Singles in the City: The association between single-person households and property transaction prices

Margreet Wiersma
19-01-2020

COLOFON

Title	Singles in the City: The association between single-person households and property transaction prices
Version	2
Author	Margreet Wiersma
Student number	S2698684
E-mail	m.wiersma.11@student.rug.nl
Supervisor	M. van Duijn

Disclaimer: "Master theses are preliminary materials to stimulate discussion and critical comment. The analysis and conclusions set forth are those of the author and do not indicate concurrence by the supervisor or research staff."

ABSTRACT

In the past years, the share of single-person households in Europe has been increasing and is expected to keep growing in the coming years. This thesis examines the association between single-person households and property prices by using an instrument variable approach within a two-stage least square regression (2SLS), using 67,524 observations of property transactions in Paris metropolitan area, France in 2015. Results show a positive association between the share of single-person households and property prices, moreover the association between single-person households and apartment prices, compared to house prices is tested and discussed. These results are relevant for housing market analysts and investors and helps in understanding market factors, as well as for policy makers and city planners for planning and provision of housing.

CONTENT

1. INTRODUCTION

1.1 Motivation

In the preceding years, the average size of households has been decreasing in Europe. There are several factors attributing to this occurrence, of which, among other things, an increasing share of people living independently, lower fertility rates, more divorces and less households living with extended family (Eurostat, 2015). As a result of this trend, the number of people living alone has increased in the past years. In European countries the share of single-person households, compared to other types of households, was the fastest growing group between 2008 and 2018, making it the most common type of household and making up one-third of the total number of households (Eurostat, 2019). Broadening the view, similar trends can also be seen outside Europe. Comparable developments are also present in the United States and Asia. Almost a third of all people in the United States lived alone in 2016, compared to a little more than 17 per cent in 1970 (U.S. Census Bureau News, 2017). Yeung and Cheung (2015) state that in general the group of single-person households in Asia is lower than in Europe and the United states, but that this group is growing faster, especially in the economically more developed societies in East-Asia.

This research is located in Paris, France. As an illustration, in metropolitan France; the share that represents single-person households in all types of households arrangements has increased from 19,6 per cent in 1962 up to 35,1 per cent in 2014 (Ined, 2019). The share is projected to reach up to 43 to 46 per cent in 2030 (INSEE, 2006). Paris is an interesting city for this research because it has a high share of people living alone. In 2011, there were five regions in the EU that had a share of single person households that were over 50 percent of the total number of households, four were German cities and the fifth region was Paris (Eurostat, 2015). Paris can be seen as an extreme case, where many people live independently, overshadowing conventional family types, like couples (with children) (Ogden and Schnoebelen, 2005).

A changing structure in family and household-types induces a change in society and policy. Single-person households use relatively more floor space and spend relatively more on household goods compared to larger households (Williams, 2007). An increasing group of single-person households would thus implicate an increase in demand for floor space, which in turn increases demand for housing. Moreover, single-person households can be expected to have different requirements for their housing than families do, so that certain dwelling types might become more popular. In addition, considered that single-person households have a sole income provider and cannot rely on income of a partner, property-owning possibilities might become more limited for a larger share of people. An increase in single-person households might thus mean an increase in demand for more affordable property types, like apartments and small houses. However, there is some adjustment time before supply catches up with an increase in demand. This is the result of several issues on the supply side, like the availability of suitable site, the duration and complexity and difficulty of the planning process, the time it takes to construct new properties and the difficulty in providing the right infrastructure and the cost of preparing undeveloped land for construction (Hsieh *et al.*, 2012). These factors are all of influence on policy for planning and provision of housing.

In line with the previous, Kohler and Van Der Merwe (2015) suggest that a declining household size could potentially affect housing price growth. Tyvimaa and Kamruzzaman (2019) tested this hypothesis in Helsinki, Finland. Their study results in a model that shows that a 1 per cent increase of young, single person households increase apartment prices by 0.51 per cent. Given the limited property-owning possibilities, due to a single income, this might lead to problems for finding suitable housing for single-person households.

According to Tyvimaa and Kamruzzaman (2019) there is not much research done into the effect of the increasing number of single-person households on property prices.

Endogeneity issues are present, as a response to this problem Tyvimaa and Kamruzzaman (2019) propose an instrument variables approach within two-stage least squares (2SLS). This research will follow, supplement and expand the research of Tyvimaa and Kamruzzaman (2019) by testing if the same results hold in Paris, France. Adding to previous research, this research will also try to find if differences among dwelling types are of influence on the association between the share of single-person households and apartment transaction prices.

1.2 Research problem statement

The aim of this study is to explore the association between single-person households and transaction prices for houses and apartments.

The main question that will be answered during the research is as follows: *'What is the association between single-person households and housing transaction prices?'* To help answering this question, three sub questions have been set up. Each question and research approach are explained in this section.

- *'What tells theory about the association between the share of single-person households and housing transaction prices?'*

The first sub question focusses on existing theoretical knowledge of the subject. It will be used for describing and comparing yet researched issues, therefor using a qualitative approach. This will be done using existing academic literature.

- *'What is the association between single-person households and housing transaction prices?'*

This question focuses on testing if single-person households are associated with transaction prices. As the aim here is to measure an association between two variables, a regression model is used, where apartment and house prices will be used as a dependent variable and share of single-person households as independent variable.

- *'How does housing type influence the association between single-person households and housing transaction prices?'*

The third sub question focusses on differences between housing types om relation to the share of single-person households. To answer this question the data will be divided into two subgroups, apartments and houses. For each subgroup a different regression will be run.

1.3 Thesis Outline

The remainder of this thesis is structured as following: chapter 2 contains a theory section in which the theoretical background of this research will be explained. The data collection and empirical approach follows in chapter 3. In chapter 4 the results are presented, followed by a conclusion and recommendations for further research in chapter 5.

2. THEORY

2.1 Transaction Price

In a general sense, a household can spend their income on two goods, housing and non-housing goods. The amount of money a household is willing to spend on a certain set of housing characteristics, while keeping the same utility level is represented by the bid-rent curve (Gross, 1988). In his research Gross (1988) estimated this willingness to pay using a discrete-choice (housing/ non-housing) bid-rent framework, where each household has its own bid-rent curve. The bid-rent curve shows the amount that a household is willing to pay for a combination of several housing characteristics, while keeping the same utility level. A household gets its maximum utility level at the point where the bid-rent curve touches the hedonic function of housing characteristics p(Z) (Gross, 1988). In figure 1 the bid-rent curves (B(z)) of household 1 and 2 are depicted. Utility maximizing amounts of attribute z are found at z_1^1 and z_2^2.

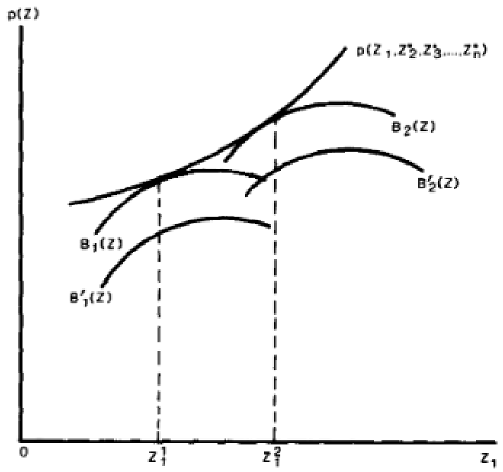

Figure 1. Household bid-rent functions for Attribute 1 (Gross, 1988).

In the owner occupier market, there are a few factors like the stock of single-family homes, the number of households and the income of households that are important in determining how much a household is willing to pay for housing. This willingness-to-pay can be depicted as an annual returning payment related to the cost of occupying a space, equivalent to 'rent' of a property (DiPasquale and Wheaton, 1992). The demand of space for households depends on the rent associated with occupying a space, relative to the cost of consuming other non-housing goods, as well as the income a household has available (DiPasquale and Wheaton, 1992).

Rent of properties is determined in the property market for space (DiPasquale and Wheaton, 1992). The four-quadrant model of DiPasquale and Wheaton (1992) represents the real estate market. The model is depicted in figure 2. When real estate is owner-occupied, there is no separation between asset and property markets. Nevertheless, the conditions of the capital market are of importance as, for example, interest rates influence on the property price a household can afford with a certain amount of rent.

The demand of space is derived from the demand of the occupiers, however the total demand for space is also dependent on other exogenous economic factors, like production levels, income and the number of households (DiPasquale and Wheaton, 1992).

The four-quadrant model in figure 2 shows the market in equilibrium. The upper-right quadrant shows the demand curve based on the willingness-to-pay. With a given, temporarily stable, housing stock this leads to a rent level. The upper-left quadrant converts this given rent level into the price actually paid for a property. This purchase price directs new construction, as can be seen in the lower-left quadrant. The curve in this quadrant represents the replacement costs of real estate. In the last quadrant (lower-right), new construction is converted in a new stock of real estate (DiPasquale and Wheaton, 1992).

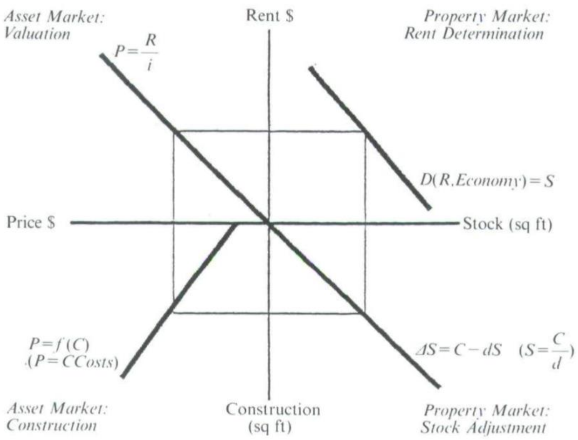

Figure 2 Four-Quadrant model (DiPasquale and Wheaton, 1992)

When factors in the model change, the equilibrium moves. When the number of households increases the demand for space increases. In the short run the stock level of real estate is stable. When all else stays equal, an increased demand will cause rents to rise, leading to increased property prices.

Eventually this will lead to new stock. However, in the new equilibrium rents will have risen and are above the previous level. This is represented in figure 3.

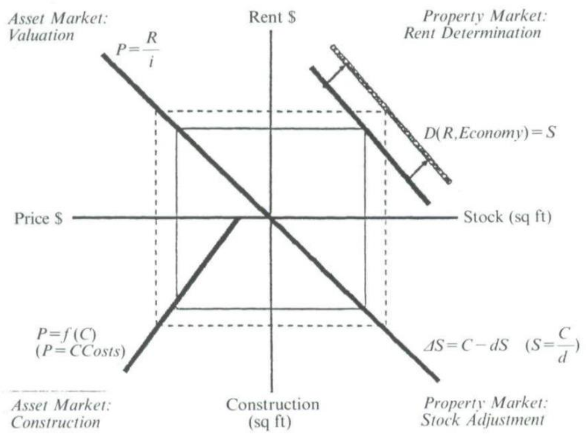

Figure 3 New Equilibrium Four-Quadrant Model (DiPasquale and Wheaton, 1992)

According to this theory, the following can be argued: When more people decide to form a single-person household, the number of households will increase. In addition single-person households use relatively more floor space (Williams, 2007). These two factors cause an increasing demand for residential space in the city. The market will transform this increased demand in increased property prices.

2.2 Singles and demand for housing

A single person has to make different decisions about living and housing than a married couple or a person with children, as the latter two formations set expectations of forming an independent family household. Single persons, however, have the choice of living with their family (remaining in the family home, or together with brothers or sisters), living with their friends or living alone (Santi, 1988). Living alone can be a choice, but it can also emerge from circumstances like divorcing or passing away from a partner.

Hall *et al*. (1997) argue that there are three factors that cause people to live alone more often. Firstly, there are compositional factors, reflecting the increasing number of older people in society, and less people that are married or have children. The second factor comprises a changing propensity to live alone. Young people are more often choosing to live alone, whereas older people are more likely to have no alternative. Thirdly the ability to live alone, where people choosing to live alone are in a more favorable position than people who are forced to live alone due to divorce or passing away of partners.

When deciding to live alone and choosing for property-ownership some affordability issues might be present. Even though homeownership is often positively associated with quality of life (Rohe and Stegman, 1994; Elsinga and Hoekstra, 2005) not everyone can afford buying a house (Withers, 1998). Relatively seen, singles have a lower purchasing power than couples, taking into account that they are the only person generating income for a household (Quintano and D'Agostino, 2006)**.** Nevertheless, single-person households need to pay for similar expenses as a couple, like kitchen utilities (oven, coffee-machine, dishwasher), bathroom, internet and tv (Tyvimaa and Kamruzzaman, 2019). This will lead to less disposable income for single-person households to spend on a property, thus leaving them to choose for more affordable properties.

Wulff (2001) finds that people living alone are more likely to live in flats or units, rather than separate dwellings. This preference is not necessarily linked to affordability only but could also be influenced by less maintenance necessary for units compared to detached houses, the view of being less isolated in a flat than in a detached dwelling. Moreover, flats are often centrally located, close to other amenities, making them more attractive. These preferences will have influence on the bid-rent curve for single-person households and consequently lead to an increase in demand for affordable apartments and houses suitable for single-person households.

However, it is found differences across types of singles lead to differences in housing preferences. Faessen (2002) studied three groups of single-person households; single-person households under 35, middle-aged single-person households and singles of 65 years and over, and found significant differences between housing preferences of these groups. People over 65 were mostly looking for multi-family buildings, whereas younger single households were preferring single-family units. In addition to these differences between highly and less urbanized areas were found.

Likewise, Wulff, (2001) divides single-person households in groups of separate life stages based on age. The supposed life stages in which a person can form a single-person household are: below 30 years (living alone before becoming a couple), 30-44 years (postponing marriage or divorce), 45-59 (mainly divorce and separation), and 60 years and older (mainly widowhood). Again, it was found that housing demand amongst single-person households was not uniform but varied across age groups. Nevertheless, single-person households are four time as likely to live in flats or units, compared to other types of households. Furthermore, differences between owner-occupier and renters and across income groups were found. Not only life-course stage was considered important, but also the timing of forming a household and anticipated time of living alone.

From theory it is expected that an increase in single-person households leads to an increase in demand for affordable apartments and houses that are fitted for single-person households. Moreover, it can be argued that single-person households should not be treated as one group, but should be seen in separate sub-groups, each with different wishes and demands for housing, with a general preference for units and apartments. It is thus expected that the association between single-person households and apartment transaction prices is stronger than the association between single-person house transaction prices.

2.3 Other determinants of transaction prices

Rosen (1974) describes that goods can be valued for their utility bearing characteristics, where hedonic prices are the implicit prices of these characteristics, that can be observed from observed prices (in relation to their amounts) of differentiated products. The house price is determined by all sets of individual characteristics.

There are many different characteristics that can be thought of. Research by Selim (2009) found that, among other things, type of house, number of rooms, house size and locational characteristics were most significant in determining house prices.

Locational characteristics can include different kinds of characteristics. For instance, research by McLeod (1984) uses the hedonic price theory on property characteristics and local amenities. In his research he finds that river views are a well valued characteristic, but that also locational characteristics like proximity to a local park or highway are of importance. There is more evidence for the importance of attractive spatial attributes. An example is the effect that natural space has on nearby property prices, as researched by Daams *et al*. (2016) who find that property buyers are willing to pay a price premium when properties are within 7 kilometer of attractive natural space.

Deducing from theory it can be expected that locational characteristics will be of influence on market price. The locational characteristics can be viewed in relation to proximity to practicalities, like highways and economic focus points, but also in relation to aesthetic appeal, like the presence of views and proximity to parks and natural space. It can be expected that when an apartment is closer to different kind of amenities, property value will increase.

However, the effect of locational characteristics is not limited to their location but can be extended to the effect of locational demographics. Past research has brought evidence that a change in demographics can lead to a change in demand and pricing of housing. Mankiw and Weil (1989) argue that a change in the demographic composition of a population can lead to change in demand for housing, which in turn creates a change in the price for housing. They state that demographics in terms of age, income and a variety of other household characteristics are important in the amount of housing that is demanded. They use age as only variable to represent the function of demand from households, looking at the size of different age groups in the population. Their findings support that a change in the number of births lead to changes in housing demand, as the number of births influence the age structure of a population. Also other factors of the demographic composition of the neighborhood seem to be of importance. Gibbons (2003) finds that the educational composition of a neighborhood influences the amount families are willing to spend to live in a neighborhood. He finds that prices increasing when the proportion of higher educated residents in a neighborhood is higher than the regional mean. This is effect is measured by using the hedonic property price model, where the educational composition is seen as an implicit price effect.

Thus, it can be argued that a changing demographic composition is of influence on house prices. On one hand, is the effect of changing demographics that change demand. A growing group of singles will increase the demand for housing, as singles use relatively more space than couples and other types of households. This increasing demand is in turn expected to increase house prices. On the other hand, the neighborhood composition will also be of importance. Households are willing to pay a price premium if the composition of a neighborhood is valued. Singles might be looking for neighborhoods where like-minded people are living and willing to pay a premium for this.

2.4 Hypotheses

This thesis aims to find the association between the share of single-person households and housing transaction prices. In order to describe this association, hypothesis based on found theory were set up and are empirically tested.

Based on the findings in theory that an increasing number of households influence on transaction prices, the same is expected for an increasing share of single person households (DiPasquale and Wheaton, 1992; Kohler and Van Der Merwe, 2015; Tyvimaa and Kamruzzaman, 2019)

Hypothesis 1: The share of single-person households in the city has a positive association with housing transaction prices.

In addition, this research aims to find differences between different housing types. As singles are found to have preference for apartments (Wulff, 2001), it is expected that the association is stronger between single-person households and apartment prices, compared to houses. This leads to the second hypothesis.

Hypothesis 2: The type of housing has an influence on the association between single-person households and transaction prices.

3. DATA & METHOD

3.1 Study Area

This research focuses on apartment and house purchases in the city of Paris (department 75) and the inner ring around Paris (departments 92, 93, 94). France is an interesting country for this study, as most households here were composed by just one or two people in 2015. The amount of people living individually in France raised from 14.9 per cent in 2007 to 16,4 per cent in 2017. The biggest share of people living alone were people over 65 years. However, the share of young single person households is increasing as well. Almost 20 per cent of French people between the age of 20 and 24 years old were living individually in 2015. The growth of this group can be attributed to extension of study time and obtaining a job at a later age (Statista Research Department, 2019).

France can be seen as a representative country for Europe, as it has had the similar demographic developments influencing household structures as the rest of Europe since 1960. These developments include an increasing number of divorces, delayed marriage, lower fertility rate and an ageing population (Hall *et al.*, 1997).

Paris, in particular, is interesting for its high share of people living alone. The share of people living alone is dominating the more conventional family types like couples and families with children (Ogden and Schnoebelen, 2005).

3.2 Property Transaction Data

Property transaction data was obtained from https://www.data.gouv.fr/fr/datasets/demandes-de-valeurs-foncieres-geolocalisees//, where datasets from a certified public service on geolocated property prices can be found. The dataset is derived from the 'Property Value Requests' dataset, which is released by DGFiP (Direction générale des Finances publiques, which is a department of the French central public administration). The dataset contains, next to property price, the date of the transaction and basic information about the property, like number of main rooms, surface of the building and X- and Y-coordinates of location of the property. The transaction price is based on the declared amount in the transfer, including VAT. The surface of the building is the surface measured on the floor between the walls. This is the sum of the actual surface area of the room and the surface area of the outbuildings. It is possible to download the property values of all over France from 2014-2018. It was chosen to work with property transaction data of only 2015, due to data on social and neighborhood characteristics being available for only this year.

3.3 Socio-economic and locational data

Information about people in the neighborhood surrounding the property were obtained from APUR (Paris Urban Agency) who have a data collection on data on Paris and the Greater Paris (Metropolitan Area) available on an open-data platform http://opendata.apur.org/. These datasets contain information of the population census, which is provided by l'INSEE, the French National Institute of Statistics and Economic Studies, and provides data on five main themes: population, household and family, housing, education and employment. INSEE developed a system for statistical purposes, for which the country is divided into units of equal size, where each basic unit contains a target size of 2000 residents. The system is called IRIS, a French acronym for 'aggregated units for statistical information'. Each IRIS area contains a population between 1800 and 5000, bordered by roads, railways and water (INSEE, 2016). The division in units allows for analysing at a small scale, where all units are having a similar size, which makes it relatively easy to make comparisons across units.

APUR has datasets and maps available containing socio-economic data about the IRIS units in the Paris area. A map with data on households in Parisian IRIS areas was used from https://carto2.apur.org/apur/rest/services/OPENDATA/RECENSEMENT_IRIS/MapServer/1, which provides data on households and their composition. It also contains the share of single-person households per IRIS area. The share of single-person households is obtained by dividing the number of households composed of one single person by all households. This data is visually displayed in Figure 4.

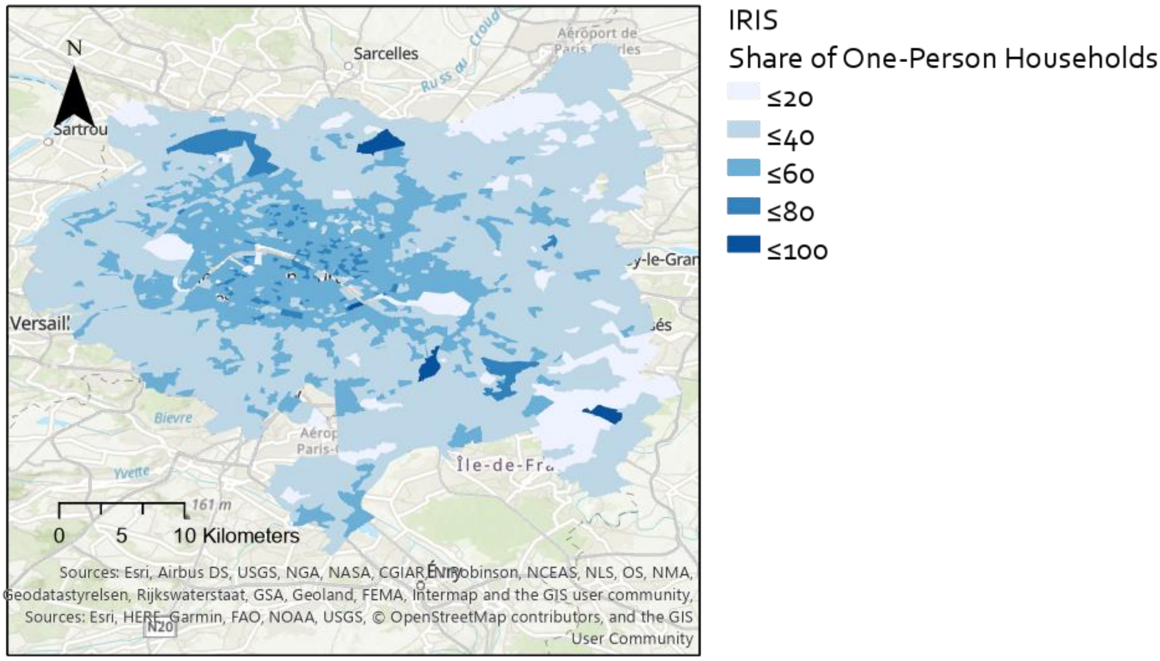

Figure 4: Share of Single-person Households 2015. The map shows the share in the city of Paris (department 75) and the inner ring around Paris (departments 92, 93, 94), This figure is based on the APUR-dataset.

In order to perform an analysis on the association between the single person households and property transaction prices, both datasets need to be combined. As both datasets contain geographic reference attributes, they were joined using the SpatialJoin operation in ArcGisPro.

In order to obtain more information on the socio-economic characteristics of the neighborhood, also data on three other themes; population, housing, education and employment were added to the dataset, by conducting a Merge operation in Stata/SE 15.0. The datasets can be found at: http://opendata.apur.org/datasets/recensement-iris-population; http://opendata.apur.org/datasets/recensement-iris-logement; http://opendata.apur.org/datasets/recensement-iris-formation.

In order to include locational characteristics, the distance to stations was measured using the Near operation in ArcGisPro. The stations were based on a map including stations and rail transport stations in Île-de-France (metro, bus, tramway, train, shuttle, RER and TER), which can be found at https://services.arcgis.com/d3voDfTFbHOCRwVR/arcgis/rest/services/emplacement_des_gares_idf/FeatureServer. The map is based on the year 2018, but as there have not been many changes since the year 2015, the map is considered to be useful.

3.4 Data Selection

The original dataset of the year 2015 of all over France consisted of 2,749,830 observations. After selecting the city of Paris (department 75) and the inner ring around Paris (departments 92, 93, 94) 2,566,414 observations were deleted.

Different types of properties are represented in the dataset; apartments, dependances (out-buildings, building located on the property separate from the main building), industrial commercial or similar property premises, and houses. In this research only apartments and houses are researched. 79,791 observations were dropped when leaving observations for dependances and industrial properties out. In order to be able to analyze the property, its location needs to be known. Therefore, each observation needs to have an X- and Y-coordinate, when dropping missing values in these variables, 1,569 observations were deleted. Each transaction has its own identification number. If a property consists of more rooms, each room gets a new transaction with the same observation number. After removing these duplicates based on the identification number, another 21,452 observations were deleted. Observations with less than 1 or more than 6 rooms were deleted, removing another 1,374 observations. The top and bottom 1 per cent of the transaction price were deleted, cutting out 1,575 observations, leaving a total of 77,655 observations.

Lastly, missing values in the control variables were found, after deleting these a dataset of 67,524 observations remained.

3.5 Data Limitations

The data is somewhat limited in the sense that the information on the share of singles is only available for the year 2015, so that an analysis over more than one year cannot be done.

Moreover, the data available on the property types is deficient of characteristics. Only the surface of the property and the number of main rooms is known, but nothing is known about the condition of the property and or different amenities within each house.

There is nothing known about the property buyers. Buyer characteristics, like gender and age, are unknown as well as the intended use. The property might be used for owner-occupying, but it could also be used for renting out.

Lastly, only the distance to public transport is calculated, but distance to other amenities is not taken into. Nevertheless, the distance to public transport also incorporates same other amenities. In previous research it was found that a new subway station contributed to the amount of openings and variety of restaurants in the neighborhood (Zheng *et al.*, 2016), similarly it can be expected that close proximity to a station also assumes close proximity to other amenities.

3.6 Descriptive Analysis

All transactions that are in the dataset are pictured in figure 5. It can be seen that the inner-city of Paris almost all transactions were of apartments, whereas towards the outer city more houses were sold.

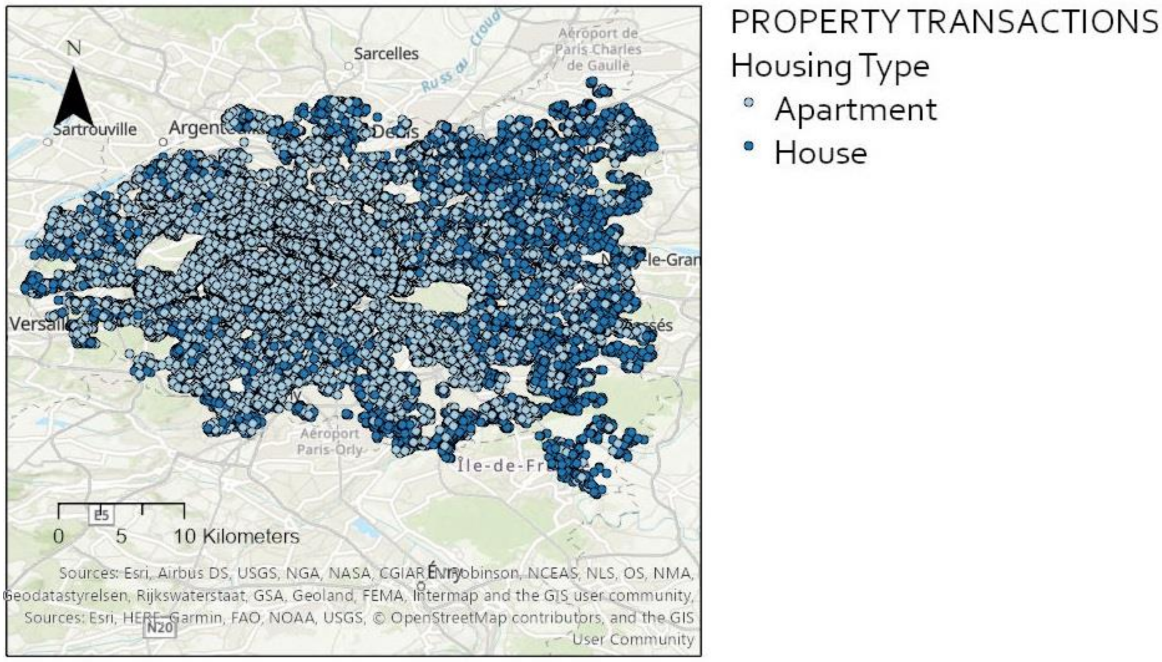

Figure 5: Property Transactions in 2015. The map shows the property transaction in the city of Paris (department 75) and the inner ring around Paris (departments 92, 93, 94), This figure is based on the DGFiP-dataset.

The descriptive statistics are shown in table 1. The complete tables can be found in Appendix B. A definition of the variables can be found in Appendix A.

The datasets consist of transactions of apartments, as well as houses, however the amount of house transactions is notably smaller than apartment transactions, 7,698 and 59,826 transactions respectively. This can be related to the knowledge that capital cities in Europe have the highest share of flats and the lowest share of houses in the total number of dwellings. The share of flats represented 99 percent of all dwelling types in Paris in 2012, leaving just one per cent for houses, business and commercial property (Eurostat, 2016). From this perspective it is logical that there are way more transactions in apartments than that there are in houses.

Property transactions range from €22000 up to €1550000, with a mean of €330729.88. The mean of property transaction prices is lower for apartments (€320849.95), than it is for houses (€407565.69). As a house usually consists of a larger amount of rooms and has a bigger surface, this is not surprising.
The share of singles has a minimum of 11.126 per cent and a maximum of 70.651 per cent, with an average of 42.631. It is remarkable that the average share of singles is lower when looking at apartments only, compared to apartments. This could be due to the fact that singles have a, relatively seen, lower purchasing power compared to couples (Quintano and D'Agostino, 2006), as apartments have a lower purchasing price. However, it could also be related to a preference for flats of single-person households. As described earlier, flats can be viewed as less isolated than a detached dwelling. In addition to this, flats are more often central located and closer to other amenities (Wulff, 2001). Moreover, this might imply that the share of singles is higher in the inner city (where most apartment transactions were) compared to the outer city.

Table 1: Descriptive Statistics

	Apartments and Houses			Apartments Only			Houses Only		
Variables	Mean (Std. Dev)	Min	Max	Mean (Std. Dev.)	Min	Max	Mean (Std. Dev.)	Min	Max
Price	330,729.88 (229,960.89)	22,000	1,550,000	320,849.95 (227,452.41)	22,000	1,550,000	407,565.69 (234,839.42)	22,877	1,55,0000
Singles	42.631 (10.502)	11.126	70.651	44.006 (9.991)	11.126	70.651	31.936 (7.92)	11.126	67.863
Surface	57.768 (29.882)	4	370	53.791 (26.779)	4	370	88.692 (34.367)	6	280
Rooms	2.7 (1.208)	1	6	2.532 (1.11)	1	6	4.01 (1.136)	1	6
Density	236.957 (168.906)	1.611	1333.624	256.046 (168.737)	1.611	1333.624	88.505 (64.043)	1.611	701.254
No Diploma	21.876 (9.672)	6.94	64.761	21.125 (9.342)	6.94	64.761	27.714 (10.194)	7.518	64.761
Families with Children	30.377 (9.799)	0	72.077	29.222 (9.498)	0	72.077	39.352 (7.088)	14.177	70.526
Principal Residences	88.373 (6.183)	40.744	98.822	87.839 (6.285)	40.744	98.822	92.529 (2.977)	73.756	98.822
Occasional Dwellings	4.333 (4.784)	.059	43.42	4.711 (4.929)	.059	43.42	1.397 (1.49)	.059	18.043
Vacant Dwellings	7.294 (2.842)	.092	27.53	7.45 (2.859)	.092	27.53	6.074 (2.373)	.092	17.369
Owner-Occupied	43.76 (15.789)	.162	91.066	41.97 (14.673)	.162	91.066	57.667 (17.188)	2.187	91.066
Private Tenants	36.597 (14.121)	1.198	88.135	38.208 (13.674)	1.198	88.135	24.071 (10.906)	1.649	72.367
Social Tenants	16.181 (16.635)	.067	97.546	16.209 (16.653)	.067	97.546	15.966 (16.499)	.067	90.882
Free Housing	3.463 (2.507)	.105	66.218	3.613 (2.49)	.105	66.218	2.296 (2.323)	.105	66.218
Public Transport	533.142 (508.453)	9.226	6357.073	479.06 (445.171)	9.226	5745.646	953.735 (726.81)	34.202	6357.073
	N = 67,524			N = 59,826			N = 7,698		

The average building was 57.768 m^2, with 2.7 rooms. The share of principal residences in the neighborhood was on average 88.373 per cent, and the share of vacancy had a mean of 7.294 per cent. On average, 43.76 per cent of properties were owner-occupied, 16.181 per cent tenant households were housed in Social Housing, 3,463 per cent of households were housed free and the average distance to public transport was 533.142 meter. After doing a multicollinearity analysis, using a correlation table and calculating variance inflation factors, share of families with children, the share of occasional dwellings, share of tenants in private housing were left out, as described in Appendix E.

3.7 Analytical Strategy

The empirical research to answer the second and third research sub questions will follow the methodology taken by Tyvimaa and Kamruzzaman (2019) who researched the effect of young, single person households on apartment prices in Helsinki, Finland using an instrument variable approach within a two stage least square regression (2SLS). The latter is executed to oppose a possible endogeneity bias, due to reverse causality.

The transaction price of apartments in Paris is used as an outcome variable. The simple OLS method would regress the transaction prices in Paris on the percentage of young singles located within small areas of Paris. This follows the hedonic model. In 1974 Rosen introduced a theoretical framework supporting the hedonic pricing method. It is stated that hedonic prices are the construct of implicit prices of different attributes. By pricing several housing attributes, a total house price can be constructed, for example Sirmans *et al.* (2005, p3) denote "*A house is made up of many characteristics, all of which may affect its value. Hedonic regression analysis is typically used to estimate the marginal contribution of these individual characteristics*".

In this research characteristics of the property itself, as well as neighborhood and geographical characteristics, as described in the data section, will be used in the model. The property transaction price is seen as a sum of all these characteristics.

Due to choosing only the year 2015 to work with, the dataset is somewhat limited in controlling for time dependent influences. To overcome this shortfall somewhat, dummy variables for the quarters in the year were created, so that time of sale can be taken into account in the analysis in order to include the effect of possible time-dependent factors that influenced the property price in 2015. However, as the transaction price is rightly skewed, this variable was transformed into a log. Other independent variables were also transformed, if this made them look more normally distributed. This leads to the following equation:

$$\begin{aligned} \mathrm{LN}(Price) = \beta_0 + \ & \beta_1\, Singles_{ia} + \beta_2\, LN(Surface_i) + \beta_3\, Rooms_i + \beta_4\, Density_{ia} \\ & + \beta_5\, LN(No\ Diploma_{ia}) + \beta_7\, Principal\ Residences_{ia} + \beta_9\, Vacant\ Dwellings_{ia} \\ & + \beta_{10}\, OwnerOccupied_{ia} + \beta_{12}\, Social\ Tenants_{ia} + \beta_{13}\, Free\ Housing_{ia} \\ & + \beta_{14}\, LN(Public\ Transport_i) + D_1 \sum_{j=2}^{4} Quarter_{ij} + \gamma \sum_{f=2}^{143} Postal\ Code_{if} + \varepsilon_1 \end{aligned} \tag{1}$$

Where LN($Price$) represents the natural logarithm of transaction price of property i; $Singles_{ia}$ refers to the proportion of single-person households in IRIS area a in which the property i is located; $Surface_i$ represents the natural logarithm of surface of property i; $Rooms_i$ refers to the number of rooms of property i; $Density_{ia}$ refers to the density of the population in IRIS area a where apartment i is located; $No\ Diploma_{ia}$ is the variable for the natural logarithm of the share of the population aged 15 or over, who are out of school without a diploma in IRIS area a where property i is located; $Principal\ Residences_{ia}$ refers to the share of principal residences in IRIS area a where property i is located; $Vacant\ Dwellings_{ia}$ represents the share of vacant dwellings in IRIS area a where property i is located; $OwnerOccupied_{ia}$ refers to the share of owner-occupied households in IRIS area a where property i is located; $Social\ Tenants_{ia}$ refers to the share of tenant households in social housing in IRIS area a where property i is located; $Free\ Housing_{ia}$ represents the share of households housed free in IRIS area a where property i is located; $Public\ Transport_i$ refers to natural logarithm of the distance in meters to public transport for property i; $Quarter_{ij}$ is a dummy variable for the quarters of the year 2015; $Postal\ Code_{if}$ is a dummy variable for fixed effects based on postal code for property i; ε_1 represents the residual error.

Tyvimaa and Kamruzzaman (2019) argue that using this equation for the basic OLS method is problematic, because the dependent variable transaction prices and the percentage of singles as independent variable, are connected to each other. For one, an increasing proportion of singles might increase apartment prices, due to an increase in demand. However, singles will be more attracted to a neighborhood where apartment prices are cheaper (as a single person alone can be assumed to have less money available for housing, compared to a couple). This means that the proportion of singles and apartment prices have an influence on each other, which means that the proportion of singles can be considered to have the characteristics of an endogenous variable, and is thus related to the error term, therefore violating basic OLS assumptions. In order to confirm if this variable is indeed prone to endogeneity, a heteroskedasticity-robust version of the Hausman test is conducted. This tests the hypothesis that the instrument variables used are exogenous, if the test statistic is significant, the variable must indeed be treated as endogenous.

As response to this problem Tyvimaa and Kamruzzaman (2019) propose an instrument variables approach within two-stage least squares (2SLS). 2SLS is done in two stages. In the first stage reduced-form equations are obtained and estimated using OLS. The values found are necessary for the next step of 2SLS. In the second step endogenous variables are replaced by the fitted values that are obtained in stage one, after which the structural equations are estimated using OLS (Brooks and Tsolacos, 2010).

Instrument variables can represent the variables in the reduced form in step one. In an instrument variable approach the variables that lead to endogeneity are replaced with ‘new’ variables that are highly correlated to the original endogenous variables, but not to the error term. These variables are known as instruments The fitted values of the instruments replace the original values in the structural equation (Brooks and Tsolacos, 2010).

Instrument variables can be considered valid if they adhere to the requirement of being uncorrelated with the error term and highly correlated with endogenous regressors (Schmidheiny, 2019). This means that a variable that is exogenous and unrelated to the error term but related to the proportion of singles must be chosen. In line with Tyvimaa and Kamruzzaman (2019) the variable 'Proportion of Singles' will be replaced by a predicted proportion based on all exogenous variables and an instrument variable. The chosen instrument variable is 'Spatially Lagged Singles'. Tyvimaa and Kamruzzaman (2019) quote the First Law of Geography by Tobler (1970, p.236) "*everything is related to everything else, but near things are more related than distant things*" explaining that it is likely that the proportion of singles in an area is most probably related to singles in surrounding areas. However, it is less likely that the transaction price of a certain area is affected by the number of singles in surrounding areas. Therefore, Spatially Lagged Singles of an IRIS area is calculated as the average proportion of singles of its neighboring areas.

The calculation of this spatially lagged variable is based on the concept of a spatial weight matrix (Anselin and Rey, 2014). The spatial weights matrix (W) is a $n \times n$ matrix, which contains elements w_{ab} that represent the neighbor structure between observations, so that:

$$\boldsymbol{W} = \begin{bmatrix} \boldsymbol{w_{11}} & \cdots & \boldsymbol{w_{1n}} \\ \vdots & \ddots & \vdots \\ \boldsymbol{w_{n1}} & \cdots & \boldsymbol{w_{nn}} \end{bmatrix} \tag{2}$$

The elements w_{ab} are 0 when i and j are not neighbors and non-zero when they are neighbors. Neighbors are defined on basis of contiguity. Contiguity exists when two spatial units share a border, that has a length larger than zero (Anselin and Rey, 2014). There is a difference between rook- and queen contiguity, where in rook-contiguity neighbors need to *have* a common edge, while for queen-contiguity the neighbors only need to *share* a common edge (Anselin and Rey, 2014). In this paper queen-contiguity is used. Moreover, only neighbors of the first order are included, neighbors of neighbors are not considered. In its most simple form, the spatial weight matrix is binary, where a one indicates a neighbor and a zero indicates non-neighbors. An observation cannot be a neighbor of itself, so that the elements on the diagonal of the matrix are equal to zero, $w_{aa} = 0$ (Anselin and Rey, 2014).

Row-standardized weights are used, so that each row-sum equals to 1 ($\sum_b w_{ab} = 1$). This is calculated by:

$$w_{ab(row-standardized)} = w_{ab} / \sum_b w_{ab} \tag{3}$$

The spatial lag of singles is notated as *SpatiallyLaggedSingles*. For observation in area a, the spatial lag of *Singles*$_a$ is denoted as *SpatiallyLaggedSingles*$_a$, the spatially lagged variable is the weighted average of the neighboring values, as described in equation X.

$$SpatiallyLaggedSingles_a = \sum_{b=1}^{n} w_{a,b} * Singles_a \tag{4}$$

The weights $w_{a,b}$ consist of the elements of the a-th row of the matrix W, that are matched up with the corresponding elements of the vector Singles. This equation represents a weighted sum of the observed values for neighbouring areas, where non-neighbours are not included (the case that $w_{ab} = 0$) (Anselin and Rey, 2014). The use of row-standardized weights leads to *SpatiallyLaggedsingles* being average of the values for *Singles* at neighboring areas.

This variable was created by importing the APUR-dataset into GeoDa (a software program that is available online and can be used for free, which is designed as an easily operated and graphical introduction to spatial analysis (Anselin, Syabri and Kho, 2006), and calculating lags based on the average of the values for *Singles* at neighboring areas. See Appendix J for figures related to this calculation.

Operationalizing the instrument variable approach as described above, using *SpatiallyLaggedSingles* the following equations are established:

$$\begin{aligned}\widehat{Singles} = \beta_0 + \ & \beta_1\, Spatially\ Lagged\ Singles_{ia} + \beta_2\, LN(Surface_i) + \beta_3 \sum_{r=2}^{6} Rooms_i \\ & + \beta_4\, Density_{ia} + \beta_5\, LN(No\ Diploma_{ia}) + \beta_7\, Principal\ Residences_{ia} \\ & + \beta_9\, Vacant\ Dwellings_{ia} + \beta_{10}\, OwnerOccupied_{ia} + \beta_{12}\, Social\ Tenants_{ia} \\ & + \beta_{13}\, Free\ Housing_{ia} + \beta_{14}\, LN(Public\ Transport_{ia}) + D_1 \sum_{j=2}^{4} Quarter_{ij} + \varepsilon_2\end{aligned} \tag{5}$$

$$\begin{aligned}\mathrm{LN}(Price) = \beta_0 + \ & \beta_1\, \widehat{Singles}_{ia} + \beta_2\, LN(Surface_i) + \beta_3 \sum_{r=2}^{6} Rooms_i + \beta_4\, Density_{ia} \\ & + \beta_5\, LN(No\ Diploma_{ia}) + \beta_7\, Principal\ Residences_{ia} + \beta_9\, Vacant\ Dwellings_{ia} \\ & + \beta_{10}\, OwnerOccupied_{ia} + \beta_{12}\, Social\ Tenants_{ia} + \beta_{13}\, Free\ Housing_{ia} \\ & + \beta_{14}\, LN(Public\ Transport_{ia}) + D_1 \sum_{j=2}^{4} Quarter_{ij} + \varepsilon_3\end{aligned} \tag{6}$$

Where $\widehat{Singles}_{ia}$ is the predicted share of single-person households in IRIS area a where property i is located. $Spatially\ Lagged\ Singles_{ia}$ refers to the average share of single-person households in the neighboring areas of IRIS area a where property i is located, which was used as an instrument variable.

In general the exogeneity of instruments cannot be tested for, however the validity of instruments on basis of their correlation with endogenous regressors can be calculated after the first stage using a of a joint F-test, that tests if the excluded instruments are significantly different from zero (Schmidheiny, 2019).

Even though regression analysis incorporates the dependence of one variable on other variables, it does not necessarily suggest that there is causation between the variables (Gujarati, 2003). This thesis therefor does not argue for an effect between single-person households and transaction prices, but only explores the association between the two variables.

Standard errors are clustered as solution for spatial autocorrelation, as described in appendix F. These clusters are based on the IRIS areas, assuming that the transaction prices in each area are comparable and impacted in the same manner, as well as that the measurement on the proportion of singles is on this same area size.

4. RESULTS & DISCUSSION

This chapter will describe the results of the empirical study. It contains a presentation of most important coefficients and standard errors of the estimated models. Results on the association between the share of single-person households and transaction prices can be found in table 4. Table 6 represents the results of splitting the type of housing into apartments and houses.

In this research, the spatially lagged variable of single household was used as an instrument to address these endogeneity issues. Moreover, it was found that there was a less than 1% likelihood that clustered patterns of transaction prices could be the result of random chance, as determined by a Moran's I test, see table 2. As the Moran's I test shows clustered patterns amongst transaction prices standard errors are clustered on the IRIS-area are used.

Table 2: Moran's I Test

Variables	Moran's Index	z-score
Price	0,988247	750,667657***

*** p<0.01, ** p<0.05, * p<0.1
Note: the Moran's I Test measures spatial correlation in the dataset.

4.1 The association between the share of single-person households and housing transaction prices.

The results of testing the first hypothesis with 2SLS are shown in table 4. Three different models are used for this hypothesis. First a base-line model is run with just one independent variable, being the proportion of singles (based on the proportion on singles in neighboring areas), the second model includes all independent variables, the third model includes all independent variables and 143 fixed effects based on postal code.

The outcomes of the OLS regression can be found in Appendix G. The findings from the first 2SLS model are coinciding with the outcomes from the OLS findings of the first models, however for the second and the third model the results deviate. In the first OLS model the coefficient for the share of single-person households is 0.0101 (with a standard error of 0.000647, $p<0.01$), which has the same direction and similar magnitude of the coefficient for the share of single-person households in the 2SLS regression (coefficient: 0.0184, standard error: 0.000865, $p<0.01$). However in model 2, where more independent variables are included, the direction and significance of the coefficient for the share of single-person households are still similar but differ in magnitude between OLS and 2SLS (Model 2 OLS: coefficient: 0.00628, standard error, 0.000708, $p<0.01$ versus Model 2 2SLS: coefficient: 0.0327, standard error: 0.00286 $p<0.01$). In the third model, with all independent variables and fixed effects included the share of single-person households is not significant in the OLS regression. The magnitude of the coefficient in OLS and 2SLS regression differs here as well (Model 3 OLS: coefficient: 0.000773, standard error: 0.000487, not significant at the 10% level, versus Model 3 2SLS: coefficient 0.0234, standard error: 0.00530, $p<0.01$).

A heteroskedasticity-robust version of the Hausman test was conducted, after which the null-hypothesis that the percentage of single-person households is an exogenous variable is rejected, as can be seen from table 3. Simply using OLS regression would therefor give biased results, from which it is concluded that the use of instrument variable regression is justified.

Table 3: Heteroskedasticity-robust version of Hausman Test Model 1, 2 & 3

Model	F-statistic
Model 1	299.942***
Model 2	194.158***
Model 3	38.1974***

*** p<0.01, ** p<0.05, * p<0.1
Note: the Heteroskedasticity-robust version of Hausman Test measures if the endogenous regressors that are used in the model are in fact exogenous.

When looking at the final model including all independent variables and fixed effects, a strength test was conducted, reaffirming that the used instrument was strong (with $R^2 = 0.776$, adjusted $R^2 = 0.776$, partial $R^2 = 0.018$ and $F = 31.575$, $p = 0.000$). The F-statistic of a joint test (which tests if the excluded instruments are significantly different from zero) should be bigger than 10 in case of a single endogenous regressor, which is a rule of thumb for indicating the relevance of instruments (Schmidheiny, 2019). This is the case for the used instrument in this research ($F = 31.575$).

Continuing with the results from the regression in table 4. In the first column (1) the most basic regression is presented, including only the effect of single-person households (measured from the instrument variable Spatially Lagged Singles). The adjusted R-squared of this model is very low (0.9 per cent), meaning that only 0.9 per cent of the variability in the data can be explained by this baseline model. However, the proportion of single-person households is significantly different from 0 at the 1 per cent level and is positive, meaning that there is a positive association between single-person households and housing transaction prices.

The second column (2) represents an expansion of this model by including control variables for property, neighborhood and geographical characteristics and dummy control variables for time. The adjusted R-squared has substantially increased, up to 66.4 per cent for this model. This means that the model has a better fit, where 66.4 per cent of the variability in the data can be explained. Almost all variables for this model are significantly different from zero at the 1 per cent (or higher) level, except for Density and Quarter 2. The coefficient for the proportion of singles has stayed significant and positive and increased in strength compared to the first model.

The final model for this hypothesis is shown in column (3). This model shows an improved model fit, with an adjusted R-squared of 73.1 per cent, which is a relatively high adjusted R-squared compared to model 1 and 2. This means that 73.1 per cent of the variability in the data can be explained by this model. The coefficient for density of the population has turned significant for this model, however some other variables, principal residences, vacant dwellings, public transport and quarter 2, are not significantly different from zero at the 1 per cent level anymore. For all significant variables, the signs have stayed the same. The proportion of single-person households is significantly different from zero at the 1 per cent level. The coefficient has a positive sign, meaning that the association between single-person households is positive. These results are in line with the expectations from the first hypothesis that share of single-person households has a positive association with housing transaction prices. The strength of the association has somewhat declined compared to the second model.

Both property characteristics, surface of the property and the number of rooms, are significantly different from zero at the 1 per cent level, and have a positive effect, so that a bigger building and a greater number of rooms are increasing property prices.

The density of the population is significantly different from zero at the 1 per cent level, but has a negative coefficient, which implies that a higher density leads to lower transaction prices. Moreover, the share of the population aged 15 or over out of school without a diploma is also significantly different from zero at the 1 per cent level.

The neighborhood characteristics share of owner-occupied households, share of tenant households in social housing and share of households housed free are all three significantly different from zero at the 1 per cent level and have a positive coefficient, meaning that all three are positively association with housing transaction prices.

Table 4: Regression Results Hypothesis 1 (2SLS)

Variables	(1) Apartments & Houses	(2) Apartments & Houses	(3) Apartments & Houses
Singles (%)	0.0184***	0.0327***	0.0234***
	(0.000865)	(0.00286)	(0.00530)
LN Surface (m^2)		0.935***	0.915***
		(0.00847)	(0.00682)
Rooms (#)		0.0166***	0.0194***
		(0.00367)	(0.00305)
Density (%)		5.28e-05	-0.000243***
		(6.21e-05)	(5.15e-05)
LN No Diploma (%)		-0.309***	-0.131***
		(0.0209)	(0.0189)
Principal Residences (%)		-0.00559***	0.00242
		(0.00190)	(0.00184)
Vacant Dwellings (%)		-0.00551*	-0.00228
		(0.00286)	(0.00224)
Owner-Occupied (%)		0.00815***	0.00700***
		(0.00133)	(0.00188)
Social Tenants (%)		0.00873***	0.00550***
		(0.00103)	(0.00153)
Free Housing (%)		0.0200***	0.00788***
		(0.00342)	(0.00252)
LN Public Transport (m)		-0.0130*	0.00144
		(0.00720)	(0.00565)
Quarter = 2		0.00648	0.00148
		(0.00460)	(0.00402)
Quarter = 3		0.0221***	0.0191***
		(0.00439)	(0.00382)
Quarter = 4		0.0203***	0.0145***
		(0.00477)	(0.00414)
Constant	11.72***	8.350***	7.959***
	(0.0374)	(0.342)	(0.501)
# of LF-Effects			143
Observations	67,524	67,524	67,524
Adjusted R-squared	0.009	0.664	0.731

Standard errors have been adjusted for 2,090 clusters in IRIS-area. Robust standard errors in parentheses. The dependent variable is transformed into a log, as well as the independent variables Surface, No Diploma and Public Transport. The reference category for Quarter is 1. A constant has been included in the regression. The third model includes Location-Fixed effects (LF-Effects), based on 143 different postal codes.
*** p<0.01, ** p<0.05, * p<0.1

Only the dummy variables for the third and fourth quarter of 2015 were found to be significantly different from zero at the 1 per cent level. The effect slightly decreased between the third and fourth quarter.

These results, implying that the association between single-person households and property transaction prices is positive is in line with expectations from theory and previous research by Tyvimaa and Kamruzzaman (2019), who find a positive association between the proportion of singles and apartment transaction prices. The researchers go even further and argue that a 1 per cent increase of young, single person households increases apartment prices by 0.51 per cent in their research. As described before, this research is more conservative and only argues for an association, as it is difficult to argue for a causal relationship from regression only. Nevertheless, increased prices, which make it less affordable to own a home, are often responded by people co-living (Maalsen, 2019). In the long run the positive association could thus lead to a less favourable choice of living alone.

Also other independent variables follow expectations from previous research and common knowledge, like bigger buildings and a greater number of rooms are increasing property prices. Gibbons (2003) argued that the housing prices increase when the proportion of higher educated residents in a neighborhood is higher than the regional mean. Reversing this arguing, it could mean that an increasing proportion on non-educated people will therefore have a negative effect.

4.2 The influence of type of housing on the association between single-person households and transaction prices.

The coefficients and robust standard errors related to the results on the second hypothesis are shown in table 6. In order to test if there is an influence of housing type on the association between single-person households and the transaction prices different groups are created within the dataset, namely apartments only and houses only. In the first column (3) the results of model tree of hypothesis one are shown as a reference for the two different groups. The two groups were both run in a different regression. The number of observations for apartments only was a lot higher than it was for houses only.

Table 5: Heteroskedasticity-robust version of Hausman test Model 3, 4 & 5

Model	F-statistic
Model 3	38.1974***
Model 4	41.0844***
Model 5	11.2702***

*** p<0.01, ** p<0.05, * p<0.1

Note: the Heteroskedasticity-robust version of Hausman Test measures if the endogenous regressors that are used in the model are in fact exogenous.

Table 6: Regression Results Hypothesis 2 (2SLS)

Variables	(3) Apartments & Houses	(4) Apartments Only	(5) Houses Only
Singles (%)	0.0234***	0.0240***	0.0448**
	(0.00530)	(0.00533)	(0.0215)
LN Surface (m^2)	0.915***	0.945***	0.548***
	(0.00682)	(0.00723)	(0.0204)
Rooms (#)	0.0194***	-0.00164	0.0240***
	(0.00305)	(0.00334)	(0.00586)
Density (%)	-0.000243***	-0.000200***	-3.62e-05
	(5.15e-05)	(5.17e-05)	(0.000293)
LN No Diploma (%)	-0.131***	-0.154***	-0.152***
	(0.0189)	(0.0202)	(0.0495)
Principal Residences (%)	0.00242	0.00213	0.00614
	(0.00184)	(0.00182)	(0.00801)
Vacant Dwellings (%)	-0.00228	-0.00226	-0.00601
	(0.00224)	(0.00230)	(0.00801)
Owner-Occupied (%)	0.00700***	0.00548***	0.0197**
	(0.00188)	(0.00178)	(0.00943)
Social Tenants (%)	0.00550***	0.00532***	0.0142**
	(0.00153)	(0.00153)	(0.00684)
Free Housing (%)	0.00788***	0.0100***	0.00979
	(0.00252)	(0.00289)	(0.00616)
LN Public Transport (m)	0.00144	-0.00375	-0.00200
	(0.00565)	(0.00581)	(0.0192)
Quarter = 2	0.00148	-0.000123	0.000221
	(0.00402)	(0.00413)	(0.0120)
Quarter = 3	0.0191***	0.0122***	0.0325***
	(0.00382)	(0.00395)	(0.0114)
Quarter = 4	0.0145***	0.0103**	0.0159
	(0.00414)	(0.00429)	(0.0128)
Constant	7.959***	8.009***	6.999***
	(0.501)	(0.495)	(2.014)
# of LF-Effects	143	143	143
Observations	67,524	59,826	7,698
Adjusted R-squared	0.731	0.753	0.557

Standard errors have been adjusted for 2,090 clusters in IRIS-area. Robust standard errors in parentheses. The dependent variable is transformed into a log, as well as the independent variables Surface, No Diploma and Public Transport. The reference category for Quarter is 1. A constant has been included in the regression. All models include Location-Fixed effects (LF-Effects), based on 143 different postal codes. Model 3 from hypothesis 1 has been included for reference as (1) Apartments and Houses.

*** $p<0.01$, ** $p<0.05$, * $p<0.1$

A heteroskedasticity-robust version of the Hausman test was conducted, after which the null-hypothesis that the percentage of single-person households can be seen as an exogenous variable is rejected for all three specifications of the model, as can be seen from table 5.

The model fit for apartments only improved slightly compared to apartments and houses taken together. The adjusted R-squared for the model with apartments only was 75,3 per cent. The model fit for houses only was substantially lower, with an adjusted R-squared of 55,7 per cent.

The coefficient for the percentage of single-person households is significantly different from zero at the (at least) 5 per cent level for both apartments and houses only. The signs for all the significant variables are equal among the three groups, no unexpected changes have come up. A few variables are not significant in the subgroup models, while they were in the group with apartments and houses together. The number of rooms is not significant in the case of apartments only, while it is for apartments and houses together and houses only. Density, Free Housing and Quarter 4 are not significant for houses only.

The association between single-person households is stronger in the case of houses, than it is in the case of apartments. A Chow-Test has been conducted ($F(156, 67{,}524) = 20.760$, as described in Appendix H), after which the null hypothesis of parameter stability over the two sample groups is rejected on the one per cent level. This means that the independent variables have different impacts on the subgroup of apartments, than on the subgroup of houses, meaning that there can be argued for different submarkets for apartments and houses. However, when looking at the variable for single-person households only, no significant difference between the coefficients in the different subsamples is found ($Z = -0.939$, $p = .352$), see Appendix I. This leads to the observation that, even though the association between the proportion of single-person households and transaction prices is stronger in the case of houses than it is in case of apartments, it cannot be argued that the share of single-person households is of more influence in one of the two markets. It implicates that an increasing share of single-person households will increase property prices for apartments as well as for houses, which suggests that an increasing share of singles increases demand for both property types. Even though single-person households might have a preference for living in flats, like researched by Wulff (2001), they might be willing to neglect their preferences and move to houses, if apartments are not available due to high demand. This means that the second hypothesis that the type of housing has an influence on the association between single-person households and transaction prices cannot be rejected.

5. CONCLUSIONS

5.1 Conclusion

The proportion of single-person households has increased in preceding years and is expected to continue to grow in the coming years. This thesis has researched the relation between single-person households and property transaction prices. There have not yet been done many studies on this topic and research on differences between property types in this association had not yet been done before. This research has tried to fill the gap in the literature by using 2SLS in combination with an instrument variable to explore the association between the proportion of single-person households and property transaction prices in Paris, France with a sample of 67,524 observations.

Based on theory, it was expected that an increase in the number of households increases the demand for space, increasing property prices through rent. It was found that demand across singles is not uniform, therefor differences in the association between the proportion of singles and apartments and the proportion of singles and houses were expected.

A heteroskedasticity-robust version of the Hausman test was conducted, which confirmed that the percentage of single-person households is an endogenous variable, thus reaffirming the necessity of a method taking this into account. The findings from the 2SLS regression indicate that there is a positive significant effect between the proportion of single-person households and property transaction prices.

Moreover, an independent regression for the subgroups apartments and houses was run. The association between the proportion of single-person households and transaction prices was found to be stronger in the case of houses than it is in case of apartments. A chow-test confirmed a structural difference between these two groups, however the association of the share of single-person households was not found to be significantly different for both groups, so that the influence of the type of housing cannot be proven.

The findings of this research show that there are different factors affecting the market. It shows that, in addition to housing characteristics, locational characteristics also demographic characteristics are of influence. These results are important for housing market analysts and investors and helps in understanding market factors, as well as for policy makers and city planners for planning and provision of housing. Housing possibilities might become more limited for a larger share of people, due to increased property prices. This results in a need for stabilizing housing affordability policies.

5.2 Limitations and recommendations for further research

This study contains some limitations related to the used data, the data is collected from open-data platforms, which may be unreliable or incomplete. Moreover, the property and population data are somewhat basic, given that the data used is only of the year 2015, property characteristics are limited (only number of rooms and surface) and personal characteristics of the buyer, like age, gender are unknown. The research would enhance if more property characteristics and buyer characteristics were known. In addition, the data represents a metropolitan area, more consistent external evidence is needed from different cities and non-urban regions to validate the results of this research.

The study most likely reflects the association for relatively wealthy single-person households who can afford to live in central Paris. As house prices are relatively high in Paris (compared to non-urban regions, smaller cities outside of Paris) it is not possible for every single-person household to afford housing in Paris. Single-person households with less income will have a large probability to be outbid by small families with a double income.

Thirdly, this study only contains the property market based on owner-occupying. However, there is a fast majority of people that does not live in their own property, but lives in a rental house. Future studies could look into the association between single-person households and property rents.

The possible presence of omitted variables makes it difficult to argue for a causal relationship. However, in the case of a well-designed experiment the chance of omitting variables disappears so that there results always describe a causal relationship. Further research could us a quasi-experimental approach, like difference-in-differences, where the difference-in-differences notes the similarity between a change in property prices close to areas with a high proportion of singles (within a set area) compared to other houses in the same neighborhoods that are outside the catchment area. The difference-in-differences estimation can be seen like the equivalent of an experiment, so that a causal relationship can be described.

REFERENCES

Anselin, L. and Rey, S. J. (2014) *Modern Spatial Econometrics in Practice: A Guide to GeoDa, GeoDaSpace and PySAL*. GeoDa Press LLC.

Anselin, L., Syabri, I. and Kho, Y. (2006) 'GeoDa - An Introduction to Spatial Data Analysis', *Geographical Analysis*, 38, pp. 5–22. doi: https://doi.org/10.1111/j.0016-7363.2005.00671.x.

Brooks, C. and Tsolacos, S. (2010) *Real Estate Modelling and Forecasting*. 1st edn. United States of America: Cambridge University Press.

Daams, M. N., Sijtsma, F. J. and van der Vlist, A. J. (2016) 'The effect of natural space on nearby property prices: Accounting for perceived attractiveness', *Land Economics*, 92(3), pp. 389–410. doi: 10.3368/le.92.3.389.

DiPasquale, D. and Wheaton, W. C. (1992) 'The Markets for Real Estate Assets and Space: A Conceptual Framework', *Real Estate Economics*, 20(2), pp. 181–198. doi: 10.1111/1540-6229.00579.

Elsinga, M. and Hoekstra, J. (2005) 'Homeownership and housing satisfaction', *Journal of Housing and the Built Environment*, 20(4), pp. 401–424. doi: 10.1007/s10901-005-9023-4.

Eurostat (2015) *People in the EU: who are we and how do we live? 2015, Office of the European Union*. doi: 978-92-79-50328-3.

Eurostat (2016) *Urban Europe, The European Territory*. Edited by M. Kotzeva. Luxembourg: Publications office of the European Union. doi: 10.4324/9781315777962-2.

Eurostat (2019) *Household composition statistics*. Available at: https://ec.europa.eu/eurostat/statistics-explained/index.php/Household_composition_statistics#Household_size (Accessed: 28 September 2019).

Faessen, W. B. M. (2002) 'Present and preferred housing of one-personhouseholds in the Netherlands, 1998', *Journal of Housing and the Built Environment*, 17(3), pp. 337–350. doi: 10.1023/A:1020221800785.

Gibbons, S. (2003) 'Paying for Good Neighbours: Estimating the Value of an Implied Educated Community', *Urban Studies*, 40(4), pp. 809–833. doi: 10.1080/0042098032000065317.

Gillen, K., Thibodeau, T. and Wachter, S. (2001) 'Anisotropic Autocorrelation in House Prices', *Journal of Real Estate Finance and Economics*, 23(1), pp. 5–30. doi: 10.1023/A:1011140022948.

Gross, D. J. (1988) 'Estimating willingness to pay for housing characteristics: An application of the Ellickson bid-rent model', *Journal of Urban Economics*. doi: 10.1016/0094-1190(88)90049-6.

Gujarati, D. N. (2003) *BASIC ECONOMETRICS*. 4th edn. New York: McGraw-Hill/lrwin.

Hall, R., Ogden, P. E. and Hill, C. (1997) 'The pattern and structure of one-person households in England and Wales and France', *International Journal of Population Geography*, 3(2), pp. 161–181. doi: 10.1002/(sici)1099-1220(199706)3:2<161::aid-ijpg64>3.3.co;2-u.

Hsieh, W., Norman, D. and Orsmond, D. (2012) 'Supply Side Issues in the Housing Sector', *RBA Bulletin*, pp. 11–19. Available at: https://www.rba.gov.au/publications/bulletin/2012/sep/2.html.

Ined (2019) *Household size*. Available at: https://www.ined.fr/en/everything_about_population/data/france/couples-households-families/households/ (Accessed: 17 December 2019).

INSEE (2006) 'Projection de menages pour la France Metropolitaine a l'horizon 2030', pp. 1–18.

INSEE (2016) *IRIS*. Available at: https://www.insee.fr/en/metadonnees/definition/c1523 (Accessed: 15 October 2019).

Kohler, M. and Van Der Merwe, M. (2015) 'Long-run Trends in Housing Price Growth', *RBA Bulletin*, (2012), pp. 21–30. Available at: https://www.rba.gov.au/publications/bulletin/2015/sep/pdf/bu-0915-3.pdf.

Maalsen, S. (2019) 'I cannot afford to live alone in this city and I enjoy the company of others: why people are share housing in Sydney', *Australian Geographer*. Routledge, 50(3), pp. 315–332. doi: 10.1080/00049182.2019.1613026.

Mankiw, N. G. and Weil, D. N. (1989) 'The baby boom, the baby bust, and the housing market', *Regional Science and Urban Economics*. North-Holland, 19(2), pp. 235–258. doi: 10.1016/0166-0462(89)90005-7.

Marquaridt, D. W. (1970) 'Generalized Inverses, Ridge Regression, Biased Linear Estimation, and Nonlinear Estimation', *Technometrics*. Taylor & Francis, 12(3), pp. 591–612. doi: 10.1080/00401706.1970.10488699.

McLeod, P. B. (1984) 'The Demand for Local Amenity: An Hedonic Price Analysis', *Environment and Planning A: Economy and Space*, 16(3), pp. 389–400. doi: 10.1068/a160389.

Ogden, P. E. and Schnoebelen, F. (2005) 'The rise of the small household: Demographic change and household structure in Paris', *Population, Space and Place*, 11(4), pp. 251–268. doi: 10.1002/psp.370.

Quintano, C. and D'Agostino, A. (2006) 'Studying Inequality in Income Distribution Of Single-Person Households in Four Developed Countries', *Review of Income and Wealth*, 52(4), pp. 525–546. doi: 10.1111/j.1475-4991.2006.00206.x.

Rohe, W. M. and Stegman, M. A. (1994) 'The Effects of Homeownership: on the Self-Esteem, Perceived Control and Life Satisfaction of Low-Income People', *Journal of the American Planning Association*. Routledge, 60(2), pp. 173–184. doi: 10.1080/01944369408975571.

Rosen, S. (1974) 'Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition', *Journal of Political Economy*. University of Chicago Press, 82(1), pp. 34–55. doi: https://doi.org/10.1086/260169.

Santi, L. L. (1988) 'The demographic context of recent change in the structure of American households', *Demography*, 25(4), pp. 509–519. doi: 10.2307/2061318.

Schmidheiny, K. (2019) ‘Short Guides to Microeconometrics - Instrument Variables’. Available at: https://www.schmidheiny.name/teaching/iv2up.pdf.

Selim, H. (2009) ‘Determinants of house prices in Turkey: Hedonic regression versus artificial neural network’, *Expert Systems with Applications*. Pergamon, 36(2), pp. 2843–2852. doi: 10.1016/J.ESWA.2008.01.044.

Sirmans, G. S., Macpherson, D. A. and Zietz, E. N. (2005) ‘The composition of hedonic pricing models’, *Journal of Real Estate Literature*, 13(1), pp. 3–43. Available at: www.jstor.org/stable/44103506.

Statista Research Department (2019) *Number of households in France in 2016, by size (in millions)*. Available at: https://www.statista.com/statistics/512974/number-households-by-size-france/ (Accessed: 20 November 2019).

Tobler, W. R. (1970) ‘A Computer Movie Simulating Urban Growth in the Detroit Region’, *Economic Geography*. [Clark University, Wiley], 46, pp. 234–240.

Tyvimaa, T. and Kamruzzaman, M. (2019) ‘The effect of young, single person households on apartment prices: an instrument variable approach’, *Journal of Housing and the Built Environment*. Springer Netherlands, 34(1), pp. 91–109. doi: 10.1007/s10901-018-9618-1.

U.S. Census Bureau News (2017) ‘Unmarried and Single Americans Week’, pp. 1–4.

Williams, J. (2007) ‘Innovative solutions for averting a potential resource crisis - The case of one-person households in England and Wales’, *Environment, Development and Sustainability*, 9(3), pp. 325–354. doi: 10.1007/s10668-006-9068-x.

Withers, S. D. (1998) ‘Linking Household Transitions and Housing Transitions: A Longitudinal Analysis of Renters’, *Environment and Planning A: Economy and Space*, 30(4), pp. 615–630. doi: 10.1068/a300615.

Wulff, M. (2001) ‘Growth and change in one person households: Implications for the housing market’, *Urban Policy and Research*, 19(4), pp. 467–489. doi: 10.1080/08111140108727894.

Yeung, W. J. J. and Cheung, A. K. L. (2015) ‘Living alone: One-person households in Asia’, *Demographic Research*, 32(1), pp. 1099–1112. doi: 10.4054/DemRes.2015.32.40.

Zady, M. F. (2000) *Z-12: Correlation and Simple Least Squares Regression*. Available at: https://www.westgard.com/lesson42.htm (Accessed: 5 November 2019).

Zheng, S. *et al.* (2016) ‘Transit development, consumer amenities and home values: Evidence from Beijing’s subway neighborhoods’, *Journal of Housing Economics*, 33, pp. 22–33. doi: https://doi.org/10.1016/j.jhe.2016.05.003.

Appendix A. Definition of Variables

Table A1: Definition of Variables

Variables	Definitions	Source
Price	Property Transaction Price in Euros	DGFiP
Singles	Share of Single-person Households	APUR
Lagged Singles	Spatially Lagged Share of Single-person Households	Calculated
Surface	Surface of Property in m^2	DGFiP
Rooms	Number of Main Rooms	DGFiP
Density	Density of Population in percentage	APUR
Postal Code	Postal Code	APUR
No Diploma	Share of the Population Age 15 or over out of School without Diploma	APUR
Family with Children	Share of Families with Child(ren) under 25	APUR
Principal Residences	Share of Principal Residences	APUR
Occasional Dwellings	Share of Second Homes and Occasional Dwellings	APUR
Vacant Dwellings	Share of Vacant Dwellings	APUR
Owner-Occupied	Share of Owner-Occupied Households	APUR
Private Tenants	Share of Tenant Households in Private Housing	APUR
Social Tenants	Share of Tenant Households in Social Housing	APUR
Free Housing	Share of Households Housed Free	APUR
Public Transport	Distance to Public Transport in Meters	APUR

Appendix B. Descriptive Statistics

Table A2: Descriptive Statistics Apartments and Houses

Variables	Mean	Std. Dev.	Min	Max
Price	330729.88	229960.89	22000	1550000
LN Price	12.507	.636	9.999	14.254
Singles	42.631	10.502	11.126	70.651
Lagged Singles	41.893	9.047	11.689	65.361
Surface	57.768	29.882	4	370
Rooms	2.7	1.208	1	6
Density	236.957	168.906	1.611	1333.624
No Diploma	21.876	9.672	6.94	64.761
Families with Children	30.377	9.799	0	72.077
Principal Residences	88.373	6.183	40.744	98.822
Occasional Dwellings	4.333	4.784	.059	43.42
Vacant Dwellings	7.294	2.842	.092	27.53
Owner-Occupied	43.76	15.789	.162	91.066
Private Tenants	36.597	14.121	1.198	88.135
Social Tenants	16.181	16.635	.067	97.546
Free Housing	3.463	2.507	.105	66.218
Public Transport	533.142	508.453	9.226	6357.073

N = 67,524

Table A3: Descriptive Statistics Apartments Only

Variables	Mean	Std. Dev.	Min	Max
Price	320849.95	227452.41	22000	1550000
LN Price	12.473	.639	9.999	14.254
Singles	44.006	9.991	11.126	70.651
Lagged Singles	43.074	8.649	11.689	65.361
Surface	53.791	26.779	4	370
Rooms	2.532	1.11	1	6
Density	256.046	168.737	1.611	1333.624
No Diploma	21.125	9.342	6.94	64.761
Families with Children	29.222	9.498	0	72.077
Principal Residences	87.839	6.285	40.744	98.822
Occasional Dwellings	4.711	4.929	.059	43.42
Vacant Dwellings	7.45	2.859	.092	27.53
Owner-Occupied	41.97	14.673	.162	91.066
Private Tenants	38.208	13.674	1.198	88.135
Social Tenants	16.209	16.653	.067	97.546
Free Housing	3.613	2.49	.105	66.218
Public Transport	479.06	445.171	9.226	5745.646

N = 58,826

Table A4: Descriptive Statistics Houses Only

Variables	Mean	Std. Dev.	Min	Max
Price	407565.69	234839.42	22877	1550000
LN Price	12.773	.539	10.038	14.254
Singles	31.936	7.92	11.126	67.863
Lagged Singles	32.701	6.427	11.689	58.067
Surface	88.692	34.367	6	280
Rooms	4.01	1.136	1	6
Density	88.505	64.043	1.611	701.254
No Diploma	27.714	10.194	7.518	64.761
Families with Children	39.352	7.088	14.177	70.526
Principal Residences	92.529	2.977	73.756	98.822
Occasional Dwellings	1.397	1.49	.059	18.043
Vacant Dwellings	6.074	2.373	.092	17.369
Owner-Occupied	57.667	17.188	2.187	91.066
Private Tenants	24.071	10.906	1.649	72.367
Social Tenants	15.966	16.499	.067	90.882
Free Housing	2.296	2.323	.105	66.218
Public Transport	953.735	726.81	34.202	6357.073

N = 7,698

Appendix C. Transformation of Dependent Variable

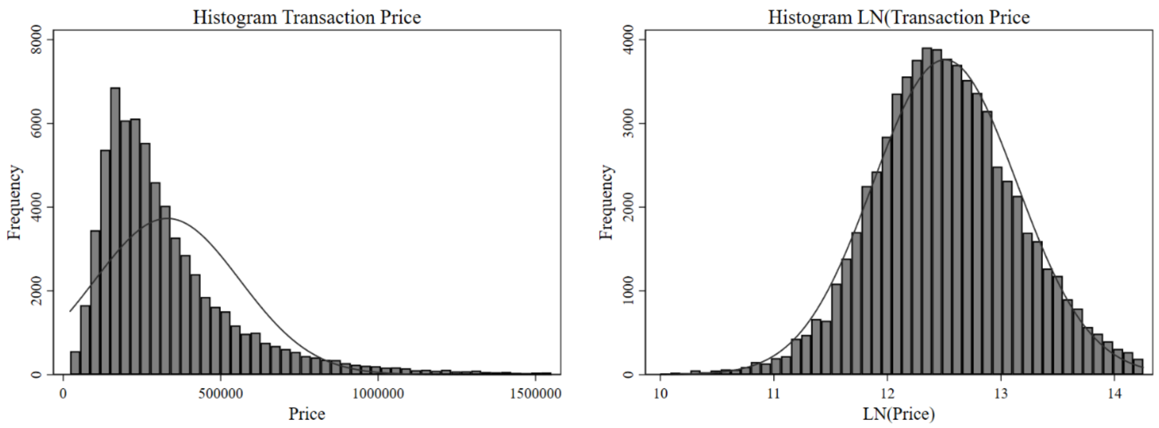

Figure A.1: Transformation of Transaction Price variable

Appendix D. Transformation of Independent Variables

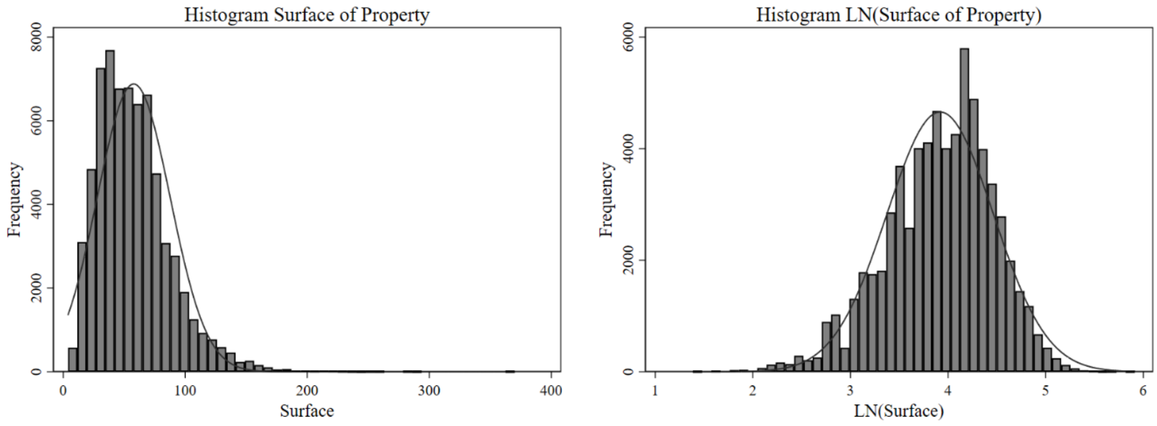

Figure A.2: Transformation of Surface variable

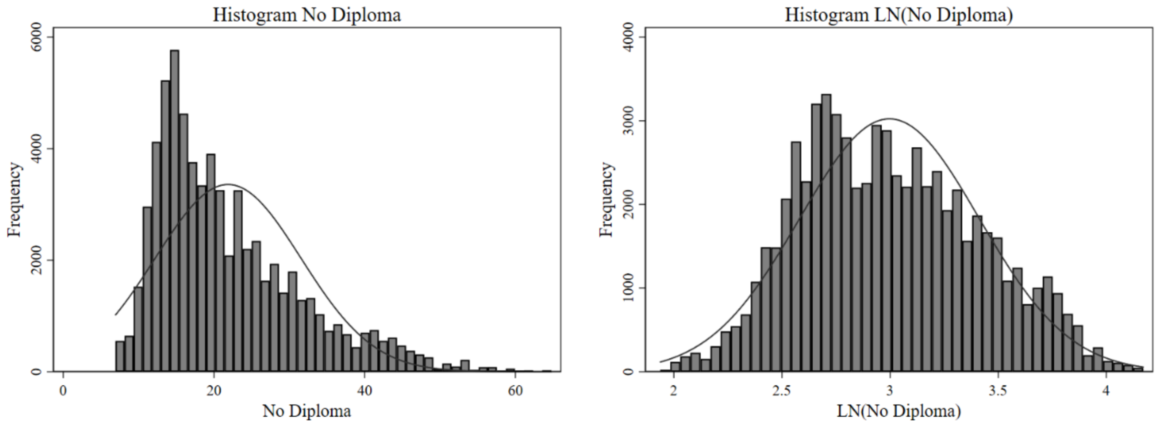

Figure A.3 Transformation of No Diploma variable

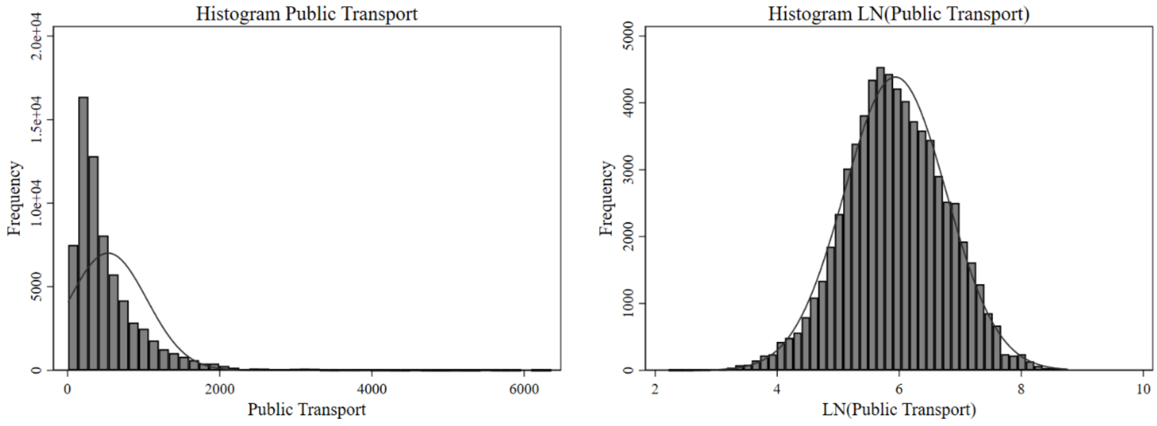

Figure A.4: Transformation of Public Transport variable

Appendix E. Data Preparation

In Stata the correlations between variables was checked, using the function correlate. There is a simple rule of thumb for evaluating correlation coefficients, where a correlation coefficient between 0.00 and 0.29 gives little to no correlation, 0.30 to 0.49 gives low correlation, 0.50 to 0.69 gives moderate correlation, 0.70 to 0.89 gives high correlation and 0.90 to 1.00 gives very high correlation (Zady, 2000).

In table A5 the matrix of correlation is shown. Some of the variables show correlation between them. To determine if these are problematic a Variance Inflation Test was conducted.

As a rule thumb for the amount of bias allowed in the data, the Variance Inflation Factor should be between 1 and 10 (Marquaridt, 1970).

The first regression was done including all variables, according to the equation:

$$\begin{aligned}\mathrm{LN}(Price) = {} & \beta_0 + \beta_1\, Singles_{ia} + \beta_2\, Surface_i + \beta_3 Rooms_i + \beta_4\, Density_{ia} + \beta_5\, No\ Diploma_{ia} \\ & + \beta_6\, Family\ with\ Children_{ia} + \beta_7\, Principal\ Residences_{ia} \\ & + \beta_8\, Occasional\ Dwellings_{ia} + \beta_9\, Vacant\ Dwellings_{ia} \\ & + \beta_{10}\, OwnerOccupied_{ia} + \beta_{11}\, Private\ Tenants_{ia} + \beta_{12}\, Social\ Tenants_{ia} \\ & + \beta_{13}\, Free\ Housing_{ia} + \beta_{14}\, Public\ Transport_{ia} + D_1 \sum_{j=2}^{4} Quarter_{ij} \\ & + f_i\, Postal\ Code_i + \varepsilon_0\end{aligned} \qquad \text{(A1)}$$

Where $\mathrm{LN}(Price)$ represents the natural logarithm of the transaction price of property i; $Singles_{ia}$ refers to the proportion of single-person households in IRIS area a in which the property i is located; $Surface_i$ represents the surface of property i; $Rooms_i$ refers to the number of rooms of property i; $Density_{ia}$ refers to the density of the population in IRIS area a where apartment i is located; $No\ Diploma_{ia}$ is the variable for the share of the population aged 15 or over, who are out of school without a diploma in IRIS area a where property i is located; $Family\ with\ Children_{ia}$ refers to the share of families with child(ren) under 25 in IRIS area a where property i is located; $Principal\ Residences_{ia}$ refers to the share of principal residences in IRIS area a where property i is located; $Occasional\ Dwellings_{ia}$ refers to the share of second homes and occasional dwellings in IRIS area a where property i is located; $Vacant\ Dwellings_{ia}$ represents the share of vacant dwellings in IRIS area a where property i is located; $OwnerOccupied_{ia}$ refers to the share of owner-occupied households in IRIS area a where property i is located; $Private\ Tenants_{ia}$ refers to the share of tenant households in private housing in IRIS area a where property i is located; $Social\ Tenants_{ia}$ refers to the share of tenant households in social housing in IRIS area a where property i is located; $Free\ Housing_{ia}$ relates to the share of households housed free in IRIS area a where property i is located; $Public\ Transport_i$ refers to the distance in meters to public transport for property i; $Quarter_{ij}$ is a dummy variable for the quarters of the year 2015; $Postal\ Code_{if}$ is a dummy variable for fixed effects based on postal code for property i; ε_0 represents the residual error.

When running this regression, the share of vacant dwellings and the share of households housed free were omitted due to collinearity. In addition, when calculating the Variance Inflation Factors, some variables showed Factors over 60, highly exceeding the permitted value. Therefor a new regression was run, without Share of Families with Children under 25, the Share of Second Homes and Dwellings, Share of Tenants in Private Housing. The Number of Main Rooms was left in as well as Postal Code (absorbed).

$$\begin{aligned} \mathrm{LN}(Price) = {} & \beta_0 + \beta_1\, Singles_{ia} + \beta_2\, Surface_i + \beta_3\, Rooms_i + \beta_4\, Density_{ia} + \beta_5\, No\ Diploma_{ia} \\ & + \beta_7\, Principal\ Residences_{ia} + \beta_9\, Vacant\ Dwellings_{ia} \\ & + \beta_{10}\, OwnerOccupied_{ia} + \beta_{12}\, Social\ Tenants_{ia} + \beta_{13}\, Free\ Housing_{ia} \\ & + \beta_{14}\, Public\ Transport_i + D_1 \sum_{j=2}^{4} Quarter_{ij} + \gamma \sum_{f=2}^{143} Postal\ Code_{if} + \varepsilon_1 \end{aligned} \quad \text{(A2)}$$

No variables were omitted by regressing, and the Variable Inflation Factors of this regression are all between 1 and 10, as can be seen in Table A6.

Table A5: Matrix of Correlations

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)
(1) Price	1.000																
(2) Singles	0.174	1.000															
(3) Surface	0.682	-0.269	1.000														
(4) Rooms	0.543	-0.294	0.856*	1.000													
(5) Density	0.110	0.627	-0.250	-0.252	1.000												
(6) Postal Code	-0.228	-0.688	0.212	0.241	-0.744*	1.000											
(7) No Diploma	-0.337	-0.467	0.030	0.094	-0.278	0.393	1.000										
(8) Families with Children	-0.201	-0.934**	0.244	0.275	-0.617	0.714*	0.523	1.000									
(9) Principal Residences	-0.280	-0.605	0.129	0.182	-0.362	0.609	0.461	0.626	1.000								
(10) Occasional Dwellings	0.303	0.542	-0.095	-0.156	0.306	-0.585	-0.453	-0.561	-0.897*	1.000							
(11) Vacant Dwellings	0.100	0.404	-0.120	-0.134	0.273	-0.341	-0.240	-0.418	-0.666	0.267	1.000						
(12) Owner-Occupied	-0.008	-0.465	0.222	0.220	-0.438	0.402	-0.138	0.319	0.208	-0.213	-0.095	1.000					
(13) Private Tenants	0.142	0.743*	-0.243	-0.262	0.530	-0.564	-0.377	-0.724*	-0.601	0.478	0.502	-0.435	1.000				
(14) Social Tenants	-0.154	-0.249	0.003	0.028	-0.070	0.168	0.506	0.377	0.394	-0.288	-0.372	-0.563	-0.489	1.000			
(15) Free Housing	0.272	0.392	-0.048	-0.095	0.235	-0.469	-0.365	-0.439	-0.542	0.558	0.239	-0.115	0.351	-0.339	1.000		
(16) Public Transport	-0.149	-0.545	0.185	0.201	-0.470	0.472	0.278	0.523	0.389	-0.341	-0.271	0.404	-0.507	0.090	-0.283	1.000	
(17) Quarter	0.026	0.020	0.004	0.003	0.024	-0.036	-0.019	-0.023	-0.018	0.018	0.009	-0.008	0.013	-0.006	0.014	-0.011	1.000

* High Correlation
** Very High Correlation

Table A6: Variance Inflation Factors

Variables	VIF	1/VIF
Share of Single-person Households	3.470	0.288
Surface of Property	3.920	0.255
Number of Main Rooms = 2	2.190	0.457
Number of Main Rooms = 3	3.250	0.308
Number of Main Rooms = 4	3.870	0.258
Number of Main Rooms = 5	3.050	0.328
Number of Main Rooms = 6	2.020	0.494
Density of Population	1.820	0.551
Share of the Population Aged 15 or over out of School without a Diploma	1.830	0.546
Share of Principal Residences	3.150	0.317
Share of Vacant Dwellings	2.050	0.488
Share of Owner-Occupied Households	4.370	0.229
Share of Tenant Households in Social Housing	3.810	0.263
Share of Households Housed Free	1.610	0.621
Distance to Public Transport	1.590	0.630
2^{nd} quarter of 2015	1.640	0.611
3^{rd} quarter of 2015	1.700	0.589
4^{th} quarter of 2015	1.690	0.591
Mean	2.610	

Appendix F. Assumption Testing

Assumption 1: The average value of errors is zero

This assumption will not be violated when including a constant term in the regression equation (Brooks and Tsolacos, 2010). As in this study a constant is included in the regression equation, it is assumed that the average value of errors is indeed zero.

Assumption 2: Homoscedasticity

The second assumption refers to the requirement of a constant variance of the error terms.

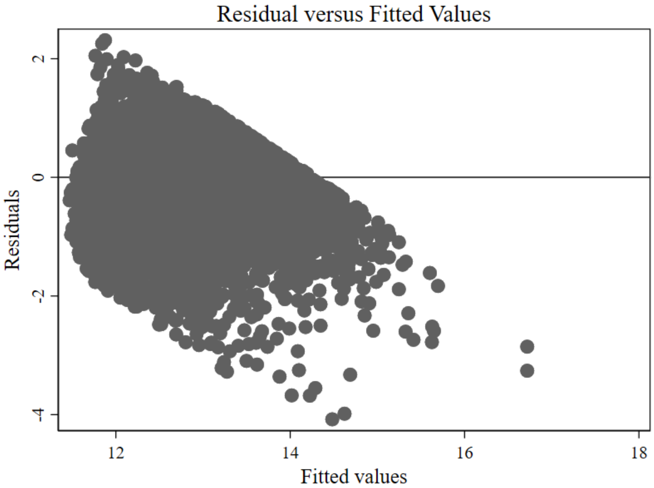

Figure A.5: Residual-versus-Fitted Values

From this graph it cannot be said that the data is homoscedastic. In addition, two tests have been done to check homoscedasticity, the Breusch-Pagan / Cook-Weisberg test and Cameron & Trivedi's decomposition of the IM-test.

Table A7: Breusch-Pagan / Cook-Weisberg test for heteroskedasticity

Ho: Constant variance
Variables: fitted values of LN Transaction Price
chi2(1) = 8572.30
Prob > chi2 = 0.0000

Table A8: Cameron & Trivedi's decomposition of IM-test

	chi2	df	p
Heteroskedasticity	8740.250	113	0.000
Skewness	1596.300	14	0.000
Kurtosis	140.050	1	0.000
Total	10476.600	128	0.000

These results show that the hypothesis of constant variance should be rejected, which leads to the conclusion that the data is heteroscedastic. The regression will still give unbiased and consistent results, however there is no longer minimum variance of the estimators (Brooks and Tsolacos, 2010).

In order to overcome this problem, regressions will be run with help of the robust regression in Stata, which is an solution proposed by Brooks & Tsolacos (2010).

Assumption 3: No covariance between error terms over time / space

The third assumption premises that no covariance over time and/or space should exist. The dataset contains cross-sectional data, therefor covariance over time is not present. However, as formulated in the first law of geography: "Everything is related to everything else, but near things are more related than distant things" (Tobler, 1970). In addition Gillen, Thibodeau, & Wachter (2001) argue that house prices are usually spatially correlated, because buildings that are close to each other are often built in the same period, resulting in similar factors. Moreover, buildings in the same area, have access to the same amenities.

Spatial autocorrelation has been calculated for the transaction price variable, as well as for the variable for the share of single-person households using the Global Moran's I statistic. The results in table A9 show that there is a less than 1% likelihood that these clustered patterns could be the result of random chance.

Table A9: Moran's I Test

Variables	Moran's Index	z-score
Price	0,988247	750,667657***
Singles	0,587268	51,866103***

*** p<0.01, ** p<0.05, * p<0.1

As a solution for spatial autocorrelation standard errors are clustered on IRIS area.

Assumption 4: No relationship between variables and error term

The Durbin-Wu-Hausmann test can be done to check the relationship between variables and the error term.

At first the normal regression equation if performed, which is similar to equation A2.

$$\begin{aligned} \mathrm{LN}(Price) = {} & \beta_0 + \beta_1\, Singles_{ia} + \beta_2\, Surface_i + \beta_3\, Rooms_i + \beta_4\, Density_{ia} + \beta_5\, No\ Diploma_{ia} \\ & + \beta_7\, Principal\ Residences_{ia} + \beta_9\, Vacant\ Dwellings_{ia} \\ & + \beta_{10}\, OwnerOccupied_{ia} + \beta_{12}\, Social\ Tenants_{ia} + \beta_{13}\, Free\ Housing_{ia} \\ & + \beta_{14}\, Public\ Transport_i + D_1 \sum_{j=2}^{4} Quarter_{ij} + \gamma \sum_{f=2}^{143} Postal\ Code_{if} + \varepsilon_1 \end{aligned} \tag{A2}$$

The residuals of this equation or stored. A new regression equation is run, including the results for residuals.

$$\begin{aligned}\text{LN}(Price) = \beta_0 + \varepsilon_1 + \beta_1\, Singles_{ia} + \ \beta_2\, Surface_i + \beta_3\, Rooms_i \ + \beta_4\, Density_{ia} \\ + \beta_5\, No\ Diploma_{ia} + \beta_7\, Principal\ Residences_{ia} + \beta_9\, Vacant\ Dwellings_{ia} \\ + \beta_{10}\, OwnerOccupied_{ia} + \beta_{12}\, Social\ Tenants_{ia} + \beta_{13}\, Free\ Housing_{ia} \\ + \beta_{14}\, Public\ Transport_i + D_1 \sum_{j=2}^{4} Quarter_{ij} + \gamma \sum_{f=2}^{143} Postal\ Code_{if} + \varepsilon_2\end{aligned} \tag{A3}$$

In table A10 the results of this regression are shown. It is tested if the residuals are statistically different form zero. As can be seen in the results, the residuals from the first regression are significantly different from zero at the 1 per cent level, meaning that the variables and the error in the model are correlated. This means that endogeneity is present in the model. This problem is resolved running a 2SLS regression, using an instrument variable.

Assumption 5: Normality of Residuals

The normality of residuals has been tested using a Kernel Density Plot, standardized normal probability plot and the inverse standardized normal probability plot. The inverse standardized probability plot shows a slight deviation from the normal distribution towards the end of the tails (See figure A6, A7 & A8). These seem to be small, therefor normality of the residuals is assumed.

Table A10: Durbin-Wu-Hausman Test

Variables	Price
Residuals	0.439***
	(0.0941)
Singles	-0.438***
	(0.0941)
Surface	0.00942***
	(0.000768)
Rooms	0.0301***
	(0.0114)
Density	0.00166***
	(0.000385)
No Diploma	-0.0420***
	(0.00742)
Principal Residences	-0.102***
	(0.0213)
Vacant Dwellings	-0.0135***
	(0.00224)
Owner-Occupied	-0.150***
	(0.0322)
Social Tenants	-0.113***
	(0.0244)
Free Housing	-0.121***
	(0.0264)
Public Transport	-0.000461***
	(9.36e-05)
Quarter = 2	-0.0330***
	(0.00653)
Quarter = 3	0.0202***
	(0.00353)
Quarter = 4, omitted	-
Constant	49.28***
	(8.004)
Observations	67,524
R-squared	0.695

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

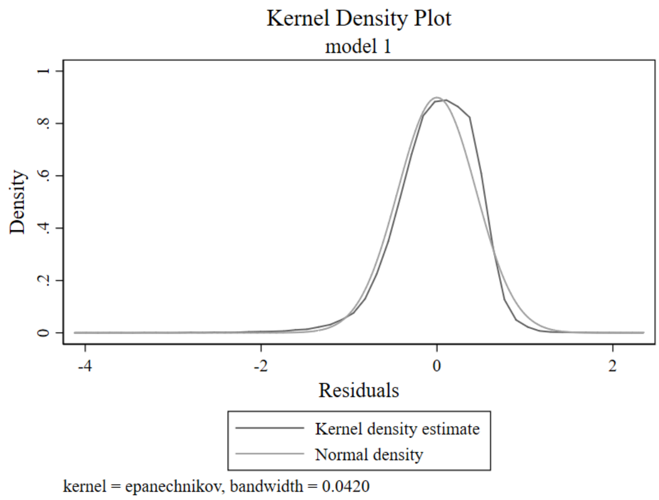

Figure A.6: Kernel Density Plot

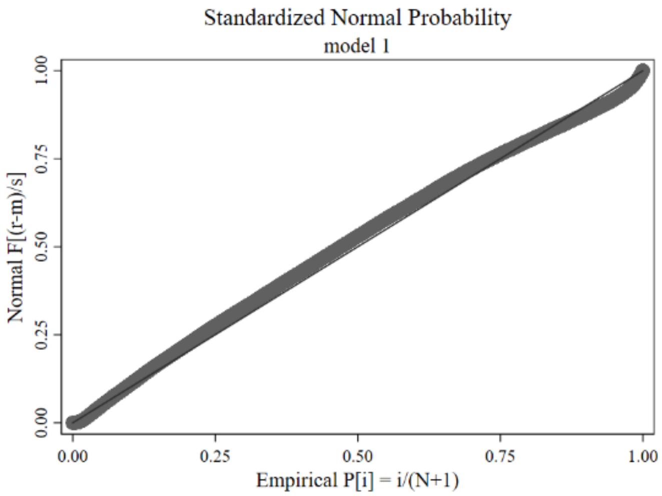

Figure A.7: Standardized Normal Probability Plot

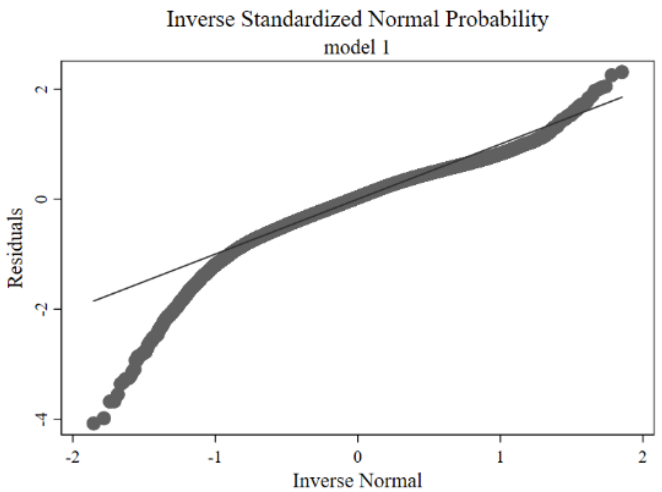

Figure A.8: Inverse Standardize Normal Probability Plot

Appendix G. OLS Results

Table A11: Regression Results Hypothesis 1 (OLS)

	(1)	(2)	(3)
Variables	Apartments & Houses	Apartments & Houses	Apartments & Houses
Singles (%)	0.0101***	0.00628***	0.000773
	(0.000647)	(0.000708)	(0.000487)
LN Surface (m^2)		0.916***	0.908***
		(0.00720)	(0.00602)
Rooms (#)		0.0111***	0.0154***
		(0.00308)	(0.00264)
Density (%)		0.000410***	-0.000152***
		(2.96e-05)	(2.82e-05)
LN No Diploma (%)		-0.420***	-0.138***
		(0.0141)	(0.0121)
Principal Residences (%)		-0.0133***	-0.00276
		(0.00198)	(0.00170)
Vacant Dwellings (%)		-0.00867***	-0.00318
		(0.00254)	(0.00204)
Owner-Occupied (%)		-0.00157***	-0.000799**
		(0.000574)	(0.000378)
Social Tenants (%)		0.00193***	-0.000749**
		(0.000476)	(0.000328)
Free Housing (%)		0.0157***	0.00154
		(0.00343)	(0.00126)
LN Public Transport (m)		-0.0324***	-0.00354
		(0.00537)	(0.00394)
Quarter = 2		0.00343	-0.000397
		(0.00412)	(0.00380)
Quarter = 3		0.0215***	0.0190***
		(0.00397)	(0.00361)
Quarter = 4		0.0211***	0.0134***
		(0.00430)	(0.00389)
Constant	168,191***	11.19***	10.05***
	(9,718)	(0.207)	(0.147)
# of LF-Effects			143
Observations	67,524	67,524	67,524
R-squared	0.030	0.721	0.763

Standard errors have been adjusted for 2,090 clusters in IRIS-area. Robust standard errors in parentheses. The dependent variable is transformed into a log, as well as the independent variables Surface, No Diploma and Public Transport. The reference category for Quarter is 1. A constant has been included in the regression. The third model includes Location-Fixed effects (LF-Effects), based on 143 different postal codes.

*** p<0.01, ** p<0.05, * p<0.1

Table A12: Regression Results Hypothesis 2 (OLS)

Variables	(3) Apartments & Houses	(4) Apartments Only	(5) Houses Only
Singles (%)	0.000773	0.00123**	0.00469***
	(0.000487)	(0.000504)	(0.00106)
LN Surface (m^2)	0.908***	0.936***	0.558***
	(0.00602)	(0.00623)	(0.0182)
Rooms (#)	0.0154***	-0.00455	0.0214***
	(0.00264)	(0.00296)	(0.00521)
Density (%)	-0.000152***	-0.000118***	0.000402***
	(2.82e-05)	(2.81e-05)	(0.000139)
LN No Diploma (%)	-0.138***	-0.154***	-0.197***
	(0.0121)	(0.0126)	(0.0263)
Principal Residences (%)	-0.00276	-0.00306*	-0.00127
	(0.00170)	(0.00164)	(0.00368)
Vacant Dwellings (%)	-0.00318	-0.00342*	-0.00408
	(0.00204)	(0.00201)	(0.00452)
Owner-Occupied (%)	-0.000799**	-0.00188***	0.00245***
	(0.000378)	(0.000399)	(0.000867)
Social Tenants (%)	-0.000749**	-0.000903***	0.00165**
	(0.000328)	(0.000345)	(0.000782)
Free Housing (%)	0.00154	0.00315**	4.08e-05
	(0.00126)	(0.00152)	(0.00165)
LN Public Transport (m)	-0.00354	-0.00699*	-0.0250***
	(0.00394)	(0.00409)	(0.00861)
Quarter = 2	-0.000397	-0.00183	-0.00137
	(0.00380)	(0.00388)	(0.0109)
Quarter = 3	0.0190***	0.0122***	0.0360***
	(0.00361)	(0.00371)	(0.0101)
Quarter = 4	0.0134***	0.00926**	0.0183
	(0.00389)	(0.00401)	(0.0116)
Constant	10.05***	10.07***	11.35***
	(0.147)	(0.145)	(0.342)
# of LF-Effects	143	143	143
Observations	67,524	59,826	7,698
Adjusted R-squared	0.763	0.785	0.653

Standard errors have been adjusted for 2,090 clusters in IRIS-area. Robust standard errors in parentheses. The dependent variable is transformed into a log, as well as the independent variables Surface, No Diploma and Public Transport. The reference category for Quarter is 1. A constant has been included in the regression. All models include Location-Fixed effects (LF-Effects), based on 143 different postal codes. Model 3 from hypothesis 1 has been included for reference as (1) Apartments and Houses.

*** p<0.01, ** p<0.05, * p<0.1

Appendix H. Chow-Test

Table A13: Chow-Test inputs

	N	*RSS*	k
Apartments & Houses	67,524	7332.762	
Apartments	59,826	6023.296	
Houses	7,698	972.377	
Number of estimated parameters			156

Chow-Test:

H_0: The coefficients estimated over Apartments are equal to the coefficients estimated over Houses.

$$\frac{(RSS_{Apartments\,\&\,Houses} - (RSS_{Apartments} + RSS_{\,Houses}))/k}{(RSS_{Apartments} + RSS_{\,Houses})/N_{Apartments} + N_{\,Houses} - 2k} = \qquad \text{(A4)}$$

$$\frac{(7332.762) - (6023.296 + 972.377))/156}{(6023.296 + 972.377)/(59{,}826 + 7{,}698 - 2 * 156)} = \frac{2.161}{0.104} = 20.760$$

Where $RSS_{Apartments\,\&\,Houses}$, $RSS_{Apartments}$, $RSS_{\,Houses}$ represents sum of squared residual for Apartments and houses, apartments, houses, respectively; k denotes the number of regressors in (each) ‘unrestricted’ regression, including a constant and $N_{\,Apartments}$, $N_{\,Houses}$ are the number of observations for apartments and houses respectively.

F(156, 67,524) = 20.760 is bigger than the critical value of 1.294 on the 1 per cent significance level, meaning that the null hypothesis of parameter stability over the two sample groups is rejected.

Appendix I. Test of Equality

H_0: The coefficient for share of single-person households estimated over Apartments are equal to the coefficients estimated over Houses.

$$z = \frac{\beta_1\,(Apartments) - \beta_1\,(Houses)}{\sqrt{{SE_{Apartments}}^2 + {SE_{Houses}}^2}} = \frac{0.0240 - 0.0448}{\sqrt{0.00533^2 + 0.0215^2}} = -0.9393 \qquad \text{(A5)}$$

Where $\beta_1\,(Apartments), \beta_1\,(Houses)$ are the estimated coefficients for the variable share of single-person households for apartments only and houses only and $SE_{Apartments}, SE_{Houses}$ are the estimated standard errors belonging to this coefficient.

Z = -0.9393, p = .352371

The result is not significant at the 1 per cent level, meaning that the null-hypothesis cannot be rejected, meaning that the coefficient for the share of single-person households is not significantly different for apartments than it is for houses.

Appendix J. Figures considering Spatially Lagged Singles

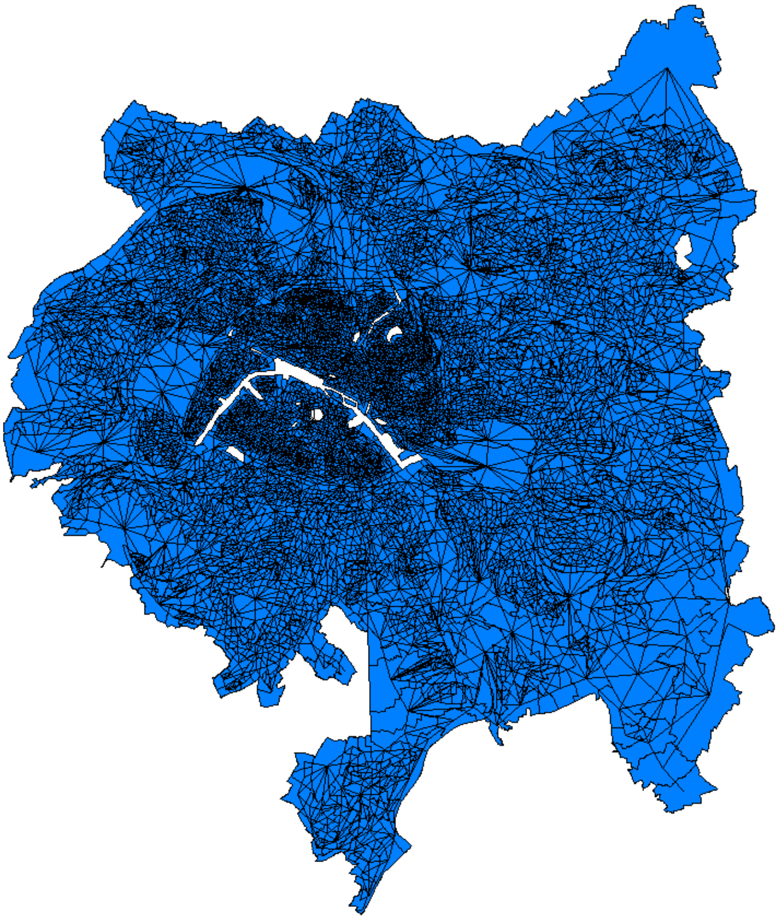

Figure A.9: Connectivity Map of Neighbors

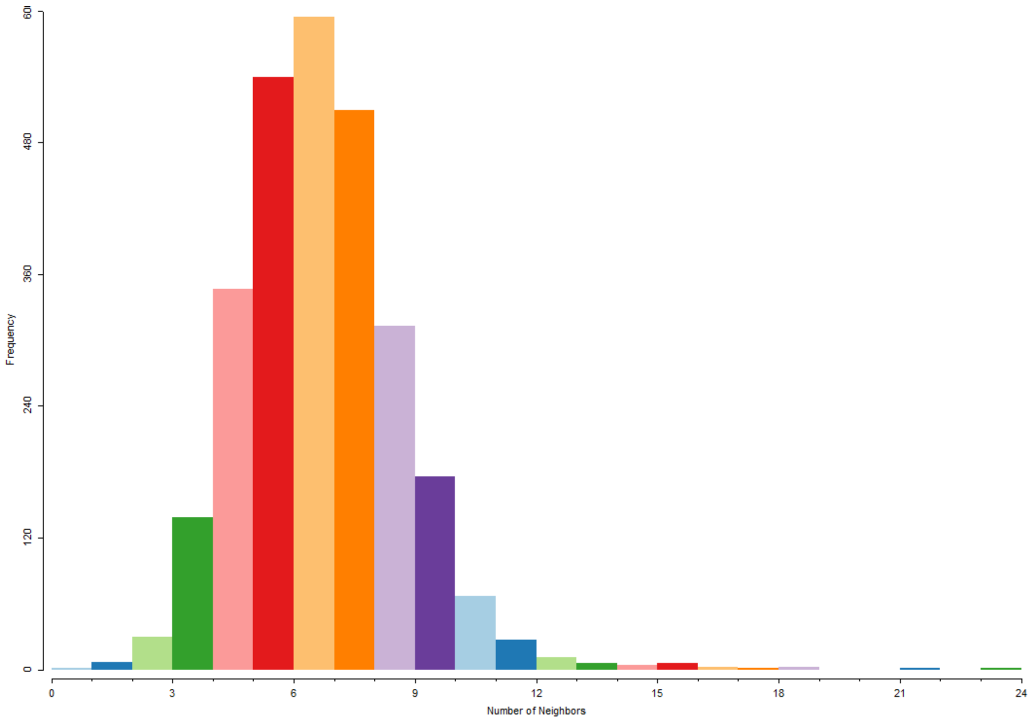

Figure A.10: Histogram with Distribution of Neighbors

Appendix K. Stata Syntax

1. Transforming downloaded files of DGFiP (Property Transaction Prices) into usable .dta files for STATA in RStudio

*Import dataset of France

France2015geo <- read.csv("D:/France2015geo.csv")

View(France2015geo)

*Get a package that allows a transformation to .dta (STATA) files

install.packages("foreign")

Library(foreign)

*Export the dataset into a .dta file

foreign::write.dta(logementvars, "D://France2015geo.dta")

2. Select Paris Urban Area and Houses and Apartments Only in STATA

use "D:\France2015geo.dta"

*select paris metropolitan area

keep if code_departement == "75" | code_departement == "92" | code_departement == "93" | code_departement == "94"

table type_local

*from Appartement, DÃ©pendance, Local industriel. commercial ou assimil� and Maison use only appartments and houses

keep if type_local == "Appartement" | type_local == "Maison"

3. SpatialJoin and Near in ArcGisPro and Moran's I

A SpatialJoin operation was performed between the APUR dataset (https://carto2.apur.org/apur/rest/services/OPENDATA/RECENSEMENT_IRIS/MapServer/1) and the selected Paris Urban Area and Houses and Apartments Only in STATA, as described in 2, where data on IRIS areas were added to the x- and y-coördinates for apartments.

The distance to Public Transport was calculated using the Near operation calculating the distance between the x- and y-coördinates on which the property was located and the x- and y-coördinates for public transport stations.

Lastly, a Morans's I was calculated using the Spatial Autocorrelation tool (Global Moran's I) for the variables price and singles, where the spatial relationship was based on inverse distance and Euclidian length.

4. **Transforming downloaded .CSV files of APUR (Socio-Economic) into usable .dta files for STATA in RStudio**

```
library(foreign)
*Import Housing dataset
RECENSEMENT_IRIS_LOGEMENT <read.csv("D:/RECENSEMENT_IRIS_LOGEMENT.csv")
View(RECENSEMENT_IRIS_LOGEMENT)
*Create a subset of the data with only necessary variables
logementvars = subset(RECENSEMENT_IRIS_LOGEMENT, select = c(c_ir, l_ir, pct_rp,pct_rsecocc, pct_logvac, pct_prop , pct_loc_prive, pct_loc_social, pct_gratuit))
View(logementvars)
*Export the new dataset into a .dta file
foreign::write.dta(logementvars, "D://logementvars.dta")
*Import Education dataset
RECENSEMENT_IRIS_FORMATION <- read.csv ("D:/RECENSEMENT_IRIS_FORMATION.csv")
View(RECENSEMENT_IRIS_FORMATION)
*Create a subset of the data with only necessary variables
formationvars = subset(RECENSEMENT_IRIS_FORMATION, select = c(c_ir, l_ir, pct_nsco_nondipl))
*Export the new dataset into a .dta file
foreign::write.dta(formationvars, "D://formationvars.dta")
*Import Population dataset
RECENSEMENT_IRIS_POPULATION <- read.csv("D:/RECENSEMENT_IRIS_POPULATION.csv")
View(RECENSEMENT_IRIS_POPULATION)
*Create a subset of the data with only necessary variables
populationvars <- subset(RECENSEMENT_IRIS_POPULATION, select = c(c_ir, l_ir, nb_densite))
View(populationvars)
*Export the new dataset into a .dta file
foreign::write.dta(populationvars, "D://populationvars.dta")
```

5. **Merge APUR Datasets in STATA**

```
use "D:\populationvars.dta"
*Merge Population and Housing datasets
merge 1:1 c_ir using "D:\logementvars.dta"
*Drop merge variable, in order to merge again
```

```
drop _merge
*Merge the dataset of Population and Housing with the Education dataset
merge 1:1 c_ir using "D:\formationvars.dta"
drop _merge
*Export into a new dataset called APURdata
save "D:\APURdata.dta"
```

6. Create lagged value in GeoDa

Importing the APUR-dataset into GeoDa (a software program that is available online and can be used for free, which is designed as an easily operated and graphical introduction to spatial analysis (Anselin, Syabri and Kho, 2006) and create a new variable using the spatial lag calculator (row-standardized).

7. Transform dataset of lagged values to STATA

```
library(foreign)
*import dataset with lagged value
laggedvalue <- read.csv("D:/laggedvalue.csv")
View(laggedvalue)
*write into a stata file
foreign::write.dta(laggedvalue, "D://laggedvalue.dta")
```

8. Analysis in STATA

```
clear all
set excelxlsxlargefile on
import excel "D:\Data\Paris2015ApartmentMaisoncomplete.xlsx",
sheet("Paris2015ApartmentMaisoncomplet") firstrow

*Drop variables that are not needed from DGFiP dataset
drop numero_dis
drop adresse_nu
drop adresse_su
drop code_type_
drop code_commu
drop nom_commun
drop ancien_cod
drop ancien_nom
drop id_parcell
drop ancien_id_
drop numero_vol
drop lot1_numer
drop lot1_surfa
drop lot2_numer
drop lot2_surfa
drop lot3_numer
```

drop lot3_surfa
drop lot4_numer
drop lot4_surfa
drop lot5_surfa
drop nombre_lot
drop nature_cul
drop code_natur
drop surface_te
drop Join_Count
drop TARGET_FID
drop nature_mut
drop lot1_num_1
drop code_nat_1
drop nature_c_1

*Drop variables that are not needed from APUR dataset

drop m2_ip m2_pop m2_emp nb_menage_ nb_menage1 nb_menag_1 nb_menag_2 nb_popmen nb_popmen_ nb_tmm nb_tmm_n5 nb_evo_tmm pct_evo_m_ nb_fam_1en nb_fam_2en nb_fam_3en nb_fam_4en nb_fam_enf pct_fam_1e pct_fam_3e pct_fam_2e pct_fam_4e nb_fam_1_1 nb_fam_2_1 nb_fam_3_1 nb_fam_4_1 nb_fam_e_1 pct_fam__1 pct_fam__2 pct_fam__3 pct_fam__4 pct_fam__5 pct_evo_fa pct_evo__1 pct_evo__2 pct_evo__3 pct_evo__4 n_sq_epci l_epci NEAR_FID

*Label variables
label variable id_mutatio "ID of Mutation"
label variable date_mutat "Date of Mutation"
label variable valeur_fon "Price"
rename valeur_fon transprice
label variable code_depar "Department Code"
label variable type_local "Type of Property"
label variable surface_re "Surface"
label variable nombre_pie "Rooms"
label variable c_ir "IRIS Code"
label variable l_ir "IRIS Name"
label variable pct_m_1p "Singles"
label variable pct_m_1p_n "Share of One-Person Households 5 Years Ago"
label variable NEAR_DIST "Public Transport"
label variable code_posta "Postal Code"

*Drop all duplicates
duplicates report id_mutatio
duplicates tag id_mutatio, generate(duplicatesidmutation)
keep if duplicatesidmutation == 0

*Drop less than one and more than six rooms
table nombre_pie
drop if nombre_pie < 1 | nombre_pie > 6

*Summary transaction price
sum transprice, detail
*Drop the values below 1% and above 99% of the transaction price
drop if transprice < 22000 | transprice >1550000

*Merge other dataset from APUR (soci-economic)
merge m:1 c_ir using "D:\Data\APURdata.dta", force

```
drop if _merge==2
drop _merge

*Drop missing values
drop if code_posta == 0
drop if pct_m_1p == 0
drop if nb_densite == 0 | pct_rp == 0 | pct_rsecocc == 0 | pct_logvac == 0 | pct_prop == 0 | pct_loc_prive == 0 | pct_loc_social == 0 | pct_loc_social == 0 | pct_gratuit == 0 | pct_nsco_nondipl == 0
drop if nb_densite == . | pct_rp == . | pct_rsecocc == . | pct_logvac == . | pct_prop == . | pct_loc_prive == . | pct_loc_social == . | pct_loc_social == . | pct_gratuit == . | pct_nsco_nondipl == .

*Merge dataset with lagged value
merge m:1 c_ir using "D:\Data\laggedvalue.dta", force
drop if _merge==2
*Delete missing values from lagged variable
drop if lagsingle == .

*Generate variable for quarter of the year
gen new_date = dofc(date_mutat)
format new_date %td
gen quarter = quarter(new_date)
drop if quarter == 0

*Label new variables
label variable nb_densite "Density"
label variable pct_fam_en "Families with Children"
label variable pct_rp "Principal Residences"
label variable pct_rsecocc "Occasional Dwellings"
label variable pct_logvac "Vacant Dwellings"
label variable pct_prop "Owner-Occupied"
label variable pct_loc_prive "Private Tenants"
label variable pct_loc_social "Social Tenants"
label variable pct_gratuit "Free Housing"
label variable pct_nsco_nondipl "No Diploma"
label variable lagsingle "Lagged Singles"
label variable quarter "Quarter"

*Set place to save results
cd "D:\Results"

*Install program to create publication tables
sysdir set PLUS "D:\results"
ssc install asdoc, replace
net install asdoc, from(http://fintechprofessor.com) replace
ssc install outreg2, replace
cd "D:\results"

*Analyze the dependent variable
sum transprice
hist transprice, frequency normal title("Histogram Transaction Price")
graph export transactionprice.png, replace

*Generate a log of the transaction price (as the original variable is skewed)
gen lntransprice = ln(transprice)
```

```
label variable lntransprice "LN(Price)"
hist lntransprice, frequency normal title("Histogram LN(Transaction Price")
graph export lntransprice.png, replace
sum transprice lntransprice

*Order variables
order transprice lntransprice pct_m_1p lagsingle surface_re nombre_pie code_depar code_posta
nb_densite pct_nsco_nondipl pct_fam_en pct_rp pct_rsecocc pct_logvac pct_prop pct_loc_prive
pct_loc_social pct_gratuit NEAR_DIST quarter

*Transformation of independent variables
hist pct_m_1p, frequency normal title ("Proportion of Singles")
graph export single.png, replace
*Surface
hist surface_re, frequency normal title("Histogram Surface of Property")
graph export surface_re.png, replace
gen lnsurface_re =ln(surface_re)
label variable lnsurface_re "LN(Surface)"
hist lnsurface_re, frequency normal title("Histogram LN(Surface of Property)")
graph export lnsurface_re.png, replace
*Density
hist nb_densite, frequency normal title("Density")
graph export nb_densite.png, replace
gen lnnb_densite = ln(nb_densite)
label variable lnnb_densite "LN(Density)"
hist lnnb_densite, frequency normal title ("LN(Density)")
graph export nb_densite.png, replace
gen recnb_densite = 1/(nb_densite)
label variable recnb_densite "1/Density"
hist recnb_densite, frequency normal title ("1/Density")
graph export recnb_densite.png, replace
gen sqnb_densite = nb_densite^2
hist sqnb_densite, frequency normal
*No Diploma
hist pct_nsco_nondipl, frequency normal title("Histogram No Diploma")
graph export pct_nsco_nondipl.png, replace
gen lnpct_nsco_nondipl =ln(pct_nsco_nondipl)
label variable lnpct_nsco_nondipl "LN(No Diploma)"
hist lnpct_nsco_nondipl, frequency normal title("Histogram LN(No Diploma)")
graph export lnpct_nsco_nondipl.png, replace
*Principal residences
hist pct_rp, frequency normal title ("Principal Residences")
graph export pct_rp.png, replace
gen sqrtpct_rp = sqrt(pct_rp)
label variable sqrtpct_rp "sqrt(Principal Residences)"
hist sqrtpct_rp, frequency normal
*Share of vacant dwellings - is almost normally distributed
hist pct_logvac, frequency normal title("Share of Vacant Dwellings")
graph export pct_logvag.png, replace
**Share of Owner-Occupied Households
hist pct_prop, frequency normal title ("Share of Owner-Occupied Households")
graph export pct_prop.png, replace
*Share of Tenants in Social Housing
hist pct_loc_social, frequency normal title("Share of Tenant Households in Social Housing")
graph export pct_loc_social.png, replace
```

gen lnpct_loc_social =ln(pct_loc_social)
hist lnpct_loc_social, frequency normal title("LN Share of Tenant Households in Social Housing")
graph export lnpct_loc_social.png, replace
gen sqrtpct_loc_social =sqrt(pct_loc_social)
hist sqrtpct_loc_social
gen sqpct_loc_social =(pct_loc_social)^2
hist sqpct_loc_social
*Housed Free
hist pct_gratuit, frequency normal title("Share of Households Housed Free")
graph export pct_gratuit.png, replace
*Distance to Public Transport
hist NEAR_DIST, frequency normal title("Histogram Public Transport")
graph export NEAR_DIST.png, replace
gen lnNEAR_DIST =ln(NEAR_DIST)
label variable lnNEAR_DIST "LN(Public Transport)"
hist lnNEAR_DIST, frequency normal title("Histogram LN(Public Transport)")
graph export lnNEAR_DIST.png, replace

*Summarize results
summarize lntransprice pct_m_1p lagsingle surface_re nombre_pie nb_densite pct_nsco_nondipl pct_fam_en pct_rp pct_rsecocc pct_logvac pct_prop pct_loc_prive pct_loc_social pct_gratuit NEAR_DIST i.quarter
outreg2 using sumstats.doc, sum(log) replace label

summarize lntransprice pct_m_1p lagsingle surface_re nombre_pie nb_densite pct_nsco_nondipl pct_fam_en pct_rp pct_rsecocc pct_logvac pct_prop pct_loc_prive pct_loc_social pct_gratuit NEAR_DIST i.quarter
outreg2 using sumstantsjoined.doc, sum(log) replace ctitle(Apartments and Houses) label
summarize lntransprice pct_m_1p lagsingle surface_re nombre_pie nb_densite pct_nsco_nondipl pct_fam_en pct_rp pct_rsecocc pct_logvac pct_prop pct_loc_prive pct_loc_social pct_gratuit NEAR_DIST i.quarter if type_local == "Appartement"
outreg2 using sumstantsjoined.doc, sum(log) append ctitle(Apartments Only) label
summarize lntransprice pct_m_1p lagsingle surface_re nombre_pie nb_densite pct_nsco_nondipl pct_fam_en pct_rp pct_rsecocc pct_logvac pct_prop pct_loc_prive pct_loc_social pct_gratuit NEAR_DIST i.quarter if type_local == "Maison"
outreg2 using sumstantsjoined.doc, sum(log) append ctitle(Houses Only) label

asdoc sum transprice lntransprice pct_m_1p lagsingle surface_re nombre_pie nb_densite pct_nsco_nondipl pct_fam_en pct_rp pct_rsecocc pct_logvac pct_prop pct_loc_prive pct_loc_social pct_gratuit NEAR_DIST quarter, label abb(.)
asdoc sum transprice lntransprice pct_m_1p lagsingle surface_re nombre_pie nb_densite pct_nsco_nondipl pct_fam_en pct_rp pct_rsecocc pct_logvac pct_prop pct_loc_prive pct_loc_social pct_gratuit NEAR_DIST quarter if type_local == "Appartement", label abb(.)
asdoc sum transprice lntransprice pct_m_1p lagsingle surface_re nombre_pie nb_densite pct_nsco_nondipl pct_fam_en pct_rp pct_rsecocc pct_logvac pct_prop pct_loc_prive pct_loc_social pct_gratuit NEAR_DIST quarter if type_local == "Maison", label abb(.)

*See if there is correlation among the variables
cor transprice pct_m_1p lagsingle surface_re nombre_pie code_depar nb_densite pct_nsco_nondipl pct_fam_en pct_rp pct_rsecocc pct_logvac pct_prop pct_loc_prive pct_loc_social pct_gratuit NEAR_DIST quarter
cor transprice pct_m_1p surface_re nombre_pie nb_densite code_posta pct_nsco_nondipl pct_fam_en pct_rp pct_rsecocc pct_logvac pct_prop pct_loc_prive pct_loc_social pct_gratuit NEAR_DIST quarter

asdoc cor transprice pct_m_1p surface_re nombre_pie nb_densite code_posta pct_nsco_nondipl pct_fam_en pct_rp pct_rsecocc pct_logvac pct_prop pct_loc_prive pct_loc_social pct_gratuit NEAR_DIST quarter

cor transprice pct_m_1p surface_re nombre_pie nb_densite code_posta pct_nsco_nondipl pct_fam_en pct_rp pct_rsecocc pct_logvac pct_prop pct_loc_prive pct_loc_social pct_gratuit NEAR_DIST quarter

asdoc cor transprice pct_m_1p surface_re nombre_pie nb_densite code_posta pct_nsco_nondipl pct_fam_en pct_rp pct_rsecocc pct_logvac pct_prop pct_loc_prive pct_loc_social pct_gratuit NEAR_DIST quarter, label abb(.)

First regression

reg lntransprice pct_m_1p surface_re i.nombre_pie nb_densite i.code_posta pct_nsco_nondipl pct_fam_en pct_rp pct_rsecocc pct_logvac pct_prop pct_loc_prive pct_loc_social pct_gratuit NEAR_DIST i.quarter

reg lntransprice pct_m_1p surface_re i.nombre_pie nb_densite pct_nsco_nondipl pct_fam_en pct_rp pct_rsecocc pct_logvac pct_prop pct_loc_prive pct_loc_social pct_gratuit NEAR_DIST i.quarter, absorb(code_posta)

estat vif

*Regression without Share of Families with Children under 25 and Share of Tenants in 2nd

reg lntransprice pct_m_1p surface_re i.nombre_pie nb_densite pct_nsco_nondipl pct_rp pct_logvac pct_prop pct_loc_social pct_gratuit NEAR_DIST i.quarter, absorb(code_posta)

estat vif

asdoc estat vif, label abb(.)

*Check OLS assumptions

reg lntransprice pct_m_1p surface_re nombre_pie nb_densite pct_nsco_nondipl pct_rp pct_logvac pct_prop pct_loc_social pct_gratuit NEAR_DIST i.quarter, absorb(code_posta)

**Checking Homoscedasticity of Results

rvfplot, yline(0) name(rvfplot) title("Residual versus Fitted Values")

graph export rvfplot.png, replace

estat hettest

asdoc estat hettest

estat imtest

asdoc estat imtest

**No relationship between variables and error term

reg lntransprice pct_m_1p surface_re nombre_pie nb_densite pct_nsco_nondipl pct_rp pct_logvac pct_prop pct_loc_social pct_gratuit NEAR_DIST i.quarter, absorb(code_posta)

reg pct_m_1p surface_re nombre_pie nb_densite pct_nsco_nondipl pct_rp pct_logvac pct_prop pct_loc_social pct_gratuit NEAR_DIST i.quarter, absorb(code_posta)

predict r_singles, resid

reg lntransprice r_singles pct_m_1p surface_re nombre_pie nb_densite pct_nsco_nondipl pct_rp pct_logvac pct_prop pct_loc_social pct_gratuit NEAR_DIST i.quarter, absorb(code_posta)

outreg2 using dwh.doc, replace ctitle(Price) label

test r_singles

*Check normality of residuals

reg lntransprice pct_m_1p surface_re nombre_pie nb_densite pct_nsco_nondipl pct_rp pct_logvac pct_prop pct_loc_social pct_gratuit NEAR_DIST i.quarter, absorb(code_posta)

predict r, resid

**Kernel density plot

kdensity r, normal name(kdensity1, replace) title("Kernel Density Plot") subtitle("model 1")

graph export kdensity1.png, replace

**Standardized normal probability plot

pnorm r, name(pnorm1, replace) title("Standardized Normal Probability") subtitle("model 1")

graph export pnorm1.png, replace

**Inverse Standardized normal probability plot

qnorm r, name(qnorm1, replace) title("Inverse Standardized Normal Probability") subtitle("model 1")
graph export qnorm1.png, replace
sktest r

*Robust regression OLS
*Model 1
reg lntransprice pct_m_1p, cluster (c_ir)
outreg2 using reg.doc, replace ctitle(Apartments and Houses model 1) label adjr2
*Model 2
reg lntransprice pct_m_1p lnsurface_re nombre_pie nb_densite lnpct_nsco_nondipl pct_rp pct_logvac pct_prop pct_loc_social pct_gratuit lnNEAR_DIST i.quarter, cluster(c_ir)
outreg2 using reg.doc, append ctitle(Apartments and Houses model 2) label adjr2
*Model 3
reg lntransprice pct_m_1p lnsurface_re nombre_pie nb_densite lnpct_nsco_nondipl pct_rp pct_logvac pct_prop pct_loc_social pct_gratuit lnNEAR_DIST i.quarter i.code_posta, cluster(c_ir)
outreg2 using reg.doc, append ctitle(Apartments and Houses model 3) label adjr2

*Model 3
reg lntransprice pct_m_1p lnsurface_re nombre_pie nb_densite lnpct_nsco_nondipl pct_rp pct_logvac pct_prop pct_loc_social pct_gratuit lnNEAR_DIST i.quarter i.code_posta, cluster(c_ir)
outreg2 using reg2.doc, replace ctitle(Apartments and Houses) label adjr2
*Model 4
reg lntransprice pct_m_1p lnsurface_re nombre_pie nb_densite lnpct_nsco_nondipl pct_rp pct_logvac pct_prop pct_loc_social pct_gratuit lnNEAR_DIST i.quarter i.code_posta if type_local=="Appartement", cluster(c_ir)
outreg2 using reg2, append ctitle(Apartments Only) label adjr2
*Model 5
reg lntransprice pct_m_1p lnsurface_re nombre_pie nb_densite lnpct_nsco_nondipl pct_rp pct_logvac pct_prop pct_loc_social pct_gratuit lnNEAR_DIST i.quarter i.code_posta if type_local=="Maison", cluster(c_ir)
outreg2 using reg2.doc, append ctitle(Houses Only) label adjr2

*Amount of categories
codebook code_posta

*Instrument variable regression 2SLS
*Hypothesis 1
*Model 1
ivregress 2sls lntransprice (pct_m_1p = lagsingle), first cluster(c_ir)
outreg2 using ivregresultcirapartmenthouses.doc, replace ctitle(Apartments and Houses model 1) label adjr2
estat endog
estat firststage
*Model 2
ivregress 2sls lntransprice lnsurface_re nombre_pie nb_densite lnpct_nsco_nondipl pct_rp pct_logvac pct_prop pct_loc_social pct_gratuit lnNEAR_DIST i.quarter (pct_m_1p = lagsingle), first cluster(c_ir)
outreg2 using ivregresultcirapartmenthouses.doc, append ctitle(Apartments and Houses model 2) label adjr2
estat endog
*Model 3
ivregress 2sls lntransprice lnsurface_re nombre_pie nb_densite lnpct_nsco_nondipl pct_rp pct_logvac pct_prop pct_loc_social pct_gratuit lnNEAR_DIST i.quarter i.code_posta (pct_m_1p = lagsingle), first cluster(c_ir)

```
outreg2 using ivregresultcirapartmenthouses.doc, append ctitle(Apartment and Houses model 3) label adjr2
estat endog
estat firststage

*Hypothesis 2
*Model 3
ivregress 2sls lntransprice lnsurface_re nombre_pie nb_densite lnpct_nsco_nondipl pct_rp pct_logvac pct_prop pct_loc_social pct_gratuit lnNEAR_DIST i.quarter i.code_posta (pct_m_1p = lagsingle), first cluster(c_ir)
outreg2 using ivregresultcirseperated.doc, replace ctitle(Apartments and Houses) label adjr2
Model 4
ivregress 2sls lntransprice lnsurface_re nombre_pie nb_densite lnpct_nsco_nondipl pct_rp pct_logvac pct_prop pct_loc_social pct_gratuit lnNEAR_DIST i.quarter i.code_posta (pct_m_1p = lagsingle) if type_local=="Appartement", first cluster(c_ir)
outreg2 using ivregresultcirseperated, append ctitle(Apartments Only) label adjr2
estat endog
*Model 5
ivregress 2sls lntransprice lnsurface_re nombre_pie nb_densite lnpct_nsco_nondipl pct_rp pct_logvac pct_prop pct_loc_social pct_gratuit lnNEAR_DIST i.quarter i.code_posta (pct_m_1p = lagsingle) if type_local=="Maison", first cluster(c_ir)
outreg2 using ivregresultcirseperated.doc, append ctitle(Houses Only) label adjr2
estat endog

*chow test inputs
ivregress 2sls lntransprice lnsurface_re nombre_pie nb_densite lnpct_nsco_nondipl pct_rp pct_logvac pct_prop pct_loc_social pct_gratuit lnNEAR_DIST i.quarter i.code_posta (pct_m_1p = lagsingle), first cluster(c_ir)
ereturn list
ivregress 2sls lntransprice lnsurface_re nombre_pie nb_densite lnpct_nsco_nondipl pct_rp pct_logvac pct_prop pct_loc_social pct_gratuit lnNEAR_DIST i.quarter i.code_posta (pct_m_1p = lagsingle) if type_local=="Appartement", first cluster(c_ir)
ereturn list
ivregress 2sls lntransprice lnsurface_re nombre_pie nb_densite lnpct_nsco_nondipl pct_rp pct_logvac pct_prop pct_loc_social pct_gratuit lnNEAR_DIST i.quarter i.code_posta (pct_m_1p = lagsingle) if type_local=="Maison", first cluster(c_ir)
ereturn list
```