Residential mobility and housing search in the digital age: A quantitative multivariate estimate of the effects of online housing search on residential mobility in the Netherlands

Bakker, Kevin

Leiden University

Residential mobility and housing search in the digital age

A quantitative multivariate estimate of the effects of online housing search on residential mobility in the Netherlands

Public Administration - Economics and Governance

Capstone Empirical housing market research: Applications of (aggregated) big data

Student: Kevin Bakker (s2475588)
Supervisor: Dr. J. Steegmans

Universiteit Leiden

3-9-2022

Abstract

This study focuses on the effect of online housing search on residential mobility while taking into account the seriousness of the search. The analysis is done using a cross-sectional, multivariate linear regression analysis on online housing market platform search data and real residential relocations data. This study builds upon online search and residential mobility literature while using a relatively novel dataset that is based upon search on an online housing market platform. Using multivariate Ordinary Least Squares regression (OLS), it is shown that online housing search, as well as the commitment or seriousness of the search, have a significant effect on residential mobility. Nevertheless, other variables are relevant as well, in particular the number of commuters, the distance of the relocation, housing price and the relative distance of the origin or home municipality to others and utilities. The results of this study indicate that search data can contribute to policy design (specifically mobility policy), spatial planning and the understanding of the search and matching process in the housing market.

Contents

Tables and Figures

Tables

Figures

Introduction

As a result of digitalization, many platforms have online been extended to the internet. These platforms digitally enable business, political and social activities and form a so-called digital platform economy (Song, Szerb, Audretsch, Komlósi, & Acs, 2021). These platforms increase the efficiency of trade by facilitating search and matching and improving the quality of matches (Goldfarb & Catherine, 2019, pp. 9-10). Braesemann & Baum (2020) and Shaw (2020) find that property technology, or PropTech, is becoming increasingly important in the real estate industry, although not much research has been done in this field. Nevertheless, online housing market platforms can be used to generate data, such as user data, which, in turn, can be used for scientific analysis. Therefore, online housing market platforms can offer insights into local housing submarkets and how the housing market search operates spatially Rae (2015). This study aims to contribute to these insights.

While digital market platforms have surged only relatively recently and large knowledge gaps exist about their influence and underlying mechanisms. Several have aimed to close the knowledge gap by investigating the influence of: transaction costs on Airbnb (Fradkin, 2017); platform design on both consumer and seller behavior on eBay (Dinerstein, Einav, Levin, & Sundaresan, 2018) and; barriers of entry on the quality of the content on a crowdfunding platform (Pu & Gaessler, 2019). Similarly, this study focuses on online housing market platforms and residential mobility.

While residential mobility has been researched extensively, online housing market platform data is novel and still emerging and therefore few studies have been done about the link between these two. This paper incorporates quantitative analysis of a large dataset whereas other studies focused on behavioral models (Bruch & Swait, 2019; Rashidi & Mohammadian, 2015), the role of kinship, social networks and structural forces affecting supply and demand

(Coulter, Ham, & Findlay, 2016), the difference between separated men and women compared with cohabiting and married individuals (Kulu, et al., 2021) and the role of housing conditions and structural policies (Causa & Pichelmann, 2020). Furthermore, most research so far concentrates on the effects on housing prices (Zhang, Zhang, Zhang, Zheng, & Lin, 2021; Zhang, Chen, Guo, & Li, 2019) or other market aspects, such as the effect of multiple online listing services on success rates in sales and performance of real estate agents (Sing & Zou, 2021). Cooke & Shuttleworth (2017) come close by studying the role of information and communication technologies (ICTs) on residential mobility and migration. Nonetheless, this study focusses is specifically on the effect of online housing market platforms on residential mobility, leaving out other forms of ICT's. The aim of this study is to tests whether online housing markets platform search influences real residential mobility to the extent that there is a causal effect using online housing market platform search data and real residential mobility data. Therefore, the research question is:

"What is the effect of online housing search on residential mobility?".

The effect of online housing market platform search on residential mobility is studied using Ordinary Least Squares (OLS) via a multivariate linear regression using the statistical program R. OLS is applied using a dataset derived from Funda, a large Dutch housing market platform (Steegmans & De Bruin, 2021a). Only one other study has been done using these data: Steegmans & De Bruin (2021b) use it to demonstrate that the gravity framework is relevant for both recreational and serious home searchers as physical location is related to the online behavior on the platform. The Funda data are combined with actual residential moving data, available in the Statistics Netherlands national database (CBS). While online housing market platforms are one of the multiple routes through which housing search – often simultaneously – takes place, this study focusses solely on the role of one large online housing market platform, leaving out smaller platforms, real estate agents, newspaper advertisements, and other mediums.

This paper contributes to the existing scientific literature regarding the housing market and digital data, in specific to establishing the links between virtual and physical behavior by analyzing the effects of digital platform data on residential mobility. Additionally, it contributes by testing the usability and limitations of these digital data. The results of this study indicate these data prove to be useful in housing policy design such as spatial planning, since they quite accurately indicate residential mobility. Consequently, the use of these data could contribute in tackling the problems surrounding the housing market in the Netherlands, which is getting increasing attention (Kanne, 2020; Fang & Liempt, 2020; Boelhouwer, 2019; Cooper & Kurzer, 2020; Gent & Hochstenbach, 2019).

The structure of this paper is described in this paragraph. The next section contains the theoretical framework, in which related theory and literature is discussed. From the literature study a hypothesis is derived. Thereafter, the relevance of this study to the public domain is explained in the public policy context section. Next both datasets are introduced, along with descriptive statistics. Then the empirical model is provided in which the theory is translated to a model that is used for the analysis. Hereafter the results are presented. Next are some robustness checks. After this the study is concluded. In the appendix some more explanatory figures can be found related to the datasets.

Theoretical framework

The two main concepts are discussed in this section: online housing search and residential mobility. First, relevant academic literature regarding both concepts is provided and combined. Due to the novelty of online housing search via housing market platforms (hereafter:

online housing search) not much research has been done into this topic. Therefore, the main conceptualization are derived from using Rae's (2015) existing literature; supplemented by literature about housing search, search and matching and digital platforms. Moreover, a broader scope is used involving the combination of multiple concepts. Residential mobility, on the other hand, has been researched extensively and many factors of influence have been identified, requiring the opposite approach.

<u>Online housing search</u>

Han and Strange (2015) state that housing is a unique good and, as a consequence, analysis requires a different approach than regular markets. They state the process of buying and selling houses consists of search, matching and bargaining. They also find that information issues play a central role in this process. Among all they emphasize housing transactions take place under uncertainty: buyers and sellers have to search for each other and there is uncertainty in both parties about the price. Similar to Blanchard & Diamond's (1989) findings, they find the matching process entails a stochastic, time consuming process in which both the supply and demand side are continuously waiting for and searching for an appropriate match. Anglin (1997) finds that a house buyer's behavior is influenced by the provision of information, the number of neighborhoods inspected and whether the buyer thinks about buying a new house. Likewise, Han & Strange (2015) find that the type of house buyer or seller is not fully random: buyers for a certain house are filtered via pre-search, based on information from advertisement, which they define as "pre-search". Moreover, Elder, Zumpano, Baryla, & Baryla (1999) find that an increase in the length of time a buyer searches on the market sampling costs increase as well. Kohn & Shavell (1974) show that during decision-making with an uncertain distribution of samples, buyers revise their criteria based on each sample that is drawn –thus influencing the duration of search – until they find the house that provides them with a certain level of utility.

Han and Strange (2015) underline the unique importance of institutions and economic agents involved in that market, who functions as an additional actor (besides buyers and sellers) in this market: the agent or real estate broker who intermediates transactions. Real estate agents often bring buyers and sellers together and therefore act as an intermediary. With the emergence of digital platforms, real estate agents have extended their methods beyond the traditional routes such as advertisements on print media, to online listing portals. These portals, part of the emerging property technology (or PropTech), allow real estate agents to efficiently and effectively list property information (Sing & Zou, 2021). Dearsley & Baum (2020) describe PropTech as "one small part of the wider digital transformation of the property industry. It describes a movement driving a mentality change within the real estate industry and its consumers regarding technology-driven innovation in data assembly, transactions, and the design of buildings and cities" (Baum and Dearsley, 2019, reported in Davenport)." In brief, Baum (2017) states PropTech consists of the combination of information provision, transactions and management and control regarding real estate in the digital context. Moreover, PropTech involves technology-based companies in the real estate sector and encompass more than multiple listing systems and market advertising websites (Facchinetti, 2021; Shaw, 2020). Braesemann & Baum (2020) and Shaw (2020) find that PropTech – or the term that Shaw prefers: "digital real estate technology" (p. 1046-1047) – is becoming increasingly important in the real estate industry.

Shaw perceives these digital real estate platforms as performative market devices that intervene in the social construction of markets. More specific, these platforms' business is to facilitate multiple parties to interact on its digital infrastructure, enabling value-creating interactions among them. He considers these platforms as theoretical and practical transformations of urban real estate markets. Braesemann & Baum prove PropTech is turning real estate into a data-driven market. Shaw shows that the role of data has become important, since users generate data on the platform and user data itself can offer previously un-seeable insights into market processes and actors. Moreover, similar to Goldfarb & Catherine (2019), Shaw argues these platforms allow for a new perspective of the market for the users and could therefore significantly influence the markets' mechanisms and the whole market itself. While Braesemann & Baum prove PropTech fuels the globalization of real estate and the involvement of large organizations, this paper is delimited to the Dutch national real estate market among residents.

Rae's (2015) framework for understanding housing search in the digital age proves very relevant for this study. In this framework he focuses on search activity of potential house buyers, making some revisions to earlier literature about housing search, due to the significant change the internet has had on this. He builds upon Maclennan (1980) about staged search and Marsh & Gibb (2011) their six stage search process while taking into account the additional digital information channels that have emerged. Rae focuses on extensive search, which is the first stage in the housing search process. He argues that the previously regarded consecutive components in the extensive housing search phase – search strategy, area orientation, establish vacancies – now occur randomly due to the online dimension. Moreover, he argues that the digital component in the housing market made the housing search process qualitatively and quantitatively different than a decade ago, and addresses the need to revise the conceptual foundations of housing market search. Thus, he finds housing search to have radically changed and regards the viewpoints such as those of Wheaton (1990) outdated.

One aspect is that, traditionally, house buyers have a limited amount of time and resources available for the search of a house. Rae finds search to be a complex process on information gathering, which has now been shifted primarily to online platforms. These platforms provide digital services for users and their business model consists of the interaction of these users within their platform (Braesemann & Baum, 2020). Consequently, Rae argues that the internet lowered search costs (the costs of looking for information) by enabling the possibility for searching online. Moreover, Song, Szerb, Audretsch, Komlósi, & Acs (2021) find that digitization of markets into online platforms lowered five types of costs that affect economic activities: search, replication, transportation, tracking, and verification. They also find that platforms facilitate matching and increase the efficiency of trade by serving as intermediaries. Additionally, Han & Strange (2015) add that the internet enabled the provision of more rich information and agree it increased the efficiency of housing search. Goldfarb & Catherine (2019) also find that the digitization of markets lower search costs and add it also increases the quality of matches between buyers and sellers by serving as intermediaries and enabling exchange between parties.

Rae (2015) also identifies the emergence of recreational search as a result of the internet, in which the search is less serious and extensive than the search activity of market entrants and motivated by other reasons, such as curiosity. Consequently, he warns that housing search data should be approached carefully since this is partly built up from search out of curiosity and therefore could dilute serious search data. Steegmans & De Bruin (2021a) take this into account and find recreational search to be more centered around the searcher's municipality and neighbor municipality while serious searchers are have a relative preference for other municipalities within the province border. However, they find no differences between the distance effects between both types of searchers. Possibly because recreational search could also influence households' preference and therefore perhaps not differ that much from serious search, like Rae considers. Even though, Rae finds that there is often noise in search data itself. The new online dimension makes it difficult to determine when search has ended, in contract to traditional theories, such as Kohn & Shavell (1974) who argue that, in general, search will end when a sample is found that exceeds a certain point of utility (the "switchpoint level"). Nevertheless, Rae thinks online housing search portals could significantly facilitate evaluating spatially dispersed housing vacancies and reduce psychical costs of housing search.

In order to cope with the challenge of distinguishing empirically meaningful housing search from online search data, Rae (2015) formulates five types of online housing search. First, there is recreational search, which is done by those that are not looking to buy a house, but are just looking out of curiosity. Often, they look for expensive houses and share these on social media. Second, inquisitive search, consists of those who are not looking to buy a house, but are interested in what is available on the market. While not directly linked to housing transactions, these data can provide potential buyers with a greater understanding of the market. Third, aspirational search, consists of those who are either looking to buy their first house or move to another one. This type of searchers use the online listing markets to gather information and get an idea what their position on the market is and what they have to do to reach a certain other position. Fourth, active search consists of those who are actively searching because they want to move and have the financial means to do so. From the perspective traditional search theory, these searchers would be the only group that would be considered in housing search. Often, these searchers use online listing portals at the start of extensive search for optimizing their search. Lastly, he mentions professional search, where property professionals search for multiple motives, such as comparing their listings to competitors, in order to compete in the market. Considering these types of search he stresses the importance of filter online search data in order to keep any analysis of search representative, since serious searchers have the intention of buying a house while recreational searchers don't. Though, he notes that recreational search could have an important role in the long-term aspiration formation (p. 460 – 461).

Residential mobility

Residential mobility, also known as intraurban migration, is defined in this study as close distance residential relocations within urban areas, that not significantly impact people's daily activities. In contrast to residential mobility, migration generally involves unsettling relocations over longer distances (Short, 1978; Niedomysl, 2011). Nevertheless, there is an overlap between residential mobility and migration in terms of distance and impact on daily life activities where the distinction is difficult to assess. In other words, there is an obscurity about the demarcation where residential mobility ends and migration starts (Coulter, Ham, & Findlay, 2016). Considering this, residential mobility is approached as relocations within the Netherlands as a whole, since this country is relatively small compared to others and residential relocations is therefore relatively close distance and considerably non-impacting on daily life activities. Similar to traditional literature, it is also assumed that space is homogeneous, so that points of interest are equally likely to be located anywhere (Gibbons, Overman, & Patacchini, 2015).

There is a lot of discussion about measuring residential mobility. Rashidi, Auld, & Mohammadian (2012) discuss three methods: including all plausible alternatives, randomly selecting a finite number of alternatives from the universal choice set and combining disaggregate alternatives into more aggregated sets which results in choice set size reduction. The first option, they argue, can be unrealistic as it assumes perfect knowledge about all alternatives and could cause unrealizable results. The second can result in biased and possibly inaccurate parameter estimation. The third option they find to be satisfactorily studied. Schirmer, Axhausen, & Eggermond (2014) consider the discrete choice framework to play an important role in residential location choice. Here, an actor has to choose one option from a set of mutually exclusive alternatives, each providing a certain level of utility and it is presumed the option with the most utility is chosen. They review the literature on locational attributes in the choice model and classify location variables. They distinguish four categories. First, there is the build environment, referring to how the area is constructed with regards to infrastructure. Specifically, these consists of the volumes of buildings, parcels, blocks, and connecting networks (for example both road and public transport networks). They then define socioeconomic environment as the social aspects of the area, which includes the population size, income level, ethnic distribution, age, and education level. Thereafter they look at points of interest, consisting of locations that provide utilities (such as the city center, hospitals and supermarkets). Lastly, they construct access and accessibility, which refers to the level of difficulty in travelling, for example to work and friends and family.

In general, Schirmer, Axhausen, & Eggermond (2014) find that population density, housing costs, crime rate and local taxes have a negative relationship to residential mobility. Moreover, they find a positive relationship between previous location, unit size, employment density, employment opportunities, access to work, commuting time, distance to work, distance to nearby supermarket and previous location. Likewise, Rashidi, Auld, & Mohammadian (2012) find that the household's current average work distance plays a significant role in the household's moving decision. In addition, Baraklianos, Bouzouina, Bonnel, & Aissaoui (2020) find accessibility measures to positively influence residential location choice. These measures consist of distance to schools, highways and public transport. Cooke & Shuttleworth (2017) find that information and communication technologies (ICTs) reduce residential mobility, especially on short distance moves (below 20 km). They hypothesize that ICTs allow people to arrange their affairs such as employment, education and contact from a distance, thus lowering the need for moving.

Section summary

In summary, housing transactions take place under uncertainty in both the matching process and the price. Thus, search – which includes the provision of information, the number of neighborhoods inspected and the level of willingness of the buyer for buying a new house – plays an important role. Nevertheless, real estate agents have an arguably important role in the search process as well, considering they often act as an intermediary third party during transactions, bringing buyers and sellers together. This intermediary aspect distinguishes the housing market from typical markets. While more search is likely to lead to better matches, it isn't possible to search continuously due to the restrains that influence search costs. The introduction of the internet and the subsequent rise it gave to PropTech or digital real estate technology – and, specifically, online listing portals –, have lowered search costs and improved matches significantly. Online search data is likely to constitute actual residential relocations behavior, since those who are looking at houses are likely to be interested in buying one. However, due to the low search costs of online housing search recreational housing search emerged, which could dilute research into serious search and should therefore be approached carefully.

Hypotheses

Considering the framework above, two hypotheses are derived. The first one takes into account the search and matching process. In search, uncertainty, utility and the provision of information play an important role, while the matching process is stochastic and time consuming. Online listing portals are supposed to have increased the efficiency of search and effectivity of matching. Therefore the first hypothesis is as follows:

H1: *Online housing search intensity influences residential mobility.*

Next, Rae (2015) underlines the importance of the type of searcher and warns that recreational search could hinder research into serious search. On the other hand, Han & Strange (2015) find that the type of house buyer or seller is not completely random since they are filtered through pre-search: those who are searching could be considered potential buyers since they already went through a filtering process or an incentive that motivated them to be searching, enabling them to inherently be considered as serious. Considering the potential effects of the determination of the search for finding a match and the methods for differentiating between the seriousness of searchers, the second hypothesis is as follows:

H2: The level of seriousness of the searcher influences residential mobility.

Public policy context

The Dutch housing market is been receiving increasing attention due to problems that surround it. In their survey, Kanne & Engeland (2020) find that nearly all Dutch citizens perceive problems with rising housing and rent prices all over the country have resulted into a shortage of affordable housing. Similarly, Fang & Liempt (2020) observe that international students experience troubles as well with regards to the Dutch housing situation. They find many consider this to be unfavorable and to be the result of discrimination, financial issues and bureaucracy. Additionally, they have the feeling not to be taken seriously about the housing situation. As a result, they find these aspects to cause poor mental health among them.

Gent & Hochstenbach (2019) look at the developments in the Dutch housing system from the post-war era until the present day. Aside from this being an inconsistent and partial process, they, like Cooper & Kurzer (2020), mainly attribute the developments in the housing market to neo-liberal rules and regulation which, in turn, created social and spatial implications that constrained the social renting sector. Additionally, the construction and availability of (affordable) houses was frustrated by new taxes, a landlord levy and the 2015 Housing Act. Munuel, Hochstenbach, Bosma, & Fernandez (2020) state that private landlordism, which they conclude is the outcome of both the promotion of homeownership, fueling the private rental sector, and that the increasing economic, monetary and regulatory efforts of governments to perceive the housing as an asset class. They, similar to Cooper & Kurzer (2020) argue that the the private rental sector grew due to the conflicting push for financialized homeownership, where they define financialization as “the increasing dominance of financial actors, markets, practices, measurements, and narratives at various scales, resulting in a structural transformation of economies, firms (including financial institutions), states, and households” (p. 4). They also warn that the conflicting push for financialized homeownership via mortgage debt to fund housing is coming near its end, with prices surpassing income for a concernable long time, raising the need for a new approach**.** Therefore, they identify new carriers of financialization as a result of these inflated prices. In the same way as Boelhouwer (2019), they find the global financial crisis, as well as the subsequent government policies to have aggravated the situation. He also finds the situation to increase social inequalities and argues that the latest policy that has been imposed do not solve the problems and may even be counterproductive.

This study contributes to the understanding of the housing market and online housing search. Housing market platform data enable the discovery of housing market trends even before the actual residential relocations has taken place and therefore could be valuable for policymakers, by providing guidance in land-use and spatial planning. Additionally, these data can complement supply and demand data which could help determine where to build and how to design policy. Lastly, adding the theoretical framework, this study contributes in understanding online housing search and the consequent matching process, thus contributing to the understanding of this process and providing added value for further increasing this process.

Data

Multiple data sets are used to answer the research question. First, the 'Online housing search dataset: Information flows of real estate platform users', which dataset consists of user-generated data from the platform Funda (Steegmans & De Bruin, 2021a). Funda is the largest housing market platform in the Netherlands (Obbink, 2020). Second, data from Statistics Netherlands (hereafter the Dutch abbreviation is used: CBS), which provides various statistical data about the Dutch economy, demography and urbanity. The CBS publishes data about the actual residential relocations as well as descriptive data that is used as control variables.

CBS dataset

From CBS multiple data is used. First and foremost, the dataset 'Verhuisbewegingen' or relocations is used to measure the dependent variable residential mobility. The residential mobility consists of all residential relocations between owner-occupied housing from 77 municipalities to 343 other municipalities in the year 2019. The relocations, as acquired from the source, are aggregated into units of five (1 is rounded to 0 and 4 to 5, et cetera). The total number of relocations between municipalities in 2019 the dataset is 172 640. The reason that there is only data about these 77 municipalities is because they applied for participation of data collection while other municipalities did not (CBS, 2020).[1] Moreover, control variables are

[1] The data originates from 154 datasets: 77 of movements within municipalities and 77 of movements between (the same as the other) municipalities. The movements within the municipalities had to be aggregated to

added that correspond with the concepts mentioned in the theoretical framework and are available through the CBS's database. First, the dataset 'Nabijheid voorzieningen; afstand locatie, regionale cijfers' (translated to 'The average distance to utilities for households in each municipality in 2019') is used (CBS, 2021a).[2] Continuingly, descriptive data on municipalities regarding population density, the average housing price, the average housing stock, and the average living space per house is used, available in the dataset 'Voorraad woningen; gemiddeld oppervlak; woningtype, bouwjaarklasse,regio' (CBS, 2021b). Next, the public disturbance and crime datasets, known as 'Geregistreeerde overlast; soort overlast, gemeentelijke indeling 2022' and 'Geregistreeerde misdrijven en aangiften; soort misdrijf, gemeente 2022' from the Dutch National Police are used (Dutch National Police, 2022a; Dutch National Police, 2022b). Lastly, 'Banen van werknemers naar woon- en werkregio' or the number of commuters between home municipality and work municipality is used (CBS, 2022).

Funda dataset

Funda is a real estate platform that connects both the supply and the demand side of the housing market. It facilitates transactions between house buyers, sellers, renters, landlords and real estate agents (Funda, 2022). The Funda dataset consists of user-generated data from the housing market platform Funda which is generated between January 1, 2018 to June 30, 2018 (Steegmans & De Bruin, 2021a). These data were generated using Google Analytics, a web analytics tool that registers website information and activity data from users (Google Analytics, 2022). Individual user's website interaction information on the Funda platform was sent to Google Analytics. These data regards website activity, registered as the number of mouse clicks

municipal level and merged with the datasets with the movements between the municipalities. See the appendix for the 77 participating origin municipalities.

[2] Unfortunately, the CBS distance data contains average distances and not median distances. While median distances are more reliable, this data is not available.

or 'hits' that users generate on the platform. Hits thus involve registered user's website activity on the Funda platform. Hit data further contains the municipality of the property that is viewed by an individual, location information of the user's municipality and details that can help indicate their level of commitment of finding a house (such as whether the user viewed the real estate agent's phone number). This information is stored on an aggregate level and doesn't include timestamps. Of all hits, approximately 95 percent of the hits are related to owner-occupied housing, leaving five percent for rental objects (Steegmans & De Bruin, 2021a).

Crucial in this research is the definition of search in the context of the internet, where search is low-costs and therefore high-volume. Similar to Steegmans and De Bruin (2021b) and as stressed upon by Rae (2015), the seriousness of the searchers is taken into account. This is possible because the Funda data comes with the subsets that differentiate between different user groups based on their activities on the platform. Considering Han and Strange's (2015) notion that real estate agents have an important role in the housing market, related subsets are used to determine the seriousness of the searchers. Besides the full sample, subsets are used that contain solely the hits that were generated exclusively by either the buyer of the property, users that messaged the real estate agent, users that viewed their phone number and users that scheduled a viewing. They offer adequate proxies for different degrees of serious search. Table 1 provides more in-depth information about these subsets.

Table 1: Funda dataset and subsets

Dataset	Description
Flows_2018H1	Full sample
Flows_2018H1_telephone_true	Subsample of users who viewed a real estate agent's telephone number
Flows_2018H1_message_true	Subsample of users who digitally contacted the real estate agent
Flows_2018H1_viewing_true	Subsample of users who scheduled a listing with the online tool
Flows_2018H1_buyer_true	Subsample of users who registered themselves as the buyer of a property
Source: Steegmans & De Bruin, Online housing search dataset: Information flows of real estate platform users, 2021a	

The Funda data, like the CBS data, are aggregated on municipal level and contains the absolute amount of hits between all municipality combinations. Additionally, it contains variables such as the distance between municipalities and dummy variables – taking either the value 0 or 1 to indicate whether something is present or not – that indicate whether each unique municipality combination is within the same municipality (or not), a neighboring one or within the same province. Table 2 contrasts the number of hits recorded in each dataset in millions. The first row contains the full sample with 100 percent of the data. Thereafter is a subset that contains only the amount of hits generated by those who indicated to have bought the property, which is the least of all subsets in the table. Next is the subset that contains the hits of those who have message the real estate agent, which contains nearly as much hits as those who applied for a viewing (8.29 versus 8.80 percent). The largest subset consists of those who have viewed the number of the real estate agent (21.72 percent).

Table 2: the total number of hits generated on the Funda platform

Number of hits on the Funda platform (in millions)					
Datasets	***Variables***	***Yes***	***%***	***No***	***Total***
Flows_2018H1	All hits	418m	100 %	-	418m
Flows_2018H1_buyer_true	Confirmed purchase?	1.7m	0.42 %	416.2m	418m
Flows_2018H1_message_true	Messaged agent?	34.7m	8.29 %	383.3m	418m
Flows_2018H1_telephone_true	Viewed number?	90.8m	21.72 %	327.2m	418m
Flows_2018H1_viewing_true	Scheduled viewing	36.8m	8.80 %	381.2m	418m
Source: Steegmans & De Bruin, Online housing search dataset: Information flows of real estate platform users, 2021a					

As mentioned earlier, the subsets are used to determine the seriousness of the searchers. As can be observed from the table, the full dataset contains far more hits than the subsets. The subsets consist of hits that were generated by users that have taken actions that indicate a certain level of seriousness. The total amount of hits seems to decrease as the level of seriousness increases. For instance, the buyer subset, assumed the most serious one, has less hits than the subset of those who scheduled a viewing. Moreover, the subset of those who have digitally messaged the real estate agent is larger than those who solely viewed – and possibly didn't even call – the real estate agent's number. This is used for the second hypothesis, which is about whether this pattern reflects the level of seriousness.

Combining both datasets

While the available residential dataset contains the data of 77 municipalities, the Funda dataset, on the other hand, covers data of all municipalities. To enable both datasets for analysis, they are merged based on corresponding municipalities, resulting into one dataset with unique municipality combinations. In the attachments are multiple tables that provide further details about both datasets. Table 14 contains a list of the total number of home or origin municipalities in the CBS dataset. Table 15 shows which municipalities are only in the CBS dataset and not in the Funda dataset. Lastly, table 16 indicates which municipalities are in the Funda dataset but not in the CBS dataset. Figure 4 and 5 in the attachments provide further illustration of the control variables. Figure 4 provides the distance to certain utilities and figure 5 details the absolute number variables, such as the population density and the housing stock. Since some variables are related to the origin municipalities, data on these variables is available on 77 municipalities, which explains the difference in missing values in the figures (the grey area's). By analyzing only 77 of the total 355 Dutch municipalities sampling bias is likely to arise. To decrease this bias, the population of all participating municipalities is added in the dataset. Nevertheless, this bias can't be completely eradicated.

Figure 1 below contains a map of the total number of relocations. 'DeparturesCBS' (left image) indicates the municipality from which the people leave, their home or origin municipality. The municipality to which they relocate is denoted as 'ArrivalsCBS' (right image).[3] As can be derived from the image, a lot of relocations take place between three large cities: Rotterdam (lower dark brown on the left), The Hague (upper dark brown on the left) and Utrecht (orange in the middle). The grey areas indicate the missing municipalities. The missing municipalities that occur in both figures are missing due to municipal relocations, such as the merge of multiple municipalities into one or the dissolvement of a town into a larger municipality. The other missing municipalities in the left image are missing due to the absence of data about these municipalities. The dataset does not cover the relocations towards all of the other municipalities: only to 339 of 355. This is because the CBS dataset had to be merged with the Funda dataset, which is based on 2017 municipality border data (consisting of 388 municipalities) while the CBS dataset is based on 2019 data, consisting of 355 municipalities. The appendix contains two tables which indicate which municipalities are missing in the Funda dataset (table 15) and which are missing in the CBS dataset (table 16).

[3] These, and the following maps and municipal figures in this document, have been generated using shape files available on Statistics Netherlands (CBS, 2021c). They have been constructed using the 'sf' and 'tmap' packages in R, using the aforementioned CBS, Funda and Dutch National Police data (CRAN, 2022; CRAN, 2021).

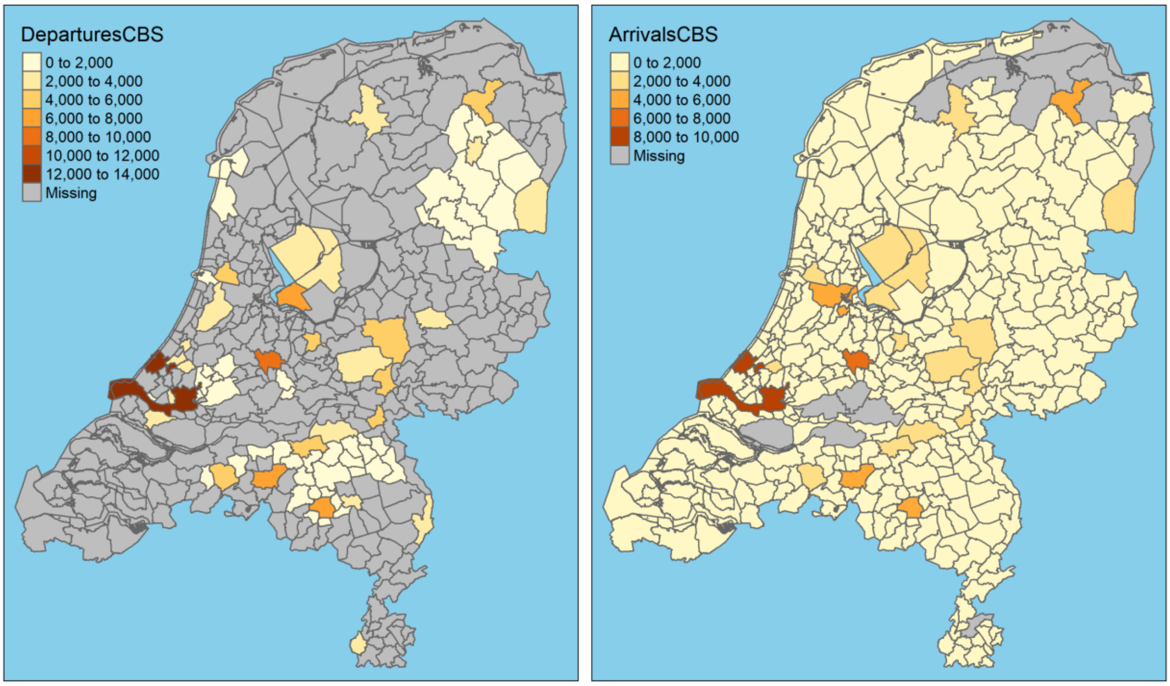


Figure 1: The total number of registered hits between municipalities in the CBS dataset

Figure 2 is a map with the total number of hits in the Funda dataset, similar to the CBS figure above. 'FundaHitsHome' (left image) indicates the municipality from which the people generate the hits, in other words their home or origin municipality. The municipality in which they search, or generate the hits, is denoted as 'FundaHitsGoal' (right image). Again, the missing municipalities that occur in both figures are missing due to municipality relocations. The other missing municipalities in the left image are missing due to the absence of data about these municipalities. The number of hits are distributed among the municipalities and show considerable differences. Like the CBS data, a lot of relocations take place between two large cities: Rotterdam (lower dark brown) and The Hague (upper dark brown). However, in contrast to CBS' relocations, in the Funda hit data Amsterdam[4] (orange in the middle on the left side) seems to be more popular than Utrecht.

[4] Missing as origin or home municipality due to missing data in the CBS dataset.

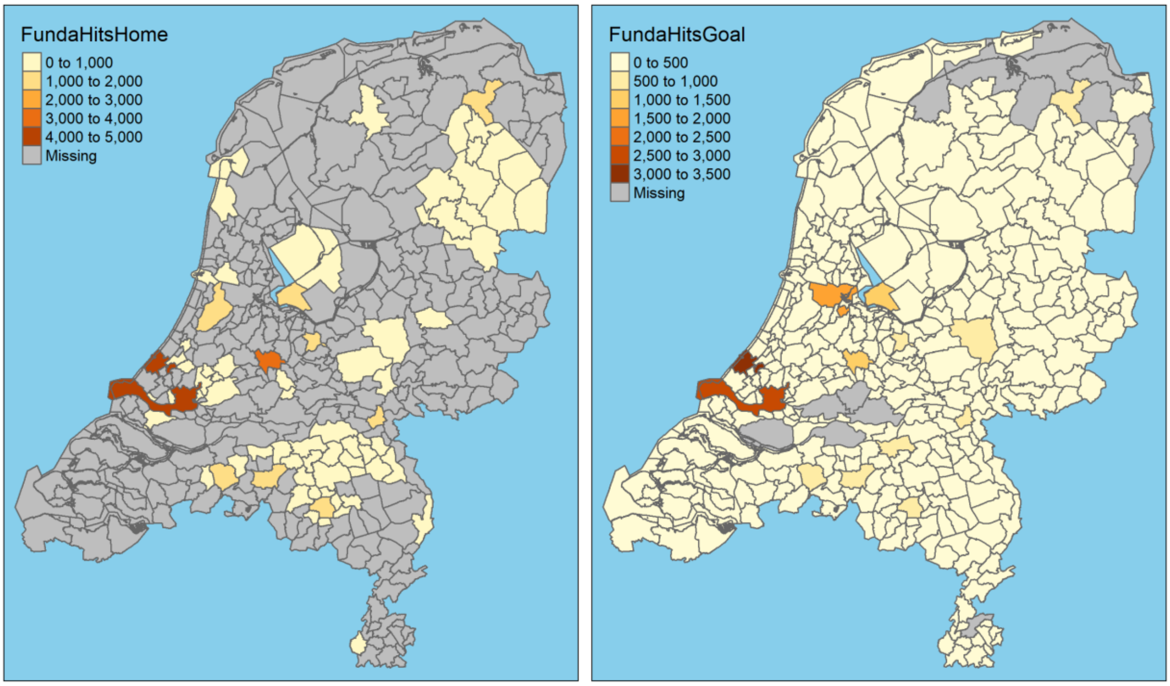


Figure 2:The total number of registered hits between municipalities on the Funda platform

<u>Descriptive statistics</u>

The dataset consists of 26 441 combinations between municipalities with the actual relocations and digital hits between them. Table 3 contains a description of all variables used in this study, including the dependent variable. The variables related to distance have been put in logarithmic format, since it is likely that the distance is highly skewed. As Steegmans & De Bruin (2021b), who use nearly similar data, found: most relocations involve shorter distances. Moreover, it is arguable that one would perceive every increase in moving distance as more drastic until the distance is such large that one would become increasingly indifferent. The same goes for the variable housing price. Next to these variables all other variables remain relatively similar to how they were imported. However, the Funda data has been rescaled in units of 10 000 to increase the interpretation of these variables, since it contains millions of observations. The commuter data was already rescaled in units of 1 000 when imported and has been kept in this format. The dummy variables for the location of the target or goal municipality is in the table as well. Furthermore, the other control variables – housing stock, population density, registered crimes, registered disturbance, population home and population goal – are expressed in the (unedited) absolute values as imported. Lastly, at the bottom of the table is the dependent variable, residential relocations, which is denoted as Y2019, which is also used as imported (in absolute values, rounded by five).

Table 3: Description of all the variables used in this study

Variables	Description
LogDistance	The logarithm of the Euclidean distance between municipality centroids of the origin municipality and goal municipality
WithinMunicipality	A dummy variable indicating 1 when the relocations take place within the same municipality and 0 if otherwise
NeighborMunicipality	A dummy variable indicating 1 when the relocations take place between a neighboring municipality and 0 if otherwise
WithinProvince	A dummy variable indicating 1 when the relocations take place between a municipality in the same province and 0 if otherwise
AllFundaHits (x 10 000)	All the mouse clicks between municipalities generated by the Funda platform users, multiply by 10 000 to get the total amount (the main independent variable of interest).
IsBuyer (x 10 000)	The mouse clicks between municipalities between municipalities generated by the Funda platform users that indicated to have bought the platform, multiply by 10 000 to get the total amount (the main independent variable of interest).
SendMessageToAgent (x 10 000)	The mouse clicks between municipalities generated by the Funda platform users that messaged a real estate agent, multiply by 10 000 to get the total amount (the main independent variable of interest).
ViewedPhoneNumberAgent (x 10 000)	The mouse clicks between municipalities generated by the Funda platform users that viewed a real estate agent's phone number, multiply by 10 000 to get the total amount (the main independent variable of interest).
ScheduledViewing (x 10 000)	The mouse clicks between municipalities generated by the Funda platform users that scheduled a viewing via the online tool in Funda, multiply by 10 000 to get the total amount (the main independent variable of interest).
LogDistanceToSupermarket	The logarithm of the average distance to a supermarket in the goal municipality
LogDistanceToHighway	The logarithm of the average distance to a highway in the goal municipality
LogDistanceToTrain	The logarithm of the average distance to a train in the goal municipality
LogDistanceToElementarySchool	The logarithm of the average distance to an elementary in the goal municipality
LogAvHousingPrice	The logarithm of the average housing price in the goal municipality
HousingStock	The available supply of housing in the goal municipality
PopulationDensity	The population density in the home municipality
NumberOfCommuters (x 1 000)	The amount of commuters between municipalities of which the flows are sorted in the same manner as the mouse clicks on Funda and the actual relocations on CBS, multiple by 1 000 to get the total amount
RegisteredCrimes	The number of registered crimes in the home municipality
RegisteredPublicDisturbance	The number of registered public disturbance in the home municipality
PopulationHome	The total population count in the origin municipality
PopulationGoal	The total population count in the goal municipality
Y2019	The total number of relocations in 2019, which is the main dependent variable of this study

Table 4 provides more statistical information about all the variables. It provides the minimum value, maximum value, mean, median standard deviation (SD) and total of each variable. For readability purposes, the variables have been put in their absolute values. Thus, the logarithmic format is not applied in this table yet but is in the analysis. Smaller values have been rounded to two decimals and larger values have been rounded to zero decimals. Regarding the interpretation of these values, it has to be noted that the variable “Within Municipality”, “Neighbor Municipality” and “Within Province” are dummy variables that take either the value 0 or 1, which explains the low values in the table. The Funda and commuter data has been decreased to 10 000 and 1 000 for readability. Moreover, there is a difference in amount of generated hits per Funda dataset. For example, the set containing all hits has a total of 416 529 253 while the subset containing only the buyers has a total of 1 737 332 hits, which is significantly less: only 0.42 percent of the full sample (as seen in table 2). Moreover, it is worth noting that the actual relocations from CBS (Y2018 and Y2019) have a median value of zero, meaning most municipality combinations don’t have relation traffic. This is because there are a lot of combinations between municipalities that did not have any residential traffic. It is interesting that the median values of all Funda sets have low medians as well, though higher than zero, which could be because of the low search costs of online housing search. Furthermore, the dependent variable Y2019 shows a similar distribution to the Funda data compared to the median. With regards to the distance to utilities, the distance tot the supermarkets and (elementary) schools are very low and the distances to the highway and the train are higher. Moreover, the housing prices remain in the same range, seeing all values are relatively close. The housing stock, like the population density, on the other hand, varies hugely, with a minimum of 564 and a maximum of 441 490. Both also have high standard deviations. Like one would intuitively predict, registered crimes and public disturbance show a similar distribution of numbers. The last two rows display the population statistics. The differences in amounts can be explained using the distribution of municipalities, which is uneven, as explained in the previous section.

Table 4: descriptive statistics about all the variables in this study.
Mind that some values are rounded to one while some do contain digits after de the decimal point

Variables	Min	Max	Mean	Median	SD	Total
Distance	0.00	296	98	92	53	2 605 857
WithinMun	0.00	1	0.00	0.00	0.05	77
NeighborMun	0.00	1	0.02	0.00	0.13	449
WithinProvince	0.00	1	0.13	0.00	0.33	3 303
All Funda Hits (x 10 000)	0.00	2 256	1.58	0.21	22	41 796
IsBuyer (x 10 000)	0.00	7.48	0.01	0.00	0.08	174
SendMessage (x 10 000)	0.00	266	0.13	0.01	2.27	3 465
ViewedPhone (x 10 000)	0.00	640	0.34	0.04	5.49	9 076
ScheduledViewing (x 10 000)	0.00	237	0.14	0.01	2.30	3 676
Y2019	0.00	6 685	6.54	0.00	90	172 640
DistanceSupermarket	0.50	2.50	1.04	0.90	0.37	27 551
DistanceHighway	0.40	38	1.89	1.50	2.96	49 850
DistanceTrain	1	49	6.90	4.40	7.14	182 359
DistanceSchool	0.50	1.60	0.77	0.70	0.19	20 413
AverageHousingPrice	142 173	902 214	291 213	279 230	84 064	99 886 345
HousingStock	564	441 490	22 000	12 905	35 890	7 510 902
PopulationDensity	69	6 459	1 258	763	1 373	96 053
Commuters (x 1 000)	0.00	153	0.12	0.00	1.79	3 178
Crimes	189	49 105	4 695	1 726	7 602	361 549
Disturbance	79	18 033	2 220	814	3 321	170 939
PopulationHome	10 502	638 712	88 681	55 147	103 383	6 828 465
PopulationGoal	932	854 047	48 156	29 445	72 597	16 517 526

In table 5 is a correlation matrix with the main variables of interest: the Funda dataset and subsets and the actual relocations from CBS. As can be derived from the table, all variables show a high and significant correlation with each other. The highest correlation with the dependent variable is with the Funda buyer subset, which seems logical, since this subset consists of users that have stated to have bought the property, whereby it should resemble the actual residential mobility the closest, in theory. The second highest correlation with the dependent variable is with the whole Funda dataset, which contains a lot of recreational search data. While this seems counterintuitive, a possible explanation could be the size of this dataset: it has far more observations than the others, which could result in a higher correlation. The lowest correlation with the dependent variable is the Funda subset consisting of the users that have messaged the real estate agent. The highest correlation within the Funda data itself is between the subset consisting of those who have messaged the real estate agent and those who have viewed the phone number of the agent. A likely reason for this is that users that wanted to contact the real estate agent, considered both options.

Table 5: Correlation between the dependent variable and main independent variables

Correlation between main independent variables and dependent variable						
	All Funda	***Is Buyer***	***Send Message***	***Viewed Phone***	***Scheduled Viewing***	***Y2019***
All Funda						
Is Buyer	0.956***					
Send Message To Agent	0.978***	0.916***				
Viewed Phone Number Agent	0.988***	0.930***	0.997***			
Scheduled Viewing	0.989***	0.945***	0.989***	0.991***		
Y2019 (dependent variable)	0.899***	0.905***	0.811***	0.835***	0.854***	
Computed correlation used pearson-method. P-value: *** p<0.01, ** p<0.05, * p<0.1.						

While table 5 shows the correlation between the main independent variables and the dependent variable, the correlation between the control variables and independent variable can be seen in table 6. A few interesting observations can be concluded from this table. First, most outcomes are statistically significant at the highest level. Exceptions are the registered disturbance, registered crimes, the population of the home municipality, the three dummy variables that indicate the relative location of the municipalities (within province to a lesser degree), the housing price and the population density. These are significant to a lesser degree and/or not often significant.

Second, among the independent variables there is a strong correlation (>0.7) between the distance to a school and distance to a supermarket (0.73), the population goal and housing stock and (1.00) registered crimes and both registered disturbance (0.98) and population (0.99) and, lastly, registered disturbance and the population in the home municipality (0.97). A moderate correlation (0.5-0.7) is found between distance and dummy within province (-0.56), commuters and the dummy within municipality (0.55) and population density and both registered crimes (0.56), registered disturbance (0.59) and population home (0.57).

Finally, a weak correlation (0.3-0.5) is found between distance and both the dummies within municipality (-0.34) and neighboring municipality (-0.38), the dummies within province and neighboring municipality (0.30), the distance to the supermarket and both distance to train (0.40), housing stock (-0.34) and population goal (-0.34), and, lastly, distance to school and both distance to train (0.36), housing stock (-0.31) and population goal (-0.32). Moreover, the dummy within municipality has a moderate correlation with the independent variable. Furthermore, as found in the theoretical framework, the number of commuters could prove to be a very relevant variable, showing a high correlation of 0.94 with the dependent variable.

Due to the high correlation between the control variables, number of commuters, and relocations in 2019, the dependent variable, the correlation between the number of commuters, the Funda datasets and the dependent variable are further explored in table 7. All correlations are found to be statistically significant at the highest level. The correlation between the commuters and the dependent variable is 0.937, which is higher than all correlations between the Funda data and the number of relocations. Furthermore, both independent variables have a high correlation with each other. In the robustness check section this is addressed again.

Table 6: Correlation between independent control variables

Correlation between main control variables and dependent variable

	Log Distance	*Within Municipality*	*Neighbor Municipality*	*Within Province*	*Y2019 (dependent variable)*	*Log Distance Supermarket*	*Log Distance Highway*	*Log Distance Train*	*Log Distance School*	*Log Housing Price*	*Housing Stock*	*Population Density*	*Commuters*	*Registered Crimes*	*Registered Disturbance*	*Population Home*	*Population Goal*
LogDistance																	
WithinMun	-0.34***																
NeighborMun	-0.38***	-0.01															
WithinProv	-0.56***	-0.02***	0.30***														
Y2019 (dep. variable)	-0.28***	0.62***	0.12***	0.04***													
LogDistanceSupermarket	0.08***	-0.01	0.01	-0.01	-0.05***												
LogDistance Highway	0.04***	0.01*	0.00	0.06***	0.03***	-0.18***											
LogDistanceTrain	0.05***	-0.01	0.01	0.06***	-0.03***	0.40***	0.21***										
LogDistanceSchool	0.11***	-0.01	0.01	-0.03***	-0.05***	0.73***	-0.09***	0.36***									
LogHousingPrice	-0.19***	-0.01	0.02***	0.07***	-0.00	-0.12***	-0.02**	-0.01	-0.11***								
Housing Stock	-0.02***	0.03***	0.02**	-0.01	0.13***	-0.34***	0.11***	-0.21***	-0.31***	0.00							
Population Density	-0.07***	-0.00	0.00	0.02***	0.05***	-0.00	0.00	0.00	-0.00	0.00	-0.00						
Commuters	-0.26***	0.55***	0.10***	0.05***	0.94***	-0.06***	0.03***	-0.03***	-0.05***	0.00	0.16***	0.04***					
RegisteredCrimes	-0.02***	0.00	0.03***	-0.00	0.08***	0.00	-0.00	-0.00	0.00	-0.00	-0.00	0.56***	0.08***				
Registered Disturbance	-0.00	-0.00	0.02***	-0.02**	0.08***	-0.00	-0.00	0.00	0.00	0.00	-0.00	0.59***	0.08***	0.98***			
PopulationHome	-0.02***	0.00	0.03***	-0.01	0.08***	0.00	-0.00	-0.00	0.00	0.00	-0.00	0.57***	0.08***	0.99***	0.97***		
PopulationGoal	-0.03***	0.03***	0.02**	-0.00	0.13***	-0.34***	0.11***	-0.21***	-0.32***	0.00	1.00***	0.00	0.16***	-0.00	0.00	-0.00	

Computed correlation used pearson-method. P-value: *** p<0.01, ** p<0.05, * p<0.1. Names of the variables are abbreviated to fit contents

Table 7: Correlation between dependent variable and multiple independent variables

Correlation between highly correlated variables

	All Funda	*Is Buyer*	*Send Message*	*Viewed Phone*	*Scheduled Viewing*	*Y2019*	*NumberOfCommuters*
All Funda							
Is Buyer	0.956***						
Send Message	0.978***	0.916***					
Viewed Phone	0.988***	0.930***	0.997***				
Scheduled Viewing	0.989***	0.945***	0.989***	0.991***			
Y2019 (dependent variable)	0.899***	0.905***	0.811***	0.835***	0.854***		
NumberOfCommuters	0.924***	0.897***	0.858***	0.872***	0.901***	0.937***	

Computed correlation used pearson-method. P-value: *** p<0.01, ** p<0.05, * p<0.1.

Empirical model

OLS multivariate linear regression is used to estimate the average effect of the Funda data. Linear regression, as described by Gelman & Hill (2007, p. 31) is "a method that summarizes how the average values of a numerical outcome variable vary over subpopulations defined by linear functions of predictors." It is used to compare measure the effect between groups that have the same observed characteristics and consists of the following components: the dependent variable (denoted by y_i); the treatment variable (denoted by P_i); and a set of control variables. With a single control variable (A_i) the related formula is $y_i = \alpha + \beta P_i + \beta A_i + \varepsilon_i$. Here the regression coefficients are α, is the intercept, βP_i the average effect of treatment, β A_i the effect of the dummy variable and $e_i$is the residual or error term (Angrist & Pischke, 2015). It is important to realize that linear regression has its limitations for estimating the average effect, since only the alternative outcome is known for each unit: we cannot observe both outcomes for one single unit (Gelman & Hill, 2007, p. 171). Therefore, the effect cannot be measured directly.

In multivariate regression the average effect is estimated by first measuring the confounding variable(s) and adjusting for their effects, followed by calculating the residual association between the main explanatory variable and the outcome for each value of the confounder(s), which provides a weighted average of these conditional effects (Toshkov, 2016, p. 224). The effect Funda on relocations is estimated using a multivariate linear regression analysis in which multiple variables are included. These variables, as discussed in the descriptive statistics, are gathered using both the CBS datasets as well as the Funda data(sub)set(s). Using these variables, the following model is composed:

Equation 1: Main model (1)

$$\begin{aligned} Y2019_{o,g} = \alpha + \beta_1 HitsFunda_{F,o,g} + \beta_2 LogDistance_{o,g} + \beta_3 WithinMun_{o,g} \\ + \beta_4 NeighborMun_{o,g} + \beta_5 WithinProv_{o,g} \\ + \beta_6 LogDistanceSupermarkt_g + \beta_7 LogDistanceHighway_g \\ + \beta_8 LogDistanceTrain_g + \beta_9 LogDistanceSchool_g \\ + \beta_{10} HousingPrice_g + \beta_{11} HousingStock_g + \beta_{12} PopulationDensity_o \\ + \beta_{13} Commuters_{o,g} + \beta_{14} Crime_o + \beta_{15} Disturbance_o \\ + \beta_{16} PopulationHome_o + \beta_{17} PopulationGoal_g + \varepsilon_{o,g} \end{aligned}$$

In this model, α is the intercept. The dependent variable is the actual relocations in 2019, is denoted as $Y2019_{o,g}$, where "o" stands for the home or origin municipality and "g" stands for the target or goal municipality. When both are after the variable, it means this is a unique value of a combination of two municipalities. Otherwise, one value is used for the entire municipality of origin or goal, regardless of the combination. This is due to the findings of other studies and reasonability; while some variables could function as a magnet (e.g. distance to utilities), others could push people away (e.g. disturbance). As can be seen in the formula, the Funda sets, LogDistance, the location dummies and Commuters contain this unique combination whereas the log distances to utilities, HousingStock and PopulationGoal have the same value when the same goal municipality and PopulationDensity, Crime, Disturbance and PopuationHome have the same origin municipality. The treatment variable is denoted as $HitsFunda_F$ and its effect β, in which $HitsFunda$ is the amount of hits in each Funda dataset $_F$. Then there are the control variables (17 in total) and their effect β. Lastly there is an error term to denote the residual.

This study takes Schirmer's, Axhausen's, & Eggermond's (2014) local attributes into account, as well as Rae's (2015) considerations about the seriousness of searchers. There are, however, some limitations to this model. First, though both datasets contain many unique data, both consist of aggregated data, which has consequences for the effectiveness of the analysis. Moreover, as mentioned before, the CBS data is rounded to units of five, causing the lower end of the values to completely disappear or appear somewhat distorted. Another shortcoming is the lack of time effects, which could capture, for example, housing market conditions, making it impossible to do fixed effects analysis, which has shown to play a role (Steegmans & Hassink, 2018). Thence, life course events and changes in individual and household circumstances are not observed while these factors are known to be important in the mobility literature when considering the household level of satisfaction with an existing housing situation (Clark and Huang 2003; Li 2004). It is also worth mentioning that there are other platforms besides Funda that can differ in supply, which could also distort the data. The Funda dataset also does not include population characteristics, causing the data to be rather general, while individual characteristics pose great value as well. Also, the distance between municipalities that is used is based on the distance between both centers and therefore isn't fully accurate. In addition, the Funda platform predominantly deals with owner-occupied housing and has about 5% private rental properties, which is different than the actual housing market data and could cause some inconsistencies since the CBS data consists for 100% of owner-occupied housing. Another inconsistency exists due to the implied method of aggregation of the housing market, leaving out most of the texture of the actual housing market; the model assumes equal (owner-occupied) houses, making no distinction in price, seasonality, individual characteristics or whatsoever. To somewhat counter this, a control variable is added with the average housing prices per municipality. Another issue with the data can occur in the proportion of hits. This may not properly reflect the proportion of the users of the platform. In other words, there may be users that generate relatively many hits compared to other users. Assuming these users have specific interests, they could contest an equal proportion of hits and therefor taking this indicator could misrepresent the population as a whole. However, using the controls the Funda datasets has to offer, these effects can be countered to some degree, while the fact remains that the proportion of hits varies among users.

Results

The results are presented in table 8. It contains six models. The models use 26 441 observations to estimate the effects of the variables. The first one does not include the Funda data. The others include the full Funda data and the subsets. The R-squared of all models is between 0.895 (model 1) to 0.921 (model 6), which would suggest around 90% of variation in residential relocations data can be explained by all models. This high value is likely the result of overfitting, however, the adjusted R-squared does not drop in all models, suggesting all models contain relevant variables. The residual standard error in all models is around 29. This means that the observed relocations in 2019 are 29 relocations away from the predicted or fitted values for 2019. The variable of interest (Funda) seems to be highly significant at the 0.01 level in all models. Moreover, there seems to be an increase in the effect of the variables that indicate serious search: while the effect of each hit on the whole Funda dataset is estimated at 1.307, that of the Funda subsets is higher (with the buyer subset being significantly higher). Considering that this variable has been increased by 10 000 – thus actually being 0.0001307 –, the average effect of every 7 651 hits are equal to one relocation. Those who messaged the real estate agent and those who viewed the phone of the real estate agent have an effect of 26.576 and 11.385, accounting each 376 and 878 hits for one relocation. The effect of those who applied for a viewing is 6.962 or 0.0006962 (about six times as much), with each 1 436 hits explaining one relocation. Lastly, the effect of the buyer data is 394.147 or 0.0394147 (302 times as much as the whole dataset), equaling every relocation to 25 hits. Notably, the effect of the commuters is 426,[5] having 23.5 commuters equal one relocation, which is larger than all Funda effects.

Besides the variables of interest, there are also control variables that are observed to be significant. In general, these are relocations within the municipality, relocations to a neighboring municipality, the distance to the supermarket and train station, the average housing price, the housing stock, the number of commuters, crime, public disturbance and the population in the home and origin municipality. The housing stock, however, seems to be negatively related, which doesn't appear logical. This variable is discussed again after the robustness checks. Furthermore, there are some variables that not to have a significant effect on residential mobility: relocations within the province, the distance to both highways and schools and the population density. Lastly, there are some control variables that have a large effect: the location dummies for within the municipality and to neighboring municipalities, the number of commuters and, to a lesser degree, the distance to the supermarket and the average housing price. The dummy indicating relocations within municipalities has a large effect ranging between around 250 and 303 in all models. The dummy for relocations to neighboring municipalities has a smaller but still relatively large effect ranging between around 27 and 30. The effect of number of commuters ranges between around 26 and 43.

Table 9 contains the same models in which insignificant variables are eliminated using backward elimination on p-values below 0.05. This removed the following variables from most models: the dummy for relocations within the province, the distance of the relocations, the distance to both highways, train stations and elementary schools and the population density. Most remarkable is the disappearance of the distance of the relocations, which is found to have an effect in most other studies. Continuing, the effect of the Funda subsets message, viewing

[5] When put in the same scale (from x1 000 to x10 000)

and buyer barely decreased while the effect of the within municipality dummy increased very slightly. Generally, the effect of the distance to the supermarket and the housing price increased moderately. The significance for the variable distance to supermarket. Lastly, the variables distance to train only remained in the second model, while the average housing price and housing stock only disappeared from the sixth model.

Table 8: Main models- using the different Funda datasets

	Dependent variable: Y2019					
	No Funda	Funda all	Funda message	Funda phone	Funda viewing	Funda buyer
	(1)	(2)	(3)	(4)	(5)	(6)
LogDistance	0.158	-0.088	0.013	-0.032	-0.044	0.185
	(0.374)	(0.349)	(0.368)	(0.363)	(0.365)	(0.325)
WithinMunicipality	249.977***	297.883***	294.016***	301.734***	303.252***	285.404***
	(4.333)	(4.116)	(4.476)	(4.376)	(4.461)	(3.784)
NeighborMunicipality	26.686***	29.681***	29.420***	29.622***	29.802***	28.191***
	(1.536)	(1.434)	(1.510)	(1.489)	(1.500)	(1.335)
WithinProvince	-0.846	-0.056	-0.474	-0.408	-0.367	-0.500
	(0.690)	(0.644)	(0.678)	(0.669)	(0.673)	(0.600)
AllFundaHits (x 10 000)		1.307***				
		(0.021)				
SendMessageToAgent (x 10 000)			5.080***			
			(0.161)			
ViewedPhoneNumberAgent (x 10 000)				2.847***		
				(0.068)		
ScheduledViewing (x 10 000)					6.962***	
					(0.187)	
IsBuyer (x 10 000)						394.147***
						(4.254)
LogDistanceToSupermarket	-1.294	-1.829**	-1.554*	-1.585*	-1.522*	-1.483**
	(0.848)	(0.791)	(0.832)	(0.821)	(0.826)	(0.736)

LogDistanceToHighway	0.295	0.393	0.351	0.354	0.379	0.293
	(0.426)	(0.398)	(0.418)	(0.413)	(0.415)	(0.370)
LogDistanceToTrain	-0.380	-0.543**	-0.476*	-0.498**	-0.497**	-0.423*
	(0.257)	(0.240)	(0.252)	(0.249)	(0.250)	(0.223)
LogDistanceToElementarySchool	0.116	0.112	0.139	0.116	0.199	-0.437
	(1.212)	(1.131)	(1.190)	(1.174)	(1.182)	(1.053)
LogAvHousingPrice	-1.409*	-2.122***	-1.680**	-1.701**	-1.605**	-0.845
	(0.758)	(0.708)	(0.744)	(0.734)	(0.739)	(0.659)
HousingStock	-0.0002**	-0.0002***	-0.0002***	-0.0002***	-0.0002**	0.00001
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)
PopulationDensity	0.0001	0.0001	0.0001	0.0001	0.0001	0.0003**
	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0001)
Commuters (x 1 000)	42.598***	27.347***	36.330***	34.124***	33.641***	25.565***
	(0.123)	(0.269)	(0.232)	(0.235)	(0.269)	(0.213)
RegisteredCrimes	-0.001***	-0.001***	-0.001***	-0.001***	-0.001***	-0.001***
	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0001)
RegisterdPublicDisturbance	0.001**	0.001**	0.001***	0.001***	0.001***	0.001***
	(0.0003)	(0.0002)	(0.0003)	(0.0003)	(0.0003)	(0.0002)
PopulationHome	0.00005***	0.00005***	0.00005***	0.00005***	0.0001***	0.00003***
	(0.00001)	(0.00001)	(0.00001)	(0.00001)	(0.00001)	(0.00001)
PopulationGoal	0.0001**	0.0001***	0.0001***	0.0001**	0.0001**	-0.00002
	(0.00004)	(0.00003)	(0.00003)	(0.00003)	(0.00003)	(0.00003)
Constant	16.077	26.568***	20.360**	20.901**	19.700**	9.523
	(10.021)	(9.353)	(9.837)	(9.706)	(9.769)	(8.705)
Observations	26,411	26,411	26,411	26,411	26,411	26,411

R^2	0.895	0.909	0.899	0.902	0.900	0.921
Adjusted R^2	0.895	0.909	0.899	0.901	0.900	0.921
Residual Std. Error	29.074 (df = 26394)	27.134 (df = 26393)	28.540 (df = 26393)	28.158 (df = 26393)	28.342 (df = 26393)	25.256 (df = 26393)

P-value: **** p<0.01, ** p<0.05, * p<0.1

Table 9: Main model using the different Funda datasets after backward elimination on p-values (only significant results)

	Dependent variable: Y2019					
	No Funda	Funda all	Funda message	Funda phone	Funda viewing	Funda buyer
	(1)	(2)	(3)	(4)	(5)	(6)
WithinMunicipality	249.599***	298.297***	294.159***	302.034***	303.587***	284.856***
	(4.000)	(3.813)	(4.171)	(4.072)	(4.160)	(3.495)
NeighborMunicipality	25.707***	29.813***	29.014***	29.353***	29.588***	27.349***
	(1.395)	(1.304)	(1.374)	(1.354)	(1.364)	(1.212)
AllFundaHits (x 10 000)		1.307***				
		(0.021)				
SendMessageToAgent (x 10 000)			5.081***			
			(0.161)			
ViewedPhoneNumberAgent (x 10 000)				2.847***		
				(0.068)		
ScheduledViewing (x 10 000)					6.961***	
					(0.187)	
IsBuyer (x 10 000)						394.136***
						(4.252)
LogDistanceToSupermarket	-1.621***	-1.945***	-1.992***	-2.060***	-1.965***	-2.002***
	(0.576)	(0.575)	(0.566)	(0.558)	(0.562)	(0.496)
LogDistanceToTrain		-0.468**				
		(0.226)				
LogAvHousingPrice	-1.626**	-2.122***	-1.811**	-1.806**	-1.704**	
	(0.738)	(0.689)	(0.724)	(0.715)	(0.719)	

HousingStock	-0.0002**	-0.0002***	-0.0002***	-0.0002***	-0.0002***	
	(0.0001)	(0.0001)	(0.0001)	(0.0001)	(0.0001)	
NumberOfCommuters (x 1 000)	42.587***	27.349***	36.323***	34.120***	33.638***	25.557***
	(0.123)	(0.269)	(0.232)	(0.235)	(0.269)	(0.212)
RegisteredCrimes	-0.001***	-0.001***	-0.001***	-0.001***	-0.001***	-0.001***
	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0001)
RegisterdPublicDisturbance	0.001**	0.001***	0.001***	0.001***	0.001***	0.001***
	(0.0003)	(0.0002)	(0.0003)	(0.0003)	(0.0003)	(0.0002)
PopulationHome	0.0001***	0.00005***	0.00005***	0.00005***	0.0001***	0.00003***
	(0.00001)	(0.00001)	(0.00001)	(0.00001)	(0.00001)	(0.00001)
PopulationGoal	0.0001**	0.0001***	0.0001***	0.0001***	0.0001**	-0.00001***
	(0.00003)	(0.00003)	(0.00003)	(0.00003)	(0.00003)	(0.000002)
Constant	18.923**	26.210***	21.340**	21.334**	20.012**	-0.663**
	(9.269)	(8.651)	(9.099)	(8.977)	(9.035)	(0.316)
Observations	26,411	26,411	26,411	26,411	26,411	26,411
R^2	0.895	0.909	0.899	0.902	0.900	0.921
Adjusted R^2	0.895	0.909	0.899	0.901	0.900	0.921
Residual Std. Error	29.074 (df = 26400)	27.132 (df = 26398)	28.539 (df = 26399)	28.158 (df = 26399)	28.341 (df = 26399)	25.258 (df = 26401)

P-value: *** p<0.01, ** p<0.05, * p<0.1

Robustness checks

In this section the models are slightly changed by evaluating the models in the previous section and provides a check for multicollinearity. First, as is noted in the descriptive statistics in the data section, a high correlation exists between both the independent variables Funda and the commuters with the dependent variable. This could indicate multicollinearity, which occurs when two or more independent variables are correlated with each other, negatively affecting the accuracy of the results of the analysis. Therefore, the multicollinearity is checked using Variance Inflation Factors score (or VIF) as seen in Table 10. The VIF score helps find variables that cause multicollinearity by qualifying the intensity of variance inflation. A score of one indicates that there is no correlation, any score above four should be investigated whether the variable should be kept and a score above ten signals serious multicollinearity (Pennysulvania State University, 2018). Also, the high R-squared could be the result of overfitting. Therefore, some variables are dropped considering the literature and their estimates in the initial model.

The VIF scores of the models are presented in table 10. A high VIF score is observed in the Funda-variables, housing stock, commuters and registered crimes, registered public disturbance, population home and population goal. Notably, the VIF score of Funda and commuters is lower when they are not in the same model, as seen in the first row (1.52 versus around 6 to 8 in the other models). Therefore, these variables are not combined again, but are not excluded from this study. Furthermore, registered crimes and registered public disturbance also show a high VIF score (50 and 24.7). Moreover, intuitively, it is arguable that these variables are related to each other. Since registered crimes has a negative estimate and registered public disturbance has a positive estimate, registered crime is removed, because it is unlikely, considering the literature, that people are less likely to move away from places with higher crime rates. The estimate of registered disturbance does correspond with the literature.

Continuing, besides having a very high VIF score of 207, the variable housing stock also produces a nearly impossible negative estimate. It seems illogical that a higher housing stock is correlated with less relocations and therefore, together with the high VIF score, this variable is removed. The VIF score of population goal will remain relatively high, but won't be removed since it isn't collinear with the dependent variable, but with registered disturbance. Seeing these variables are used as control variables and they are not collinear with the variables of interest, they are not removed (Allison, 2012). Table 11 contains the net VIF scores of the refined model. As mentioned above, registered disturbance and population home have high VIF scores (19 and 17.6) while the rest is within acceptable boundaries.

Table 10: Multicollinearity using Variance Inflation Factors (VIF) on base models

	No Funda	***Funda all***	***Funda message***	***Funda phone***	***Funda viewing***	***Funda buyer***
LogDistance	2.18	2.18	2.18	2.18	2.18	2.18
WithinMunicipality	1.71	1.77	1.89	1.85	1.90	1.72
NeighborMunicipality	1.23	1.23	1.24	1.23	1.24	1.23
WithinProvince	1.63	1.63	1.63	1.63	1.63	1.63
Funda data	0.00	7.23	4.33	4.66	6.10	5.24
LogDistanceTo Supermarket	2.49	2.49	2.49	2.49	2.49	2.49
LogDistanceToHighway	1.18	1.18	1.18	1.18	1.18	1.18
LogDistanceToTrain	1.36	1.36	1.36	1.36	1.36	1.36
LogDistanceTo ElementarySchool	2.27	2.27	2.27	2.27	2.27	2.27
LogAvHousingPrice	1.08	1.08	1.08	1.08	1.08	1.08
HousingStock	207.33	207.36	207.40	207.34	207.33	207.54
PopulationDensity	1.59	1.59	1.59	1.59	1.59	1.59
NumberOfCommuters	1.52	8.38	5.62	5.92	7.65	6.02
RegisteredCrimes	50.07	50.07	50.08	50.08	50.10	50.07
RegisterdPublicDisturbance	24.71	24.71	24.71	24.71	24.72	24.71
PopulationHome	38.72	38.72	38.72	38.72	38.72	38.73
PopulationGoal	208.37	208.39	208.43	208.37	208.37	208.57

Table 11: Multicollinearity using VIF scores on final models

	No Funda	Funda all	Funda message	Funda phone	Funda viewing	Funda buyer
LogDistance	2.15	2.15	2.15	2.15	2.15	2.15
WithinMunicipality	1.70	1.51	1.38	1.41	1.42	1.52
NeighborMunicipality	1.23	1.23	1.22	1.23	1.23	1.23
WithinProvince	1.62	1.62	1.62	1.62	1.62	1.62
Funda/Commuters	1.52	1.31	1.17	1.19	1.21	1.32
LogDistanceToSupermarket	2.46	2.46	2.46	2.46	2.46	2.46
LogDistanceToHighway	1.17	1.17	1.17	1.17	1.17	1.17
LogDistanceToTrain	1.35	1.35	1.35	1.35	1.35	1.35
LogDistanceTo ElementarySchool	2.25	2.25	2.25	2.25	2.25	2.25
LogAvHousingPrice	1.07	1.07	1.07	1.07	1.07	1.08
PopulationDensity	1.54	1.54	1.54	1.54	1.54	1.54
RegisterdPublicDisturbance	18.00	18.01	18.00	18.00	18.00	18.00
PopulationHome	17.61	17.61	17.61	17.61	17.61	17.62
PopulationGoal	1.20	1.19	1.19	1.19	1.19	1.19

New models

The new refined models are presented below in equation 2 and 3. Following the changes above, the equation has been split up in two, with the model having the data of either the commuters (equation 2) or Funda (equation 3). Moreover, housing stock and registered crime are removed from both equations. As a result, the models contain 14 variables instead of the initial 17. The new models are shown in table 12 and 13.

Equation 2: New model with only commuter data, excluding Funda data (2)

$$
\begin{aligned}
Y2019_{o,g} = {} & \alpha + \beta_1 LogDistance_{o,g} + \beta_2 WithinMun_{o,g} + \beta_3 NeighborMun_{o,g} \\
& + \beta_4 WithinProv_{o,g} + \beta_5 LogDistanceSupermarkt_g \\
& + \beta_6 LogDistanceHighway_g + \beta_7 LogDistanceTrain_g \\
& + \beta_8 LogDistanceSchool_g + \beta_9 HousingPrice_g \\
& + \beta_{10} PopulationDensity_o + \beta_{11} Commuters_{o,g} + \beta_{12} Disturbance_o \\
& + \beta_{13} PopulationHome_o + \beta_{14} PopulationGoal_g + \varepsilon_{o,g}
\end{aligned}
$$

Equation 3: New model with only Funda data, excluding commuter data (3)

$$
\begin{aligned}
Y2019_{o,g} = {} & \alpha + \beta_1 HitsFunda_{F,o,g} + \beta_2 LogDistance_{o,g} + \beta_3 WithinMun_{o,g} \\
& + \beta_4 NeighborMun_{o,g} + \beta_5 WithinProv_{o,g} \\
& + \beta_6 LogDistanceSupermarkt_g + \beta_7 LogDistanceHighway_g \\
& + \beta_8 LogDistanceTrain_g + \beta_9 LogDistanceSchool_g \\
& + \beta_{10} HousingPrice_g + \beta_{11} PopulationDensity_o + \beta_{12} Disturbance_o \\
& + \beta_{13} PopulationHome_o + \beta_{14} PopulationGoal_g + \varepsilon_{o,g}
\end{aligned}
$$

Table 12: New model after robustness checks using the different Funda datasets

	Dependent variable: Y2019					
	No Funda	Funda all	Funda message	Funda phone	Funda viewing	Funda buyer
	(1)	(2)	(3)	(4)	(5)	(6)
LogDistance	0.106	-0.899**	-1.427***	-1.272***	-1.106**	-0.049
	(0.372)	(0.410)	(0.507)	(0.483)	(0.458)	(0.402)
WithinMunicipality	249.798***	456.528***	655.450***	612.971***	583.109***	441.036***
	(4.333)	(4.491)	(5.326)	(5.116)	(4.873)	(4.426)
NeighborMunicipality	26.686***	39.448***	51.624***	47.953***	46.549***	37.118***
	(1.536)	(1.689)	(2.089)	(1.991)	(1.887)	(1.658)
WithinProvince	-1.032	1.415*	1.934**	1.688*	1.629*	0.661
	(0.689)	(0.759)	(0.940)	(0.895)	(0.848)	(0.745)
Commuters (x 1 000)	42.604***					
	(0.123)					
AllFundaHits (x 10 000)		3.226***				
		(0.010)				
SendMessageToAgent (x 10 000)			26.577***			
			(0.116)			
ViewedPhoneNumberAgent (x 10 000)				11.379***		
				(0.046)		
ScheduledViewing (x 10 000)					27.925***	
					(0.105)	
IsBuyer (x 10 000)						836.590***
						(2.658)
LogDistanceToSupermarket	-1.066	-2.457***	-2.611**	-2.553**	-2.334**	-2.277**

	(0.843)	(0.928)	(1.149)	(1.095)	(1.038)	(0.911)
LogDistanceToHighway	0.376	0.687	0.789	0.661	0.710	0.198
	(0.425)	(0.468)	(0.579)	(0.552)	(0.523)	(0.459)
LogDistanceToTrain	-0.440*	-0.868***	-0.982***	-0.900***	-0.870***	-0.364
	(0.256)	(0.281)	(0.349)	(0.332)	(0.315)	(0.276)
LogDistanceToElementarySchool	-0.204	-0.205	-0.004	0.075	0.471	-0.420
	(1.207)	(1.328)	(1.645)	(1.567)	(1.485)	(1.304)
LogAvHousingPrice	-1.383*	-3.262***	-3.041***	-2.760***	-2.313**	-0.329
	(0.758)	(0.835)	(1.034)	(0.985)	(0.934)	(0.820)
PopulationDensity	0.0003*	0.0002	0.00004	0.0001	0.0002	0.001***
	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)
RegisterdPublicDisturbance	-0.0001	-0.0002	0.0002	0.0001	0.0002	0.001**
	(0.0002)	(0.0003)	(0.0003)	(0.0003)	(0.0003)	(0.0002)
PopulationHome	0.00001	-0.000004	-0.000002	-0.000002	-0.00001	-0.00003***
	(0.00001)	(0.00001)	(0.00001)	(0.00001)	(0.00001)	(0.00001)
PopulationGoal	-0.00001***	-0.000005	0.00001***	0.00001***	0.00001**	0.00001***
	(0.000003)	(0.000003)	(0.000004)	(0.000003)	(0.000003)	(0.000003)
Constant	17.232*	45.791***	44.688***	40.400***	34.473***	3.978
	(10.019)	(11.031)	(13.663)	(13.016)	(12.333)	(10.831)
Observations	26,411	26,411	26,411	26,411	26,411	26,411
R^2	0.895	0.873	0.805	0.823	0.841	0.877
Adjusted R^2	0.895	0.873	0.804	0.823	0.841	0.877
Residual Std. Error (df = 26396)	29.089	32.025	39.669	37.789	35.805	31.444

P-value: *** p<0.01, ** p<0.05, * p<0.1

Table 13: New model using the different Funda datasets after robustness checks and backward elimination on p-values (only significant results)

	Dependent variable: Y2019					
	No Funda	Funda all	Funda message	Funda phone	Funda viewing	Funda buyer
	(1)	(2)	(3)	(4)	(5)	(6)
LogDistance		-0.882^{**}	-1.361^{***}	-1.228^{***}	-1.054^{**}	
		(0.402)	(0.498)	(0.475)	(0.450)	
WithinMunicipality	249.392^{***}	456.980^{***}	655.853^{***}	613.035^{***}	583.316^{***}	441.153^{***}
	(4.004)	(4.470)	(5.303)	(5.090)	(4.849)	(4.039)
NeighborMunicipality	26.495^{***}	39.347^{***}	51.665^{***}	48.013^{***}	46.575^{***}	37.681^{***}
	(1.459)	(1.683)	(2.083)	(1.985)	(1.881)	(1.505)
WithinProvince	-1.130^{**}	1.510^{**}	2.050^{**}	1.754^{**}	1.716^{**}	
	(0.568)	(0.755)	(0.935)	(0.891)	(0.844)	
Commuters (x 1 000)	42.602^{***}					
	(0.123)					
AllFundaHits (x 10 000)		3.224^{***}				
		(0.010)				
SendMessageToAgent (x 10 000)			26.576^{***}			
			(0.116)			
ViewedPhoneNumberAgent (x 10 000)				11.385^{***}		
				(0.046)		
ScheduledViewing (x 10 000)					27.933^{***}	
					(0.105)	
IsBuyer (x 10 000)						836.689^{***}
						(2.655)

LogDistanceToSupermarket	-1.634***	-2.537***	-2.934***	-2.789***	-2.411***	-2.833***
	(0.576)	(0.651)	(0.841)	(0.801)	(0.759)	(0.618)
LogDistanceToTrain		-0.708***	-0.836**	-0.775**	-0.728**	
		(0.266)	(0.330)	(0.315)	(0.298)	
LogAvHousingPrice	-1.505**	-3.274***	-3.091***	-2.809***	-2.372**	
	(0.739)	(0.833)	(1.033)	(0.984)	(0.932)	
PopulationDensity						0.001***
						(0.0002)
PopulationHome	0.00001***					-0.00003***
	(0.000002)					(0.00001)
PopulationGoal	-0.00001***		0.00001***	0.00001***	0.00001***	0.00001***
	(0.000003)		(0.000004)	(0.000003)	(0.000003)	(0.000003)
RegisterdPublicDisturbance		-0.0002***	0.0002***			0.001**
		(0.0001)	(0.0001)			(0.0002)
Constant	18.942**	45.803***	45.096***	41.069***	35.046***	-0.697**
	(9.283)	(10.980)	(13.616)	(12.970)	(12.289)	(0.341)
Observations	26,411	26,411	26,411	26,411	26,411	26,411
R^2	0.895	0.873	0.805	0.823	0.841	0.877
Adjusted R^2	0.895	0.873	0.805	0.823	0.841	0.877
Residual Std. Error	29.089 (df = 26402)	32.026 (df = 26401)	39.667 (df = 26400)	37.788 (df = 26401)	35.805 (df = 26401)	31.442 (df = 26402)

P-value: *** p<0.01, ** p<0.05, * p<0.1

Results final models

Table 12 and 13 contain the new models in which some variables are removed as discussed above. As can be derived from the tables, the overall fit of the new models has impaired because the residual standard error increased from around 25 to 29 to around 29 to nearly 40. The R-squared, also slightly dropped from around 0.90 to between 0.80 to nearly 0.90. With regards to the variables of interest, all remain significant at the highest level, while an increase in their effects is observed. Most notably, the independent variable of interest, the Funda data, which is due to the exclusion of the commuter data. In the new models before backward elimination (table 12), the effect of the whole Funda dataset increased about 2.5 times (from 1.307 to 3.226), that of the Funda message subsample around five times (5.080 to 26.557), the Funda phone subsample around four times (2.847 to 11.379), the Funda viewing subsample about five times (6.962 to 27.925) and the Funda buyer subsample about two times (394.174 to 836.590). In the whole Funda dataset, each 3 100 hits are correlated with one relocation. In the subsample of the users that viewed the phone number of the real estate agent and those who messaged the real estate agent, each 879 and 380 hits are correlated with one relocation. In the subsample with those who applied for a viewing of the property and those who registered themselves as buyer of the property, each 358 and 12 hits are correlated with one relocation.

Since these effects are also significant, the first hypothesis is confirmed: online housing search intensity is found to influence residential mobility since more hits are associated with more relocations. Though, the biggest effects are still observed in the dummy that indicates whether the goal municipality is the same municipality and, to a smaller degree, the dummy that indicates whether the goal municipality is the neighboring municipality. In the new models, the effect of the commuters (426)[6] is still smaller than nearly all Funda sets, except here the model with the buyer data is larger (837). The effect of the dummy that indicates whether the municipality is the same municipality ranges between 249 and 656 relocations if true and the dummy for neighboring municipality ranges between 26.5 and 51.6 if true. The number of commuters between municipalities correlates each 23.5 commuters with 1 relocation. Nevertheless, it appears that the Funda subsets do indicate seriousness, since the effects seem to increase as the level of seriousness increases. This confirms the second hypothesis which states that the level of seriousness influences residential mobility, which aligns with Rae's (2015) reasoning that recreational search data could dilute serious search data. Steegmans & De Bruin (2021) find similar results while also concluding no significant differences exist between the distance effects between both types of searchers.

After backward elimination on p-values (table 13), all control variables are not (completely) removed in the new models, however they are dropped in some of the models. Compared to the base models, the population of the home municipality has been dropped from five of the six models and registered public disturbance has been dropped from half of the models. On the other hand, the distance of the relocation remains in four of the six models compared to none in the base models. Moreover, the dummy for relocations within the province remains as well, except for the model with the Funda buyer subset. Lastly, the distance to the train station variable is now included in four models instead of one. Furthermore, the distance variable and the distance to a train station variable are dropped in the model with the commuter data and the Funda buyer subset. The housing price variable is dropped in the model with the Funda buyer subset. The population density variable only remains in the model with the Funda buyer subset. The population in the home municipality variable only remains in the model with

[6] When put in the same scale (from x 1 000 to x 10 000)

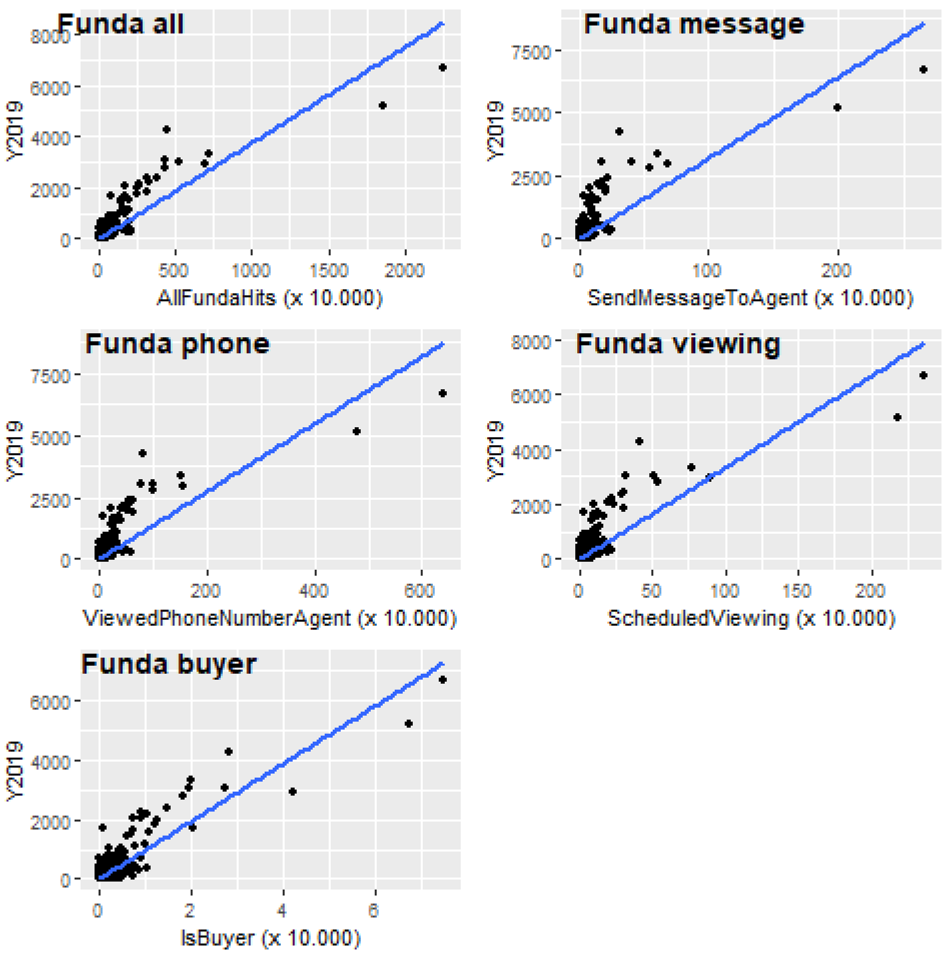
the commuter data. The population in the goal municipality only drops in the model with all Funda data. The registered public disturbance only remains in the models with all Funda data and the Funda message subset. The effects of all Funda sets, together with the effects of the distance to a train station, very slightly decrease after backward elimination while the (negative) effects of the distance to a supermarket increases (table 13).

While most effects seem logical apart from being significant, some effects appear counterintuitive or questionable. First of all, the dummy variable for relocations within the province. While the model with the commuter data provides a negative effect for this variable, the other models provide a positive effect and the model with the Funda buyer data drops this variable. Moreover, the population density, which is only significant in the model with the Funda buyer subset, may have a positive effect, buy it remains only in one of the six models. Moreover, the population in the home municipality remains only in two models between which the effects are both positive and negative. Considering it is expected that a higher population in the home municipality is correlated with a higher effect while this is conflicting between the two models, the effect is questionable. Lastly, the public disturbance in the home municipality is expected to positively influence relocations. However, next to this variable being dropped from three of the six models, the effects between the remaining models conflict with each other, making this effect questionable as well. The effects of the other variables – the Funda sets, commuters, distance, within municipality dummy, neighboring municipality dummy, distance to supermarket, distance to train, housing price, population goal (except for commuter set) – are in line with expectations. Most notably, the distance variable, which was dropped in the base models after backward elimination, does remain in four of the six new models.

Figure 3 plots the correlation between the main variable of interest and the dependent variable. While the scale of the y-axis remains relatively equal, the scale of the x-axis varies.

This is due to the difference in observations between the Funda datasets. which occurs because the complete Funda dataset contains far more hits than the buyer dataset (see table 4). All tables have a similar regression line and distribution of points. Though, as seen on the right end of all graphs, the observations of a few large municipalities could disproportionately affect the results and should therefore be approached carefully. This observation is taken into account in the conclusion.

Figure 3: Correlation between main independent variables and dependent variable



Conclusion

The aim of this study is to research the effects of online housing search on residential mobility while considering the degree of seriousness of the search. Building upon theoretical constructs and Rae's conceptual framework, a framework for online housing search and residential mobility is established, along with two hypotheses which state that residential mobility is influenced by (1) online housing search and (2) the level of searcher seriousness. To enable the analysis of the effects of online housing search on residential mobility multiple datasets are used. First, CBS' residential relocations data from 77 municipalities is combined with online housing market platform search data from Funda. Additionally, control variables available on CBS are added. These data are analyzed via OLS multivariate regression analysis using the statistics program R. After some robustness checks on the base models, the final models are presented.

After applying OLS on the data, online housing search is found to have a significant effect on residential mobility. This confirms the first hypothesis. Moreover, the level of searcher seriousness is also found to have an effect on residential mobility, as an increase in effect is observed in more serious search data, confirming the second hypothesis. In the whole Funda dataset each 3,100 hits is correlated with one relocation. In the subsample of the users that viewed the phone number of the real estate agent and those who messaged the real estate agent, each 879 and 380 hits are correlated with each relocation. In the subsample with those who applied for a viewing of the property and those who registered themselves as buyer of the property, each 358 and 12 hits are correlated with one relocation. Nevertheless, other control variables are found to have a significant effect as well. Most notably, the dummies that indicate whether the municipality is either the same municipality or a neighboring one, ranging between 249 and 656 relocations and 26.5 and 51.6 if true. Lastly, each 23.5 commuters correlate with 1 relocation.

This research does have limitations. Most importantly, this study uses aggregate data, leaving out most of the context of the actual housing market: the model assumes homogeneous (owner-occupied) housing, making no distinction in price, seasonality, actor characteristics or whatsoever. This makes it impossible to analyze individual characteristics, which many authors find to influence residential mobility as well (Clark, Huang, & Withers, 2003; Li, 2004). As noted by Steegmans & De Bruin (2021b), it is unlikely that the aggregate data within the dataset flawlessly portrays individual user intentions while it is likely to have considerable correlation. Related to this, the CBS data, is aggregated into units of five, possibly distorting the results for smaller places. Furthermore, time effects are not taken into account either due to the absence of timestamps in the Funda data. Another remark about the Funda data is that about 5% of the data is generated by private rental data while the CBS data consists solely of owner-occupied housing. Another issue of concern within the Funda data is that the proportion of hits may not properly reflect the proportion of the users of the platform. In other words, there may be users that generate relatively many hits compared to other users. Lastly, there are more accurate research methods than multivariate linear regression that perform better in estimating causality. Moreover, regarding the control variables, there are several things to note. First, the distance is calculated from municipality's center to center and does not indicate the actual (mean or median) distance. Next, the control variables regarding distance include the mean distance whereas the median distance would provide a more accurate estimation. Furthermore, as discussed in the final results section, the observations of a few large municipalities could disproportionately affect the results. Lastly, there are other variables next to the main independent variable that can explain the correlation with the relocations in 2019. The first variable is the number of commuters, which has a correlation of 94%. Last are the relocations within the municipality, which has a correlation of 61%. These variables, mainly the previous year, are likely to have a large influence as well.

Further research should focus on mover's characteristics, the effect of larger municipalities and/or the differences within provinces and municipalities, which have not been examined separately in this paper. By subsetting the data, provinces can be compared based on the number of habitants and population density (taking into account cities and rural areas) which could add more accuracy to the differences on moving distance. Moreover, further research via fixed effects for example, would provide further insights and include time effects, and, due to the distribution of the relocation data, involve negative binomial regression or similar method which could potentially further increase the understanding of moving and housing market platform data. Another contribution in further research could be added by applying the gravity framework on the CBS data and compare results with Steegmans & De Bruin (2021b) and even adding a time dimension, since the CBS data covers multiple years.

This study contributes to the understanding of housing (sub)markets by identifying relevant variables that are correlated with residential relocations and thus relevant for residential mobility. By analyzing novel type data, this study also contributes to the analysis of user-generated search data in the housing market and thus new related insights. Moreover, the size of the data and the various variables provide many insights into the housing market. Simultaneously, this study provides angles for further research and simultaneously contributes to real world applications of these data. Ultimately, this study contributes to mobility policy by unveiling important factors in relocations. Apart from indicating the relevance of online housing search data in residential mobility, other relevant indicators are identified as well, such as housing prices and the relative location to other municipalities, utilities, and work. Altogether, these outcomes can inform policymakers and therefore assist them in optimizing mobility policy and tackling the issues that are currently surrounding the Dutch housing market.

Bibliography

Allison, P. (2012, September 10). *When Can You Safely Ignore Multicollinearity?* Retrieved from https://statisticalhorizons.com/: https://statisticalhorizons.com/multicollinearity#some%20can%20be%20ignored

Anglin, P. M. (1997). Determinants of Buyer Search in a Housing Market. *Real Estate Economics*, 567-589.

Angrist, J. D., & Pischke, J.-S. (2015). *Mastering Metrics.* New Jersey: Princeton University Press.

Baraklianos, I., Bouzouina, L., Bonnel, P., & Aissaoui, H. (2020). Does the accessibility measure infuence the results of residential location choice modelling? *Transportation*, 1147 - 1176.

Baryla, E. A., & Zumpano, L. V. (1995). Buyer Search Duration in the Residential Real Estate Market: The Role of the Real Estate Agent. *The Journal of Real Estate Research*, 1-13.

Baum, A. (2017). *PropTech3.0.* Oxford: University of Oxford.

Baum-Snow, N., & Ferreira, F. (2015). Causal Inference in Urban and Regional. In G. Duranton, V. Henderson, & W. Strange, *Handbook of Regional and Urban Economics* (pp. 4-63). Amsterdam: Elsevier B.V.

Blanchard, O. J., & Diamond, P. (1989). *The aggregate matching function.* Cambridge: National Bureau of Economic Research.

Boelhouwer, P. (2019). The housing market in The Netherlands as a driver for social inequalities: proposals for reform. *International Journal of Housing Policy* , 447-456.

Braesemann, F., & Baum, A. (2020). PropTech: Turning real estate into a data-driven market? *Oxford Future of Real Estate Initiative*, https://ssrn.com/abstract=3607238.

Bruch, E., & Swait, J. (2019). Choice Set Formation in Residential Mobility and Its Implications for Segregation Dynamics. *Demography*, 1665–1692.

Bruch, E., & Swait, J. (2019). Choice Set Formation in Residential Mobility and Its Implications for Segregation Dynamics. *Demography*, 1665-1692.

Burdett, K., & Mortensen, D. (1980). Search, Layoffs, and Labor Market Equilibrium. *Journal of Political Economy*, 652-672.

Bürgke, M. (2006). Residential location choice model for the Greater Zurich area conference paper STRC 2006. *Swiss Transport Research Conference*, https://doi.org/10.3929/ethz-a-005228652.

Causa, O., & Pichelmann, J. (2020). Should I stay or should I go? Housing and residential mobility across OECD countries. *OECD Economics Department Working Papers No. 1626*, https://dx.doi.org/10.1787/d91329c2-en.

CBS. (2020, June 9). *Verhuisbewegingen*. Retrieved from https://dashboards.cbs.nl/: https://dashboards.cbs.nl/v3/verhuizingendashboard/

CBS. (2021a, Oktober 15). *Nabijheid voorzieningen; afstand locatie; regionale cijfers*. Retrieved from https://opendata.cbs.nl: https://opendata.cbs.nl/statline/portal.html?_la=nl&_catalog=CBS&tableId=80305ned&_theme=413

CBS. (2021b, Oktober 27). *Voorraad woningen; gemiddeled oppervlak; woningtype; bouwjaarklasse, regio*. Retrieved from https://opendata.cbs.nl:

https://opendata.cbs.nl/statline/portal.html?_la=nl&_catalog=CBS&tableId=82550NED&_theme=271

CBS. (2021c, May 6). *Wijk- en buurtkaart 2019*. Retrieved from https://www.cbs.nl/: https://www.cbs.nl/nl-nl/dossier/nederland-regionaal/geografische-data/wijk-en-buurtkaart-2019

CBS. (2022, Januari 21). *Banen van werknemers naar woon-en werkregio*. Retrieved from https://opendata.cbs.nl: https://opendata.cbs.nl/statline/portal.html?_la=nl&_catalog=CBS&tableId=83628NED&_theme=242

Clark, W. A., Huang, Y., & Withers, S. (2003). Does commuting distance matter?: Commuting tolerance and residential change. *Regional Science and Urban Economics*, 199-221.

Cooke, T., & Shuttleworth, I. (2017). The effects of information and communication technologies on residential mobility and migration. *John Wiley & Sons, Ltd*, https://doi.org/10.1002/psp.2111.

Cooper, A., & Kurzer, P. (2020). Similar Origins – Divergent Paths: The Politics of German and Dutch Housing Markets. *German Politics*, DOI: 10.1080/09644008.2020.1764541.

Coulter, R., & Scott, J. (2015). What Motivates Residential Mobility? Reexamining Self-Reported Reasons for Desiring and Making Residential Moves. *POPULATION, SPACE AND PLACE*, 354-371.

Coulter, R., Ham, M. v., & Findlay, A. M. (2016). Re-thinking residential mobility: Linking lives through time and space. *Progress in Human Geography*, 352–374.

CRAN. (2021, June 16). *tmap: Thematic Maps*. Retrieved from https://cran.r-project.org/: https://cran.r-project.org/web/packages/tmap/index.html

CRAN. (2022, Februari 4). *sf: Simple Features for R*. Retrieved from https://cran.r-project.org/: https://cran.r-project.org/web/packages/sf/index.html

Dana, J., & Orlov, E. (2014). Internet Penetration and Capacity Utilization in the US Airline Industry. *American Economic Journal: Microeconomics 6*, 106-37.

Dearsley, J., & Baum , A. (2020, 5 5). *What is PropTech?* Retrieved from https://www.unissu.com/: https://www.unissu.com/proptech-resources/what-is-proptech

Dinerstein, M., Einav, L., Levin, J., & Sundaresan, N. (2018). Consumer Price Search and Platform Design in Internet Commerce. *American Economic Review*, 1820-1859.

Dutch National Police. (2022a, Februari 12). *Geregistreerde misdrijven en aangiften; soort misdrijf, gemeente*. Retrieved from https://data.politie.nl/: https://data.politie.nl/portal.html?_la=nl&_catalog=Politie&tableId=47013NED&_theme=97

Dutch National Police. (2022b, Februari 15). *Geregistreerde overlast; soort overlast, gemeente*. Retrieved from https://data.politie.nl: https://data.politie.nl/portal.html?_la=nl&_catalog=Politie&tableId=47021NED&_theme=98

Elder, H. W., Zumpano, L. V., Baryla, E. A., & Baryla, E. A. (1999). Buyer Search Intensity and the Role of the Residential Real Estate Broker. *Journal of Real Estate Finance and Economics*, 351-368.

Ellison, G., & Ellison, S. F. (2014). Match Quality, Search, and the Internet Market for Used Books. *Massachusetts Institute of Technology*.

Facchinetti, M. (2021, Februari 12). *The second phase of the PropTech movement - view from Switzerland*. Retrieved from https://www.unissu.com/: https://www.unissu.com/proptech-resources/proptech-view-from-switzerland

Fang, C., & Liempt, I. v. (2020). ‘We prefer our Dutch’: International students’ housing experiences in the Netherlands. *Housing Studies*, 822-842.

Fradkin, A. (2017, March 21). Search, Matching, and the Role of Digital Marketplace Design in Enabling Trade: Evidence from Airbnb. SSRN.

Funda. (2022, Accessed Januari 29). *https://www.funda.nl/*. Retrieved from Over funda: https://www.funda.nl/over-funda/

Gelman, A., & Hill, J. (2007). *Data analysis using regression and multilevel/hierarchical models.* New York: Cambridge University Press.

Gelman, A., & Hill, J. (2007). *Data analysis Using regression and Multilevel/Hierarchical Models.* New York: Cambridge University Press.

Gent, W. v., & Hochstenbach, C. (2019). The neo-liberal politics and socio-spatial implications of Dutch post-crisis social housing policies. *International Journal of Housing Policy*, 156-172.

Gibbons, S., Overman, H. G., & Patacchini, E. (2015). Spatial Methods. In G. Duranton, V. Henderson, & W. Strange, *Handbook of Regional and Urban Economics* (pp. 116-164). Amsterdam: Elsevier.

Goldfarb, A., & Catherine, T. (2019). Digital Economics. *Journal of Economic Literature*, 3-43.

Google Analytics. (2022, Accessed Februari 15). *Google Analytics Help*. Retrieved from https://support.google.com/: https://support.google.com/analytics#topic=9143232

Habib, M. A., & Miller, E. J. (2009). Reference-Dependent Residential Location Choice Model Within a Relocation Context. *Journal of the Transportation Research Board*, 92-99.

Han, L., & Strange, W. C. (2015). The Microstructure of Housing Markets: Search, Bargaining and Brokerage. In G. Duranton, V. Henderson, & W. Strange, *Handbook of Regional and Urban Economics* (pp. 815-879). Amsterdam: Elsevier.

Hochstenbach, C., Aalbers, M. B., Wind, B., & Arundel, R. (2020). Resurgent landlordism in a student city: urban dynamics of private rental growth. *Urban Geography*, DOI: 10.1080/02723638.2020.1741974.

Hu, L., & Wang, L. (2019). Housing location choices of the poor: does access to jobs matter? *Housing Studies*, 1721-1745.

Jin, J., & Lee, H. (2018). Understanding residential location choices: an application of the UrbanSim residential location model on Suwon, Korea. *International Journal of Urban Sciences*, 216-235.

Kanne, P. (2020, September 21). *WONINGMARKT EEN VAN DE GROTE VRAAGSTUKKEN VAN DEZE TIJD*. Retrieved from https://www.binnenlandsbestuur.nl/: https://www.binnenlandsbestuur.nl/bestuur-en-organisatie/kennispartners/i-o-research/woningmarkt-een-van-de-grote-vraagstukken-van.14531875.lynkx

Kanne, P., & Engeland, W. v. (2020, 9 21). *WONINGMARKT EEN VAN DE GROTE VRAAGSTUKKEN VAN DEZE TIJD.* Amsterdam: I&O Research. Retrieved from https://www.binnenlandsbestuur.nl/: https://www.binnenlandsbestuur.nl/bestuur-en-organisatie/kennispartners/i-o-research/woningmarkt-een-van-de-grote-vraagstukken-van.14531875.lynkx

Kassambara, A. (2022, Januari 31). *http://www.sthda.com/*. Retrieved from Regression Model Accuracy Metrics: R-square, AIC, BIC, Cp and more: http://www.sthda.com/english/articles/38-regression-model-validation/158-regression-model-accuracy-metrics-r-square-aic-bic-cp-and-more/

Kenton, W. (2021, August 4). *https://www.investopedia.com/*. Retrieved from Durbin Watson Statistic Definition: https://www.investopedia.com/terms/d/durbin-watson-statistic.asp#:~:text=What%20Is%20the%20Durbin%20Watson%20Statistic%3F&text=The%20Durbin%2DWatson%20statistic%20will,to%204%20means%20negative%20autocorrelation.

Kohn, M., & Shavell, S. (1974). The Theory of Search. *JOURNAL OF ECONOMIC THEORY*, 93-123.

Kulu, H., Mikolai, J., Thomas, M., Vidal, S., Schnor, C., Willaert, D., . . . Mulder, C. (2021). Separation and Elevated Residential Mobility: A Cross-Country Comparison. *European Journal of Population*, 121–150.

Li, S.-M. (2004). Life Course and Residential Mobility in Beijing, China. *Environment and Planning: A Economy and Space*, 27-43.

Luccas Jr, R. E., & Prescott, E. C. (1974). Equilibrium search and unemployment. *Journal of Economic Theory*, 188-209.

Maclennan, D. (1980). *Housing Economics: An Applied Approach.* London: Longman.

Marsh, A., & Gibb, K. (2011). Uncertainty, Expectations and Behavioural Aspects of Housing Market Choices. *Housing, Theory and Society*, 215-235.

Micheal, R., & Becker, G. (1973). ON THE NEW THEORY OF CONSUMER BEHAVIOR. *The Scandinavian Journal of Economics*, 378-396.

Munuel, A. B., Hochstenbach, C., Bosma, J., & Fernandez, R. (2020). The Death and Life of Private Landlordism: How Financialized Homeownership Gave Birth to the Buy-To-Let Market. *Theory and Society*, 1-23.

Niedomysl, T. (2011). How Migration Motives Change over Migration Distance: Evidence on Variation across Socio-economic and Demographic Groups. *Regional Studies*, 843-855.

Obbink, H. (2020, November 12). *'Oude dame' Funda komt met iets nieuws*. Retrieved from https://www.trouw.nl/: https://www.trouw.nl/economie/oude-dame-funda-komt-met-iets-nieuws~b3446aee/?referrer=https%3A%2F%2Fwww.google.com%2F

Palma, A. d., Picard, N., Motamedi, K., & Wedell, P. (2005). A model of residential location choice with endogenous housing prices and traffic for the Paris region. *European Transport*, 67-82.

Pennysulvania State University. (2018). *https://online.stat.psu.edu*. Retrieved from Multicollinearity: https://online.stat.psu.edu/stat462/node/180/

Pu, Z., & Gaessler, F. (2019). Openness as Platform Strategy Evidence from Quasi-Experiment in Crowdfunding. *Max Planck Institute for Innovation and Competition Research Paper Series*, Paper No. 18-05.

R statistics for Political Science. (2020, November 23). *Choose model variables by AIC in a stepwise algorithm with the MASS package in R*. Retrieved from https://rforpoliticalscience.com/: https://rforpoliticalscience.com/2020/10/23/choose-model-variables-by-aic-in-a-stepwise-algorithm-with-the-mass-package-in-r/

Rae, A. (2015). Online Housing Search and the Geography of Submarkets. *Housing studies*, 453-472.

Rashidi, T. H., & Mohammadian, A. K. (2015). Behavioral Housing Search Choice Set Formation: A Spatial Hazard-based Screening Model. *International Regional Science Review*, 151-170.

Rashidi, T. H., Auld, J., & Mohammadian, A. K. (2012). A behavior housing search model: Two-stage hazard-based and multinonial logit approach to choice-set formation and location selection. *Transportation Research Part A*, 1097-1107.

Saadatmand, F., Lindgren, R., & Schultze, U. (2019). Configurations of platform organizations: Implications for complementor. *Research Policy*, https://doi.org/10.1016/j.respol.2019.03.015.

Salomon, I., & Ben-Akiva, M. (1983). The use of the life-style concept in travel demand models. *Environment and Planning A*, 623-638.

Schirmer, P. M., Axhausen, K. W., & Eggermond, M. A. (2014). The role of location in residential location choice models: a review of literature. *THE JOURNAL OF TRANSPORTATION AND LAND USE*, 3-21.

Shaw, J. (2020). Platform Real Estate: theory and practice of new urban real estate markets. *Urban Geography*, 1037-1064.

Short, J. R. (1978). Residential mobility. *Progress in Human Geography*, 419-447.

Sing, T. F., & Zou, Y. (2021). Improving Information Efficiency in Housing Markets Using Online Listings. *SSRN*, http://dx.doi.org/10.2139/ssrn.3772412.

Song, A., Szerb, L., Audretsch, D., Komlósi, É., & Acs, Z. (2021). The evolution of the global digital platform economy: 1971 - 2021. *Small Business Economy*, 1629 - 1659.

Steegmans, J., & de Bruin, J. (2021, Februati 22). *Online housing search dataset: Information flows of real estate platform users*. Retrieved from https://data.mendeley.com/: https://data.mendeley.com/datasets/rtn8t47t6j/1

Steegmans, J., & de Bruin, J. (2021, Februari 22). *Online housing search dataset: Information flows of real estate platform users*. Retrieved from Mendeley Data: https://data.mendeley.com/datasets/rtn8t47t6j/1

Steegmans, J., & de Bruin, J. (2021). Online housing search: A gravity model approach. *PLoS ONE*, 1-31.

Steegmans, J., & De Bruin, J. (2021a, Februari 22). *Online housing search dataset: Information flows of real estate platform users*. Retrieved from Mendeley Data: https://data.mendeley.com/datasets/rtn8t47t6j/1

Steegmans, J., & De Bruin, J. (2021b). Online housing search: A gravity model approach. *PLoS ONE*, 1-31.

Steegmans, J., & Hassink, W. (2018). Decreasing house prices and household mobility: An empirical study on loss aversion and negative equity. *Journal of Regional Science*, 611-634.

Toshkov, D. (2016). *RESEARCH DESIGN IN POLITICAL SCIENCE*. London: PALGRAVE.

Vyvere, Y. v., Oppewal, H., & Timmermans, H. (1998). The Validity of Hierarchical Information Integration Choice Experiments to Model Residential Preference and Choice. *Geographical Analysis*, 254-272.

Wheaton, W. C. (1990). Vacancy, Search and Prices in a Housing Market Matching Model. *Journal of Political Economy*, 1270-1292.

Williams, J. T. (1995). Pricing Real Assets with Costly Search. *The Review of Financial Studies*, 55-90.

Zhang, W., Chen, S., Guo, D., & Li, B. (2019). The Impact of Internet Real Estate Intermediary Platform on the Real Estate Market. *Proceedings of the 4th International Conference on Crowd Science and Engineering*, 132-139.

Zhang, X., Zhang, J., Zhang, Y., Zheng, Y., & Lin, Z. (2021). Online property brokerage platform and prices of second-hand houses: Evidence from Lianjia's entry. *Electronic Commerce Research and Applications*, https://doi.org/10.1016/j.elerap.2021.101108.

Zhou, B., & Kockelman, K. M. (2008). Microsimulation of Residential Land Development and Household Location Choices. *Journal of the Transportation Research Board*, 106-112.

Appendix

Origin municipalities in CBS dataset

Table 14: List of origin municipalities that are in the CBS dataset

Participating origin municipalities	#	Participating origin municipalities	#
's-Gravenhage	1	Leeuwarden	40
's-Hertogenbosch	2	Leiden	41
Aa en Hunze	3	Leidschendam-Voorburg	42
Almere	4	Lelystad	43
Amersfoort	5	Maassluis	44
Apeldoorn	6	Maastricht	45
Arnhem	7	Meierijstad	46
Assen	8	Meppel	47
Bernheze	9	Midden-Drenthe	48
Best	10	Mill en Sint Hubert	49
Bodegraven-Reeuwijk	11	Nijmegen	50
Boekel	12	Nissewaard	51
Borger-Odoorn	13	Noordenveld	52
Boxmeer	14	Nuenen, Gerwen en Nederwetten	53
Boxtel	15	Oirschot	54
Breda	16	Oosterhout	55
Capelle aan den IJssel	17	Oss	56
Coevorden	18	Rotterdam	57
Cuijk	19	Schagen	58
Culemborg	20	Schiedam	59
De Wolden	21	Sint-Michielsgestel	60
Den Helder	22	Sint Anthonis	61
Deventer	23	Son en Breugel	62
Ede	24	Tilburg	63
Eindhoven	25	Tynaarlo	64
Emmen	26	Uden	65
Etten-Leur	27	Utrecht	66
Geldrop-Mierlo	28	Veldhoven	67
Gouda	29	Velsen	68
Grave	30	Venlo	69
Groningen	31	Vught	70
Haarlemmermeer	32	Waalre	71
Hardenberg	33	Waalwijk	72
Helmond	34	Waddinxveen	73
Heusden	35	Westerveld	74
Hoogeveen	36	Zaanstad	75
Houten	37	Zoetermeer	76
Krimpenerwaard	38	Zuidplas	77
Landerd	39		

Missing municipalities in CBS dataset (only in Funda dataset)

Table 15: List of total municipalities that are only in the Funda dataset

Municipalities only in Funda data	
1. Aalburg	2. Marum
3. Bedum	4. Menameradiel
5. Bellingwedde	6. Menterwolde
7. Binnenmaas	8. Molenwaard
9. Cromstrijen	10. Neerijnen
11. De Marne	12. Noordwijkerhout
13. Dongeradeel	14. Nuth
15. Eemsmond	16. Onderbanken
17. Ferwerderadiel	18. Oud-Beijerland
19. Franekeradeel	20. Rijnwaarden
21. Geldermalsen	22. Schinnen
23. Giessenlanden	24. Slochteren
25. Grootegast	26. Strijen
27. Haarlemmerliede en Spaarnwoude	28. Ten Boer
29. Haren	30. Vianen
31. Hoogezand-Sappemeer	32. Vlagtwedde
33. Kollumerland en Nieuwkruisland	34. Werkendam
35. Korendijk	36. Winsum
37. Leek	38. Woudrichem
39. Leerdam	40. Zederik
41. Leeuwarderadeel	42. Zuidhorn
43. Lingewaal	44. het Bildt
45. Littenseradiel	

Missing municipalities in Funda dataset (only in CBS dataset)

Table 16: List of total municipalities that are only in the CBS dataset

Municipalties only in CBS dataset
Altena
Beekdaelen
Het Hogeland
Hoeksche Waard
Midden-Groningen
Molenlanden
Noardeast-Fryslân
Vijfheerenlanden
Waadhoeke
West Betuwe
Westerkwartier
Westerwolde

Control variables details

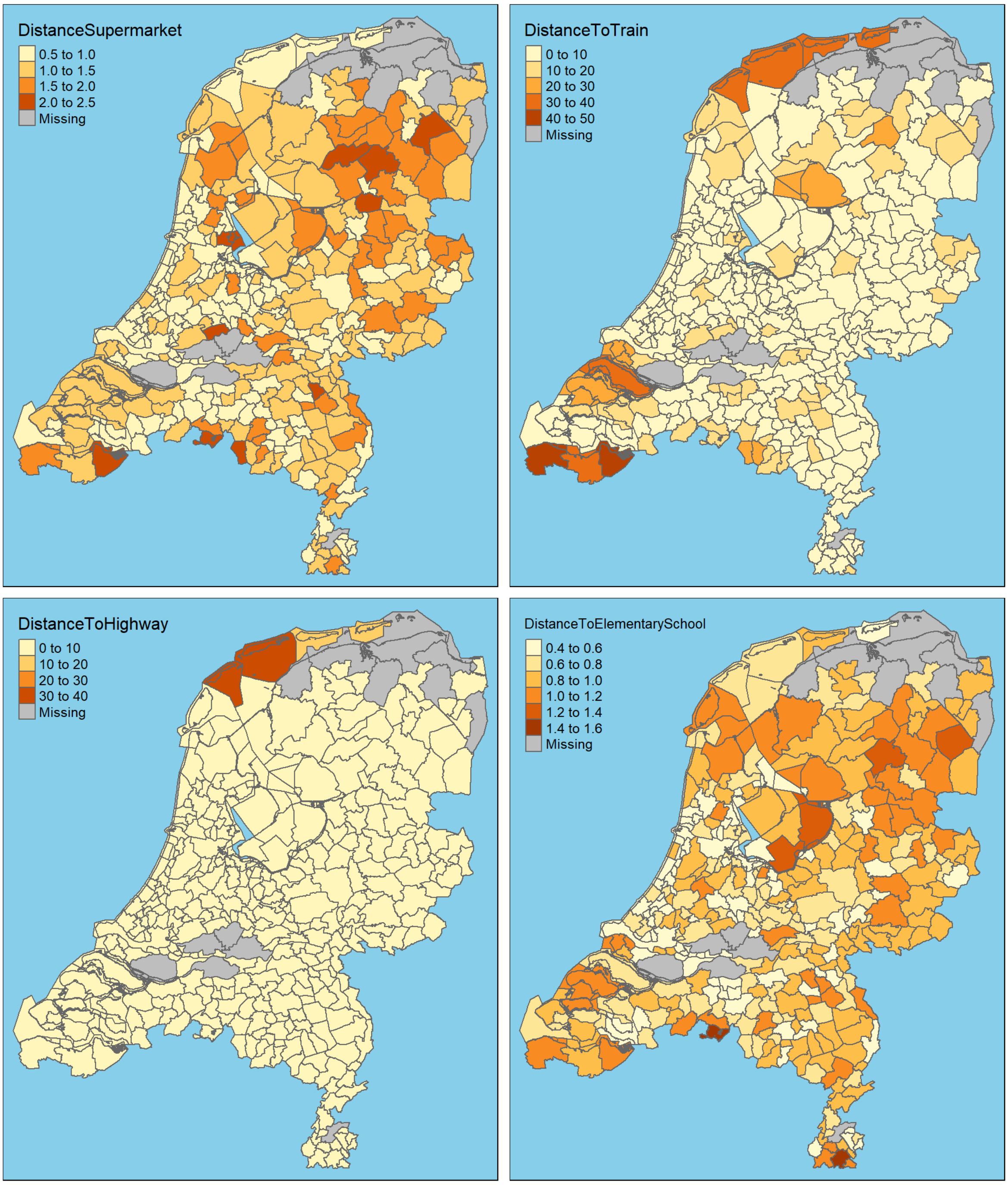


Figure 4: Mean distance to control variables

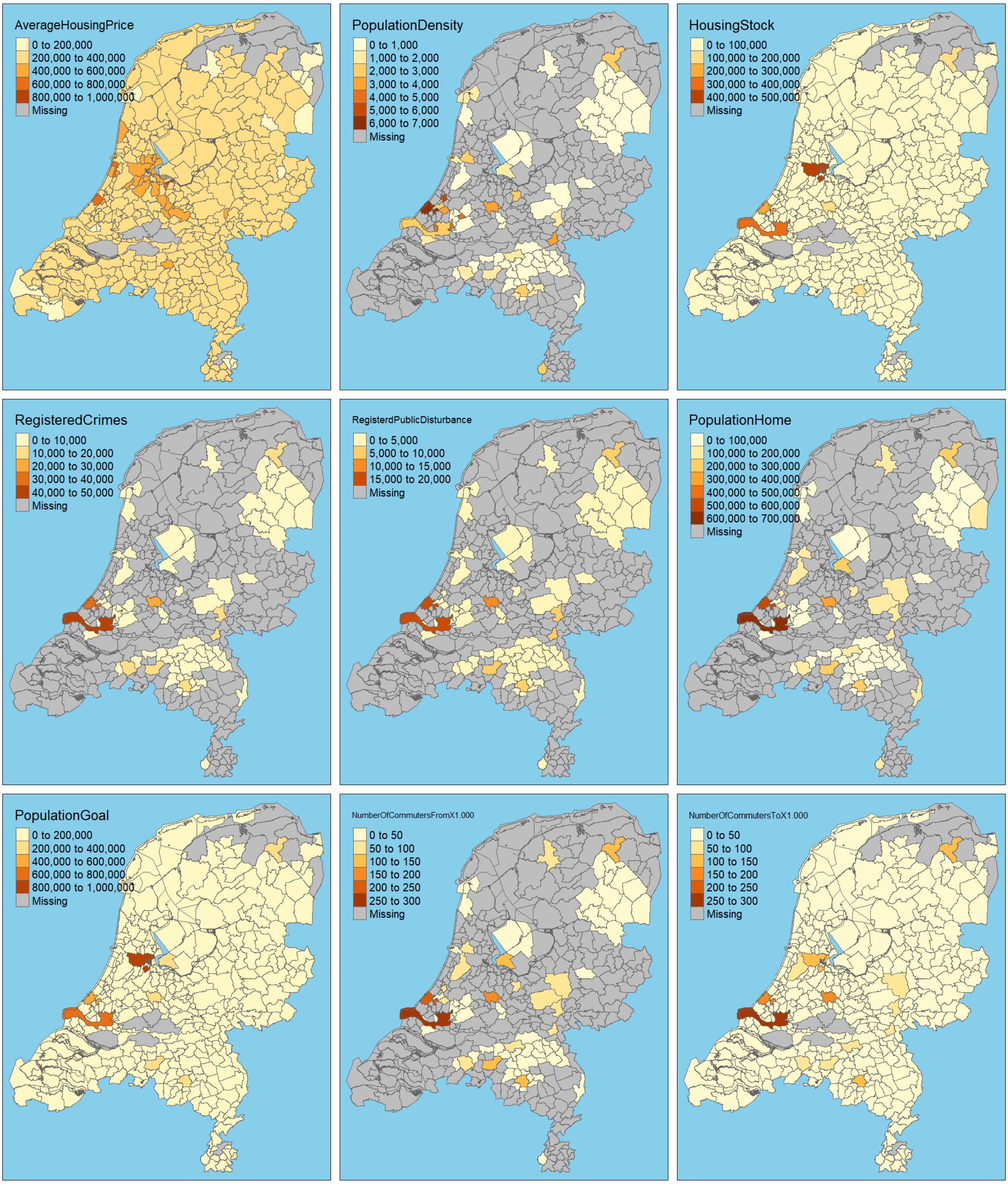


Figure 5: Absolute number control variables