
&\quad + \sum_{i\in[0,6]\setminus 2} \beta_{3,i} \mathbb{1}_{floor=i} + \sum_{i\in[1,2]} \beta_{4,i} \mathbb{1}_{nbParkings=i} + \beta_5 noElevator \\
&\quad + \beta_6 outdoor + \beta_7 noCellar + \sum_{i\in[1,9]} \beta_{8,i} \mathbb{1}_{builtPeriod=i} \\
&\quad + \sum_{i\in[2012,2019]} \beta_{9,i} \mathbb{1}_{saleYear=i} + \sum_{department\in IdF} \beta_{10,i} \mathbb{1}_{deptartment=i} \Big)
\end{aligned}
$$

Results are displayed in appendix B, for the whole dataset in columns (1) together with subsets of the evaluations of, respectively, owners, non-owners, sellers and buyers. They show our datasets present selection biases. It confirms that, at the departmental level, the four departments of the great suburb and the socially disadvantage Seine-Saint-Denis are underrepresented. Our decision to consider only the transactions for which we know the previous transaction price disadvantages the most recent constructions. On a temporal perspective, recent sales have a lower probability of presence in our listing, especially for the year 2019. This is due to design choices of MA, which have harmed the use of the "feedback" feature in recent years (it used to be directly under the result of the estimation computed by MA, and is now displayed lower). From one type of user to another, biases are fairly stable. Still, estimations by buyers, are more evenly distributed among all recorded sales.

All things consider, note that the pseudo-R^2 of each model remains low, indicating the "selection" process is reasonably random. Yet, to avoid our result to be distorted by these biases, we employ the Heckman (1979) correction for the rest of our analysis. It consists in adding the inverse Mills' ratio computed from the above probit models in all our regressions, to control for selection biases.

III. Explaining the Error

In this third part, we try to assess what influences real estate market participants price beliefs and the error they make predicting the transaction values. We estimate the following equation (2) with an OLS regression for three subsets of the dataset presented in the previous section: Owners, Sellers and Buyers; which respectively correspond to the users who have declared that they own the apartment, that they own and have the intention to sell it or that are looking to buy one. Note that the Sellers dataset is included in Owners. Results are presented in Table 4.

$$\ln\left(\frac{user\ estimate}{transaction\ price}\right)$$
$$= \beta_0 + \beta_1 vacation\ home + \beta_2 rental\ investment$$
$$+ \beta_3 \ln\left(\frac{MA\ estimate}{transaction\ price}\right) + \beta_4 \ln\left(\frac{purchasing\ price}{transaction\ price}\right)$$
$$+ \beta_5 \ln(holding\ time) + \beta_6 \ln(elapse\ time)$$
$$+ \beta_7\ price\ variation\ during\ elpase\ time$$
$$+ \beta_8\ price\ variation\ last\ year + \beta_9\ inverse\ Mills\ ratio$$
$$+ \sum_i \beta_i\ hedonic\ feature_i + \varepsilon$$

For Owners and Sellers, we test if there is a difference between the majority that sold their primary residence and other kinds of residence through the dummies *vacation home* and *rental investment*. We also investigate if the effect of possible reference point on the user beliefs. First, we consider the log of the relative error of the MA estimate that the user reacts to. The second reference point considered is the purchasing price through the log of the ratio of the previous transaction price, recorded in the notarial database, and the present transaction value. As in Van der Cruijsen et al. (2018), we want to measure the impact of the tenure, i.e. how long the present owner has held the apartment, through the log of the holding time. The user does not appraise at the exact same date the sale occurs. To control for it, we use two variables: the log of the elapsed time, which represents how many days passed between the user estimates and the sale agreement date, and the price variation during this elapse time (computed with the price indices produces by INSEE, see section II). Following part III and Goodman and Ittner (1992), we have also introduced the log of the variation of the market price during the year before the estimate. We employ the Heckman correction (Heckman 1979) by introducing the Inverse Mills ratio obtained from the probit regression of equation (1) (see section II and appendix B). Finally, we make the most of the rich description of the apartments in the dataset, controlling for all the hedonic features. The complete list of the characteristics considered and the measures of their effects on the users' prediction errors is available in appendix C.

Before analyzing table 4 figures, let's first highlight the main learning from the regressions of the prediction errors by the sole hedonic features, presented in appendix C. The most noticeable result is that the users' errors are only modestly correlated to the characteristics of the apartment. Indeed, the explanatory power of the hedonic model remains low with R^2 going from 5%, for buyers, to 8% for sellers. Moreover, we measure statistically significant

effects for only a limited number of features. In the matter, we come to a similar conclusion than Goodman and Ittner (1992). They use a result close to ours to claim that owners' estimates can be used in aggregate studies as measures of their home values, because they appear to be mostly unbiased from that perspective.

Table 4 presents the regressions performed on three different sub-datasets that gather users of different profile, with and without the Heckman correction for each. The impact of the inverse Mills ratio is limited, with low magnitudes and no statistical significance. Moreover, the effects of the other variables are left almost unchanged by the correction. This reassure us on the generality of our results, obtained through data collected on the Internet.

While hedonic characteristics of the apartments are not determinant in explaining the users' errors, the introduction of elements that interfere with their thought process makes the explanatory power grows. From the purely hedonic regression, the R^2 is multiplied by more than 3, going from a 5% to 8% to a range of 18% to 35%. This increase is coherent with the results of Nothcraft and Neale (1987). They measure the differences in the anchor value, a listing price in their case, explain between 17% and 40% of the variance in the estimates of the participants in their experiment. Not only do these variables explain the variance but also the bias of the owners' estimates. Indeed, table 4 intercepts are all statistically identical to 0, whereas they are positive and significant in the regressions using hedonic regressions only, with and without Heckman correction.

The dominant effect, that explains this increase in the explanatory power, is the one related to the relative error of the AVM. Indeed, in the sense of the anchoring and adjustment heuristic, the MA appraisal is a very plausible candidate for a reference point. The positive, large and statistically significant correlation between the error of the MA model and the user's estimate shows that facing an ambiguous situation they have limited knowledge of, the users grab this appraisal-like value and use it as an anchor to build up their own estimate. The magnitude of the effect, from 0.17 to 0.36 for 1, is to be considered together with the large variance of the AVM relative error itself (standard deviation of 15% against 10% for the user error), which emphasizes the importance of this factor. The clear difference between the impact on owners' and sellers' opinion, on the one hand, and the one of buyers, on the other hand, is also meaningful. Asking price has been repeatedly shown to be an anchor for the housing market participants (Nothcraft and Neale 1987, Black and Diaz 1996, Bucchianeri and Minson 2013). Hence, we interpret the reduced impact on buyers' estimates of the AVM results, which

is half of the one on sellers', as the result of the availability of such an alternative reference point. Indeed, buyers gave their opinion on apartments that most probably are already on the market. On the other hand, most sellers (78% according to table 2), use the services of MA before they put their property on the market, probably to help them to choose an asking price. Therefore, the influence of the model output on the formers is mitigated by their awareness of the asking price, unavailable to us in this study. We see this as a proof of an updating mechanism of their beliefs about the housing price by market participants. We exanimate this issue more carefully in section V bellow.

Dependent variable	$\ln(user\ estimate/transaction\ price)$					
	Owners (1)	Sellers (2)	Buyers (3)	Owners (4)	Sellers (5)	Buyers (6)
intercept	0.0128 (0.0219)	0.0226 (0.0233)	0.0085 (0.0658)	0.0097 (0.0114)	0.0017 (0.0124)	0.0237 (0.0198)
vacation home	0.0047 (0.0040)	0.0065 (0.0043)		0.0047 (0.0040)	0.0065 (0.0043)	
rental investment	0.0036 (0.0031)	0.0075* (0.0034)		0.0036 (0.0031)	0.0074* (0.0034)	
Log relative error AVM	0.3574*** (0.0069)	0.3465*** (0.0075)	0.1734*** (0.0123)	0.3574*** (0.0069)	0.3465*** (0.0075)	0.1736*** (0.0122)
Log purchase price diff.	0.0047* (0.0022)	0.0049* (0.0024)	-0.0029 (0.0038)	0.0047* (0.0022)	0.0048* (0.0024)	-0.0029 (0.0038)
ln(holding time)	0.0043*** (0.0013)	0.0045** (0.0014)	-0.0021 (0.0023)	0.0042*** (0.0013)	0.0044** (0.0014)	-0.0021 (0.0023)
ln(elapse time)	0.0065*** (0.0009)	0.0077*** (0.0010)	0.0015 (0.0016)	0.0065*** (0.0009)	0.0078*** (0.0010)	0.0015 (0.0016)
price variation during elapse time	-0.4452*** (0.0739)	-0.3626*** (0.0824)	-0.7186*** (0.1680)	-0.4485*** (0.0712)	-0.3880*** (0.0788)	-0.7120*** (0.1657)
price variation of last year	-0.2262*** (0.0457)	-0.2445*** (0.0501)	0.0163 (0.0759)	-0.2288*** (0.0429)	-0.2635*** (0.0468)	0.0231 (0.0706)
Inv. Mills ratio	-0.0013 (0.0078)	-0.0086 (0.0081)	0.0052 (0.0214)			
Hedonic control	Yes	Yes	Yes	Yes	Yes	Yes
Nb Observations	7463	6042	1899	7463	6042	1899
R^2	0.349	0.353	0.179	0.349	0.353	0.179
Adj-R^2	0.345	0.348	0.159	0.345	0.348	0.159
F-Statistic	79.4	65.4	8.6	81.1	66.7	8.8

Standard errors between parenthesis
*** p<0.001, ** p<0.01, * p<0.05, ° p<0.1

Table 4-4 : Explaining the users' prediction errors (OLS estimations of equation 2)

The other possible anchor in equation (2), the purchasing price of the apartment, has almost no effect on the users' estimates. This is in line with the literature on internal reference prices for durable goods, which rely more on the present context than historical prices (see

Mazumdar et al. 2005 for a survey). Our real-life measure contradicts the results Corina and Chenavaz (2011) got from a laboratory experiment in which both acting sellers and acting buyers were proven to be influenced by the previous transaction price. This result may also seem to be in contradiction with the results of Genesove and Mayer (2001) or Van der Cruijsen et al. (2018). However, these studies make the distinction between decreasing and increasing price to underline the effect of the owners' loss aversion. We make such a distinction in the next section. The owner history with its apartment is not limited to the price he or she paid for it, years ago. These "years" matter too. The length of the tenure has a positive effect on how much owners and sellers value their property. The magnitudes of the coefficients we measure imply an overestimation of +3.6% for a ten-year tenure. These figures are consistent with Van der Cruijsen et al. (2018) whose estimates for the impact on an owner's estimate of long holding times (over 10 years or over 20 years) range from +2% to +5%. As we control for a large number of characteristics and because buyers do not exhibit such a positive correlation between their estimates and the length of tenure by the current owner, this is proof of homeowners being subject to endowment effect. Strahilevitz and Loewenstein (1998) show the effect of a few minutes ownership on something as trivial as a mug. In a housing context, Nash and Rosenthal (2014) measure that college students value more the hall to which they were randomly assigned by a lottery. Thus, we would have been a surprise that divestiture aversion does not affect the value of something so personal and with such a massive financial importance for most households as an apartment.

The fair price of a real estate asset is a moving target and time passes between the estimate and the time the sale is concluded. If not entirely, MA users are aware of the price evolution. A third to three quarters of the department price index variation between those two dates is reflected in the error they made. Given that the indices are computed on a quarterly basis and that each department host heterogenous housing areas, these figures demonstrate a good understanding of the market dynamics, in line with the conclusions made in Henrique (2013) or Windsor et al. (2014). However, a difference emerges between the two sides of the transaction, as buyers' predictions follow the market trends more closely. First, the magnitude of the coefficient related to the price variation during elapse time is significantly larger for buyers. But mostly, their estimates are free of the lag the ones of owners and sellers suffer from. Indeed, we measure a negative effect of the previous year's variations on the errors made by homeowners, similar to the one highlighted in Goodman and Ittner (1995), but not by buyers. Once again, the fact that the latter are in contact with several different sellers, through listings

or property tours, right from the beginning of their search seems to improve their understanding of the market reality, while sellers hold on to their "dreamed" value longer. In a PVAR analysis of the Australian homeowners' self-estimates, Windsor et al. (2014) point that they are "sticky" and "backward looking". This would explain why, owners show a tendency overestimate their property more during busts and less during booms. The consequences of such a lag of the sellers' valuations on the matching process are explored in Genesove and Han (2012).

Beyond the market trends, the elapsed time that separates the estimation and the sale has an effect on the opinion of sellers. In detail, our results presented in table 4 show that a seller values the same apartment for +4% to +5% more a year from the sale than at the transaction date, everything else equal. Goodman and Ittner (1992) have already measured such an effect. Everything happens as if, the closer we get from the transaction, the less upwardly biased sellers are; as if the search for a counterpart slowly reduces their overly optimistic expectations (see section II for evidence of the generally positive bias of owners and sellers). We do not measure such an effect for buyers. Two reasons for that. First, because, in our measures, buyers do not exhibit any a priori bias that would need to be compensated by a learning mechanism. Second, because if this *elapse time* is a good measure of the progression of the seller in the search, it does not tell us much about the buyer's. It is not because someone visits or is interested in a property that will be sold the next day that this person is the one that has actually bought it. A short *elapse time* for a seller means he or she is getting close to the conclusion of the process. For a buyer, it may be correlated with a search that just began as much as with a search that is about to end. In section VI, we use the declarative search stages instead of the sole elapse time to investigate how the progression in the search impact the internal reference price of both sellers and buyers.

IV. Loss-Aversion

The loss aversion of homeowners has been documented in different settings. Genesove and Mayer (2001) show it influences their marketing strategy, Einiö et al. (2008) see evidence of it in the transaction prices and Van der Cruijsen et al. (2018) blame it for the owners' overly optimistic sale price predictions. Section III results indicate owners' and sellers' beliefs are much more influence by the AVM result than by the price they pay for their apartment. However, we did not differentiate the majority whose property values increase and the minority who face a nominal loss. In this section, we investigate how the threat of a nominal loss influences the sellers' estimates. In particular, we are interested in the effect it has on the weight

they give to the two reference points available, the AVM result and the purchase value. An OLS regression of equation (3) is performed and the results are presented in table 5.

$$\ln\left(\frac{user\ estimate}{transaction\ price}\right) = \beta_0 + \beta_1 vacation\ home + \beta_2 rental\ investment + \beta_3 loss\ dummy + \beta_4 \ln\left(\frac{MA\ estimate}{transaction\ price}\right) + \beta_{4'} \ln\left(\frac{MA\ estimate}{transaction\ price}\right) * loss\ dummy + \beta_5 \ln\left(\frac{purchasing\ price}{transaction\ price}\right) + \beta_{5'} \ln\left(\frac{purchased\ price}{transaction\ price}\right) * loss\ dummy + \beta_6 \ln(holding\ time) + \beta_7 \ln(elapse\ time) + \beta_8\ price\ variation\ during\ elpase\ time + \beta_9\ price\ variation\ last\ year + \beta_{10}\ inverse\ Mills\ ratio + \sum_i \beta_i\ hedonic\ feature_i + \varepsilon$$

In equation (3), we add to equation (2) a dummy variable indicating that the value of the apartment decreased since the previous sale and we use it in interaction with the AVM error and the relative difference between the purchase and sale prices. In table 5, this dummy variable is computed in two different ways. First, the *value lost* variable, which takes 1 if the "true value" of the apartment, measure by its sale prices, decreases between the moment the owners purchased it and the moment he or she sells it, and 0 otherwise. As the seller can't possibly know the realization price at the time he or she estimates the apartment, we also compute the *disappointing estimate* variable. It takes 1 if the AVM result is inferior to the purchase price, announcing to him or her a nominal lost, and 0 if it is good news. Evaluation of equation (3) is performed only for the Owners and Sellers dataset because buyers most probably ignore the price that the current owner pay for an apartment. Indeed, sales records are only public in France as of April 2019, with a lag of at least 12 months and a depth five years[29]. Some sellers might disclose this information to prospective buyers (see Corina and Chenavaz 2011 and Shavell

[29] See decree n°2018-1350 (JORF n°0302) for the exact condition of the publication of the *Demandes de Valeurs Foncières* dataset.

1994 for discussion about when to disclose this information is profitable), but as we have no way to control for it, we prefer to discard the Buyers dataset.

First, let us notice that all the results discuss in the previous section hold after the introduction of the dummies indicating a loss. However, in contrast with what table 4 results, we measure that the price they paid to acquire their property impacts owners' and sellers' opinion about its value, but only in a loss situation. As in Van der Cruijsen et al. (2018), who measure an average +6% overestimation bias among owners facing a nominal loss, owners and sellers whose apartment decrease in value estimate it for +2% more on average, whatever the way we measure this drop.

In Corina and Chenavaz (2011), participants of a design experiment asked to act as sellers show the tendency to anchor their estimates to the purchase price of the property, no matter the price dynamics. However, this tendency is stronger in the case of a nominal value lost. Results of table 5, we have drawn from the responses of actual sellers, are even more extreme. In the vast majority of the cases in which the value of the apartment has increased since it was first acquire, the sellers' estimates are free of the influence of the purchase price. The coefficient related to the *log purchase price difference* is not statistically different from 0 in all four regressions of table 5. On the other hand, in the case of a loss, it has a significant and positive impact of the opinion sellers have about their property value. The same coefficient, but in interaction with one of the loss dummies, ranges from 0.06 to 0.14 with p-values below the 0.1% level. Such an anchoring, in a loss situation, is typical of the loss aversion phenomenon theorized in Kahneman and Tversky (1979) through the convexity of the "value function" in the loss domain (at the opposite of its concavity for gains). This cognitive bias has been blamed for the reluctance to realize losses among real estate sellers (see Genesove and Mayer 2001, Einio et al. 2008 and Anenberg 2011) as well as among amateur and professional stock market traders (Odean 1998, Lock and Mann 2000). It has also been one of the reasons invoked in the literature to explain why volumes drop when prices decrease (Stein 1995).

Dependent variable	$\ln(user\ estimate/transaction\ price)$			
	Owners (1)	Sellers (2)	Owners (3)	Sellers (4)
intercept	0.0272	0.0382°	0.0181	0.0280
	(0.0218)	(0.0232)	(0.0218)	(0.0232)
vacation home	0.0031	0.0050	0.0034	0.0050
	(0.0040)	(0.0043)	(0.0040)	(0.0043)

rental investment	0.0032 (0.0031)	0.0065° (0.0034)	0.0030 (0.0031)	0.0063° (0.0034)
lost value	0.0258*** (0.0038)	0.0238*** (0.0042)		
disappointing estimate			0.0193*** (0.0039)	0.0163*** (0.0044)
Log relative error AVM	0.3607*** (0.0073)	0.3500*** (0.0079)	0.3748*** (0.0077)	0.3663*** (0.0083)
Log relative error AVM *lost value=1	-0.0833*** (0.0196)	-0.0924*** (0.0209)		
Log relative error AVM * disappointing estima=1			-0.0444° (0.0240)	-0.0884** (0.0282)
purchasing price rel. diff	-0.0012 (0.0023)	-0.0013 (0.0025)	-0.0017 (0.0023)	-0.0015 (0.0025)
purchasing price rel. diff *lost value=1	0.0612*** (0.0168)	0.1105*** (0.0230)		
purchasing price rel. diff * disappointing estimate=1			0.0867*** (0.0158)	0.1369*** (0.0215)
ln(holding time)	0.0041** (0.0013)	0.0043** (0.0014)	0.0046*** (0.0013)	0.0050*** (0.0014)
ln(elapse time)	0.0060*** (0.0009)	0.0071*** (0.0010)	0.0061*** (0.0009)	0.0073*** (0.0010)
price variation during elapse time	-0.4267*** (0.0734)	-0.3502*** (0.0817)	-0.4398*** (0.0734)	-0.3653*** (0.0817)
price variation of last year	-0.2068*** (0.0454)	-0.2186*** (0.0498)	-0.2082*** (0.0454)	-0.2165*** (0.0498)
inv. Mills ratio	-0.0079 (0.0078)	-0.0150° (0.0080)	-0.0063 (0.0078)	-0.0134° (0.0080)
Hedonic control	Yes	Yes	Yes	Yes
Nb Observations	7463	6042	7463	6042
R^2	0.358	0.365	0.357	0.365
Adj-R^2	0.353	0.359	0.353	0.359
F-statistic	77.9	64.9	77.8	64.8

Standard errors between parenthesis
*** p<0.001, ** p<0.01, * p<0.05, ° p<0.1

Table 4-5 : Impact of nominal value loss on the estimates made by Owners and Sellers (OLS estimations of equation 3)

In addition to considering their purchase price as a relevant reference point for assessing the current value of their apartment, owners facing a nominal loss also decrease the credence they give to the AVM result. The coefficient that measure the impact of the *log relative error of the AVM* in interaction with the loss dummies are all negative and are statistically significant in all regressions. Overall, the correlation between the users' prediction errors and the ones made by the AVM remain positive, but it is reduced by up to -25%. Hence in a loss situation, sellers still update their reference point with this up-to-date but disappointing information but less than when it announces a gain. Arkes et al. (2008) point out a similar tendency of security

traders to adjust their reference points upward (i.e. in a gain situation) more easily than downward. In the context of real estate Havard (1999) shows valuers have the same tendency to increase valuation when they have knowledge it was too low previously, but rarely do the opposite. A first partial conclusion would be that our results show the anchoring and adjustment process of housing market participants implies the weighting of different possibly contradicting information. In the next section, we take a dynamic perspective to account for the evolution of this weighting along the duration of the search.

Two arguments might be put forward to call for caution about our results on the sellers' loss aversion. First, only a minority of our user are in a loss situation. Yet, these 10% for the *lost value* dummy and 15% for *disappointing estimate* actually have to be considered with the rather large size of datasets. They correspond to hundreds of observations, which allows for the statistical significance of our result. Second, the possibility of reverse causality exists. As proposed by Taylor (1999), a negative herding mechanism could lead sellers that overestimate their property to finally sell at a discount because of a longer time-on-market. The fact that our results are similar with the two dummies, while the variable *disappointing estimate* is immune against such mechanism, let us consider them with confidence.

V. Search Stage

The sales and the buying of a property takes time, usually several weeks or even months. Along this search for a counterpart, the beliefs about the prices are likely to change as buyers and sellers interact with intermediaries or each other, and gather signals from different sources. From a search theory perspective Kohn and Shavell (1974) and Rothchild (1974) model the updating procedure of the price beliefs and its consequences. However, the internal reference price of an economic agent being hardly observable during the decision process, there is no, to the author's knowledge, evidence of its evolution along the search. That is the gap that the present section aims at filling. Users of the MA's AVM are in different stages of their process. As table 2 indicates, we have gathered estimates from users in all degrees of advancement.

In this section we enrich our model to account for this temporal dimension of the search. For buyers we rely on its answer to the question: "Where are you with your search?" on a four-level scale that goes from "I am wondering" to "I have just made an offer". For sellers, two different measures are proposed. First, a similar six-level categorical scale that corresponds to the user answer to the question: "When do you plan to sell your apartment?" which goes from

"In more than six months" to "It is already for sale". Second, the *elapse time* variable, already describe above, which is continuous. We introduce these variables in the equation (4) in interaction with the *log relative error of the AVM* variable to measure the change in the importance of the MA appraisal along the search. For categorical variables, we leave the most common and central values out as references (respectively "I wish to sell it soon" and "I am actively searching").

The following model, equation (4), is evaluated through an OLS regression. The results are presented in table 6. Let us recall that for the buyers, the search stages have only been surveyed from 2014, resulting in a reduction in the number of observations in comparison with previous regressions in column (4).

$$\begin{aligned}\ln\left(\frac{user\ estimate}{transaction\ price}\right) = {} & \beta_0 + \beta_1 vacation\ home + \beta_2 rental\ investment + \\ & \beta_3 search\ progression + \beta_4 \ln\left(\frac{MA\ estimate}{transaction\ price}\right) + \\ & \beta_{4'} \ln\left(\frac{MA\ estimate}{transaction\ price}\right) * search\ progression + \\ & \beta_5 \ln\left(\frac{purchasin\ \ price}{transaction\ price}\right) + \beta_6 \ln(holding\ time) + \\ & \beta_7\ price\ variation\ during\ elpase\ time + \\ & \beta_9\ price\ variation\ last\ year + \beta_{10}\ inverse\ Mills\ ratio + \\ & \sum_i \beta_i\ hedonic\ feature_i + \varepsilon\end{aligned}$$

Dependent variable	$\ln(user\ estimate/transaction\ price)$			
	Owners (1)	Sellers (2)	Sellers (3)	Buyers (4)
intercept	0.0103	0.0179	0.0677**	0.0022
	(0.0217)	(0.0231)	(0.0229)	(0.0867)
vacation home	0.0049	0.0065	0.0061	
	(0.0040)	(0.0043)	(0.0043)	
rental investment	0.0035	0.0071*	0.0082*	
	(0.0031)	(0.0034)	(0.0034)	
Seller more than 6 months			0.0033	
			(0.0036)	
Seller within 6 months			-0.0062°	
			(0.0036)	
Seller now			0.0017	
			(0.0031)	
Seller on sale			0.0043	
			(0.0033)	
Buyer wondering				-0.0061
				(0.0088)
buyer starting				-0.0156**
				(0.0056)
Buyer offer				0.0159**
				(0.0054)
Log relative error AVM	0.0797**	0.0623*	0.3687***	0.1708***
	(0.0282)	(0.0303)	(0.0177)	(0.0193)
Log relative error AVM *ln(elapse time)	0.0606***	0.0623***		
	(0.0060)	(0.0064)		
Log relative error AVM *Seller more than 6 months			0.0761**	
			(0.0247)	
Log relative error AVM *Seller within 6 months			0.0310	
			(0.0245)	
Log relative error AVM *Seller now			-0.0272	
			(0.0213)	
Log relative error AVM *Seller on sale			-0.1171***	
			(0.0221)	
Log relative error AVM *Buyer wondering				0.0085
				(0.0496)
Log relative error AVM *buyer starting				0.0542
				(0.0330)
Log relative error AVM *Buyer offer				-0.0322
				(0.0336)
Relative purchasing price diff	0.0044°	0.0045°	0.0057*	-0.0004
	(0.0022)	(0.0024)	(0.0024)	(0.0044)
ln(holding time)	0.0040**	0.0045**	0.0048***	-0.0001
	(0.0013)	(0.0014)	(0.0014)	(0.0026)
ln(elapse time)	0.0071***	0.0083***		
	(0.0009)	(0.0010)		
price variation during elapse time	-0.3869***	-0.3047***	-0.1955*	-0.6362***
	(0.0736)	(0.0820)	(0.0811)	(0.1790)
price variation of last year	-0.2349***	-0.2520***	-0.2798***	-0.0287
	(0.0454)	(0.0497)	(0.0496)	(0.0808)
inv. Mills ratio	-0.0009	-0.0078	-0.0136°	0.0059
	(0.0077)	(0.0080)	(0.0081)	(0.0279)
Hedonic control	Yes	Yes	Yes	Yes
Nb Observations	7463	6042	6042	1359
R^2	0.358	0.363	0.358	0.201
Adj-R^2	0.353	0.358	0.352	0.170
F-Statistic	81.0	66.9	58.4	6.3

Standard errors between parenthesis
*** p<0.001, ** p<0.01, * p<0.05, ° p<0.1

Table 4-6 : OLS estimations of equation 4

First, let us notice that once again, the results from table 6 confirm the ones of table 4. Regarding the interaction between the influence of the AVM results as an anchor and the progression in the search for a counterpart, estimations of equation 6 advocate for a decreasing influence on the sellers' beliefs but no evolution regarding the buyers' opinions. In column (1) and (2) of table 6, we considered the progress of the search process through the continuous variable of the *log of the elapsed time*. The larger it is, the further sellers were from making the sale when they made their prediction. The positive, significant at the 0.1% level, coefficients related to the interaction variable between *logged relative error AVM* and *logged elapse time* in both columns shows that the further they are from the deal, the more owners rely on the MA appraisal to assess a value to their property. It is decreasing up to the point that, the day before the sale, it has four to six times less influence than the average effect we measured in tables 4 and 5.

In column (3), we take the search progression of the seller into account through the categorical declarative degree of advancement of their project. It confirms the decreasing influence of the reference point over time. But first, let us highlight a difference. Contrary to the *elapsed time*, we do not measure a decrease in the overall positive owner bias: the fixed effects proper to the different stages are not statistically different from zero. On the other hand, the interaction variables between the search stages and the relative error of the AVM reveal a pattern perfectly in line with the previous results. From the ones who declare they do not plan to sell within the next six months, to the ones whose apartments are for sale already, the impact of the MA estimates on the sellers' error goes from 0.44 to 0.25. Once again, we see that prospective sellers, that have just started gathering information, are far more sensitive to the anchor offered by the MA website than the ones who have already advertised the sale and received potential buyers' visits.

Our results regarding buyers are totally different. Column (4) presents a similar regression than column (3), with the different declarative stage of the search for an apartment to purchase accounted for. From one stage to another, we do not measure statistically significant changes in the influence of the AVM error on the prediction error of the buyers. As the degree of progress increases, the impact we measure remains similar to the one measured in table 4. Yet we see a difference in the relative level of the prediction errors. On average they are 1.6% lower for *starting* buyers than for *active* buyers, our references, and +1.6% higher for

the ones who just made an offer. We understand these differences as an acceptance from buyers to increase their bid in order to make a deal with a seller.

The decreasing impact of the anchor provided by the AVM for sellers can be viewed as evidence of a learning mechanism of what should be a fair price for their apartment. Following the works of Kohn and Shavell (1974) and Rothschild (1974), Bikhchandani and Sharma (1996) proposed a Bayesian-like updating mechanism of beliefs. According to them, when sampling from an unknown distribution, the weight the decision maker assigns to the prior increases with the number of samples already made. As the decision maker increases the number of observations, the uncertainty is reduced and every new sample carries less information. At the limit it brings no information at all and the prior converges to the true distribution. The sellers' problem is different but we argue the learning mechanism is comparable. They do not update their beliefs by sampling price quotes but gather information to solve the problem they are facing: deciding of their asking and reservation prices (see Merlo et al. 2015). They can rely on several sources of information to do this: online AVM as the one of MA, public listings, real estate agents or recent transaction prices[30] of comparable apartments. As their decision process goes on, sellers gathered information from these different sources, as buyers in Bikhchandani and Sharma (1996) gather quotes from merchants. By doing so, they gradually decrease their uncertainty, reinforce their prior and become less sensitive to exterior estimates. Interestingly, machine learning algorithms, that aim at mimicking the human decision process such as neural networks, gradient boosting or support vector machines, use successfully a similar decreasing learning rate (Hastie et al. 2009, chapters 10, 11 and 13).

A decreasing learning rate may explain our results for sellers but fail to explain why we do not observe a similar behavior among buyers. Just as sellers, buyers should increase the degree of certitude they have on prices and be less and less biased by the AVM error in their predictions. However, the problem they face is different from the one of sellers who as to assess a fair value for only one property: their own. Buyers have to decide their willingness to pay for a different apartment every time. Hence, their ability to correct the result of the MA machine

[30] November 6th 2013, the French fiscal administration opened an online service called PATRIM that enabled private person to request for a sample of the real estate fiscal record in order to prepare the sales or the purchase of a real estate property (see https://www.service-public.fr/particuliers/vosdroits/R34630)

thanks to information they have increasingly gathered along their search is limited. Then the question that arises is to understand why their opinion is less biased by the AVM result than sellers despite this lack of learning mechanism. The answer relies, in our opinion, on the argument already made in section III and supported by the literature (Nothcraft and Neale 1987, Black and Diaz 1996, Bucchianeri and Minson 2013). Unlike most of the sellers in our sample, buyers were asked to predict a price for an apartment already on the market. Hence, they make their estimate with an available asking price at their disposal. The nature of the problem changes. Buyers choose how much credit they give to two different pieces of information relative to the exact same apartment, whereas sellers as to weight the AVM result and information relative to imperfect comparables. A result from table 6 enhances our confidence in this explanation. A minority of the sellers knows the asking price for their property at the moment they are asked to predict the transaction price, the 22% that use the MA tool while their apartment is already on sale. The coefficient of the impact of the AVM error on their predictions is of 0.25 closer to the 0.17 of the buyers, yet statistically different at the 5% level, than of the average seller. Listing prices of the apartment are not available to us in the present study. A continuation of our work would be to compare the impact these two potentially contradicting anchors on the decision maker opinion.

Overall, the results of table 6 show that the anchoring and adjustment process of buyers and sellers is sensitive to their degree of certitude of the decision maker and the availability of different reference points. The more the search goes and the more relevant sources of information are available, the less the anchoring effect biases their decision process.

VI. Conclusion

In this article, we use apartment value estimates collected on the Internet to identify the factors that impact the opinion of the housing market participants. The persons whose estimations are studied here are users of an online AVM tool, asked to give their own sentiment after receiving the results from the models. Both the human and the machine prediction are compared with actual transaction price of the subsequent sale of the apartment, which occurred within a year following the estimation. Respondents are sellers and buyers with different degrees of advancement in their search for a counterpart.

Our first original result concerns the prediction error of the buyers that we are the first to measure, from our knowledge. Unlike homeowners, buyers' opinion is not biased by their

own interest. While sellers in our sample overestimate the value of their property, buyer do not underestimate the price of the apartments they visit. They catch up with the market reality quicker than sellers who hold on to overly optimistic expectations longer. Second, our work highlights the fact that the decision process of homebuyers and homesellers follows an anchoring and adjustment scheme, when asked to predict what would be a fair sale price. On both sides of the transaction, users' prediction errors show a strong correlation with the AVM errors which act as a reference point, with a stronger sensitivity of sellers. We attribute this difference to the fact, at the moment they make the prediction, buyers can rely on another anchor, the asking price, whereas most sellers actually use the online valuation tools to choose such a first marketing price. Sellers estimation is also subject to the influence of several possible reference points. We measure that sellers, whose property values decrease since they have purchased it, use their former purchasing price as an anchor for their prediction. Moreover, the impact of the AVM result on their prediction is weaker than for those who experience a nominal gain. This phenomenon can be understood as the effect of the homeowners' loss aversion and the general behavioral bias that drive investors to more easily update the prediction about their asset values upward than downward. Finally, we bring evidence of a learning mechanism among sellers along their search. As the process goes on and the information gathered reduce the degree of uncertainty, they decreasingly rely on the reference point offer by the AVM to predict a fair price for their apartment. Overall, taking into account an easily available reference point, together with the already known effects of endowment and owner estimation lags, we explain up to 35% of the variance in the prediction error of sellers and 19% for buyers. The R^2 of previous studies only reached a maximum of 11%. We see this as an argument in favor of the anchoring and adjustment heuristic as a way to model the thought process of homebuyers and sellers.

Our work is part of the recent real estate literature that leverage dataset drawn from Internet platforms in an academic perspective. Beside the empirical results on anchoring and adjustment in the housing market in a real-life setting, the present article highlights a meaningful result on these platforms. We demonstrate that the prediction tools these platforms

increasingly offer[31] influence the opinions of market participants. A future line of work would be to investigate to what extent, these predictions influence the end result of the sale process.

[31] Web site comparable to Meilleurs Agents exist all over the word, see for example Zillow (US) or Zoopla (UK) for the most famous ones.

Appendix

Appendix A: Indices “Notaire-Insee” for the Paris Region

Indices are based on the average of the prices in the year 2015 and corrected from the seasonal variation. Source: www.insee.fr/fr/statistiques/series/105071770

Table:

	Île-de-France	Essonne	Hauts-de-	Paris	Seine-et-	Seine-Saint-	Val-de-Marne	Val-d'Oise	Yvelines
1996-Q1	36,5	47,1	38,6	31	46,7	41,8	40	43,9	46,2
1996-Q2	35,9	48,8	38,1	30,5	45,6	40,2	39,5	42,8	44,2
1996-Q3	35,3	47,9	37	30,1	44,7	40	39,3	41,8	42,8
1996-Q4	35,1	47,7	37,1	29,9	44,9	39,1	39,4	41,6	42,9
1997-Q1	34	45,6	36,1	28,9	43,5	37,7	38	40,2	41,7
1997-Q2	33,9	45,4	35,5	29	43,3	37,7	37,8	40,3	41,6
1997-Q3	33,3	44,5	35,5	28,3	43,3	36	36,9	40,1	41,1
1997-Q4	33,3	45	35,4	28,4	43	36,3	36,9	40,6	41,3
1998-Q1	33,4	44,7	35,2	28,5	43,7	35,9	36,6	40	41,6
1998-Q2	33,5	44,4	35,3	28,7	43	35,6	36,5	40,4	41,9
1998-Q3	33,6	44,5	35,3	28,9	42,8	35,3	36,5	39,8	42,3
1998-Q4	34	44,1	35,7	29,5	43,5	35,3	36,6	40	42,8
1999-Q1	34,8	45,2	36,3	30,6	42,9	35,9	37,3	40,3	43
1999-Q2	35,2	45,1	36,7	31	43,5	35,9	37,3	40,3	43,7
1999-Q3	36,3	45,8	37,8	32,2	45	36,1	38,3	41,2	45,1
1999-Q4	37,1	45,7	38,8	33,2	45,1	36,7	39	41,3	45,6
2000-Q1	38,4	46,4	39,8	34,7	46,3	36,8	39,7	43,2	47,6
2000-Q2	39,4	46,9	40,9	35,8	47	37,7	40,4	42,9	48,1
2000-Q3	40,3	47,6	42,1	36,5	46,6	38,5	41,3	43,6	49,6
2000-Q4	40,9	48,2	42,7	37,3	46,7	38,9	41,9	43,4	50,2
2001-Q1	41,9	48,7	43,5	38,5	49,1	39,3	42,5	44,8	51
2001-Q2	42,5	49,5	44,1	38,8	49,2	40,3	43,3	46,3	52,4
2001-Q3	43,5	49,9	45,2	39,9	50,2	41	44,1	46,6	52,8
2001-Q4	44,2	50,4	45,7	40,5	51,3	41,7	45	47,8	54,5
2002-Q1	44,8	51	46,5	41	51,8	42,9	46,3	48,6	54,4
2002-Q2	46,1	52,5	48,3	42	53,3	44	47,5	49,3	55,8
2002-Q3	47,4	53,1	49,3	43,5	55,1	45	48,8	50,5	57,1
2002-Q4	48,8	54,8	50,5	45	56,1	46,3	50,2	51,8	58,2
2003-Q1	50,3	56,5	52,5	46,3	57,7	47,6	51,5	53,1	60,5
2003-Q2	51,9	57,8	54,2	47,8	59,1	49,1	53,6	55	62
2003-Q3	53,3	60	55,5	48,9	60,6	51,1	55	57,9	63,6
2003-Q4	55,1	62,4	57,9	50,2	62,3	53,3	57,5	58,9	66,3
2004-Q1	57,2	64,7	59,9	52	65,2	55,8	60	61,8	69,4

2004-Q2	59,4	67,7	62	54,2	67,6	57,9	62,3	65,3	69,8
2004-Q3	61,4	72	64	55,4	71	60,7	65,7	68,3	73,4
2004-Q4	63,8	75,3	66,3	57,6	74,5	63,7	67,3	72	76
2005-Q1	66,2	78,6	68,4	59,6	76,9	67,2	70,5	75	78,9
2005-Q2	68,6	82,5	70,8	61,4	81,5	70,9	73,8	78,8	81,9
2005-Q3	71,5	86,3	73,9	63,9	85,2	75,2	76,9	82,3	84,7
2005-Q4	73,5	89,7	75,5	65,6	88,6	77,5	79	87	87
2006-Q1	75,8	94,2	77,6	67,3	93,8	81,6	82,5	90,5	88,9
2006-Q2	78	97,9	79,5	69,1	97	84,7	84,7	94	91,8
2006-Q3	79,5	99,6	81	70,5	98,6	87,1	86,3	95,8	93,3
2006-Q4	81	101,6	82,3	71,9	101,1	88,4	87,7	97,4	94,5
2007-Q1	82,6	102,7	84,2	73,6	102,8	90,5	88	100	95,9
2007-Q2	84	103,8	85,2	75	103,7	92,3	90,2	100,4	97,1
2007-Q3	85,4	103,5	86,6	76,9	104,4	93,5	90,8	101,5	97,6
2007-Q4	87,6	105	88,7	79,4	105,4	94,9	92,6	104,1	98,6
2008-Q1	88,4	106,1	89,1	80,6	106,5	95,3	92,9	104,9	99,2
2008-Q2	89,2	105,6	89,3	82,2	105,6	95,1	92,8	104,5	99,3
2008-Q3	89	104,5	89,3	82,4	104,6	94,2	92,2	103,4	99
2008-Q4	87,6	102,1	87,9	81,2	102,2	92,5	90,8	101,2	97,1
2009-Q1	84,7	98,2	84,8	78,9	98,6	89,7	87,7	97,4	93,4
2009-Q2	82,3	96,6	82,4	76,2	97,1	87,9	85,8	95,8	90,6
2009-Q3	82,7	96,5	82,7	76,7	97,1	88,2	86,1	95,9	90,8
2009-Q4	84,1	98	84,2	78,1	98,7	89,2	87,4	97,3	92,6
2010-Q1	86,8	100,1	86,8	81,1	100,3	91,8	90,1	99,4	94,7
2010-Q2	89,4	101,4	89,4	84,3	102,1	93,3	92	100,9	96,6
2010-Q3	92	103	91,6	87,5	103,8	95,6	94,1	102,3	98,6
2010-Q4	95,6	104,6	94,7	92,2	104,8	98,5	97,4	104,2	100,6
2011-Q1	99,2	106,6	97,5	97,3	106,8	100,2	99,2	105,9	103,2
2011-Q2	103,4	108,2	101,6	102,7	107,7	103,4	102,8	107,7	106,3
2011-Q3	105	109	103,6	104,5	107,3	105,1	104,6	108,6	107,4
2011-Q4	105,5	108,2	104,5	105,3	107	104,8	104,5	107,9	108
2012-Q1	105,1	108,9	104,5	104,1	108	104,9	104,8	108,3	108
2012-Q2	105,2	107,4	103,8	105,4	107	104,1	104	107,1	107,2
2012-Q3	105,2	107	104,2	105,3	106,1	104,4	104,3	106,8	106,8
2012-Q4	104,7	107,7	104	104,3	106,5	103,9	104	107,4	107,7
2013-Q1	104,6	106,6	103,7	104,5	105,8	103,5	103,8	106,8	106,7
2013-Q2	104,2	106,7	103,6	103,6	105,9	104,3	104,1	106,6	106,1
2013-Q3	103,5	106,1	102,6	103,2	105,7	103	102,8	105,8	105,5
2013-Q4	102,9	104,7	102,2	102,8	104,5	102,3	102,7	104,9	104,3
2014-Q1	103	104,4	102,8	102,7	103,7	103,2	103,1	104,3	104
2014-Q2	102,6	103,7	102,2	102,6	102,5	102,2	102,5	103,7	103,8
2014-Q3	101,5	102,5	101,1	101,2	102,5	101,5	101,7	103	102,7
2014-Q4	101,2	101,1	102	100,4	101,2	101,6	101,4	101,5	103,4
2015-Q1	100,3	100,5	100,6	99,8	100,4	100,8	100,6	100,8	100,7

2015-Q2	99,8	100	99,9	99,6	100,3	100,3	100	100,2	99,9
2015-Q3	99,9	99,6	99,7	100,1	99,6	99,8	99,7	99,6	100
2015-Q4	100,1	99,7	99,3	100,8	99,4	99,6	99,6	99,5	99,4
2016-Q1	100,4	99,3	100,1	101,2	98,6	99,8	99,9	99,4	99,7
2016-Q2	101,4	99,4	100,9	102,6	98,6	100,4	100,7	99,6	100,3
2016-Q3	102,2	99,9	101,6	103,7	99,5	101	101,4	100,2	100,8
2016-Q4	103	99,9	101,8	105,1	100	100,5	101,6	100,9	101,4
2017-Q1	104,5	101,3	103,4	106,8	101,8	102,6	103,2	101,8	101,9
2017-Q2	106,1	100,6	104,8	109,5	100,9	103	103,8	101,1	102,3
2017-Q3	108	100,7	106,9	111,8	100,8	105,1	105,6	101,3	103,4
2017-Q4	108,8	100,5	106,3	114,2	100,9	104	104,8	101,3	103,4
2018-Q1	109,8	100,9	108,1	115	100,4	106,2	106,1	101,3	103,7
2018-Q2	111,2	101	109,1	117,2	100,7	107,2	106,7	101,6	103,6
2018-Q3	112,5	101,8	110,9	118,7	101,1	108,7	107,9	102,2	104,5
2018-Q4	113,8	101,9	111,6	120,7	100,9	110,4	109,3	102,1	103,9
2019-Q1	115,1	101,4	112,8	122,6	100,9	111,8	110,3	102,5	104,5
2019-Q2	116,7	102,3	114,3	124,7	102	113,2	111,2	102,6	105,7
2019-Q3	117,8	102,9	115,8	126	101,7	114,4	112,5	103,3	105,5
2019-Q4	120,2	104,4	118,1	128,7	103,9	116,6	114,1	105,3	108,1

Table 4-7 : Notaire-Insee inidces for the Paris Region from 1996 to 2019

Graph:

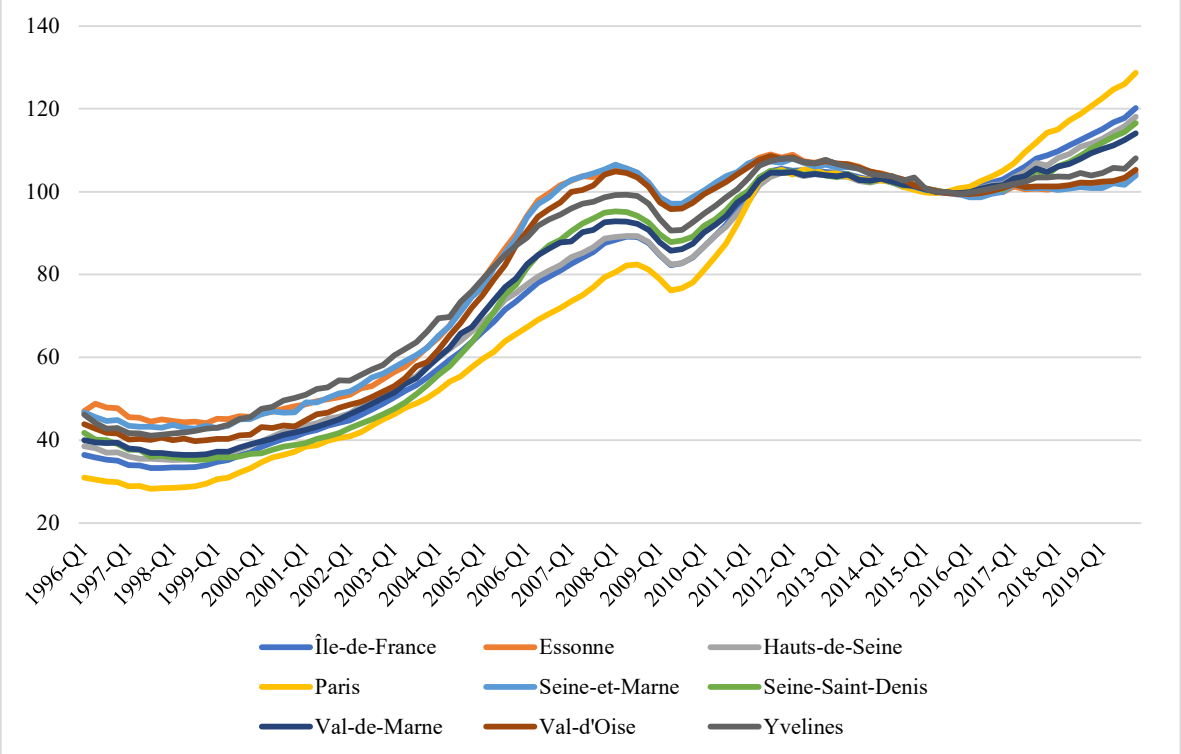

Figure 4-1 : Notaire-Insee indices for Paris Region from 1996 to 2019

Appendix B: Probit analysis

Dependent variable	$\ln(user\ estimate/transaction\ price)$				
	All users (1)	Owners (2)	Non-owners (3)	Sellers (4)	Buyers (5)
intercept	-1.8134***	-1.9211***	-2.4255***	-1.9942***	-2.5318***
	(0.0221)	(0.0247)	(0.0357)	(0.0267)	(0.0414)
1 room	-0.1553***	-0.1597***	-0.1067***	-0.1496***	-0.1031***
	(0.0131)	(0.0151)	(0.0201)	(0.0163)	(0.0235)
3 rooms	0.0496***	0.0496***	0.0339*	0.0534***	0.0289
	(0.0098)	(0.0110)	(0.0155)	(0.0119)	(0.0181)
4 rooms	-0.0527***	-0.0474***	-0.0547**	-0.0443**	-0.0579*
	(0.0127)	(0.0143)	(0.0205)	(0.0155)	(0.0240)
5 rooms	-0.1151***	-0.1373***	-0.0472	-0.1284***	-0.1092**
	(0.0200)	(0.0231)	(0.0303)	(0.0252)	(0.0372)
6 rooms	-0.3901***	-0.4639***	-0.2073***	-0.4634***	-0.2278***
	(0.0397)	(0.0504)	(0.0534)	(0.0562)	(0.0646)
no bathroom	-0.1592***	-0.1601***	-0.1185*	-0.2026***	-0.1060°
	(0.0344)	(0.0401)	(0.0509)	(0.0454)	(0.0591)
2 bathrooms	0.0417*	0.0351°	0.0458°	0.0140	0.0443
	(0.0163)	(0.0187)	(0.0252)	(0.0205)	(0.0302)
3 bathrooms	-0.0883	-0.1135	-0.0423	-0.0832	-0.0203
	(0.0599)	(0.0751)	(0.0810)	(0.0802)	(0.0980)
0 floors	-0.0451**	-0.0607***	0.0037	-0.0461**	-0.0155
	(0.0141)	(0.0158)	(0.0230)	(0.0170)	(0.0267)
1 floors	0.0012	-0.0118	0.0307	-0.0217	0.0004
	(0.0124)	(0.0139)	(0.0202)	(0.0151)	(0.0236)
3 floors	0.0192	0.0101	0.0345°	0.0030	0.0105
	(0.0130)	(0.0145)	(0.0210)	(0.0158)	(0.0244)
4 floors	0.0390**	0.0107	0.0885***	-0.0001	0.0596*
	(0.0140)	(0.0159)	(0.0218)	(0.0173)	(0.0254)
5 floors	0.0387*	0.0132	0.0818***	0.0046	0.0530°
	(0.0160)	(0.0183)	(0.0245)	(0.0200)	(0.0287)
6 floors	-0.0302*	-0.0414*	0.0051	-0.0369*	-0.0054
	(0.0152)	(0.0172)	(0.0238)	(0.0186)	(0.0277)
no elevator	0.0126	0.0166°	0.0013	0.0207°	0.0146
	(0.0086)	(0.0098)	(0.0136)	(0.0106)	(0.0160)
no cellar	-0.0574***	-0.0552***	-0.0470**	-0.0547***	-0.0625***
	(0.0093)	(0.0106)	(0.0148)	(0.0114)	(0.0174)
outdoor	0.0155	0.0137	0.0150	0.0106	0.0267
	(0.0103)	(0.0116)	(0.0166)	(0.0126)	(0.0195)
1 parking	-0.0306**	-0.0172	-0.0554**	-0.0145	-0.0644**
	(0.0106)	(0.0119)	(0.0174)	(0.0129)	(0.0204)
2 parkings	0.0295	0.0148	0.0594°	0.0171	0.0405
	(0.0192)	(0.0217)	(0.0309)	(0.0236)	(0.0369)
secondary room	-0.0346	-0.0878*	0.0541	-0.0887*	0.0227
	(0.0322)	(0.0389)	(0.0448)	(0.0421)	(0.0545)
Built date <1850	-0.1241*	-0.1247°	-0.0919	-0.1251°	-0.2357*
	(0.0587)	(0.0693)	(0.0836)	(0.0753)	(0.1157)
Built date [1850 ; 1913]	0.0068	0.0227	-0.0257	0.0127	-0.0013
	(0.0184)	(0.0212)	(0.0271)	(0.0230)	(0.0313)
Built date [1948 ; 1969]	-0.0434**	-0.0129	-0.0971***	-0.0281	-0.0689**
	(0.0141)	(0.0160)	(0.0218)	(0.0173)	(0.0252)
Built date [1970 ; 1980]	-0.1000***	-0.0783***	-0.1171***	-0.0808***	-0.1185***
	(0.0166)	(0.0188)	(0.0261)	(0.0203)	(0.0307)
Built date [1981 ; 1990]	-0.0829***	-0.0853**	-0.0498	-0.0716*	-0.0556
	(0.0230)	(0.0261)	(0.0363)	(0.0279)	(0.0428)
Built date [1991 ; 2000]	-0.0465*	-0.0309	-0.0657°	-0.0180	-0.0593
	(0.0223)	(0.0250)	(0.0364)	(0.0268)	(0.0427)

Built date [2001 ; 2010]	0.0338	0.0571*	-0.0303	0.0523*	-0.0486
	(0.0209)	(0.0232)	(0.0346)	(0.0251)	(0.0414)
Built date > 2011	-0.6995***	-0.6887***	-0.5604***	-0.7155***	-0.5745***
	(0.0242)	(0.0278)	(0.0386)	(0.0313)	(0.0476)
unknown built date	-0.0875***	-0.0838***	-0.0687***	-0.0855***	-0.0719**
	(0.0129)	(0.0149)	(0.0192)	(0.0161)	(0.0225)
Paris	0.2220***	0.1491***	0.3115***	0.1352***	0.2645***
	(0.0138)	(0.0155)	(0.0226)	(0.0168)	(0.0260)
Seine-et-Marnes	-0.2516***	-0.2093***	-0.3248***	-0.2133***	-0.3721***
	(0.0234)	(0.0253)	(0.0452)	(0.0275)	(0.0566)
Yvelines	-0.1313***	-0.1345***	-0.0897**	-0.1311***	-0.1072**
	(0.0187)	(0.0208)	(0.0317)	(0.0226)	(0.0372)
Essone	-0.3136***	-0.2606***	-0.4266***	-0.2384***	-0.3487***
	(0.0232)	(0.0249)	(0.0481)	(0.0265)	(0.0520)
Hauts-de-Seine	0.1416***	0.1091***	0.1795***	0.0951***	0.1434***
	(0.0139)	(0.0155)	(0.0230)	(0.0168)	(0.0267)
Seine-Saint-Denis	-0.1017***	-0.0875***	-0.1165***	-0.0668**	-0.1199**
	(0.0181)	(0.0199)	(0.0318)	(0.0214)	(0.0368)
Val-d'Oise	-0.1872***	-0.1400***	-0.3007***	-0.1183***	-0.2679***
	(0.0223)	(0.0240)	(0.0442)	(0.0256)	(0.0502)
sold in 2012	-0.1514***	-0.1654***	-0.0766**	-0.1438***	-0.0232
	(0.0171)	(0.0192)	(0.0274)	(0.0204)	(0.0313)
sold in 2013	0.1377***	0.1321***	0.1066***	0.1419***	0.1116***
	(0.0141)	(0.0156)	(0.0231)	(0.0166)	(0.0270)
sold in 2015	-0.0757***	-0.0913***	-0.0197	-0.0982***	-0.0294
	(0.0149)	(0.0167)	(0.0239)	(0.0180)	(0.0284)
sold in 2016	-0.1660***	-0.1965***	-0.0565*	-0.2154***	-0.0623*
	(0.0153)	(0.0173)	(0.0241)	(0.0188)	(0.0286)
sold in 2017	-0.1881***	-0.2151***	-0.0795**	-0.2475***	-0.0725*
	(0.0154)	(0.0174)	(0.0244)	(0.0191)	(0.0288)
sold in 2018	-0.1575***	-0.1722***	-0.0814***	-0.1886***	-0.0555*
	(0.0149)	(0.0167)	(0.0239)	(0.0181)	(0.0279)
sold in 2019	-0.3090***	-0.3226***	-0.2009***	-0.3436***	-0.2009***
	(0.0164)	(0.0186)	(0.0262)	(0.0203)	(0.0313)
Nb Observations	581307	581307	581307	581307	581307
Pseudo-R^2	0.057	0.051	0.058	0.051	0.055
Log-Likelihood Ratio	6950.3	4678.5	2631.8	3948.0	1733.8

Standard errors between parenthesis
*** p<0.001, ** p<0.01, * p<0.05, ° p<0.1

Table 4-8 : Result of a probit regression of the probability of presence in our dataset

Appendix C: Regression of the error with hedonic characteristics with Heckman correction

Dependent variable	$\ln(user\ estimate/transaction\ price)$			
	Owners (2)	Non-owners (3)	Sellers (4)	Buyers (5)
intercept	0.1976***	0.1142*	0.2103***	0.0795
	(0.0196)	(0.0495)	(0.0205)	(0.0596)
1 room	0.0123**	-0.0058	0.0106*	-0.0052
	(0.0043)	(0.0058)	(0.0046)	(0.0069)
3 rooms	-0.0031	-0.0050	-0.0014	-0.0060
	(0.0028)	(0.0041)	(0.0030)	(0.0048)
4 rooms	0.0047	-0.0000	0.0060	-0.0005
	(0.0037)	(0.0054)	(0.0040)	(0.0064)
5 rooms	0.0205***	-0.0202*	0.0261***	-0.0180°
	(0.0061)	(0.0079)	(0.0066)	(0.0100)
6 rooms	0.0135	-0.0032	0.0135	-0.0112
	(0.0120)	(0.0145)	(0.0133)	(0.0182)
2 bathrooms	-0.0001	0.0061	0.0058	-0.0022
	(0.0047)	(0.0065)	(0.0052)	(0.0081)
3 bathrooms	0.0129	0.0410*	0.0024	0.0228
	(0.0273)	(0.0196)	(0.0288)	(0.0241)
0 floors	0.0163***	0.0036	0.0125**	0.0137°
	(0.0041)	(0.0061)	(0.0043)	(0.0072)
1 floors	0.0030	-0.0004	0.0043	0.0067
	(0.0036)	(0.0054)	(0.0038)	(0.0065)
3 floors	0.0025	0.0045	0.0001	0.0120°
	(0.0037)	(0.0055)	(0.0040)	(0.0065)
4 floors	-0.0016	-0.0036	-0.0032	0.0099
	(0.0040)	(0.0058)	(0.0044)	(0.0069)
5 floors	0.0011	0.0059	-0.0053	0.0179*
	(0.0047)	(0.0065)	(0.0051)	(0.0076)
6 floors	-0.0018	-0.0083	-0.0007	0.0035
	(0.0047)	(0.0064)	(0.0050)	(0.0077)
no elevator	-0.0067*	-0.0048	-0.0103***	-0.0039
	(0.0026)	(0.0039)	(0.0029)	(0.0047)
no cellar	0.0095***	0.0050	0.0109***	0.0047
	(0.0029)	(0.0041)	(0.0031)	(0.0050)
outdoor	-0.0019	-0.0056	-0.0039	-0.0020
	(0.0028)	(0.0040)	(0.0030)	(0.0048)
1 parking	0.0103**	0.0102*	0.0145***	0.0090
	(0.0033)	(0.0051)	(0.0036)	(0.0062)
2 parkings	0.0142**	0.0056	0.0144*	0.0092
	(0.0055)	(0.0086)	(0.0059)	(0.0108)
Exceptional view	0.0105**	-0.0028	0.0100**	-0.0108
	(0.0033)	(0.0058)	(0.0035)	(0.0070)
Inf. relative perception	-0.0117	-0.0301***	-0.0176°	-0.0356***
	(0.0090)	(0.0062)	(0.0099)	(0.0074)
Sup. relative perception	0.0086***	0.0098*	0.0051°	0.0111°
	(0.0024)	(0.0045)	(0.0026)	(0.0057)
Renovated facade	0.0057*	0.0051	0.0045	0.0088°
	(0.0027)	(0.0041)	(0.0029)	(0.0049)
Renovated shared parts	0.0076**	0.0127**	0.0071**	0.0095*
	(0.0025)	(0.0039)	(0.0027)	(0.0047)
Secondary room]	0.0214*	0.0017	0.0176°	0.0125
	(0.0086)	(0.0136)	(0.0090)	(0.0173)
Renovation needed	-0.0158***	-0.0107*	-0.0115*	-0.0113*
	(0.0043)	(0.0043)	(0.0046)	(0.0051)
Built date <1850]	0.0077	-0.0099	0.0133*	-0.0160°
	(0.0057)	(0.0073)	(0.0061)	(0.0087)

Built date [1850 ; 1913]	-0.0038	-0.0064	-0.0010	-0.0036
	(0.0045)	(0.0057)	(0.0049)	(0.0067)
Built date [1948 ; 1969]	-0.0021	-0.0114*	0.0026	-0.0075
	(0.0041)	(0.0052)	(0.0045)	(0.0064)
Built date [1970 ; 1980]	0.0209***	-0.0048	0.0199***	-0.0015
	(0.0042)	(0.0061)	(0.0045)	(0.0073)
Built date [1981 ; 1990]	0.0255***	0.0126*	0.0231***	0.0151*
	(0.0045)	(0.0061)	(0.0048)	(0.0073)
Built date [1991 ; 2000]	0.0109°	0.0051	0.0117°	0.0105
	(0.0065)	(0.0095)	(0.0069)	(0.0113)
Built date [2001 ; 2010]	0.0072	0.0070	0.0051	0.0194°
	(0.0061)	(0.0093)	(0.0065)	(0.0114)
Built date > 2011	0.0322**	0.0071	0.0283*	0.0045
	(0.0104)	(0.0174)	(0.0116)	(0.0214)
Paris	-0.0269***	-0.0117	-0.0239***	0.0019
	(0.0042)	(0.0078)	(0.0045)	(0.0087)
Seine-et-Marnes	0.0285***	0.0102	0.0303***	0.0088
	(0.0070)	(0.0141)	(0.0074)	(0.0181)
Yvelines	0.0067	-0.0063	0.0100°	-0.0075
	(0.0056)	(0.0088)	(0.0060)	(0.0108)
Essone	0.0375***	0.0197	0.0356***	0.0219
	(0.0071)	(0.0155)	(0.0074)	(0.0164)
Hauts-de-Seine	-0.0130**	0.0006	-0.0137**	0.0078
	(0.0040)	(0.0068)	(0.0043)	(0.0078)
Seine-Saint-Denis	0.0180***	0.0114	0.0188***	0.0096
	(0.0053)	(0.0094)	(0.0056)	(0.0109)
Val-d'Oise	0.0280***	0.0279*	0.0268***	0.0349*
	(0.0066)	(0.0132)	(0.0069)	(0.0147)
Inv. Mills ratio	-0.0655***	-0.0357*	-0.0684***	-0.0286
	(0.0076)	(0.0170)	(0.0077)	(0.0199)
Nb Observations	7463	2874	6042	1899
R^2	0.078	0.054	0.083	0.070
Adj-R^2	0.073	0.041	0.076	0.049
F-Statistic	15.3	4.0	13.2	3.4

Table 4-9 : Regression of the error with hedonic characteristics with Heckman correction

Appendix D: Regression of the error with hedonic characteristics without Heckman correctio

Dependent variable	$\ln(user\ estimate/transaction\ price)$			
	Owners (2)	Non-owners (3)	Sellers (4)	Buyers (5)
intercept	0.0366***	0.0116	0.0380***	-0.0047
	(0.0060)	(0.0085)	(0.0064)	(0.0101)
1 room	0.0038	-0.0095°	0.0022	-0.0081
	(0.0042)	(0.0055)	(0.0045)	(0.0066)
3 rooms	0.0004	-0.0035	0.0022	-0.0051
	(0.0028)	(0.0040)	(0.0030)	(0.0047)
4 rooms	0.0044	-0.0004	0.0055	-0.0010
	(0.0038)	(0.0054)	(0.0040)	(0.0064)
5 rooms	0.0165**	-0.0201*	0.0222***	-0.0197*
	(0.0061)	(0.0079)	(0.0066)	(0.0099)
6 rooms	-0.0016	-0.0066	-0.0028	-0.0153
	(0.0120)	(0.0144)	(0.0133)	(0.0180)
2 bathrooms	0.0001	0.0068	0.0057	-0.0016
	(0.0048)	(0.0065)	(0.0052)	(0.0080)
3 bathrooms	0.0070	0.0392*	-0.0012	0.0223
	(0.0274)	(0.0196)	(0.0290)	(0.0241)
0 floors	0.0129**	0.0037	0.0101*	0.0133°
	(0.0041)	(0.0061)	(0.0043)	(0.0072)
1 floors	0.0029	0.0006	0.0037	0.0067
	(0.0036)	(0.0053)	(0.0039)	(0.0065)
3 floors	0.0036	0.0057	0.0009	0.0123°
	(0.0037)	(0.0055)	(0.0040)	(0.0065)
4 floors	-0.0001	-0.0006	-0.0021	0.0116°
	(0.0040)	(0.0056)	(0.0044)	(0.0068)
5 floors	0.0023	0.0087	-0.0042	0.0194*
	(0.0047)	(0.0063)	(0.0051)	(0.0076)
6 floors	-0.0027	-0.0076	-0.0013	0.0038
	(0.0047)	(0.0064)	(0.0051)	(0.0077)
no elevator	-0.0034	-0.0032	-0.0066*	-0.0027
	(0.0026)	(0.0039)	(0.0028)	(0.0046)
no cellar	0.0076**	0.0047	0.0092**	0.0041
	(0.0029)	(0.0041)	(0.0031)	(0.0050)
outdoor	-0.0017	-0.0057	-0.0038	-0.0021
	(0.0028)	(0.0040)	(0.0031)	(0.0048)
1 parking	0.0086*	0.0085°	0.0127***	0.0074
	(0.0034)	(0.0051)	(0.0036)	(0.0061)
2 parkings	0.0137*	0.0060	0.0140*	0.0096
	(0.0055)	(0.0086)	(0.0059)	(0.0108)
Exceptional view	0.0111***	-0.0026	0.0109**	-0.0109
	(0.0033)	(0.0058)	(0.0035)	(0.0070)
Inf. relative perception	-0.0118	-0.0296***	-0.0182°	-0.0354***
	(0.0091)	(0.0062)	(0.0099)	(0.0074)
Sup. relative perception	0.0078**	0.0093*	0.0041	0.0110°
	(0.0024)	(0.0045)	(0.0026)	(0.0057)
Renovated facade	0.0059*	0.0052	0.0044	0.0090°
	(0.0027)	(0.0041)	(0.0029)	(0.0049)
Renovated shared parts	0.0089***	0.0130***	0.0088***	0.0097*
	(0.0025)	(0.0039)	(0.0027)	(0.0047)
Secondary room]	0.0218*	0.0017	0.0184*	0.0121
	(0.0086)	(0.0136)	(0.0091)	(0.0173)
Renovation needed	-0.0159***	-0.0110*	-0.0121**	-0.0113*
	(0.0044)	(0.0043)	(0.0046)	(0.0051)
Built date <1850]	0.0069	-0.0096	0.0129*	-0.0158°
	(0.0057)	(0.0073)	(0.0062)	(0.0087)

Built date [1850 ; 1913]	-0.0111*	-0.0077	-0.0101*	-0.0051
	(0.0044)	(0.0056)	(0.0048)	(0.0066)
Built date [1948 ; 1969]	-0.0017	-0.0111*	0.0031	-0.0072
	(0.0041)	(0.0052)	(0.0045)	(0.0064)
Built date [1970 ; 1980]	0.0213***	-0.0057	0.0200***	-0.0020
	(0.0042)	(0.0061)	(0.0046)	(0.0073)
Built date [1981 ; 1990]	0.0251***	0.0121*	0.0225***	0.0145*
	(0.0045)	(0.0061)	(0.0048)	(0.0073)
Built date [1991 ; 2000]	0.0118°	0.0054	0.0126°	0.0102
	(0.0065)	(0.0095)	(0.0070)	(0.0113)
Built date [2001 ; 2010]	0.0133*	0.0084	0.0113°	0.0199°
	(0.0060)	(0.0093)	(0.0065)	(0.0114)
Built date > 2011	-0.0017	-0.0088	-0.0119	-0.0095
	(0.0097)	(0.0156)	(0.0107)	(0.0191)
Paris	-0.0172***	-0.0015	-0.0149***	0.0089
	(0.0041)	(0.0061)	(0.0044)	(0.0072)
Seine-et-Marnes	0.0138*	-0.0008	0.0149*	-0.0010
	(0.0068)	(0.0131)	(0.0073)	(0.0167)
Yvelines	-0.0013	-0.0097	0.0011	-0.0105
	(0.0055)	(0.0087)	(0.0060)	(0.0106)
Essone	0.0209**	0.0048	0.0193**	0.0120
	(0.0069)	(0.0138)	(0.0072)	(0.0149)
Hauts-de-Seine	-0.0065	0.0065	-0.0080°	0.0117
	(0.0040)	(0.0062)	(0.0043)	(0.0073)
Seine-Saint-Denis	0.0117*	0.0068	0.0130*	0.0055
	(0.0053)	(0.0091)	(0.0056)	(0.0105)
Val-d'Oise	0.0180**	0.0168	0.0175*	0.0269*
	(0.0065)	(0.0121)	(0.0069)	(0.0137)
Nb Observations	7463	2874	6042	1899
R^2	0.069	0.053	0.071	0.069
Adj-R^2	0.064	0.040	0.064	0.049
F-Statistic	13.7	4.0	11.4	3.4

Table 4-10 : Regression of the error with hedonic characteristics without Heckman correction

5. Conclusion

Les travaux présentés dans cette thèse se proposaient d'étudier le fonctionnement du marché immobilier à travers un prisme nouveau. Son originalité ne repose pas sur une formulation inédite du cadre théorique modélisant les mécanismes de rencontre et de recherche d'un accord entre acheteurs et vendeurs. Elle tient à la démarche empirique que nous avons entreprise pour observer ces mécanismes. Dans la continuité d'un mouvement commun à l'ensemble des domaines de recherches en sciences sociales, qui s'appuie sur les observations faites dans le monde numérique, les trois études composant cette thèse se basent sur les données issues d'une plateforme immobilière en ligne. Le postulat central soutenant la démarche est que les comportements des utilisateurs sur un tel site sont un miroir de leurs activités réelles sur le marché. Ainsi, nous avons pu mesurer des aspects des processus d'achat et de vente qui échappaient jusque-là à l'analyse empirique, faute de données. Les trois chapitres s'intéressent à trois dimensions différentes de ces processus, chacun s'insérant dans un cadre théorique qui lui est propre. Le premier quantifie les frictions dans le mécanisme d'appariement sur le marché immobilier à travers le modèle de *matching* Diamond-Mortensen-Pissarides). Le second consiste en une analyse du processus de recherche des acheteurs qui repose sur un modèle de recherche séquentielle *à la* Stigler. Enfin, le troisième est une preuve que la formation de l'opinion des acheteurs et des vendeurs sur les prix s'appuie sur une heuristique d'ancrage et d'ajustement, comme introduite par Kahneman et Tversky. Avant de revenir sur les apports respectifs de chacune de ses études, nous pouvons déjà tirer un bilan quant au bien-fondé de la démarche globale proposée dans ce travail doctoral.

Les résultats obtenus nous confortent dans nos avis quant à la pertinence de l'utilisation des données issues d'une plateforme Internet pour l'étude du marché immobilier. En premier lieu, car elle nous a permis des observations par ailleurs impossibles dans le monde analogique. Bien que la création de portails web les regroupant ait facilité l'exercice, dénombrer le stock de biens en ventes était déjà possible avant l'avènement d'Internet grâce aux annonces diffusées par les vendeurs. Mesurer l'intensité de la demande faisant face à cette offre, comme nous le faisons pour la première fois dans le premier chapitre, n'est en revanche possible qu'à travers l'analyse de l'activité des acheteurs sur une plateforme. C'est effectivement dans l'observation des comportements de ces derniers, qui laissaient jusqu'ici moins de traces que ceux des vendeurs, que l'apport de ces données nous semble le plus évident. Elles nous ont permis, dans le second chapitre, de suivre leur processus de recherche à travers un proxy des visites

successives menant à un achat pour comprendre comment des prix différents pouvaient être payés pour des biens apparemment semblables. Enfin, car les agents économiques utilisent les outils proposés par ces plateformes alors qu'ils sont actifs sur le marché, elles offrent la possibilité de les interroger in situ sur leur état d'esprit. C'est ainsi que les estimations faites par acheteurs et vendeurs sur le site nous ont permis de comprendre la formation de leurs prix de référence interne au cours de leurs recherches. À chaque fois, nous avons passé ces observations nouvelles au tamis de cadres théoriques éprouvés. Elles ont parfois confirmé des résultats établis, d'autres fois remis en question certaines hypothèses, mais à chaque fois offert la vision d'une réalité cohérente. Autant que possible, dans les premiers et derniers chapitres notamment, nous les avons confrontés à des observations venant de bases de données plus classiques et couramment utilisées en recherche immobilière comme les bases fiscales ou notariales. Non seulement nos nouvelles mesures concordent avec ces dernières, mais de plus elles permettent souvent d'en réduire la part de variance inexpliquée. En somme, nous avons acquis la conviction que l'utilisation de données issues de plateformes numériques est une opportunité pour la recherche en immobilier, mais pas seulement. Des sites similaires se développent et prennent une place toujours plus importante dans nos économies, en conséquence les secteurs où une démarche similaire à la nôtre sera possible et pertinente ne feront que grandir dans les années à venir.

Pour en revenir au seul marché immobilier, la vision d'ensemble qu'en offrent ces nouvelles mesures et qui transparait de nos travaux est celui d'un marché soumis à des frictions et des incertitudes fortes. Notre première étude en est la parfaite illustration. Nous utilisons les proportions d'acheteurs et de vendeurs utilisant un outil d'estimations immobilières en ligne, croisées au nombre exact de ventes enregistrées par l'administration fiscale, entre 2013 et 2017, dans quarante-deux des plus grandes aires urbaines françaises, pour proposer la première estimation de la fonction de *matching* du marché immobilier. Cette estimation révèle que les effets de congestion que génèrent les uns pour les autres les particuliers actifs sur ce marché sont forts. En effet, nous mesurons que l'appariement entre acheteurs et vendeurs se fait avec des rendements d'échelle décroissants, entre 0.7 et 0.8, avec des élasticités moyennes à leurs nombres respectifs de 0.4 et 0.3. Concrètement, cela signifie que si leurs nombres respectifs doublent, le nombre de ventes qui résultent de leurs rencontres n'augmente que de 60% à 75%. D'un point de vue académique, ce résultat contredit l'hypothèse de rendement d'échelles constant faite jusqu'ici dans la littérature immobilière. Rappelons que cette hypothèse se base sur les estimations faites sur le marché de l'emploi, premier champ d'application des modèles

de *matching*. D'un point de vue de la conception du marché, cela signifie que des organisations ayant intérêt à maximiser le nombre de ventes, comme les intermédiaires ou les plateformes immobilières devraient pousser à une plus grande segmentation du marché immobilier. Cette séparabilité accrue entre les segments pourrait se faire par l'entremise de moteurs de recherche d'annonces plus fins, de plus d'information disponible sur les biens en vente ou par l'assistance des acheteurs par des agents immobiliers mandatés par eux. Cette dernière solution, courant dans les pays anglo-saxons, pourrait également atténuer le problème de coordination que révèlent nos résultats. En plus des acteurs privés du marché, les pouvoirs publics pourraient également se réjouir d'une plus grande efficacité dans ce *matching* immobilier au vu du temps et des efforts qu'il consomme chaque année pour une part significative de la population.

Au-delà de la vision d'un processus de rencontre qui serait purement aléatoire, les estimations de deux extensions de cette première modélisation montrent que les comportements et les particularités des particuliers jouent un rôle dans le mécanisme d'appariement. Premièrement, sans qu'elles expliquent ces rendements d'échelles décroissants, l'impact négatif sur les volumes de ventes qu'a notre proxy du nombre de visites par acheteur suggère que deux défaillances de coordination opèrent simultanément sur le marché immobilier. La première tient au fait que certains vendeurs ne reçoivent aucune visite, quand d'autres en reçoivent de nombreuses, malgré l'impossibilité de vendre le même appartement à des personnes différentes. C'est le fameux problème urne-balles. La seconde réside dans la stratégie des acheteurs de visiter simultanément plusieurs biens, ce qui occupe plusieurs vendeurs, quand bien même ils ne feront d'offres que sur un seul bien à la fois. On peut penser qu'un « entremetteur » ayant une vision sur les deux faces du marché puisse amoindrir ces problèmes de coordination, même si la fragmentation du marché de l'intermédiation immobilière en France minimiserait l'impact de cette opportunité. Notons par ailleurs que ces modifications dans l'organisation du marché, qui pourraient augmenter le nombre de ventes, ne serviraient pas nécessairement les intérêts individuels des particuliers acheteurs ou vendeurs.

Nos derniers résultats sur l'efficacité du mécanisme d'appariement sur le marché immobilier concernent l'impact des caractéristiques des participants sur ce dernier. Tous les particuliers ne sont pas égaux dans leur capacité à générer des ventes. La première distinction est d'ordre temporel. Conformément à l'approche *stock-flow*, la proportion de biens nouvellement mis en vente influe positivement sur le volume de ventes, contrairement à ceux présents depuis un certain temps sur le marché et pour lesquels les acheteurs potentiels n'ont

pas exprimé d'intérêt. À l'inverse, les nouveaux acquéreurs, qui commencent tout juste leurs recherches, semblent moins à même d'aboutir à une transaction immédiatement. Comme s'ils avaient besoin de temps pour se décider et qu'au contraire ceux cherchant depuis longtemps perdaient patience. Nous mesurons chez les vendeurs d'autres sources de différence ayant une incidence sur le *matching*. Premièrement, les investisseurs, vendant un bien qui n'est pas leur résidence principale, mais qui était proposé à la location, sont plus susceptibles de vendre. Soit parce qu'ils se concentrent uniquement sur l'utilité monétaire de la propriété, soit parce qu'ils sont moins exposés à certains biais cognitifs comme l'effet de dotation. D'autre part, nous constatons une corrélation négative entre le nombre de vendeurs faisant appel à un agent immobilier et les volumes de ventes, à nombre d'acheteurs et vendeurs constants. Cela peut s'expliquer par des échanges, autrement mutuellement bénéfiques, empêchés par l'ajout des frais d'intermédiation au prix de vente. Le service rendu par ces « facilitateurs de marché » ne serait alors pas à la hauteur de la commission qu'ils en demandent. Cependant, dans la mesure où la majorité des vendeurs s'essaye à la vente entre particuliers avant de vendre majoritairement via agence, une causalité inverse est également possible. C'est dans les marchés plus compliqués que les particuliers font appel au service d'un agent, après avoir échoué à vendre seul.

L'ensemble de ces résultats empiriques n'ont pas de précédent à notre connaissance. Le prolongement naturel et immédiat de nos travaux serait donc bien évidemment d'essayer de les répliquer. Il a fallu plusieurs années, voire décennies, pour que s'ancre le consensus autour des rendements d'échelles constants dans l'appariement sur le marché de l'emploi. Nos seuls résultats ne sauraient être suffisants pour conclure définitivement à des rendements d'échelles décroissants pour l'ensemble des marchés immobiliers, à travers le monde et au cours du temps. Si la nature confidentielle des données à notre disposition empêche leur diffusion et donc la reproduction de nos mesures, notre démarche, elle, peut être facilement imitée. Des plateformes immobilières et des portails d'annonces en ligne existent et jouent un rôle prépondérant et croissant dans la rencontre entre acheteurs et vendeurs dans la majorité des marchés du logement à travers le monde. La réplication d'estimations semblables à la nôtre dans des contextes différents s'avèrerait très certainement riche d'enseignement et permettrait une comparaison entre des marchés aux organisations diverses. Ces mesures dans des contextes différents offriraient en premier lieu la possibilité d'identifier les raisons de la décroissance des rendements d'échelles que nous constatons. Par ailleurs, des résultats récents font état d'une différence dans la relation entre les intensités de l'offre et de la demande à l'échelle de marchés

parfaitement séparés et à l'intérieur même de ces marchés, entre leurs différents segments. La prise en compte de la segmentation des marchés immobiliers français, pris comme des touts indivisibles dans notre étude, pourrait révéler des variations entre les mécanismes d'appariement dans les segments très intégrés et dans ceux plus fortement isolés. Enfin, par analogie à l'économie du travail, nous faisons une distinction claire entre acheteur et vendeur. Dans la réalité, cette séparation n'est pas si nette, puisque de nombreux acquéreurs mettent leur propre résidence en vente alors qu'ils cherchent à acheter et inversement. La prise en compte des particuliers actifs des deux côtés du marché simultanément serait une prolongation intéressante de nos travaux.

Le chapitre suivant révèle comment frictions dans leur processus de recherche et incertitudes sur la valeur des biens immobiliers influencent le prix payé par un acheteur. La nature même du problème qui se pose aux acheteurs les entraine naturellement à débourser des montants différents pour des biens aux caractéristiques similaires. Ces derniers doivent en effet décider séquentiellement d'accepter ou non les conditions des vendeurs qu'ils rencontrent sans certitude sur le temps qui les sépare de la prochaine rencontre et sur l'offre qui leur sera faite alors. Ils doivent de plus tenir compte de coûts de recherche et de préférences temporelles propres à chacun. L'extension des modèles existant dans la littérature que nous proposons prédit, en plus des résultats déjà admis, un ajustement à la hausse du prix de réserve proportionnellement à une augmentation de l'espérance du taux de rencontre. Surtout, elle démontre l'effet ambigu de la taille du segment dans lequel se fait la recherche. L'augmentation de cette dernière s'accompagne d'un relèvement du prix de réserve si et seulement si la valeur du temps gagné par le raccourcissement des délais séparant les visites dépasse le renforcement des coûts de recherche dans une réserve de biens plus divers ou plus éloignés les uns des autres.

Encore une fois, notre apport le plus significatif est empirique plutôt que théorique. Jusqu'ici les articles traitant du processus d'achat devaient se contenter d'étudier l'impact des conditions initiales pesant sur lui pour expliquer la variance dans les prix des transactions immobilières. Ils n'existaient effectivement que peu de traces mesurables sur leur recherche elle-même, contrairement à celle des vendeurs qui eux se signalent en passant des annonces. En reconstruisant le parcours d'acheteurs parisiens utilisant un outil d'estimation en ligne au cours de leurs recherches, nous sommes en mesure de quantifier comment l'histoire d'un achat immobilier pèse sur sa conclusion. Il ressort de notre étude que l'élément qui distingue le plus un acheteur d'un autre consiste en ses croyances quant à la distribution des prix. Plus

précisément, nous montrons qu'un acheteur ayant visité des appartements en moyenne plus (respectivement moins) chers par mètre carré selon l'outil d'estimation que celui qu'il finit par acquérir, paie ce dernier plus (respectivement moins) cher, toutes choses égales par ailleurs. Une telle adaptation du prix de référence est cohérente avec une heuristique d'ancrage et d'ajustement, mais également avec un processus de recherche avec apprentissage. L'amateurisme des acheteurs face à la grande diversité dans les prix pratiqués plaide effectivement pour un tel mécanisme, où chaque visite est utilisée pour mettre à jour les croyances sur ce qu'est un prix juste et dissiper une part de cette incertitude.

Les valeurs des estimations des appartements visités le long du parcours d'achat révèlent également que les contraintes budgétaires pesant sur les acheteurs influencent le prix de la transaction. Il apparait que ceux s'écartant de leur budget initial en achetant un bien à la valeur absolue estimée plus élevée que ceux visités jusqu'alors le paient moins cher. Soit, car ils négocient plus fortement, soit, car ils ne font d'offres que si le vendeur le sous-évalue. En dehors des considérations de différence dans les niveaux de prix, nous apportons par ailleurs une première preuve que plus la recherche s'étend dans le temps moins le prix d'achat est élevé. Ce résultat en parfaite adéquation avec la théorie est le pendant de l'effet positif du temps d'écoulement d'un bien sur son prix de vente. Dans les deux cas, les particuliers les plus pressés par le temps paient ces contraintes temporelles en acceptant plus vite des propositions moins avantageuses. Enfin, nos prédictions quant à l'effet de la taille du segment envisagé sur le prix d'achat semblent se confirmer. Si des contraintes plus lâches quant aux caractéristiques, mesurées par la diversité des appartements visités, sont liées à une réduction des prix d'acquisition, l'étalement géographique de la recherche, plus à même d'accentuer les coûts de recherche, entraine une augmentation.

Ces résultats, s'ils venaient à être généralisés en étant reproduits par d'autres, pourraient avoir des conséquences concrètes sur la façon dont est organisé le processus d'achat. La prise de conscience de la façon dont la trajectoire du processus biaise la perception sur les prix plaide pour une plus grande information à la disposition des acheteurs, pour la responsabilisation des producteurs d'information immobilière et éventuellement pour un accompagnement par un tiers capable d'expliquer les différences dans les prix rencontrés. Par ailleurs, le rabais obtenu par les acheteurs de notre échantillon ayant concentré leurs recherches sur un seul et même quartier peut donner des pistes pour l'optimisation de la stratégie d'achat. Il y aurait un intérêt à mieux définir la zone d'intérêt en amont de la recherche et à s'y tenir, pour maximiser le retour sur les

coûts d'information et de déplacements liés à chaque visite plutôt que de se disperser. Les risques de ne pas être en mesure de détecter une bonne affaire à portée de main dépassant l'opportunité d'en rencontrer une ailleurs.

Notre troisième étude traite également de la capacité des particuliers à intégrer l'information à leur disposition pour se former une opinion sur le prix du bien qu'ils entendent acheter ou vendre. Cependant, plutôt que d'observer indirectement ses croyances par l'effet qu'elles ont sur les prix de transactions, comme dans le second article, nous sondons acheteurs et vendeurs directement, alors qu'ils sont engagés dans leur processus de recherche. Par l'entremise d'une fonctionnalité de l'outil d'estimation siège de nos observations, qui permet aux utilisateurs d'indiquer ce que serait un prix juste pour le bien dont ils viennent d'évaluer la valeur, nous récoltons les avis de participants actifs au marché immobilier de la région parisienne, entre 2012 et 2019. En comparant ces valorisations avec les montants des transactions effectivement conclues pour ces mêmes appartements, dans l'année qui suit, nous mesurons l'exactitude des prédictions de plus de dix mille propriétaires, vendeurs ou acheteurs. Encore une fois, les traces numériques semées par les utilisateurs de cette plateforme nous permettent des mesures inédites. Seul l'avis de propriétaires, pas nécessairement engagés dans processus de vente, via des questionnaires des administrations fiscales, avait fait jusqu'ici l'objet d'une attention dans la littérature académique. Les seuls résultats sur des particuliers actifs sur le marché viennent d'expériences menées en laboratoire où des personnes « jouaient le rôle » d'un acheteur ou d'un vendeur.

Nos résultats révèlent qu'au jeu des prédictions, les acheteurs sont bien plus performants que les vendeurs. D'abord, l'erreur absolue moyenne de leurs prédictions est de plus de deux points inférieure à celle des vendeurs, 6.2% contre 8.4% sur notre échantillon. Surtout, leur position sur le marché ne biaise pas leur perception des prix. En effet, si les vendeurs surestiment systématiquement leurs biens de plus de 5%, en conformité avec les résultats de la littérature sur le biais des propriétaires, les acheteurs de notre échantillon apparaissent comme largement non biaisés. Tout au plus, ils surestiment légèrement (1%) les biens qu'ils visitent quand une opinion perturbée par leurs intérêts aurait dû les pousser à les sous-estimer. Ils semblent également être plus au fait des évolutions des conditions de marché. Leurs estimations ne sont pas influencées par les changements de prix des mois qui précèdent alors que celles des vendeurs sont négativement corrélées à ces variations, mesurées par les indices de l'INSEE. De plus, les mouvements enregistrés par ces mêmes indices entre la date de l'estimation et la date

de vente expliquent mieux leurs erreurs. Enfin, l'évaluation faite par le moteur d'estimation en ligne de la plateforme influence moins leur idée du prix juste que celle des propriétaires vendant leur appartement.

Le second résultat significatif de cette étude est qu'elle confirme dans un contexte de marché réel ce que des expériences « en laboratoire » indiquaient. Face au problème d'estimer une quantité numérique inconnue, en l'occurrence la valeur monétaire d'un appartement, les particuliers se reposent sur une heuristique d'ancrage et d'ajustement. Les erreurs de prédiction des acheteurs comme celle des vendeurs s'expliquent en partie par l'erreur faite par l'outil d'estimation dont le résultat venait de leur être présenté. Les particuliers utilisent donc cet avis d'expert comme une valeur d'ancrage sur laquelle ils construisent leur prix de référence interne. Tous ne présentent pas la même sensibilité à ce point de référence. L'influence sur les acheteurs est moins forte que sur les vendeurs, comme noté plus haut, mais nous constatons également une hétérogénéité de l'effet parmi ces derniers. D'abord entre ceux dont la valeur du bien a baissé depuis leur acquisition initiale et les autres. Ceux-là sont sujets à une aversion à la perte et basent également leur estimation sur leur prix d'achat. Cette concurrence d'une autre valeur de référence diminue l'influence de la valorisation par le modèle du site pour ces propriétaires refusant d'accepter une vente à perte. Deuxièmement, son effet varie alors que le vendeur progresse dans son processus de vente. Plus l'issue de la recherche approche, moins il basera son opinion sur cette nouvelle information de prix, comme si un phénomène d'apprentissage le rendait de plus en plus sûr son jugement.

Par l'observation de leur prix de référence interne, ces mesures nous éclairent sur les mécanismes d'assimilation de l'information et de découverte des prix par les particuliers engagés dans un projet immobilier. Le fait que les acheteurs s'attachent moins que les vendeurs à des espérances initiales trop optimistes s'explique par les différences dans leur parcours de recherche d'une contrepartie. Dès le début, les premiers sont confrontés à la réalité du marché en prenant connaissance des divers biens en vente à travers les annonces et par leurs rencontres avec plusieurs vendeurs. Les seconds sont plus passifs et n'ont du marché qu'une vision étroite centrée sur leur bien propre. Ils ne rencontrent de contradiction qu'à travers les visites des acheteurs, qui n'arrivent qu'assez tard, à l'échelle de l'ensemble du projet de vente. Ce décalage de perception, théorisé dans plusieurs modèles décrivant la microstructure du marché immobilier, est une explication avancée par leurs auteurs à la chute des volumes et l'augmentation des délais de vente dans les marchés baissiers. La pondération des différents

points de référence selon leur disponibilité et le signal qu'ils portent montre comment, par touches successives, les particuliers ajustent leurs opinions et diminuent l'incertitude à laquelle ils font face. En la matière, la prise en compte des prix d'annonces représente une piste pour des travaux futurs. S'ils nous étaient indisponibles dans cette étude, les particuliers qui y avait accès, c'est-à-dire les acheteurs et les vendeurs ayant déjà mis leurs biens en vente, montrent une sensibilité plus faible au résultat de l'outil d'estimation. Enfin, ce travail souligne que loin de n'être que des terrains d'observation, les plateformes immobilières sont devenues des acteurs à part entière du marché. Puisque l'information qu'elle diffuse influence la perception de ses utilisateurs, elles ont de fortes chances d'avoir un impact sur le marché. La quantification de cet effet éventuel sur les comportements hors ligne des agents économiques, jusque dans les prix de transaction est un champ de recherche qu'il reste à investir.

La vision du fonctionnement du marché immobilier que nous livrent les observations nouvelles, permises par la démarche empirique déployée dans cette thèse, est donc celle d'un marché où les frictions et incertitudes sont fortes. Elles décrivent un marché où la rencontre entre acheteurs et vendeurs et leurs accords sur les prix se fait « à tâtons » et au prix de beaucoup de temps et d'efforts. Une vision en ligne avec la littérature existante, mais qui a pour elle l'avantage de pouvoir proposer une quantification des effets de ces tâtonnements. Dans l'ensemble des travaux présentés, ainsi que dans les pistes d'extensions avancées dans cette conclusion ou dans les opportunités identifiées en introduction, nous nous sommes cantonnés à traiter ces différentes sources de frictions de façon indépendante. Chacune des études ou pistes de recherche que nous avancions ne s'intéressait qu'à un seul aspect du problème. Le fait que toutes nos mesures aient déjà été possibles sur la même plateforme nous laisse penser que cette vision en silos n'est qu'une étape dans l'étude du marché immobilier.

Puisque les données d'un même site peuvent révéler à la fois l'histoire de la recherche de l'acheteur et celle de la commercialisation du bien, les informations reçues par eux, leurs interactions avec des tiers et l'état global du marché tant du point de vue de l'offre que de la demande, une compréhension holistique des phénomènes amenant à une transaction devient possible. Une telle approche permettrait d'expliquer leurs conséquences à l'échelle micro, sur le montant d'une transaction en particulier, comme macro, sur les évolutions en matière de prix et de volume. Pour se déployer, ce changement de paradigme devra certainement s'écarter des techniques économétriques classiques pour utiliser des outils propres à l'étude des systèmes complexes. Cette évolution nous semble d'autant plus probable que les plateformes elles-

mêmes, qui ont la maitrise des données la rendant possible, y trouveraient un intérêt. D'une part, car en réduisant la part d'inexpliqué dans le montant de chaque transaction, elle pourrait améliorer la qualité des informations qu'elles diffusent à leurs utilisateurs, notamment leurs outils d'estimations. D'autre part, alors que certaines plateformes entendent jouer un rôle croissant dans la transaction, cela leur offrirait un avantage compétitif certain vis-à-vis des autres acteurs du marché. Cette boucle de rétroaction entre maitrise de la donnée et renforcement d'une position dominante est bien connue dans l'économie des plateformes web et ne va pas sans poser des questions. Faudra-t-il, demain, créer de nouvelles régulations propres à ces acteurs dominants à la fois fournisseurs d'information et de services, intermédiaires et marchands de biens ?

Au-delà du seul cas immobilier, l'existence de ces plateformes doit également interroger les chercheurs quant à leur démarche de recherche. Comme nous espérons en avoir convaincu le lecteur, au moins en ce qui concerne notre champ d'études, elles représentent une formidable opportunité pour faire avancer notre compréhension des comportements humains. Cette opportunité s'accompagne de contraintes et de défis méthodologiques nouveaux, mais pas seulement. Elle crée également une tension éthique qui demande aux chercheurs de se positionner à titre individuel et en tant que communauté. Nous identifions au moins deux problèmes de natures différentes. Le premier est celui du consentement des personnes à partager les données utilisées dans nos études. Les opinions publiques, notamment européennes, montrent une sensibilité toujours croissante aux questions de respect de la vie privée en ligne et la recherche, au-delà du respect évident des dispositions légales, ne saurait être à la traine de ce mouvement. Comme le souligne Salganik (2017), il peut exister un certain fantasme chez le chercheur quant à la possibilité d'occuper la place centrale dans cet équivalent numérique du panoptique des frères Bentham. Cependant, il est nécessaire de rappeler qu'en tant que citoyen, le chercheur a également une place dans l'une des cellules soumises à cette surveillance permanente. Le fait que ces microscopes digitaux à disposition des chercheurs soient souvent créés et maintenus par des entreprises renforce la sensibilité de la question. Les bases fiscales, dont l'accès est possible dans le cadre de certaines recherches, sont un exemple de bases de données extrêmement intrusives, préexistant à l'avènement des plateformes. Pourtant, le fait que sa constitution soit le fait d'un état au service de l'intérêt au public et que son utilisation et son partage à des tiers se fassent sous un contrôle démocratique limite en partie, les risques d'un usage secondaire néfaste.

La nature privée des intérêts ayant gouverné à la construction de ces bases nous emmène naturellement vers le second problème posé par leur utilisation à des fins de recherche. Comme nous l'avons déjà souligné, ces plateformes ne sont pas seulement des terrains d'observations, mais également des acteurs des marchés sur lesquels elles opèrent. Le contrôle qu'elles ont sur les données nécessaires à ce nouveau type d'études leur donne de fait les moyens de peser sur les agendas de recherche. Au-delà du risque de censure éventuel de certains résultats, on peut craindre une prépondérance de plus en plus grande d'une vision utilitariste de la production du savoir en sciences sociales, en particulier en économie. D'autant plus que l'organisation de l'économie de plateforme pousse à la consolidation autour d'acteurs hégémoniques sur des secteurs entiers. Une recherche sur un de ces secteurs allant à l'encontre ou ne servant simplement pas les intérêts d'un tel acteur dominant serait de fait plus difficile à mettre en place. Le monde académique se trouve confronté, comme d'autres acteurs, au problème des monopoles constitués par ces entreprises du web. Il s'agit donc de mettre en regard les possibilités que ces derniers offrent avec les contraintes qui les accompagnent.

On peut voir là, pour conclure, une dernière contribution de ce travail doctoral. En plus de participer à l'effort de compréhension du fonctionnement du marché immobilier et à la littérature académique se basant sur des données issues du monde numérique, on peut aussi la considérer comme une étude de cas d'une collaboration entre le monde académique et une plateforme Internet. Loin d'être la seule ou même la première, elle demeure un exemple rare dans le contexte français tant, à l'instar des plateformes elles-mêmes, ce champ de recherche est dominé par des acteurs nord-américains. Elle démontre qu'une collaboration respectant les problèmes éthiques cités plus haut est possible et, nous espérons en avoir convaincu le lecteur, potentiellement fructueuse. Il y a tout lieu de s'en réjouir tant les horizons de recherche qu'ouvrent de telles collaborations sont riches pour les sciences sociales, en générale, et les sciences économiques, en particulier. S'il est permis de se donner une dernière fois en exemple, l'auteur de cette thèse et la société propriétaire de la plateforme en question envisagent déjà de faire perdurer cette collaboration. Cela permettra d'approfondir les résultats établis ici, en étendant par exemple les estimations des modèles de *matching* à d'autres pays européens dans lesquels l'entreprise étend ces activités. Ce prolongement de la coopération rendra également possible l'exploration de nouvelles pistes de recherche qui ne pouvaient rentrer dans le cadre du travail présenté ici. On pensera par exemple, en premier lieu, à l'étude de la concurrence sur le marché de l'intermédiation immobilière.

Bibliographie

Anenberg, Elliot (2011). Loss Aversion, Equity Constraints and Seller Behavior in the Real Estate Market. *Regional Science and Urban Economics*, 41(1), 67–76. https://doi.org/10.1016/j.regsciurbeco.2010.08.003

Albrecht, J., Anderson, A., Smith, E. and Vroman, S. (2007). Opportunistic matching in the housing market. *International Economics Review*. 48 (2), 641–664. https.doi.org/10.2307/4541983

Albrecht, J., P.A Gautier, and Vroman S. (2003). Matching with Multiple Applications. *Economics Letters* 78.1 (2003): 67-70. https://doi.org/10.1016/S0165-1765(02)00178-7

Albrecht, J., Gautier, P.A., and Vroman, S. (2006). Equilibrium directed search with multiple applications. *The Review of Economic Studies.* 73 (4), 869-891, https://doi.org/10.1111/j.1467-937X.2006.00400.x

Albrecht, J., Gautier, P.A., and Vroman, S. (2016). Directed search in the housing market. *Review of Economic Dynamics* .19, 218-231, https://doi.org/10.1016/j.red.2015.05.002

Anderson, P and Burgess, S. (2000). Empirical Matching Functions: Estimation and Interpretation Using Disaggregate Data. *Review of Economics and Statistics*. 82 (1), pp, 93-102, https://doi.org/10.1162/003465300558669

Anglin, P. M. (1997). Determinants of buyer search in a housing market. *Real Estate Economics*, 25 (4), 567–589. https://doi.org/10.1111/1540-6229.00728

Anglin, P.M., Rutherford, R. and Springer, T.M. (2003). The trade-off between the selling price of residential properties and time-on-the-market: The Impact of Price Setting. *The Journal of Real Estate Finance and Economics*, 26 (1) 95–111. https://doi.org/10.1023/A:1021526332732

Arkes, H., Hirshleifer, D., Jiang, D. and Lim, S. (2008). Reference point adaptation: tests in the domain of security trading. *Organizational Behavior and Human Decision Processes*. 105(1), 67–81. https://doi.org/10.1016/j.obhdp.2007.04.005

Arnosti, N., Johari, R., and Kanoria, Y. (2018). Managing Congestion in Matching Markets. *Manufacturing and Service Operations Management*, (Forthcoming), https://doi.org/10.2139/ssrn.2427960

Athey, S., and Luca, M. (2019). Economists (and Economics) in Tech Companies. *Journal of Economic Perspectives*, 33 (1), 209-30. https://doi.org/10.1257/jep.33.1.209

Azar, J., Marinescu, I., and Steinbaum M. (2020). Labor Market Concentration. *The Journal of Human Resources*? (Forthcoming), https://doi.org/10.3368/jhr.monopsony.1218-9914R1

Back, M.D., Küfner, A.C.P., and Egloff. B. (2010). The Emotional Timeline of September 11, 2001. *Psychological Science*, 21(10), 1417–19. https://doi.org/10.1177/0956797610382124

Bailey, M., Cao, R., Kuchler, T. and Stroebel, J. (2018). The Economic Effects of Social Networks: Evidence from the Housing Market. *Journal of Political Economy*, 126(6), 2224-2276. https://doi.org/10.1086/700073

Baker, S. and Fradkin A. (2017). The Impact of Unemployment Insurance on Job Search: Evidence from Google Search Data. *The Review of Economics and Statistics*. 99(5), 756-768, https://doi.org/10.1162/REST_a_00674

Banfi S. and Villena-Roldán B. (2019) Do High-Wage Jobs Attract More Applicants? Directed Search Evidence from the Online Labor Market. *Journal of Labor Economics*. 37 (3), 715-746, https://doi.org/10.1086/702627

Banerjee, A. V. (1992). A Simple Model of Herd Behavior. *Quarterly Journal of Economics.* 107 (3), 797- 817, https://doi.org/10.2307/2118364

Barbaro, Michael, and Tom Zeller Jr. 2006. "A Face Is Exposed for AOL Searcher No. 4417749." New York Times, August. http://select.nytimes.com/gst/abstract.html?res=F10612FC345B0C7A8CDDA 10894DE404482.

Beck, K.M., Beedle, M., Bennekum, A.V., Cockburn, A., Cunningham, W., Fowler, M., Grenning, J., Highsmith, J., Hunt, A., Jeffries, R., Kern, J., Marick, B., Martin, R.C., Mellor, S.J., Schwaber, K., Sutherland, J., & Thomas, D. (2013). Manifesto for Agile Software Development.

Benítez-Silva H., Eren S., Heiland F., Jiménez-Martín S. (2015). How well do individuals predict the selling prices of their homes? *Journal of Housing Economics*, 29, 12-25. https://doi.org/10.1016/j.jhe.2015.04.001

Bennet, R. and Pinto R. (1994). The Hiring Function in Local Labour Markets in Britain. *Environment and Planning*. 26 (12), 1957-1974, https://doi.org/10.1068/a261957

Ben-Shahar, D. and Golan, R. (2019). Improved information shock and price dispersion: A natural experiment in the housing market. *Journal of Urban Economics*.112(C), 70-84. https://doi.org/10.1016/j.jue.2019.05.008

Bell, U.-L. (1997). A Comparative Analysis of the Aggregate Matching Process in France, Great Britain and Spain. *Banco de Espafna- Servicio de Estudios, Documento de Trabajo 9721.*

Belot, M., Kircher, P., and Muller, P. (2019). Providing advice to jobseekers at low cost: An experimental study on online advice. *The Review of Economic Studies*. 86 (4), 1411-1447, https://doi.org/10.1093/restud/rdy059

Bikhchandani, S., Hirshleifer, D. and Welch, I. (1992), A Theory of Fads, Fashion, Custom, and Cultural Change in Informational Cascades. *Journal of Political Economy*. 100 (5), 992-1026, https://doi.org/10.1086/261849

Bikhchandani, S. and Sharma, S. (1996). Optimal search with learning. *Journal of Economic Dynamics and Control*, 20 (1-3), 333–59. https://doi.org/10.1016/0165-1889(94)00854-7

Black, R. and Diaz, J. (1996). The Use of Information versus Asking Price in the Real Property Negotiation Process. *Journal of Property Research,* 13(4), 287–297. https://doi.org/10.1080/095999196368808

Blake, T., and Coey, D. (2014). Why Marketplace Experimentation is Harder than It Seems: The Role of Test-Control Interference. *Proceedings of the Fifteenth ACM Conference on Economics and Computation*, 567–582. https://doi.org/10.1145/2600057.2602837

Blanchard, O. and Diamond P. (1994). Ranking, unemployment duration, and wages. *The Review of Economic Studies*. 61 (3), 417-434, https://doi.org/10.2307/2297897

Bolton, G., Greiner, B., Ockenfels, A. (2013). Engineering Trust: Reciprocity in the Production of Reputation Information. *Management Science*, 59(2), 265–85. https://doi.org/10.1287/mnsc.1120.1609

Box, G.E.P. (1976). Science and statistics. *Journal of the American Statistical Association*, 71(356), 791–799. https://doi.org/10.1080/01621459.1976.10480949

Brenčič, V. (2012). Wage posting: Evidence from job ads. *Canadian Journal of Economics*. 45 (4), 1529–1559, https://doi.org/10.1111/j.1540-5982.2012.01738.x

Bucchianeri, G. and Minson, J. (2013). A homeowner's dilemma: Anchoring in residential real estate transactions. *Journal of Economic Behavior and Organization*. 89, 76–92. https://doi.org/10.1016/j.jebo.2013.01.010

Burdett K. and Coles M. (1997). Marriage and Class. *The Quarterly Journal of Economics.* 112 (1), 141–168, https://doi.org/10.1162/003355397555154

Burdett, K. and Mortensen, D. (1998). Wage Differentials, Employer Size, and Unemployment. *International Economic Review*. 39 (2), 257- 73. https.doi.org/10.2307/2527292

Burgess, S. (1993). A Model of Competition between Unemployed and Employed Job-Searchers: An Application to the Unemployment Outflow Rate in Britain. *The Economic Journal*. 103 (420), 1190-1204, https.doi.org/10.2307/2234245

Burgess, S. and Profit, S. (2001). Externalities in the matching of workers and firms in Britain. *Labour Economics*. 8 (3), 313-333, https://doi.org/10.1016/S0927-5371(01)00036-7

Cailly C., Côte J.-F., David A., Friggit J., Gregoir S., Nobre A., Proost F., Rougerie C., Schoffit S., Tauzin N., Thélot H. (2019). Les indices Notaires-Insee des prix des logements anciens Méthodologie v4. *INSEE Méthodes*, 132.

Carrillo, P.E. (2012). An empirical stationary equilibrium search model of the housing market. *International Economic Review*, 53 (1), 203–234, https://doi.org/10.1111/j.1468-2354.2011.00677.x

Carrillo, P.E. (2013). To sell or not to sell: measuring the heat of the housing market. *Real Estate Economics*, 41 (2), 310–346. https://doi.org/10.1111/reec.12003

Case, K. and Shiller, R. (1989). The Efficiency of the Market for Single-Family Homes. *American Economic Review*, 79(1), 125-37. https://doi.org/10.1016/j.apenergy.2018.06.013

Cavallo, A. and Rigobon, R. (2016). The Billion Prices Project: Using Online Prices for Measurement and Research. *Journal of Economic Perspectives*, 30(2), 151-178. https://doi.org/10.2307/43783711

Chapelle, G. and Eyméoud, J.B. (2017). Can big data increase our knowledge of the rental market ? LIEPP Working paper.

Cherbonnier, F. and Leveque, C. (2019). The impact of competition on experts' information disclosure: the case of real estate brokers. Working paper.

Choi, H. and Varian, H. (2012). Predicting the Present with Google Trends. *Economic Record*. 88(1),2–9. https://doi.org/10.5018/economics-ejournal.ja.2018-34

Coles, M., and Petrongolo, B. (2008). A Test between Stock-Flow Matching and the Random Matching Function Approach. *International Economic Review.* 49.4 (2008): 1113-1141, https://doi.org/10.1111/j.1468-2354.2008.00508.x

Coles, M. and Smith E. (1996). Cross-Section Estimation of the Matching Function: Evidence from England and Wales. *Economica*. 63 (252), 589-98, https://doi.org/10.2307/2554997

Coles, M. and Smith E. (1998). Marketplaces and Matching. *International Economic Review*. 39 (1), 239-254. https://doi.org/10.2307/2527239

Corina, P. and Chenavaz, R. (2011). Sellers' and Buyers' Reference Point Dynamics in the Housing Market. *Housing Studies*, 26(3), 329-352. https://doi.org/10.1080/02673037.2011.542095

Courant, P.N. (1978). Racial prejudice in a search model of the urban housing market. *Journal of Urban Economics*. 5(3), 329-345. https://doi.org/10.1016/0094-1190(78)90014-1

Cox, M. and Ellsworth, D. (1997). Application-controlled demand paging for out-of-core visualization. *Proceedings. Visualization '97*, 235-244. https://doi.org/10.1109/VISUAL.1997.663888

Cubbin, J. (1974). Price, quality, and selling time in the housing market. *Applied Economics,* 6 (3), 171–187. https://doi.org/10.1080/00036847400000017

Dachis, B., Duranton, G., and Turner, M.A. (2011). The effects of land transfer taxes on real estate markets: evidence from a natural experiment in Toronto. *Journal of Economic Geography*. 12 (2), 327–354, https://doi.org/10.1093/jeg/lbr007

David, A., Dubujet, F., Gouriéroux, C., Laferrère, A., Friggit, J., Moisant, T., Le Blanc, D., Lollivier, S. (2002). Les indices de prix des logements anciens. *Insee Méthodes.* 98.

Dellarocas, C. (2003). The digitization of word of mouth: Promise and challenges of online feedback mechanisms. *Management Science*. 49(10), 1407–1424. https://doi.org/10.2307/4134013

Diamond, P. (1982a). Aggregate Demand Management in Search Equilibrium. *Journal of Political Economy*. 90 (5), 881-894, https://doi.org/10.1086/261099

Diamond, P. (1982b). Wage Determination and Efficiency in Search Equilibrium. *Review of Economical Studies*. 49 (2), 217-227, https://doi.org/10.2307/2297271

Diaz, A., Jerez, B. (2013). House prices, sales, and time-on-market: a search-theoretic framework. *International Economics Review*. 54 (3), 837–872, https://doi.org/10.1111/j.1468-2354.2007.00440.x

Drake, M.S., Roulstone, D.T. and Thornock, J.R. (2012). Investor information demand: Evidence from Google searches around earnings announcements. *Journal of Accounting Research*. 50(4), 1001-1040. https://doi.org/10.1111/j.1475-679X.2012.00443.x

Einav, L., Knoepfl, D., Levin, J. and Sundaresan, N. (2014). Sales Taxes and Internet Commerce. *American Economic Review*, 104(1). 1–26. https://doi.org/10.1257/aer.104.1.1

Einav, L., Kuchler, T., Levin, J. and Sundaresan, N. (2015). Assessing Sale Strategies in Online Markets Using Matched Listings. *American Economic Journal: Microeconomics*. 7(2), 215-247. https://doi.org/10.1257/mic.20130046

Einav, L., Farronato, C. and Levin, J. (2016). Peer-to-peer markets. *Annual Review of Economics*. 8, 615–635. https://doi.org/10.1146/annurev-economics-080315-015334

Einav, L., Farronato, C., Levin, J. and Sundaresan, N. (2018). Auctions versus Posted Prices in Online Markets. *Journal of Political Economy*. 126(1),178-215. https://doi.org/10.1086/695529

Einav, L. and Levin, J. (2014). The Data Revolution and Economic Analysis. *Innovation Policy and the Economy*. 14:1-24. https://doi.org/10.1086/674019

Einiö, M., Kaustia, M., and Puttonen, V. (2008). Price Setting and the Reluctance to Realize Losses in Apartment Markets. *Journal of Economic Psychology*, 29(1), 19–34. https://doi.org/10.1016/j.joep.2007.02.004

Elder, H.., Zumpano, L., and Baryla, E. (1999). Buyer search intensity and the role of the residential real estate broker. *Journal Real Estate Finance and Economics*. 18 (3), 351–368, https://doi.org/10.1023/A:1007737102125

Ettredge, M., Gerdes, J., and G Karuga (2005). Using web-based search data to predict macroeconomic statistics. *Communications of the ACM.* 48 (11), 87-92. https://doi.org/10.1145/1096000.1096010

Evans, D. and Schmalensee, R. (2007). The Catalyst Code: The Strategies Behind the World's Most Dynamic Companies. Boston: *Harvard Business School Press*.

Faberman, J. and Kudlyak M. (2016). What does online job search tell us about the labor market? *Economic Perspectives, Federal Reserve Bank of Chicago*. 1(1), 1-15.

Faberman, J. and Kudlyak (2019). The Intensity of Job Search and Search Duration. *American Economic Journal: Macroeconomics*. 11 (3), 327-357, https://doi.org/10.1257/mac.20170315

Ford, J.S., Rutherford, R.C. and Yavas, A. (2005). The effects of the internet on marketing residential real estate. *Journal of Housing Economics*.14(2), 92-108. https://doi.org/10.1016/j.jhe.2005.06.003

Fradkin, A. (2017a). Digital marketplaces. In *The New Palgrave Dictionary of Economics*. London: Palgrave Macmillan.

Fradkin, A. (2017b). Search, Matching, and the Role of Digital Marketplace Design in Enabling Trade: Evidence from Airbnb. *Working Paper*, https://doi.org/10.2139/ssrn.2939084

Fradkin, A., Grewal, E., Holtz, D. and Pearson, M. (2017). The determinants of online review informativeness: Evidence from field experiments on Airbnb. Working Paper. https://doi.org/10.2139/ssrn.2939064

Franke M. and van Dijk, D. (2018). Internet Search Behavior, Liquidity and Prices in the Housing Market. *Real Estate Economics*. 46 (2), 368-403, https://doi.org/10.1111/1540-6229.12187

Gatzlaff, D. and Tirtiroglu, D. (1995). Real Estate Market Efficiency: Issues and Evidence. *Journal of Real Estate Literature.*3(2),157-189. https://doi.org/10.2307/44103296

Gee, L. (2019). The More You Know: Information Effects on Job Application Rates in a Large Field Experiment. *Management Science*. 65(5), 1949-2443. https://doi.org/10.1287/mnsc.2017.2994

Ge, C., Huang, K.-W., Png, I.P.L. (2016). Engineer/scientist careers: Patents, online profiles, and misclassification bias. *Strategic Management Journal*. 37(1), 232–253. https://doi.org/10.1002/smj.2460

Geltner, D., MacGregor, B. D. and Schwann. G.M. (2003). Appraisal Smoothing and Price Discovery in Real Estate Markets. *Urban Studies.* 40(5-6),1047–1064. https://doi.org/10.2307/43084304

Genesove, D. and Han L. (2012). Search and matching in the housing markets. *Journal of Urban Economics*, 72 (1), 31–45, https://doi.org/10.1016/j.jue.2012.01.002

Genesove, D. and Mayer, C. (1997). Equity and time to sale in the real estate market. *The American Economic Review*. 87 (3), 255–269. https://doi.org/10.2307/2951345

Ginsberg, J., Mohebbi, M., Patel, R., Brammer, L., Smolinski, M.S. and Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*. 457,1012–1014. https://doi.org/10.1038/nature07634

Glower, M., Haurin, D.R. and Hendershott, P.H. (1998). Selling time and selling price: the influence of seller motivation. *Real Estate Economics*. 26 (4), 719–740, https://doi.org/10.1111/1540-6229.00763

Goodman, J.L. and Ittner, J.B. (1992), The accuracy of home owners' estimates of house value. *Journal of Housing Economics*. 2(4), 339-357. https://doi.org/10.1016/j.jhe.2019.101660

Gregg, P. and Petrongolo, B. (1997). Random or Non-Random Matching? Implications for the Use of the UV Curve as a Measure of Matching Performance. Papers 13, Centre for Economic Performance and Institute of Economics

Hall, J., Kendrick, C. and Nosko. C. (2016). The Effects of Uber's Surge Pricing: A Case Study. Working paper.

Han, L. and Strange, W. (2016). What is the role of the asking price for a house? *Journal of Urban Economics*. 93, 115-130. https://doi.org/10.1016/j.jue.2016.03.008

Han, L. and Strange, W. (2015). The Microstructure of Housing Markets: Search, Bargaining, and Brokerage. *Handbook of Regional and Urban Economics*. 5, 813-886, https://doi.org/10.1016/B978-0-444-59531-7.00013-2

Hastie, T., Tibshirani, R. and Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed. New York: *Springer.*

Haurin, D. (1988). The Duration of Marketing Time of Residential Housing. *Real Estate Economics*. 16(4), 396-410. https://doi.org/10.1016/j.regsciurbeco.2011.08.010

Havard, T. (1999). Do valuers have a greater tendency to adjust a previous valuation upwards or downwards? *Journal of Property Investment and Finance*, 17(4), 365–373. https://doi.org/10.1108/14635789910271755

Heckmann, J. J. (1979). Sample Selection Bias as a Specification Bias. *Econometrica,* 47(1), 153-161. https://doi.org/10.2307/1912352

Hendel, I., Nevo, A., Ortalo-Magné, F. (2009). The relative performance of real estate marketing platforms: MLS versus FSBO Madison.com. *The American Economic Review*. 99 (5), 1878–1898, https://doi.org/10.1257/aer.99.5.1878

Henriques, A.M., (2013). Are Homeowners in Denial about their House Values? Comparing Owner Perceptions with Transaction-Based Indexes. *Finance and Economics Discussion Series 2013-79. Board of Governors of the Federal Reserve System (U.S.)*. https://doi.org/10.2139/ssrn.2357665

Hilbert, M. and López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*. 332(6025): 60–65. https://doi.org/10.1126/science.1200970

Horowitz, J.L. (1992). The role of the list price in housing markets: theory and an econometric model. *Journal Applied Economics*. 7 (2), 115–129, https://doi.org/10.1002/jae.3950070202

Horton, J. (2017). The Effects of Algorithmic Labor Market Recommendations: Evidence from a Field Experiment. *Journal of Labor Economics*. 35(2), 345-385. https://doi.org/10.1086/689213

Horton, J. (2019). Buyer Uncertainty About Seller Capacity: Causes, Consequences, and a Partial Solution. *Management Science*. 65 (8), 3449-3947, https://doi.org/10.1287/mnsc.2018.3116

Horton, J., Rand, D.G. and Zeckhauser, R.J. (2011). The online laboratory: Conducting experiments in a real labor market. *Experimental Economics*.14(3), 399–425. https://doi.org/10.1007/s10683-011-9273-9

Horton, J. and Tambe, P. (2015) Labor Economists Get Their Microscope: Big Data and Labor Market Analysis. *Big Data*. 3(3),130-137. https://doi.org/10.1089/big.2015.0017

Howard, W., Gunnar, G., Michael, B. (1974). TREBIG: A 360/75 FORTRAN program for three-mode factor analysis designed for big data sets. *Behavior Research Methods and Instrumentation*, 6, 53-54. https://doi.org/10.3758/BF03200290

Ihlandfelt, K. R., and Martinez-Vazquez, J. (1986). Alternative Value Estimates of Owner-Occupied Housing: Evidence on Sample Selection Bias and Systematic Errors. *Journal of Urban Economics* 20(3), 356-369. https://doi.org/10.1016/0094-1190(86)90025-2

Ihlanfeldt, K. and Mayock, T. (2012). Information, search, and house prices: revisited. *The Journal of Real Estate Finance and Economics*, 44 (1), 90–115. https://doi.org/10.1007/s11146-010-9282-z

Jones, S. (1989). Job Research Methods, Intensity and Effects. *Oxford Bulletin of Economics and Statistics.* 51(3), 277-296

Jovanovic, B. (1979) Job Matching and the Theory of Turnover. *Journal of Political Economy*. 87 (5), 972-990, https://doi.org/10.1086/260808

Kabra, A., Belavina, E., and Girotra K. (2016). Designing promotions to scale marketplaces. *Working paper.*

Kahneman, D. and Amos Tversky, A. (1979). Prospect Theory: An Analysis of Decision under Risk. *Econometrica*, 47(2), 263–292. https://doi.org/10.1142/9789814417358_0006

Kain, J. F., and Quigley, J. M. (1972). Note on Owner's Estimate of Housing Value. *Journal of the American Statistical Association,* 67(340), 803-806. https://doi.org/10.1080/01621459.1972.10481296

Kholodilin, K., Podstawski, M., and Siliverstovs, B. (2010). Do Google Searches Help in Nowcasting Private Consumption?: A Real-Time Evidence for the US. *Discussion Papers of DIW Berlin*, 997, https://doi.org/10.2139/ssrn.1615453

Kish, L., and Lansing, J. B. (1954). Response Errors in Estimating the Value of Homes. *Journal of the American Statistical Association*. 49(267), 520-538. https://doi.org/10.2307/2281128

Kohavi, R., Longbotham, R., Sommerfield, D. et Henne, R. (2009) Controlled experiments on the web: survey and practical guide. *Data Mining and Knowledge Discovery*. 18, 140–181. https://doi.org/10.1007/s10618-008-0114-1

Kohn, MG. and Shavell, S. (1974). The theory of search. *Journal of Economic Theory*, 9, 93–123. https://doi.org/10.1016/0022-0531(74)90061-1

Krainer, J. (2001). A theory of liquidity in residential real estate markets. *Journal of Urban Economics*. 49 (1), 32-53, https://doi.org/10.1006/juec.2000.2180

Kroft, K., Lange, F., and Notowidigdo, M. (2013), Duration dependence and labor market conditions: Evidence from a field experiment. *Quarterly Journal of Economics*, 128 (3), 1123–1167, https://doi.org/10.1093/qje/qjt015

Krueger, A. and Mueller, A. (2011), Job search, emotional well-being, and job finding in a period of mass unemployment: Evidence from high-frequency longitudinal data. *Brookings Papers on Economic Activity*, 42 (1), 1–81. https://doi.org.10.1353/eca.2011.0001

Kuhn, P. and Shen, K. (2013). Gender discrimination in job ads: Evidence from China. *Quarterly Journal of Economics*, 128 (1), 287–336, https://doi.org/10.1093/qje/qjs046

Lambson, V.E., McQueen G.R. and Barrett A.S. (2004). Do out-of-state buyers pay more for real rstate? An Examination of Anchoring-Indexed Bias and Search Costs. *Real Estate Economics*, 32 (1), 85–126, https://doi.org/10.1111/j.1080-8620.2004.00085.x

Larceneux F., Lefebvre T., Simon A. (2015). What added value of estate agents compared to FSBO transactions? Explanation from a perceived advantages model. *Journal of Housing Economics*, 29, 72–82, https://doi.org/10.1016/j.jhe.2015.06.002

Layard, R., Nickell, S., and Jackman, R. (1991). Unemployment, Macro- economic Performance and the Labour Market. *Oxford University Press*.

Lefebvre, T. (2015). Une nouvelle ère pour l'intermédiation en immobilier résidentiel : fondements, digitalisation et limites. Université Paris Dauphine - Paris IX

Levitt, S.D. and Syverson, C. (2008). Market Distortions When Agents Are Better Informed: The Value of Information in Real Estate Transactions. *Review of Economics and Statistics*. 90(4), 599-611. https://doi.org/10.2307/40043103

Li, N. (2017). Who Are My Peers? Labor Market Peer Firms Through Employees' Internet Co-Search Patterns. *Rotman School of Management Working Paper* No. 2558271. https://doi.org/10.2139/ssrn.2558271

Li, J. and Netessine, S. (2020), Higher Market Thickness Reduces Matching Rate in Online Platforms: Evidence from a Quasiexperiment. *Management Science*, 66 (1), 1-501, https://doi.org/10.1287/mnsc.2018.3223

Ling, C.D., Naranjo, A., and Petrova, M.T. (2018). Search costs, behavioral biases, and information intermediary effects. *Journal of Real Estate Finance and Economics*, 57 (1), 114–151. https://doi.org/10.1007/s11146-016-9582-z

Loberto, M., Luciani, A. and Pangallo, M. (2020). What Do Online Listings Tell Us About the Housing Market? *Bank of Italy Working Paper* No. 1171. https://doi.org/10.2139/ssrn.3176962

Locke, P. and Mann, S.C. (2000). Do Professional Traders Exhibit Loss Realization Aversion? *Working Paper. Texas Christian University*

Luca, M. (2016). Reviews, Reputation, and Revenue: The Case of Yelp.Com. *Harvard Business School NOM Unit Working Paper* No. 12-016.

Machin, S. and Manning, A. (1999). The Causes and Consequences of Long-term Unemployment in Europe. *Handbook of Labor Economics*, Vol. 3, Ch. 47, 3085-3139. https://doi.org/10.1016/S1573-4463(99)30038-9

Marinescu, I. (2017). The general equilibrium impacts of unemployment insurance: Evidence from a large online job board. *Journal of Public Economics*, 150,14-29, https://doi.org/10.1016/j.jpubeco.2017.02.012

Marinescu, I. and Rathelot. R. (2018). Mismatch Unemployment and the Geography of Job Search. *American Economic Journal: Macroeconomics*. 10(3), 42-70, https://doi.org/10.1257/mac.20160312

Marinescu, I. and Wolthoff, R. (2020). Opening the Black Box of the Matching Function: The Power of Words. *Journal of Labor Economics,* 38(2), 535-568, https://doi.org/10.1086/705903

Mason, W., Suri, S. (2012) Conducting behavioral research on Amazon's Mechanical Turk. *Behavior Research Methods*. 44, 1–23. http://doi.org/10.3758/s13428-011-0124-6

Masterov, D.V., Mayer, U.F. and Tadelis, S. (2015). Canary in the e-Commerce Coal Mine: Detecting and Predicting Poor Experiences using Buyer-to-Seller Messages. In *EC '15 Proceedings of the Sixteenth ACM Conference on Economics and Computation*. 81–93. https://doi.org/10.1145/2764468.2764499

Mayzlin, D., Dover, Y. and Chevalier, J. (2014). Promotio.nal Reviews: An Empirical Investigation of Online Review Manipulation. *American Economic Review*. 104(8), 2421–55. https://doi.org/10.1257/aer.104.8.2421

Mazumdar, T., Raj, S. and Sinha, I. (2005). Reference price research: review and propositions. *Journal of Marketing*, 69(4), 84–102. https://doi.org/10.1509/jmkg.2005.69.4.84

Merlo, A. and Ortalo-Magné, F. (2004). Bargaining over residential real estate: evidence from England. *Journal of Urban Economics*, 56 (2), 192–216. https://doi.org/10.1016/j.jue.2004.05.004

Merlo, A., Ortalo-Magné, F., and Rust, J. (2015), The home selling problem: theory and evidence. *International Economic Review*, 56 (2), 457–484, https://doi.org/10.1111/iere.12111

Miller, N.G. (1978). Time-on-market and selling price. *Real Estate Economics*, 6 (2), 164–174. https://doi.org/10.1111/1540-6229.00174

Morgan, P.B. (1985). Distributions of the duration and value of job search with learning. *Econometrica*, 53 (5), 1199–1232. https://doi.org/10.2307/1911018

Mortensen, D. (1982a). The Matching Process as a Non- Cooperative/Bargaining Game. *The Economics of Information and Uncertainty, University Chicago Press*, 233-58,

Mortensen, D. (1982b). Property Rights and Efficiency in Mating, Racing, and Related Games. *The American Economic Review*. 72 (5), 968-79. https://doi.org/10.2307/1812016

Mortensen, D. (2011). Markets with Search Friction and the DMP Model. *The American Economic Review*. 101(4), 1073–1091, https://doi.org/10.1257/aer.101.4.1073

Mumford, K. and Smith, P. (1999). The Hiring Function Reconsidered: on Closing the Circle. *Oxford Bulletin of Economics and Statistics*. 61 (3), 343-64. https://doi.org/10.1111/1468-0084.00133

Nash, J. G., and Rosenthal, R. A. (2014). An Investigation of the Endowment Effect in the Context of a College Housing Lottery. *Journal of Economic Psychology*, 42, 74–82. https://doi.org/10.1016/j.joep.2014.01.001

Ngai, R. and Tenreyro, S. (2014). Hot and Cold Seasons in the Housing Market. *The American Economic Review.* 104 (12), 3991-4026, https://doi.org/10.1257/aer.104.12.3991

Northcraft, G.B. and Neale, M.A. (1987). Experts, amateurs, and real estate: An anchoring-and-adjustment perspective on property pricing decisions. *Organizational Behavior and Human Decision Processes*, 39 (1), 84–97. https://doi.org/10.1016/0749-5978(87)90046-X

Novy-Marx, R. (2009). Hot and cold markets. *Real Estate Economics*, 37 (1), 1–22, https://doi.org/10.1111/j.1540-6229.2009.00232.x

Odean, T. (1998). Are Investors Reluctant to Realize Their Losses? *The Journal of Finance*, 53(5), 1775-1798. https://doi.org/10.1111/0022-1082.00072

Pallais, A. and Sands, E. (2016). Why the referential treatment? Evidence from field experiments on referrals. *Journal of Political Economy*. 124 (6), 1793-1828, https://doi.org/10.1086/688850

Paolacci, G., Chandler, J. and Ipeirotis, P. (2010). Running Experiments on Amazon Mechanical. *Judgment and Decision Making*. 5,(5), 411-419.

Petrongolo, B, and Pissarides, C. (2001). Looking into the Black Box: A Survey of the Matching Function. *Journal of Economic Literature*, 39 (2): 390-431, https://doi.org/10.1257/jel.39.2.390

Piazzesi, M., Schneider, M., and Stroebel J. (2020). Segmented Housing Search. *American Economic Review*, 110 (3): 720-59, https://doi.org/10.1257/aer.20141772

Pissarides, C. (1984). Search Intensity, Job Advertising and Efficiency. *Journal of Labor Economics*. 2 (1), 128- 143, https://doi.org/10.1086/298026

Pissarides, C. (1985). Short-Run Equilibrium Dynamics of Unemployment, Vacancies, and Real Wages. *The American Economic Review,* 75 (4), 676-690. https://doi.org/10.2307/1821347

Pissarides, C. (1986). Unemployment and Vacancies in Britain. *Economic Policy*, 1 (3), 676-690, https://doi.org/10.2307/1344583

Pissarides, C. (2000). Equilibrium Unemployment Theory. 2nd ed. Cambridge: MIT Press

Preis, T. and Moat H. (2014). Adaptive nowcasting of influenza outbreaks using Google searches. Royal Society Open Science. 1 (2), https://doi.org/10.1098/rsos.140095

Pury, C. L. S. (2011). Automation Can Lead to Confounds in Text Analysis. *Psychological Science*. 22(6), 835–836. https://doi.org/10.1177/0956797611408735

Quan, D.C. and Quigley, J.M. (1991). Price formation and the appraisal function in real estate markets. *The Journal of Real Estate Finance and Economics*, 4 (2), 127–146. https://doi.org/10.1007/BF00173120

Rae, A. and Sener, E. (2016). How website users segment a city: The geography of housing search in London. *Cities*. 52, 140-147. https://doi.org/10.1016/j.cities.2015.12.002

Roche, J.-C. and Tirole, J. (2003). Platform Competition in Two-Sided Markets. *Journal of the European Economic Association*, 1(4), 990–1029, https://doi.org/10.1162/154247603322493212

Romanyuk, G. (2017). Ignorance is strength: Improving the performance of matching markets by limiting information. Working Paper.

Rosenfield, D. and Shapiro, R. (1981). Optimal adaptive price search. *Journal of Economic Theory,* 25(1), 1-20. https://doi.org/10.1016/0022-0531(81)90014-4

Roth, A. (2002). The Economist as Engineer: Game Theory, Experimentation, and Computation as Tools for Design Economics. *Econometrica*. 70(4), 1341-1378. https://doi.org/10.1111/1468-0262.00335

Rothschild, M. (1974). Searching for the lowest price when the distribution of price is unknown. *Journal of Political Economy*, 82 (4), 689–711. https://doi.org/10.1086/260229

Rutherford, R.C., Springer, T.M. and Yavas, A. (2005). Conflicts between principals and agents: evidence from residential brokerage. *Journal of Financial Economics*. 76(3), 627-665. https://doi.org/10.1016/j.jfineco.2004.06.006

De Los Santos, B., Hortacsu,A., and Wildenbeest, M. (2012). Testing models of consumer search using data on web browsing and purchasing behavior. *American Economic Review*, 102(6), 2955-80. https://doi.org/10.1257/aer.102.6.2955

Salant, S. W. (1991). For sale by owner: When to use a broker and how to price the house. *Journal of Real Estate Finance and Economics*, 4 (2), 157-74. https://doi.org/10.1007/BF00173122

Salganik, M. (2017). Bit by Bit, Social Research in the Digital Age. Princeton: *Princeton University Press*

Schnare, A. and Kulick, R.B. (2009) Do Real Estate Agents Compete on Price? Evidence from Seven Metropolitan Areas. In Housing Markets and the Economy: Risk, Regulation, and Policy, Lincoln Institute of Land Policy, Edward L. Glaeser & John M. Quigley, eds., 2009. https://doi.org/10.2139/ssrn.1501256

Shavell, S. (1994). Acquisition and disclosure of information prior to sale. *The RAND Journal of Economics*, 25(1),20–36. https://doi.org/10.2307/2555851

Shimer, R. (2004). Search Intensity. Unpublished Paper

Srnicek, N. (2017). Platform Capitalism. Cambridge: *Polity Press*

Stein, J. (1995). Prices and trading volume in the housing market: a model with down-payment effects. *The Quarterly Journal of Economics*, 110(2), pp. 379–406. https://doi.org/10.2307/2118444

Stigler, G. J. (1961). The Economics of Information. *Journal of Political Economy*. 69(3), 213-225. https://doi.org/10.1086/258464

Strahilevitz, M. A., and Loewenstein, G. (1998). The Effect of Ownership History on the Valuation of Objects. *Journal of Consumer Research*, 25(3), 276–289. https://doi.org/10.1086/209539

Taylor, C. (1999). Time-on-the-Market as a Sign of Quality. *The Review of Economic Studies*, 66 (3), 555–578, https://doi.org/10.1111/1467-937X.00098

Tambe, P., and Hitt, L. M. (2012). The productivity of information technology investments: New evidence from IT labor data. *Information Systems Research*, 23(3), 599-617. https://doi.org/10.1287/isre.1110.0398

Tucker, C., Zhang, J., and Zhu, T. (2013). Days on market and home sales. The RAND Journal of Economics, 44 (2), 337-360. https://doi.org/10.1111/1756-2171.12022

Turnbull, G.K. and C.F. Sirmans (1993). Information, search, and house prices. *Regional Science and Urban Economics*, 23 (4), 545–557. https://doi.org/10.1016/0166-0462(93)90046-H

Tversky, A., and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157), 1124-1131. https://doi.org/10.1126/science.185.4157.1124

Ugander, J., Karrer, B., Backstrom, L. and Marlow. C. (2011). The Anatomy of the Facebook Social Graph. arXiv:1111.4503

Ursu, R. M. (2018). The power of rankings: Quantifying the effect of rankings on online consumer search and purchase decisions. *Marketing Science*, 37(4), 507-684. https://doi.org/10.1287/mksc.2017.1072

Van der Cruijsen, C., Jansen, M.J. and van Rooij, M. (2018). The Rose-Tinted Spectacles of Homeowners. *Journal of Consumer Affairs,* 52(1), 61-87. https://doi.org/10.1111/joca.12134

Van Dijk, D. et Francke, M. (2018). Internet Search Behavior, Liquidity and Prices in the Housing Market. *Real estate economics*. 46(2), 368-403. https://doi.org/10.1111/1540-6229.12187

Vosen, S. and Schmidt, T. (2011), Forecasting private consumption: survey-based indicators vs. Google trends. *Journal of Forecasting*. 30(6), 565-578. https://doi.org/10.1002/for.1213

Webster, F. (2002). Theories of the Information Society, 2nd edition. Cambridge: *Routledge*.

Wheaton, W.C. (1990). Vacancy, search, and prices in a housing market matching model. *Journal of Political Economy*, 98 (6), 1270–1292, https://doi.org/10.1086/261734

Wilhelmsson, M. (2008). Evidence of buyer bargaining power in the Stockholm residential real estate market. *Journal of Real Estate Research*, 30 (4), 475–500. https://doi.org/10.5555/rees.30.4.ahh40320465881h1

Windsor C., La Cava G., Hansen, J. (2014). Home price beliefs: Evidence from Australia. *Journal of Housing Economics*, 29, 41-58. https://doi.org/10.1016/j.jhe.2015.05.002

Wu, J., Deng, Y. (2015). Intercity Information Diffusion and Price Discovery in Housing Markets: Evidence from Google Searches. *Journal of Real Estate Finance and Economics.* 50, 289–306. https://doi.org/10.1007/s11146-014-9493-9

Yavaş, A., Yang, S. (1995). The strategic role of listing price in marketing real estate: theory and evidence. *Real Estate Economics*, 23 (3), 347–368. https://doi.org/10.1111/1540-6229.0066

Zuehlke, T. (1987). Duration Dependence in the Housing Market. *The Review of Economics and Statistics*, 69(4), 701-704. https://doi.org/10.2307/1935966