Erasmus School of Economics
Urban, Port & Transport Economics
Master thesis

The Impact of the 'Noord-Zuidlijn' Metro Stations on House Prices

Name: Kjeld Snoep
Supervisor: Jeroen van Haaren

April 29, 2020

Abstract

In this paper, we explore the impact of Amsterdam's 'Noord-Zuidlijn' on house prices using house sales data of the years 1990-2019. This paper puts a special emphasis on the construction period of the Noord-Zuidlijn, exploring the combined effect of both the construction progress and distance to the metro line. The amount of information provided on the Noord-Zuidlijn is used as a proxy for this construction progress. The results show a positive effect of a closer position with respect to metro stations. This effect increases as the construction of the metro line progresses. When zooming in on the effect of potential externalities on the adjecent area of the Noord-Zuidlijn stations, this research provides inconclusive results. The outcomes of the data study using a distance variable with a construction progress component are promising, as this variable outperforms commonly used variables in price determination of traffic projects.

Keywords: Accessibility, Amsterdam, Construction progress, Noord-Zuidlijn, Hedonic pricing, House pricing, Metro line

Contents

1 Introduction

Amsterdam's population of 862,965 (2019) is served by a metro network that was extended in 2017, with the opening of the 'Noord-Zuidlijn'. This metro line is the central line of the fish-bone shaped public transportation system in Amsterdam. The quantified benefit of this metro line can be seen in the rise of travelers using public transportation, enjoying a generally faster public transport system. Economic theory suggests that this higher degree of accessibility should also be seen in a value increase in the prices of houses located in the nearby area. Empirical research into this theory is of great importance to both investors and city planners alike as they benefit from a detailed picture of the impact of a potential investment in traffic projects.

The effect of location characteristics on house prices is discussed in many research papers (Luttik, 2000; Ottensmann et al., 2008; Kiel and Zabel, 2008), often concluding significant impact of various location factors. This effect is both researched in theoretical economics (Alonso et al., 1964) and applied economics (Monson, 2009). Both approaches focus on identifying a set of factors that may influence house prices. Past research investigating the impact of accessibility opportunities on a city reveals a positive influence on house prices (Nelson, 1992). Previous studies often focus on a unique combination of time and location. This uniqueness leads to different results per study especially in significance. We notice that these studies are usually performed in the USA, focussing on a variety of transportation options. In general, these studies conclude that improved accessibility leads to higher house prices.

Our study focuses on Amsterdam, the biggest city in the Netherlands and home of the Noord-Zuidlijn. The construction of the Noord-Zuidlijn started in 2002 and was surrounded by both technical and political issues. The metro line was highly controversial before and during its construction. The controversy stemmed from high costs, much delay, fear of property damage in the old city centre of Amsterdam and mistakes made during construction. The metro line finally opened on the 22nd of July 2018, seven years later than the initially planned opening moment.

In this paper, we investigate the effect of the distance to the Noord-Zuidlijn on house prices with a special interest in the construction period. We examine the effect of the future presence of the Noord-Zuidlijn during the entire scope of the construction period. We research

whether the future value of the Noord-Zuidlijn is already incorporated in house prices from the beginning of the construction period, leading to different prices before its opening. This is done in a data study, using house sales in Amsterdam in a linear regression model.

The remainder is structured as followed: we begin with a Literature review in Section 2, in which theoretical framework is introduced and complimented with case studies. After providing a data description in Section 3 is given, we introduce the methodology in Section 4. Thereafter, Section 5 presents the results and elaborates on the numbers provided in this section. Finally, we provide our concluding observations.

2 Literature Review

The construction of the Noord-Zuidlijn addresses a problem concerning the effect of accessibility. The opening of the Noord-Zuidlijn increased the accessibility of regions around the metro stations. The gain in accessibility results in an external effect in these regions as well, such as change in travel times from and to these regions and an effect on the sight of the region. Therefore, the presence of the Noord-Zuidlijn and its externalities may have an effect on house prices in the region around the metro station.

In our case, this effect is more difficult to measure because the Noord-Zuidlijn was not yet open during a part of our period of interest. This complication makes that we should not only analyse the effect of the Noord-Zuidlijn, but also should take future value into account in the period it was not opened yet.

Many land value theories are based on the work of Von Thünen (1875), who explains differences in farmland values, using an accessibility argument. However, this model only works for land value of farmland and not the ground on which houses are placed. Brueckner et al. (1987) later on explained that the papers of Alonso et al. (1964); Mills (1967); Muth (1969) can be synthesised in one 'AMM-model', which extends this theory to an urban model. One of the conclusions of this model is that the price per square foot is higher in a business centre, based on the assumption that people living further away from the business centre have to be compensated for their commute. This suggests that land value is lower, and housing sizes are larger in the suburbs (Brueckner et al., 1987). However, the costs of this commute can also be measured in travel time, rather than distance, without the assumptions failing to hold. Note that for this time interpretation, compared to distances, a set of more complex polygons as (changed) congestion and/or public transport can make traveling time non-linear in relation to distance (Chowdhury et al., 2015). Therefore, living near stations connecting to the centre can act as a substitute for living in the centre itself, thereby taking pressure of the housing market in the centre and moving this towards areas around these stations (Fejarang, 1993).

We conclude this theoretical framework by stating that new public transportation, such as the Noord-Zuidlijn, can have an influence on the land/house prices in cities, as it changes the accessibility of certain areas along the route.
Many researchers have investigated this effect of transits. One of the first authors to examine

the effect of new transit lines on house prices was Spengler (1930), who found that new transit lines have a positive effect on the change of house prices when compared to houses in parts of the city not near to these transit lines. However, he also suggests that the role of accessibility is limited compared to other location factors, such as individuals' convenience and comfort of living in an area. The quantitative methods in this field did develop over time. Adkins (1957) was one of the first to focus on house price changes, caused by a new central expressway, using a model comparing house price changes in different sections of Dallas. After comparing these sections he concludes that this expressway has a positive effect on house prices. After this study, multiple papers used before-after comparison estimations to calculate the effects of public transportation on house prices (Gatzlaff and Smith, 1993; Damm et al., 1980; Weinstein et al., 2002). We provide an overview of these studies in Table 1, in which it is shown that these studies create similar conclusions.

The seminal works of Lancaster (1966) and Rosen (1974), result in a rich tradition of studies focused on explaining the heterogeneity of goods, such as houses, using homogeneous attributes. In the models used for these studies, each attribute has an effect on the price of the goods. Much research, using these hedonic pricing models, focuses on house prices, with a specific interest in the effect of accesibility (Sands, 1993; Forrest et al., 1996; Cervero and Duncan, 2002).
In these pricing models, several variables are taken into account in the explanation of house prices, which means that the data is not only corrected for the variable of interest, but also for other characteristics that may change per case. This aspect of the model makes it easier to compare different samples of houses. As this model has the potential to take more variables into account, it can give a more detailed explanation of variables influencing house prices.

When examining the results of the research shown in Table 1, we see that, in general, new traffic projects have a positive influence on the house prices in the region. This effect is still inconclusive as several papers in which this effect is insignificant are present.

If we compare the results, we see that the heavy rail constructions appear to have a more significant effect than light rail constructions. Papers researching the effect of busses appear to have an insignificant effect in even more cases (Yang et al., 2019). Note, the research on the effect of busses is not as extensive, as the policy relevance is less substantial. Following the trend in papers provided, we can assume that the presence of Heavy Rail transport, such as

the Noord-Zuidlijn is more likely to have a significant effect on its neighbourhood.

As these papers are studies of many different regions and their conclusions are in line with economic theory, we expect that this is also the case for the Noord-Zuidlijn. In all cases, either a comparison or hedonic pricing model is used when investigating the effect of public transportation. These two methods appear to lead to significant results and give a straightforward variable interpretation as output. However, more often than not, these studies do not have any spatial variables incorporated in their models. This might create an omitted variable bias when stations are only located in central or rural terrain. Therefore, we should take spatial variables into account for the Noord-Zuidlijn, which is partly located in the city centre.

In other fields, many alternative approaches exist when it comes to the determination of prices. In these methods, not only linear time trends are used, but other variables are considered too. These variables can potentially cover for variance over time. This might be useful as, considering the problems and procrastination over the construction period, both the expected construction time and general opinion of the Noord-Zuidlijn were probably not linear. (Kristoufek, 2013) uses both Google Trends data and Wikipedia data as a trend variable to estimate the price of the bitcoin, (Li et al., 2014) uses sentiment analysis on news articles over time to estimate stock prices, and (Schumaker and Chen, 2009) uses machine learning on financial news articles to estimate stock prices. These papers are all relatively new, perhaps because a lack of data might have formed a bottleneck in an earlier stage or due to evolution in the applicable methods. The fact that this is new also means that it has not been applied in many fields, explaining the absence of the use of such methods in estimating housing prices. However, we did find research touching upon a related subject in Palos-Sanchez and Correia (2018), who uses Google trends to estimate the Airbnb demand numbers, which give significant results in both Spain and Portugal. We believe that these trends can also potentially be used to estimate the progress of construction of the Noord-Zuidlijn, as the Airbnb demand is probably correlated with the value of the property.

Based on Fisher (1930), we find the construction period of the Noord-Zuidlijn to be important. More specifically, Fischer researches added value over time and usually documents a growth rate larger than one. This finding suggests that housing prices increased from the moment the construction of Noord-Zuidlijn was announced. Unlike Fischer, we expect this increase in house prices at a given location not to be exponential over time. Based on the progress esti-

mation of the Noord-Zuidlijn, which was, following the news, highly unstable over time, we assume that the function of prices cannot be covered in a closed-form expression. We expect the value of this increase in house prices to be correlated with the information provided about the Noord-Zuidlijn. Therefore, we assume that the total amount of information provided can be used as a proxy of the construction progress of the Noord-Zuidlijn. Using this construction progress as a method to flatten the shock of the opening of the Noord-Zuidlijn creates a combination between both the comparison and hedonic pricing models, which is not used in past empirical research of the effect of accessibility on house prices.

Furthermore, it is also noteworthy to look into Nelson (1992), as he finds that living closer to stations, for the very close range, has a negative effect on house prices. This finding is in contradiction with the other papers in which houses closer to researched transport opportunities are worth more. However, this negative effect can be explained by the research of Li and Brown (1980), which suggests that non-residential land uses can have a negative effect within close range of these uses as a result of externalities such as noise or visual changes. Nelson (1992) suggests this negative effect might only be present in high-income neighbourhoods as the people living in these areas are less likely to use public transportation. Simons and El Jaouhari (2004) investigated this phenomenon in an empirical study by focused on houses closer than 750 feet from a railroad track. They found that this proximity has a negative effect on house prices. This negative effect is potentially created by nuisance or potential occurrence of accidents. Sirmans et al. (2005) confirms this in an overview of papers, as they report that a metro station within a quarter of a mile has a negative effect on house prices. This effect may also be present in our case as crowds and the metro line itself can produce nuisance during the entire day. As a result, this can lead to a smaller increase or even decrease of the house prices within close range to the metro compared to houses further away from metro stations.

Other factors that have an influence on house prices have to be taken into account too. Sirmans et al. (2005, 2006) show that many variables have been used in hedonic pricing models, of which certain characteristics have been used very often and turn out to be very popular. Examples of these characteristics are age, size and the presence of air-conditioning in a house. An overview of the 20 most used variables used in hedonic pricing studies is given in Table 10 in Appendix A.

When comparing these variables with the literature in our research, we see that many of these

variables are used in hedonic pricing models that aim to explain accessibility. However, we also notice some differences in variable choices. It appears that many sets of variables in the papers used in our research use a more compact set of variables. Furthermore, different types of location data are used more often too.

Table 1: Overview of studies investigating the effect of accessibility.

Studies	Location	Impact of	impact on	findings	s/ns*	Methods
Debrezion et al. (2011)	Amsterdam, Rotterdam Enschede, NL	Heavy Rail(HR)	Residential property (R)	+HR	s	Hedonic Price (HP)
Nelson (1992)	Atlanta, US	HR	R	-distance to station +distance to station squared	s	HP
Bollinger et al. (1998)	Atlanta, US	HR/ Highway	Office Rent	+HR / -Highway	s	HP
Li et al. (2019)	Beijing, CN	Light Rail(LR)/ Highway	R	+LR/ +Highway	s	HP
Armstrong Jr (1994)	Boston, US	HR/Highway	R	+HR/-Highway	ns	HP
Hess and Almeida (2007)	Buffalo, US	LR	R	+LR	s	HP
Sands (1993)	California, US	HR	R	+HR	s	Comparison
Landis et al. (1994)	California, US	HR/ LR/ Highway	R	+HR/ -LR/ +Highway	ns	HP
Weinstein et al. (2002)	Dallas, US	LR	R	+LR	N/A	Comparison
Weinstein et al. (2002)	Dallas, US	LR	Commercial Property (C)	+LR	N/A	Comparison
Cervero and Duncan (2002)	Los Angeles, US	HR/ LR/ Highway	R	inconclusive	s/ns	HP
Cervero and Duncan (2002)	Los Angeles, US	HR/ LR/ Highway	C	+-HR/ +LR/ -Highway	s/ns	HP
Dorantes et al. (2011)	Madrid, ES	LR	R	+LR	s	HP
Forrest et al. (1996)	Manchester, UK	LR	R	-LR	3∗s/1∗ns	HP
Gatzlaff and Smith (1993)	Miami, US	LR	R	+LR	N/A	Comparison
Gatzlaff and Smith (1993)	Miami, US	LR	R	+LR	ns	HP
Chen et al. (1997)	Portland, US	LR	R	+LR	s	HP
Cervero (2010)	San Diego, US	HR/ LR/ Highway	R	inconclusive/ -Highway	s/ns	HP
Cervero (2010)	San Diego, US	HR/ LR/ Highway	C	+-HR/+-LR/+Highway	s/ns	HP
Lawless and Gore (1999)	Sheffield, UK	Tram	R	No negative effect	N/A	Comparison
Du and Mulley (2007)	Sunderland, UK	Metro	R	+Metro	ns	Comparison
Du and Mulley (2006)	Tyne & Wear Region, UK	Public transport/ Car accessibility/ LR	R	+public transport/ -car accessibility/+LR	s	HP
Damm et al. (1980)	Washington, US	Metro	R	+Metro	s	HP
Damm et al. (1980)	Washington, US	Metro	C	+Metro	s	HP
Debrezion et al. (2007)	several studies	HR/LR	R/ Rent	+HR / +LR	s	Comparison

* Significant/non significant result according to the study in question.

To conclude, based on both the theoretic and applied literature, we construct the following two hypotheses:

- *The effect of the Noord-Zuidlijn on house prices is negatively related to the distance from the Noord-Zuidlijn metro stations. This effect of distance becomes more clear as the information provided on the construction of the Noord-Zuidlijn increases.*
- *The Noord-Zuidlijn has a negative effect on the price of houses within a very close range of the Noord-Zuidlijn.*

3 Data

In this section, we introduce the data that is used to evaluate the hypothesis created in the previous section and we provide descriptive statistics to give clear insight of the data structure. We perform analysis on house sales data provided. The NVM [1] gives us access to sales data on houses registered in Amsterdam, containing all registered house sales from January 1, 1990 until June 28, 2019. The total sample size is $173{,}454$ data points. An overview of information of the variables is given in Table 11 in Appendix B.

3.1 Individual Data Selection

As the NVM data is based on forms, filled in by individuals and there is no check on these forms, we might have unrealistic data points in our sample. Therefore, we decide to clean the data by dropping unrealistic observations. This sequential cleaning process is described below.

3.1.1 Transaction Prices

We start our analysis of the NVM data with the most important variable, the transaction price. This variable is the dependent variable of our model. The transaction prices we have range from $-€1$ to $€999.999.999$, which seems highly unlikely. Therefore, we decide to put a threshold on both the maximum and minimum value to make the transactions in our dataset more realistic. We decide to drop prices lower than $€100,000$ or higher than $€1,000,000$, which leads to $18{,}402$ observations being dropped. Of these dropped observations, approximately 80% represents an apartment sale. This percentage is in line with the total data sample. Figure 1 shows that many of these dropped sales occur prior to 2000, more than two years before construction on the Noord-Zuidlijn started. We bear in mind that some dropped observations might represent valid data as prices in those years seem to be lower. Despite this remark, we decide not to include dropped data of this period. We do this as these sales took place in a less relevant period.

[1] NVM: Dutch association of real estate agents and appraisers

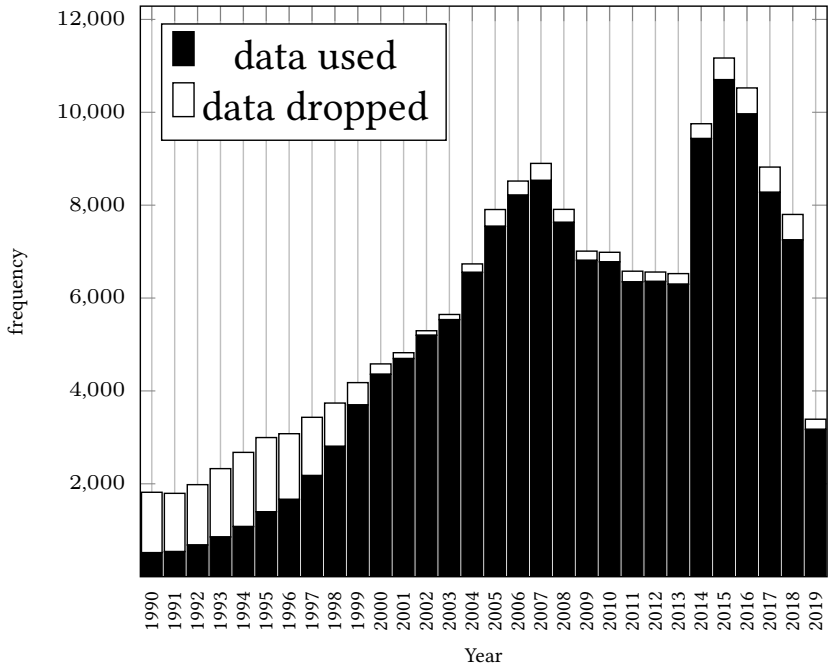


Figure 1: Distribution of values dropped because of an unrealistic transaction price over time.

3.1.2 Geographical Data

The NVM data provided contains an elaborate set of geographical data on every observation, namely an id code for different region levels, a zip code and a house number. These zip codes and house numbers are used to get an exact address for each observation. When checking for these addresses, it turns out that 546 observations no longer exist. We decide to drop these observations as it is hard to locate them. We are aware that some buildings/roads might have changed over time, but we assume the location of each address/path to the closest metro station to be roughly the same as nowadays. Furthermore, we assume that the hedonic data cover for the fact that houses may have been rebuilt or renovated. We use these addresses to calculate the distance to the closest metro station of the Noord-Zuidlijn. Note that, when using these specific geographical variables, other geographical variables can lose value. Therefore, we should be careful when adding other potential geographical variables.

3.1.3 Ceiling Height

We focus on the volume (m^3) and the living area (m^2) of every house. If we combine these variables using the volume divided by the living area, we get the average ceiling height in meters of every house. As these heights range from 0.0068027 to $99,999$ meters, we once

again expect that this is not completely realistic. These unrealistic values can either be due to the floor area, volume, or both. To cover for these unrealistic numbers, we decide to drop the observations with an average ceiling height higher than four meters and lower than two meters, as this is the range of most of the houses. This made us drop a total of $3,382$ observations. These dropped observations are distributed evenly over all house types, except for living farms. Living farms apparently have a higher ceiling significantly more often [2]. This might be because farm houses have a higher ceiling in general. However, as these ceiling heights extend to 19.5 meters, we decide to keep the same thresholds dropping observations of farm houses. We do this because many unrealistic values may still be present if we change these thresholds.

3.1.4 Number of Floors and Rooms

We check for outliers in the number of floors and rooms and notice some unrealistic numbers in both variables. We notice that buildings with up to eight floors were sold, which does not seem very realistic. Therefore, we decide to only keep observations which have up to five floors, which results in 172 observations being dropped. Most of these observations are canal houses and mansions. After this, we investigate the number of rooms in the houses. This number varies from 0 to 103, which does not seem realistic or is simply impossible. Therefore, we choose to drop the houses with less than 1 room or more than 14 rooms, resulting in $1{,}146$ observations being taken out of the used sample. Of these dropped values, only 42 are due to a to big amount of rooms. A total of 17 of these houses turn out to be mansions. However, mansions also account for 78 of the dwellings without any rooms. The combination of the number of rooms being denoted wrongly and the fact that some dwellings have insanely many rooms leads to us concluding that it is a valid choice to put a truncation at 14 rooms, even for mansions.

3.1.5 Period for Sale

Buildings can be for sale for a long period, but that can also occur due to wrong registration of data in observations. This means that we cannot trust the validity of the other data within these observations, as at least one part is probably not true. We use the difference between the date on which a house was registered and unregistered to compute the number of days

[2] A 1% significance level is used in all cases unless stated else

the house was for sale. This calculation generates values in a range from -2,111 to 9,751 days, which seems highly unlikely. We drop observations which have a selling period of less than 0 days or more than 500 days, which leads to 3,595 observations being dropped. This also leads to us excluding houses that were actually for sale for a long time. However, these houses probably had a long selling period for a reason. One reason can be overpricing. When selecting these overpriced houses, it would lead to upward bias of house prices. When we investigate the 3,276 houses that have been listed to be for sale for more than 500 days, Figure 2 shows most of these observations occur after the Financial crisis of 2007–08. This is likely a consequence of people not being able to sell their house during the economic crisis itself. Even though the dropping numbers may be skewed towards this period, the distribution of the data over the years does not appear to change its shape.

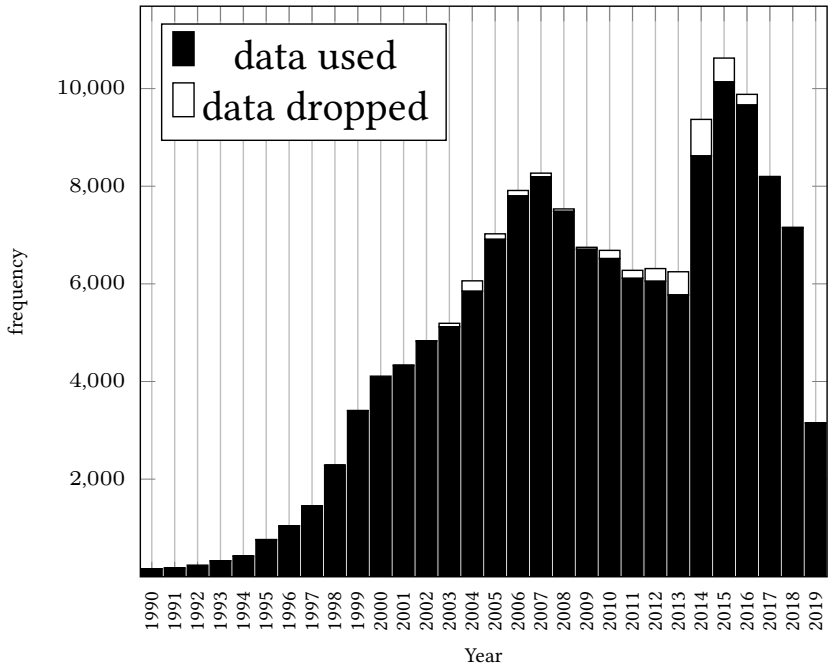


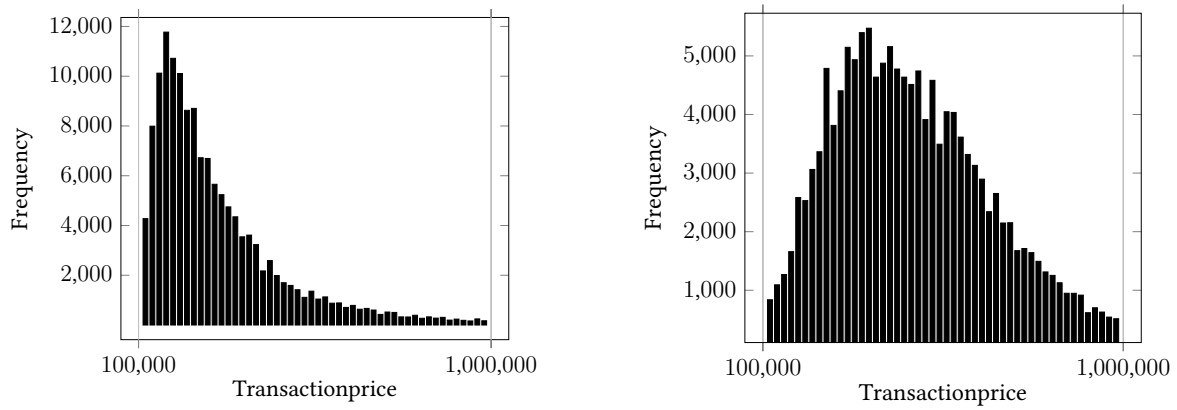
Figure 2: Distribution of sales dropped because of an unrealistic period for sale.

3.2 Data Description

After cleaning the data, we start analysing the distribution of certain variables. These distributions provide much information on the underlying variables themselves. We strive for absence of influential outliers within our variables and can potentially prevent this by modifying variables. Note that we do not discuss subjective variables, such as the garden quality, in our model, as these can vary and are not properly measurable.

3.2.1 Dependent Variable Analysis

The transaction price, which is our dependent variable, as will be explained in section 4, is preferred to be continuous and normally distributed. However, Figure 3a shows that this is not the case for the transaction price, as the histogram is very left-skewed. We deal with this skewness by taking the natural logarithm. This transformation slightly changes the definition of the dependent variable as it now shows procentual differences, as opposed to absolute differences in value.



(a) Distribution of the real transaction price.

(b) Distribution of log transaction price.

Figure 3: Histograms of trancsactionprices.

Figure 3b shows the histogram with log scaling on the x-axis. This histogram is closer to a normal distribution, which means that the model with the log of transaction prices follows the assumptions closer. We are aware that this model is still not normally distributed due to the truncation we created. This can be seen in Figure 14 in Appendix B, where the QQ-plot of this truncated log variable has small tails. This is probably due to the truncation described in Section 3.1.1. Considering this truncation, this variable's distribution is sufficiently close to a normally distributed variable. As we are striving for a normal distribution, we decide to take the log of the transaction price as the variable of interest.

3.2.2 Categorical Variables

When examining categorical variables, we aim for them to be as evenly distributed as possible, such that they cannot cover for outliers in the sample.

Table 2 shows that the building periods are sufficiently evenly distributed over time with at

least a few thousands observations in each period. This creates a good basis for potential regression analysis. Note that as unknown and very old buildings are listed under the same category, we cannot know whether the building era of houses in this category is unknown or ancient. Therefore we drop values within this category. This leads to 106 observations being dropped. We also take note of the low cardinality of some dwelling types, which has to be taken into account if 'dwelling type' is an explaining variable in our final model as it decreases explaining power. This potentially leads to the 'recreational dwelling' being dropped.

Furthermore, when we focus on the directions the gardens are situated, we notice the position of a garden is not known or not present due to absence for over 76% of the data points. The number of observations that would be dropped, if we were to drop potential unknown values, is too big. Therefore, we assume all unknown garden positions have the same effect on house prices as absent gardens. Furthermore, Table 2 shows a strong bias towards the south. Given an equal distribution over the directions we see 'South' occur significantly more often. This hints towards an unrealistic bias for the better garden positions (South and South-West) being filled in. This is another reason for us not to include this variable in our model. However, we do still have the possibility to create a dummy variable stating whether a garden is present or not.

Table 2: Information on all categorical variables.

(a)

Building period	Total
1500-1905	25036
1906-1930	41445
1931-1944	12943
1945-1960	6405
1961-1970	12999
1971-1980	5057
1981-1990	13428
1991-2000	16206
2001≤	9092
Dwelling type	**Total**
Simple house	1073
House boat	227
Recreational dwelling	1
Single-family house	11606
Canal house	462
Mansion	2935
Living farm	27
Bungalow	146
Villa	336
Countryhouse	14
Ground-floor apartment	18722
Upstairs apartment	78640
Maisonette	5195
Porch apartment	12788
Gallery flat	8772
Welfare flat	33
Ground-floor apartment with an upstairs	1636

Dwelling Type	Total
House	16827
Apartment	125784
- With elevator	26023
- Without elevator	99761
Attic stairs	**Total**
Present	816
Absent	141795
Attic	**Total**
Present	4676
Absent	137935
Loft	**Total**
Present	1547
Absent	141064
Practice Room	**Total**
Present	1676
Absent	141064
Living Room Shape	**Total**
L-room	7683
T-room	143
z-room or U-room	533
Open room	8713
Room en suite	8030
Other sort of living room	117509
Balcony	**Total**
Present	77250
Absent	65361

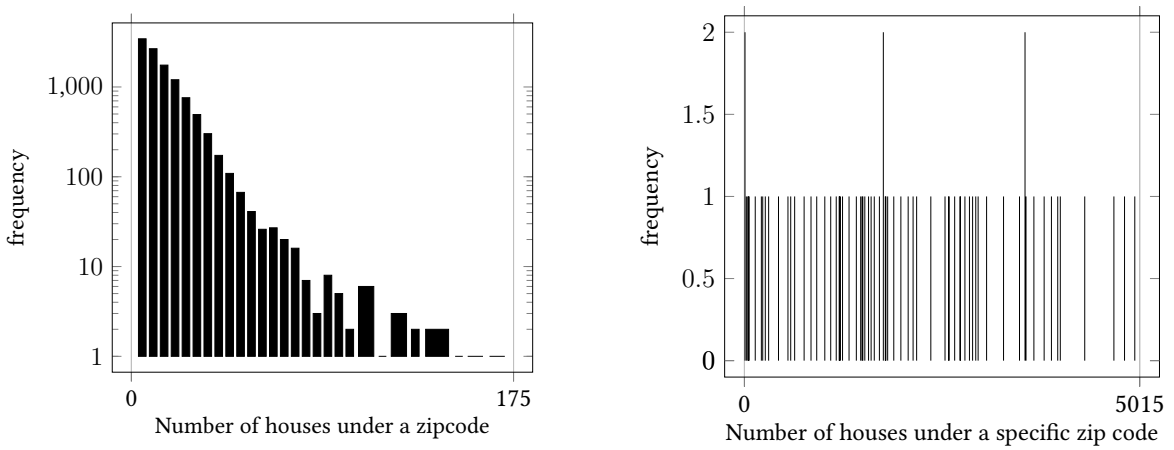
Table 2: Information on all categorical variables, continued.

(b)

Garden positioning	**Total**
North	2496
North-East	1651
East	3904
South-East	3258
South	9095
South-West	5499
West	4901
North-West	2833
No garden	108974
Insulation	**Total**
Present	99566
Absent	43045
Permanently inhabited	**Total**
Yes	141956
No	655

Ground lease construciton	**Total**
Present	78761
Absent	47844
Unknown	16006
Parking space	**Total**
Present	12512
- Indoor	3453
- Outdoor	11759
Absent	127399
Heating	**Total**
Present	133904
Absent	8707
Partially Rented out	**Total**
Yes	464
No	142147

Lastly, in the data sample used, we have all zipcodes of the sold houses, such that they can be used as geodata. However, when analysing zip codes at a 6-sign level, Figure 4a shows that many of these zip codes (3440) only have one to five observations in total. Therefore, we decide to only use the first four digits of the zipcodes. This results in a higher cardinality per region (Figure 4b), and therefore higher significance. Note that this comes with a small loss in data, but potentially leads to clearer estimators regarding location in Amsterdam.



(a) The sample size of the 6-digit specific zip-codes.

(b) The sample size of the 4-digit specific zip-codes.

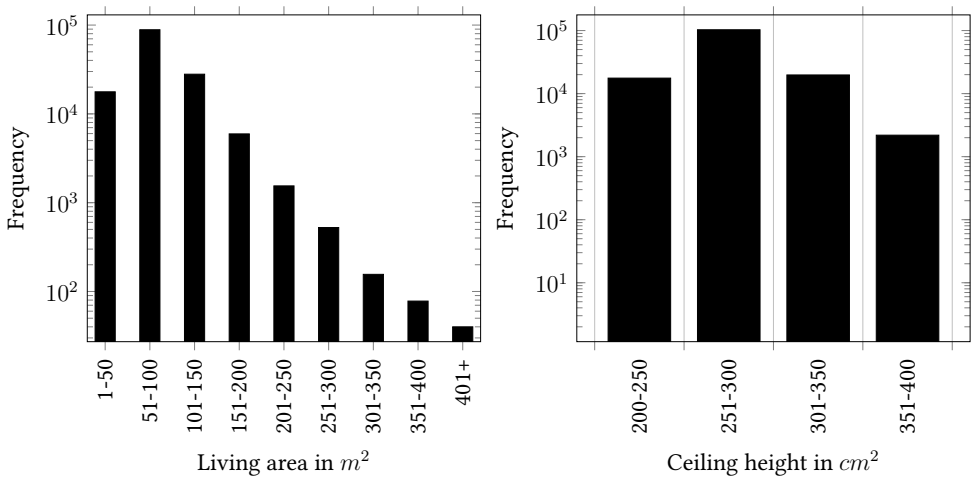
Figure 4: Distributions of zip code frequencies.Each bar represents a range of five different frequencies.

3.2.3 Continuous Variables

We also have many continuous variables given by the NVM. We aim for these variables to be evenly distributed, especially amongst the more extreme values of each variable. These extreme values have a bigger influence on the model, so when they turn out to be outliers, it will create unrealistic values. Figures 5e and 5f present the distribution of number of toilets and the number of bathrooms. These may potentially have an outlier amongst them. It turns out that the possibility for strong outliers is very unlikely for 'Number of bathrooms' as the highest number of bathrooms reported is 5. However, for the number of toilets this number is 20, which may lead to outliers having to much weight. We decide not to perform any further analysis on this high number as we do not trust the validity of the variable 'Number of toilets'. If the variable were to be valid, this would mean that we have no data of houses with one toilet and more than $10{,}000$ observations of houses with no toilets. Therefore, we decide not to use this variable at all.

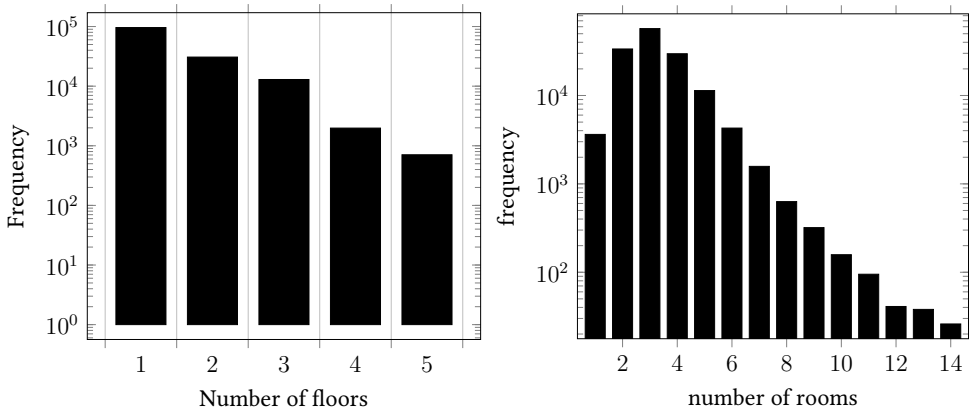
Furthermore, Figure 5 shows that all variables are quite evenly distributed, except for the living area. However, we can partly cover this by taking the logarithm of the squared area to even out the distribution a bit more. Note that we only propose this for 'squared living area', 'ceiling height', and 'volume', as some other variables might become hard to interpret after

taking the logarithm.



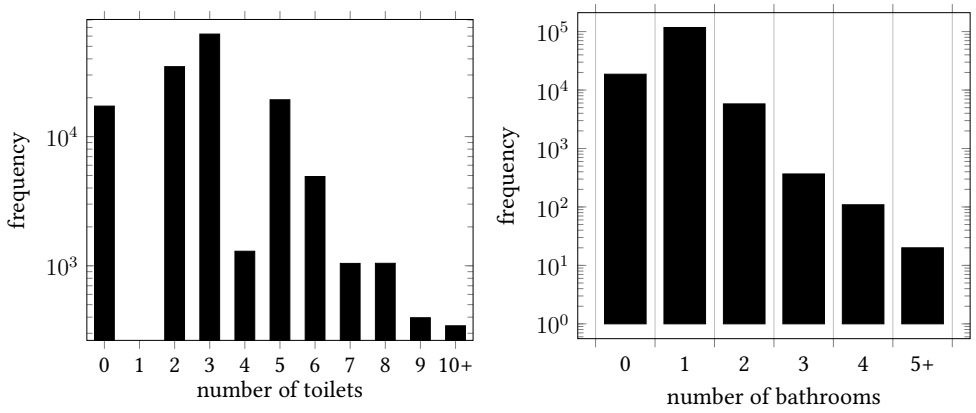
(a) Distribution of the living area of the dataset.

(b) Distribution of the Ceiling heights in the dataset.



(c) Distribution of the number of floors in the dataset.

(d) Distribution of the number of rooms in the dataset.



(e) Distribution of the number of toilets in the dataset.

(f) Distribution of the number of bathrooms in the dataset.

Figure 5: Distribution of all continuous variables.

We also received the date on which a sale is made, which we can use as a trend variable. However, we also decide to partly cover for this trend by using CBS [3] data on the house values. Figure 6 shows this trend contains some conjuncture which might explain a bigger share of the dependent variable than a linear time trend would.

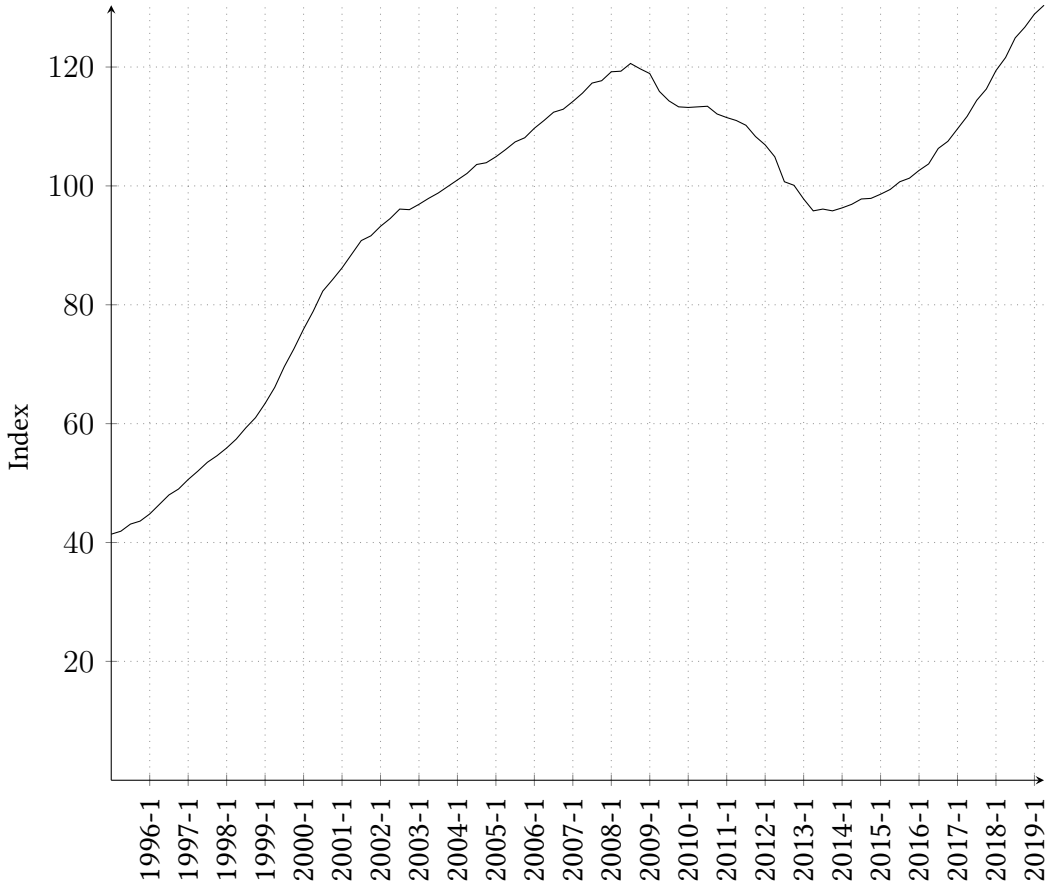


Figure 6: House price index since 1995. The basis (100) is the average over 2015 (CBS, 2020).

3.3 Inter-variable Analysis

Furthermore, we want to analyse the independent variables on their dependency, both with each other and with the dependent variable. Both these analysis are important for significance and for reliability of the effect of the independent variables.

[3]CBS: Statistics Netherlands, a Dutch governmental institution that gathers statistical information about the Netherlands.

3.3.1 Relation with the Dependent Variable

We examine the relation between the individual variables and the log of the transaction price to examine the influence of every variable on the transaction price. We do this both via correlation coefficients and regression analysis.

When we analyse the correlation coefficients, we see that many continuous variables are correlated with the log of the transaction price. The correlations can be as high as 61% for the log of the usable floor space. This correlation might hint towards a variable that is suitable to include in a regression. Following the same reasoning we assume that the presence of an elevator might not have a big influence. We do have to be careful though, as combining variables might create a new variable which does have explaining power. It might, for example, be perfectly reasonable that the presence of an elevator only has an influence when it concerns apartments. Low correlations combined with absence of the variables in Sirmans et al. (2005) hints to 'Attic' and 'Attic stairs' being bad estimators. A total overview of all correlations is shown in Table 12 in Appendix B.

Analysing the variables as regressors gives slightly more insight, as it is a way to express the explanatory power of categorical variables as well. This shows that some variables do not explain very much of the total variance. Table 3 shows that amongst other variables 'Attic', 'Attic stairs', 'Elevator', 'Partially rented out', 'Plot size', and 'Practicing room area' explain less than 1% of the variance. This leads us to choose not to use variables with this little explaining power as individual explaining variables. However, we keep in mind that the 'elevator' may perform better when combined with the House apartment dummy, as elevators do not add much value to houses but can be very appreciated for apartment dwellings. For all other variables, we decide that, as they explain very little of the variance within the dependent variable, they are not useful and will therefore not be used.

Table 3: Overview of R-squared statistics, which states how much in regression on the log of the transaction price using only a constant and one independent variable.

Variable	R-squared	Variable	R-squared
LN Transaction price	1	Attic	0
Attic stairs	0	Balcony	0.0052
Building period	0.1402	Busy road	0.0102
Dormer	0.0051	Dwelling type	0.1333
Elevator	0.0008	Garden	0.0243
Ground lease construction	0.0029	Heating type	0.0248
Indoor parking space	0.0158	Living room shape	0.0161
Location beautiful environment	0.0351	Loft	0.0003
Num bathrooms	0.0195	Num floors	0.1202
Num insulation types	0.0149	Num kitchens	0.0057
Num rooms	0.2152	Num sculleries	0.0081
Open porch	0.0155	parking	0.0253
Partially rented out	0.0001	Permanently inhabited	0.0006
Plot size	0	Practicing room	0.0055
roof terrace	0.0843	Practicing room area	0.0005
Sale condition	0.0149	Time for sale	0.0045
year	0.1336	zipcode 4 digits	0.2266
Ceiling height	0.0891	LN Ceiling height	0.0872
Usable floor area	0.3348	LN Usable floor area	0.3633
Volume	0.3575	LN Volume	0.4001

3.3.2 Relation between Independent Variables

When investigating the relations between the independent variables, we want the correlations between them to be as close to zero as possible. When this is not the case, our model can have multicollinearity issues, which can lead to an imprecise and unrealistic estimated effect of variables. Table 13 in Appendix B gives an overview of all correlation coefficients between

non-categorical variables. Every variable has at least one other variable with which it has a correlation of more than 0.10 or less than -0.10, so we do have to take this into account when explaining the effect of variables on the house price. We even have a few correlation coefficients higher than 0.90 ('Open porch'-'house apartment dummy' & 'LN volume'-'LN Usable floor area'). We decide not to use these variables together in a model as the estimates can be heavily influenced due to this high correlation. We choose not to use 'open porch', as we see that 'dwelling type', an extension of 'house apartment dummy', explains more variance. Additionally, we choose the floor area rather than the volume variable, because 'volume' is a variable which is a factor of other variables used and we expect that, when we add 'Ceiling height' to the model, 'LN volume' will only explain the effect of 'LN Usable floor area'.

3.4 Variable of Interest

Based on Section 2, we conclude it is improbable that the effect of the Noord-Zuidlijn is identical throughout the entire region of Amsterdam. To avoid this, we calculate the distance to cover for geographical differences. We also have to take into account that the Noord-Zuidlijn opened in July 2018. This means there was no added value of the location of the Noord-Zuidlijn up until that moment, but there is potential added value before that moment. To cover for this potential added value we decide to use different types of information to create an insight in the stage of the process towards opening the Noord-Zuidlijn .

3.4.1 Distance

To calculate the distance to the Noord-Zuidlijn, we use Google Maps to get the pedestrian distance to every metro station of the Noord-Zuidlijn from every address and use the closest metro station in our data. We decide not to use euclidean distance, as it is usually impossible to walk in a straight line from an address to the closest metro station. We think the pedestrian distance is most realistic as people using public transportation walk to their first stop (Daniels and Mulley, 2013; El-Geneidy et al., 2014).

Figure 7 shows most addresses in the dataset have a metro station within $4{,}500$ meters. The price of the houses further away are not likely to directly be influenced by the construction of the Noord-Zuidlijn. None of the few outliers present in the data have an extreme value. Hence, this creates little influence if these data points happen to be unrealistic.

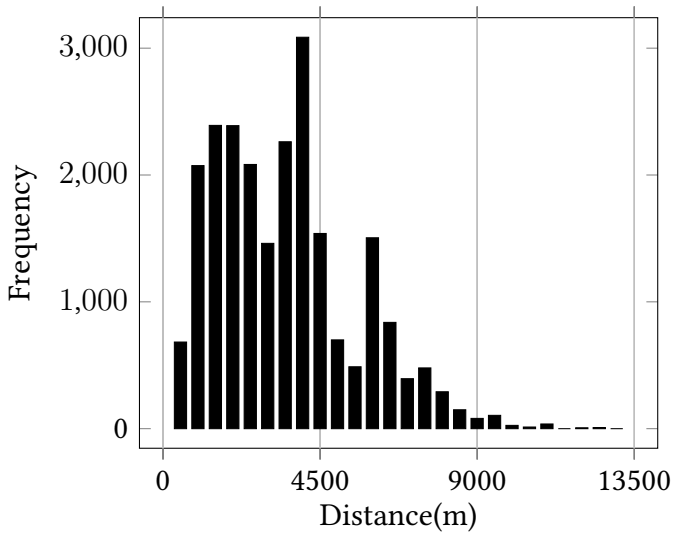


Figure 7: Distribution of the distance to the closest Noord-Zuidlijn metro station from every registered house sale that is used.

3.4.2 Progress Estimates

As listed in the Section 2, we expect the house prices to gradually increase over the period in which the Noord-Zuidlijn was created. Especially considering that the construction of the metro line came with some setbacks which had a direct influence on the region around the trail. We expect the increase in house prices due to the Noord-Zuidlijn to be partly explained in the house sales prices over the entire process of building the Noord-Zuidlijn.

However, we cannot just state that there was a linear construction time. This is unrealistic as the construction took years longer than expected, due to several problems occurring. Additionally, the information on the construction progress has to be distributed before it can be incorporated in the perception of the construction progress of the Noord-Zuidlijn. Therefore, we use other data to represent the information provided in this building process. This data may correct for potential variance in the construction process, and could easily be distributed to potential house buyers. Based on the papers of Schumaker and Chen (2009); Kristoufek (2013); Li et al. (2014) we decide to picture this process using news data and Google Trends [4] analysis. For the news articles we use Nexis Uni [5]. The google trends data report numbers since 2004 on specific phrases.

[4]Google Trends: Search trends feature that shows an indication of how frequently a given search term is entered into Google's search engine

[5]Nexis Uni: online archive for news sources with a regional, national and international character

We perform individual sentiment analysis by reading all news articles of Nexis Uni provided from January 1990 till July 2019, 17,483 in total. These articles are checked for relevance of the term 'Noord-Zuidlijn' in its content. If an article is relevant, it is categorised on positive or negative content regarding the Noord-Zuidlijn. The number of relevant articles is listed per quarter. Figure 8 shows that these numbers contain many peaks and create a volatile line. We see one clear peak in the negative news. This peak appears in the period in which the construction of the Noord-zuidlijn stumbled upon some problems, one of which being the prolapse of houses leading to damage. When these problems were solved, we notice a peak in the good news. Also, peaks in good news are noticeable in both the first and final stages of construction.

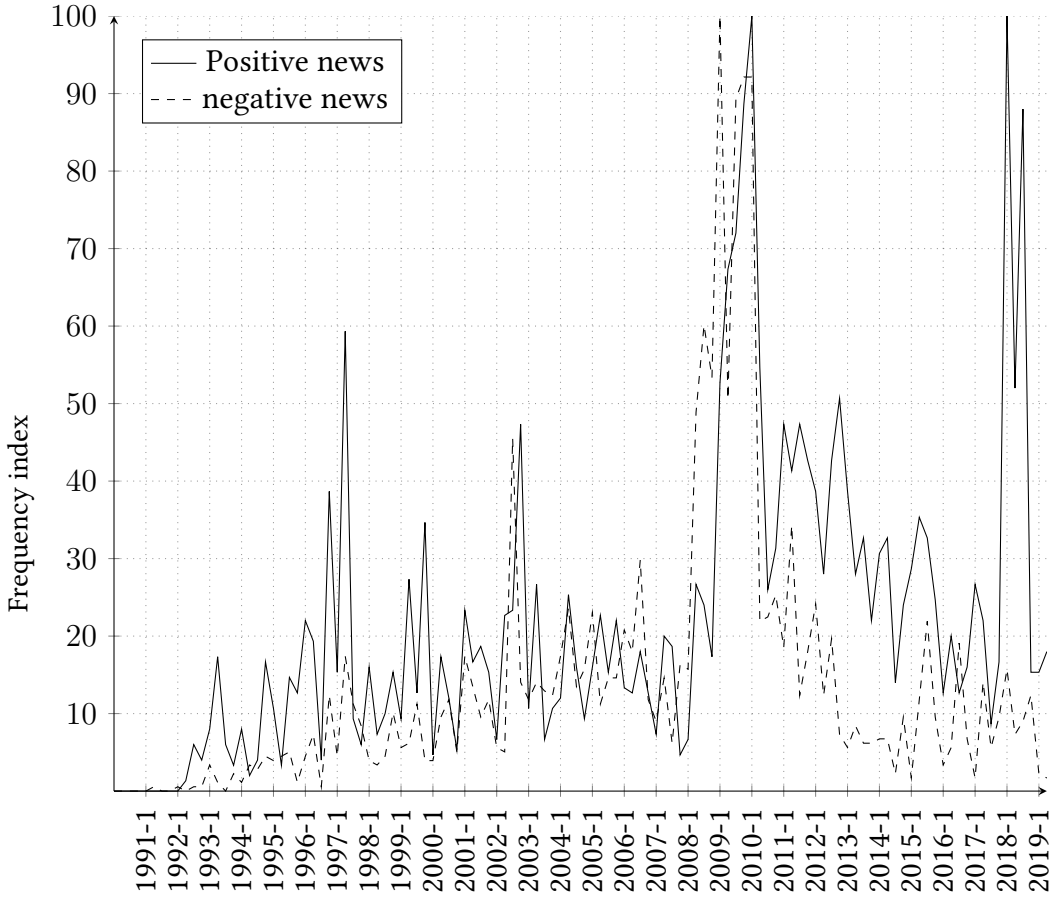


Figure 8: Relative frequency of news articles per quarter since 1990. The basis (100) is the quarter in which the frequency is maximal.

When examining the Google Trends in Figure 9, we notice less clear peaks in the frequency lines. Although we expect the lines to represent a similar pattern, this is not the case, which is odd. The Google Trends variable of 'Metrolijn 52', which is the official name of the Noord-

Zuidlijn, might have been not well known for a big period of the construction of the Noord-Zuidlijn. Hence the late peak and an improbable distribution of the actual progress of the Noord-Zuidlijn. The other two variables show more similarities. It is clear that we see peaks around the same periods. These periods are in line with the the peaks in Figure 8, which hints towards a general peak in public interest during these periods. These peaks are not as clear in the Google Trends data, which might create the idea that the value of information of news provides is not distributed instantly, but over a larger amount of time. This could be an explanation of the spread of search numbers. Therefore, the variables used in Google Trends, represent a different insight in the representation of information of the construction progress.

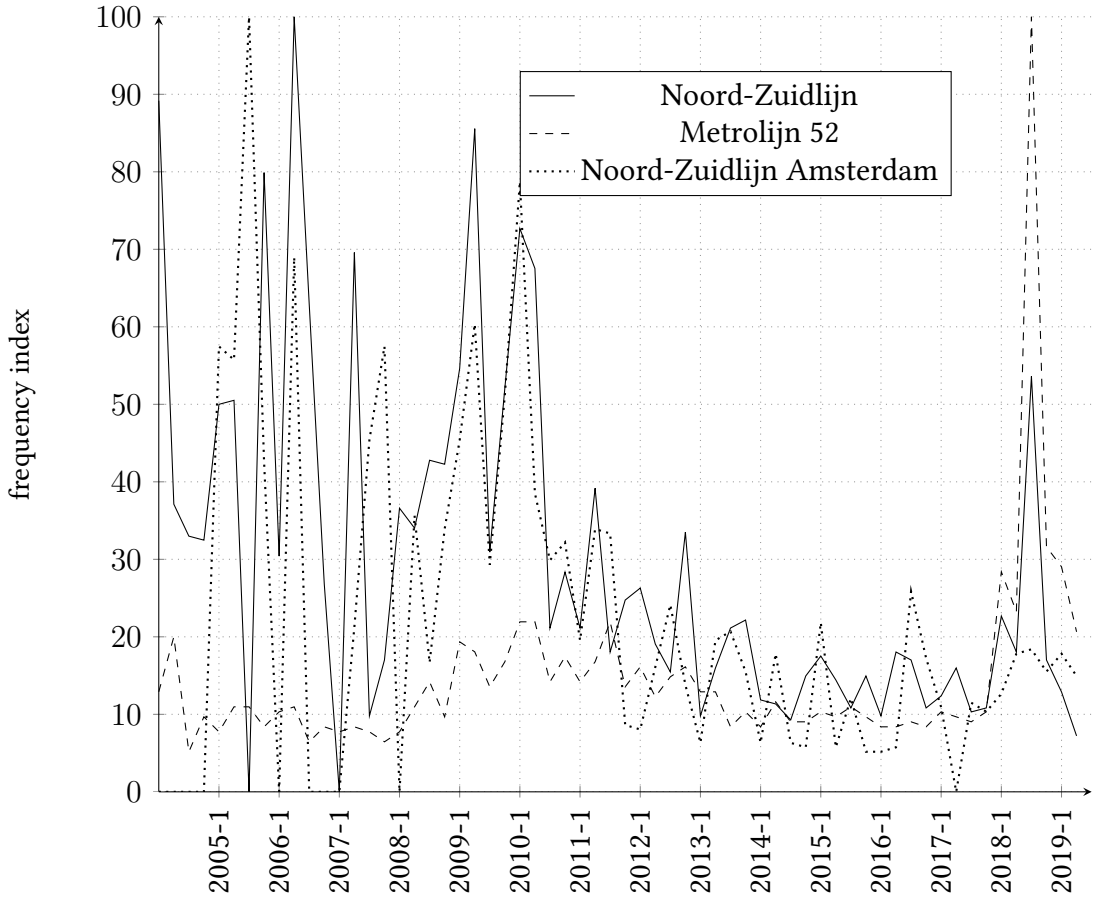


Figure 9: Relative frequency of Google Trends per quarter since 2004. The basis (100) is based quarter in which the frequency is maximal.

3.4.3 Construction of Variables of Interest

We want to create a combined effect between the proximity to the metro stations to cover the benefits of living closer and the effect of the news data or Google Trends data to cover the

construction progress of the Noord-Zuidlijn. Therefore, we have to create a variable which encapsulates the dynamics of both variables.

We decide to create this variable by considering the cumulative number of articles published prior to the quarter in which a house sale is made. We divide this sum by the total number of articles published over the time of construction of the metro line. This creates an information component which grows over time and therefore represents the effect of the house price growing as the construction of the Noord-Zuidlijn advances. We expect that this number roughly describes the stage of the construction as a percentage of the finishing progress. This is based on the assumption that media output represents the amount of information provided. The total amount of information provided is then used as a proxy for the construction progress.

We combine this information component with the distance from a house to a Noord-Zuidlijn metro station, which we will refer to as the distance component. We have to take into account that, as the information component gets larger over time, we want the distance component to be positive for all distances. Additionally, we want the value of the distance component to be larger when a house is located closer to the metro station, such that both a closer distance and more information have a positive effect on the variable of interest. Following these constraints, we decide to take the maximum distance in meters in our set, namely $13,300$, and subtract the distance to the metro station. This procedure results in the following variables:

- $Positive\ articles_i = \frac{\sum_{j=1}^{quarter_i - 1} Positive\ articles\ quarter_j}{Total\ amount\ of\ articles} * (13300 - Distance_i),$

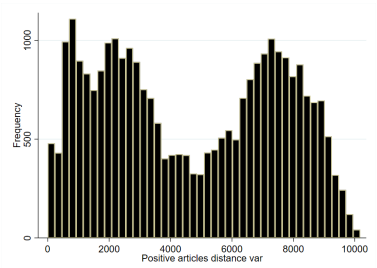
- $Total\ articles_i = \frac{\sum_{j=1}^{quarter_i - 1} All\ articles\ quarter_j}{Total\ amount\ of\ articles} * (13300 - Distance_i),$

- $Trend_i^k = \frac{\sum_{j=1}^{quarter_i - 1} Trend\ index\ quarter_j^k}{Total\ sum\ of\ trends^k} * (13300 - Distance_i).$

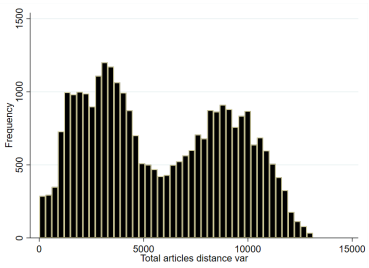
In which i represents the observation index and k represents the trend searches: 'Noord-

Zuidlijn', 'Metrolijn 52' and 'Noord-Zuidlijn Amsterdam'. This variable shows a linear relation to distance and this relation gets stronger when, depending on the variable used, more articles are published or more searches are performed, which is what we strive for.

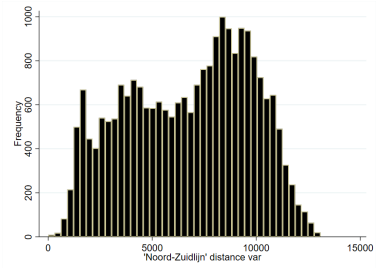
When analysing the distribution of the variables of interest in Figure 10, we notice the trends of each variable create a similar shape, except for the 'Metrolijn 52' variable. This variable has a bigger share of low values, which is probably due to the late growth of the information component. Even though the distributions have an odd shape, all variables do not appear to contain specific outliers in their distributions. Hence, we think that these variables do not have to be cleaned any further.



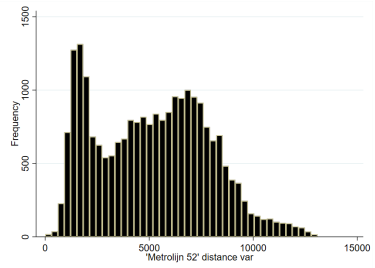
(a) Distribution of 'Positive articles'.



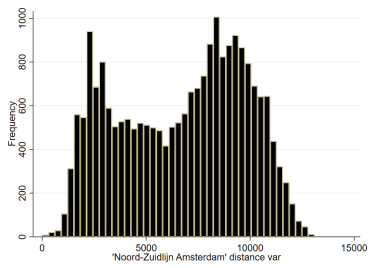
(b) Distribution of 'Total articles'.



(c) Distribution of 'Noord-Zuidlijn Trend'.



(d) Distribution of 'Metrolijn 52 Trend'.



(e) Distribution of 'Noord-Zuidlijn Amsterdam Trend'.

Figure 10: Distribution of the variables of interest created.

The correlations coefficients of the variable of interest and other variables, show that correlations for most variable combinations are low. The only variable that is highly correlated with the variables of interest, is 'year'. The correlation of this variable with the variables of interest is more than 0.62 in all cases. This high correlation can lead to multicollinearity, with a higher variance of the variable as a result. Furthermore, it is possible that strong multicollinearity causes variables to switch signs. Therefore, we decide to exclude 'year' from the variables in our regression analysis. An overview of correlations is presented in Table 15 in Appendix B

We are also interested in an effect for the houses within a close range of the metro stations to test for our second hypothesis. We test for a potential negative effect of externalities within the nearby area of metro stations. Based on the literature we decide to use a distance of 100 meters and 200 meters from the metro stations to create this area. In order be able to test this hypothesis, we create a variable similar the variable of interest mentioned earlier in this section. We create a variable which uses the same information component. This component is combined with a dummy variable stating whether a house is closer than 100 or 200 meters from the metro stations. We will refer to these variables as 'Dist$<$200m' and 'Dist$<$100m'. These variables can be used in the models, in which the variable of interest indicates which information component is used for the construction of this variable.

4 Methodology

From Section 2, we learn that extra transport opportunities can lead to an increase in house prices in general. However, houses that are very close to stations can form an exception on this finding (Sirmans et al., 2005). Additionaly, we created a hypothesis which incorporates the value of information in that section. For this hypothesis, the constructed variable of interest will be used to test this hypothesis for both validity and significance.

We create a linear regression model, using house transaction prices as the dependent variable and several control variables as independent variables. These control variables do not have to explain the entire value of the transaction cost but should describe the main determinants. These variables are partially based on the choices made by Sirmans et al. (2006) and other closely related literature. We expect the independent variables used to be separated in several groups, such as house characteristics ('building period', 'LN usable floor area', 'number of bathrooms'), location characteristics ('zip code'), and external factors ('House value index'). These variables already have to explain a sufficient part of the transaction costs.

When taking a data sample of n observations into account using m variables plus a constant, we can formulate the following linear program.

$$\min_{\beta} \quad \sum_{i=1}^{n} \varepsilon_i^2 \tag{1}$$

$$\text{s.t.} \quad Y_i = \beta_0 + \beta_1 \times x_{1,i} + \cdots + \beta_m \times x_{m,i} + \varepsilon_i \qquad \text{for } i = 1 \ldots n \tag{2}$$

$$\beta_j \in \mathbb{R} \qquad \text{for } j = 0 \ldots m \tag{3}$$

In this program, i represents the observation index and j represents the variable index used.β_j's represent the weight given to each variable j and ε_i represents the error term of observation I. The Y's and X's represent the dependent and independent variables.

We solve this linear regression model using the ordinary least squares (OLS) method. This method describes the value of the dependent variable as a linear combination of the independent variables. We use an error term that is normally distributed with an expected value of zero. Each of these independent variables is given a weight, such that the sum of squared

errors is minimised.

This OLS method is based on seven assumptions. If these assumptions hold, the OLS estimator is the best linear unbiased estimator (BLUE) (Heij et al., 2004). These assumptions are the following:

Firstly, the independent variables have to be non-stochastic. As a result of this, we cannot have more variables than observations. We expect that none of the variables we use are stochastic as we decide to erase the subjective variables, such that every variable is quantifiable.

Secondly, the expected value of the error term has to be zero for every of the n error terms. We cannot be sure whether this is true or not. However, as we dropped most of the unrealistic observations, we expect every observation to be realistic. Therefore, we expect that each observation can be estimated equally well, leading to an expected error term of zero.

Thirdly, the expected variance of every ε has to be equal to a fixed σ. Note that σ does not have to be known. We assume that σ might change slightly over, for example, the usable floor area, as the variance of the value of big houses might be bigger. The same thing might apply to older buildings, but in general, we think these differences will not be big and will therefore not affect the model specification to a large extent.

Fourthly, the expected correlation between two error terms has to be zero for every combination, that is $E[\varepsilon_i \varepsilon_j] = 0$ for all $i \neq j$. We assume this is true, as every sale is a sale on its own. Obviously, there can be some exceptions. For example, there may be big construction projects, which lead to house sales for around the same price range or people owning multiple apartments and valuating them at a specific price. However, we do not think that these cases appear often. Therefore, this will probably not be an impeding factor in the final results.

Fifthly, the model has to be linear. We do expect that this is the case for all variables. The variables consisting of a number can have a small non-linearity because the addition of the first asset (e.g. kitchen) has a bigger added value than the second asset. However, this effect is assumed to be negligible as the extra assets do still add a significant amount of value.

Sixthly, each β is fixed. We assume this is true as we assume that the percentage additional value of each variable to be a rather constant element.

Lastly, the error terms are jointly normally distributed. There is reason to believe this is not true due to the truncation used in our dependent variable. This bound on the variable creates an error term which is not normally distributed for OLS estimates (Cragg, 1971). However, we decide to keep the truncation as unrealistic values would not be in line with the third assumption.

Not all of these assumptions have to hold as strongly in order for the method to be robust. As these assumptions do not have to be too stringent, the OLS regression creates a method that gives a clear insight in the effect of parameters while giving reliable results.

Given that these assumptions hold, we obtain a solution set of $\hat{\beta}$'s based on the solution of the model described in 1-3. In this solution, the $\hat{\beta}$'s represent the weight given to each variable and ε represents the error term. These $\hat{\beta}$'s all have a variance, which is used to investigate whether the true parameter values of the β's are significantly different from zero, using a t-test. If a $\hat{\beta}$ is significantly different from zero, the effect of the variable is considered significant.

When the OLS method is performed on a model with all potential variables of the base model included, we check for significance of each variable using t-tests. We use the General-to-specific selection method (Herwartz, 2010) to generate a model in which only significant variables are present. This method iteratively drops the least significant variable from the model and performs a new OLS regression until all variables in the model are significant. We do this as it provides an efficient way to generate a model which gives significant results. The resulting model is used as our base model.

After we create a base model, we add the variables of interest. We do this by creating an individual model per variable of interest, which consists of the variables of the base model, a constant and that variable of interest. We compare these models by analysing the R-squared statistics and the significance of the variable of interest in each model.

Based on these models, we select a variable of interest that will be used for further analysis. We check individual components of this variable of interest. We do this by comparing the model used with a model that uses either only the information component or the distance component of the variable of interest. Both these models are solved using OLS, which gives us results we can use to compare the performance of different models. When checking for the distance component, we investigate three scenarios: one in which the distance factor of

the Noord-Zuidlijn is added over the entire time span used, one in which we only use the extra value of the Noord-Zuidlijn when construction started, and one where we start using distance factor at the time of the opening. The final two options are considered because these are two easily designated moments at which the public gained valuable knowledge about the accesibility of the region. We create a variable for each of these three cases: 'Distance, always','Distance, construction', and 'Distance, opening'. The latter variable is mostly used in previous research that concerns new construction projects.

Afterwards, we aim to test the hypothesis concerning the effect of externalities of the metro line within a close region of the metro station. We do this by creating two separate models in which 'dist$<$100' and 'dist$<$200' are added to the variable set. We then check for significance of the variable and the sign of the variable. When analysing the results we should bear in mind that the variable of interest that covers all distances is still present in the model too and we should add the effect of this variable to get the effect of the 100 meter or 200 meter range.

Additionally, we can test our models for some of the seven assumptions. We can test for heteroscedasticity using a Breusch-Pagan test (Breusch and Pagan, 1979), which is a Lagrange multiplier test. We apply this test to the variables in our base model as well as any additional variables. We also propose an analysis of the residuals combined with all independent variables and the dependent variable to look for clear non-linearities in the residual pattern, for example, a potential overlooked quadratic term.

5 Results

After transforming all variables, we create linear estimation models. In this section we test the hypotheses formed in Section 2 using the OLS method described in Section 4. We first start by checking for the hypothesis concerning the accessibility effect which means that the distance from houses to the Noord-Zuidlijn should have a negative influence on house prices, this effect should grow as the information provided grows. This can be can be seen by investigating the sign and significance of our variable(s) of interest. Thereafter, we check for the externalities hypothesis, we do this by investigating the effect of adding the 'Dist$<$200m' and 'Dist$<$100m' variables in one of the models.

5.1 Base Model

First, we start by constructing a base model using all variables we consider good estimators. A list of these variables is given in Table 16 in appendix B, This list includes all variables which are accepted in Section 3 as potential input. We apply the General-to-specific selection method, which results in 'Indoor parking space' being dropped. The full specification the resulting base model can be found in Table 17 in Appendix C.

This model has an R-squared value of 0.8242, meaning that it roughly explains 82.42% of the total variance of the dependent variable. The effect of most coefficients on the house price, is expected. However, 'number of floors' and 'number of bathrooms' have a negative effect on the house price, which seems counter intuitive. We think that the coefficient of 'number of floors' is negative because, as the floor space is a variable in our set too, people prefer to have as few floors as possible given the size of the floor area, such that stairs do not take up space. The effect of the number of bathrooms might be explained similarly, as the dwelling types 6-10[6] only have an average of more than one bathroom. These dwelling types are the, in general, more expensive dwelling types. Therefore, this variable might cover some variance of these dwelling types, in which dwellings with more bathrooms usually are less expensive. This is also shown when we regress the number of bathrooms on different dwelling types. We see that is average amount of bathrooms is significantly different for these house types. We

[6]Dwelling types: type 6=Canal house, type 7=Mansion, type 8= Living farm, type 9=Bungalow, and type 10=villa

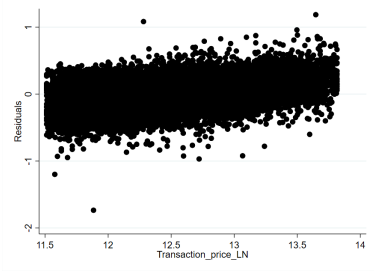
decide to keep this variable as it is easy to determine where its sign switch comes from and therefore actually gives an expected result. However, we do have to keep in mind that the variable explanation might be slightly off for this variable.

Figure 11 shows us that, in general, there is no reason to assume non-linearity in the model. We notice that the trend over the House Price index changed slightly, but the pattern is not extremely non-linear. As all the other variables are categorical, the error terms is always distributed normally around every category of the variable. Therefore, we did not take these variables into consideration. Furhtermore, we conclude that, as the the inclusion of the number of bathrooms does not lead to non-linearity, the inclusion of bathroom does not influence our model.

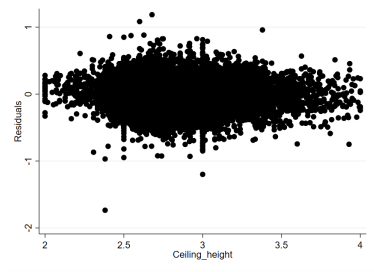
Furthermore, Figure 11 does not show clear hints of heteroscedasticity, as the size of the error terms appear to be equally distributed over the scope of every variable. We do notice some differences, but these bigger outliers often are present within groups or ranges that do have more observations. This number of observations leads to a higher chance of bigger error term showing in the scatter plot. When using the Breusch-Pagan test we do find reason to reject the assumption of homoscedasticity. However, this regression model contains many factors which are low, leading to all variable having an explaining power of less than 7% of the dynamics of the error term, and therefore we are not concerned about heavy differences compared to correcting for these changes.

5.2 Trend Monitored by News Articles

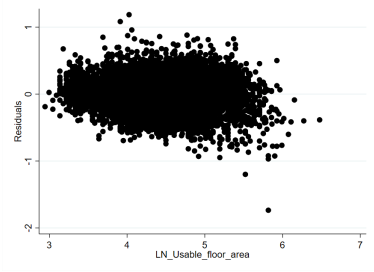
When analysing the variables of interest we investigate a model with the 'positive articles' variable and one with the 'total articles' variable as variable of interest. Conform our hypothesis of positive effect, the coefficients of both estimators are significant and positive. This means our first hypothesis cannot be rejected. When we focus on the significance in Table 4, we see that both estimators are similar, the same thing applies to the number of observations and the R-squared statistic. However, we see a difference in the t-values. These t-values show that although both significant, the significance of 'Total articles' is higher. This, in combination with the slightly higher R-squared value, is reason for us to choose for 'Total articles' as a better estimator of news.



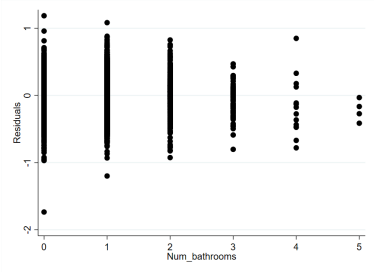
(a) Scatterplot of residuals and the LN transactionprice.



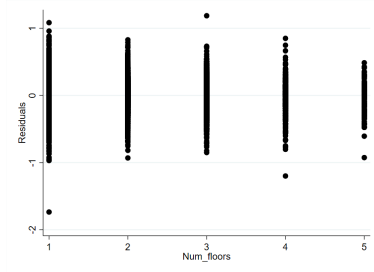
(b) Scatterplot of residuals and the Ceiling height.



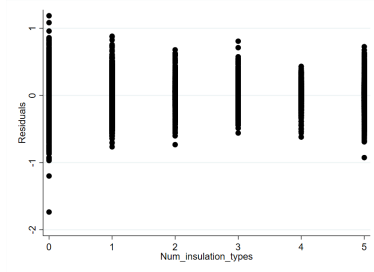
(c) Scatterplot of residuals and the LN Usable floor area.



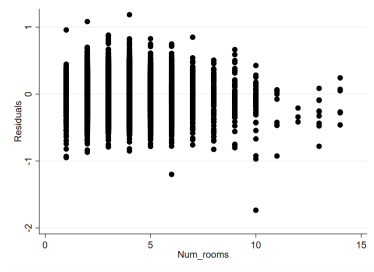
(d) Scatterplot of residuals and the number of bathrooms.



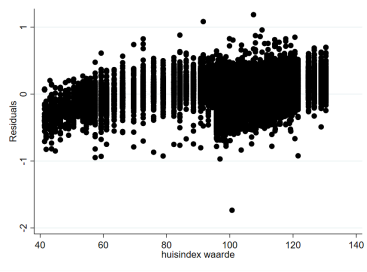
(e) Scatterplot of residuals and the number of floors.



(f) Scatterplot of residuals and the number of insulation types.



(g) Scatterplot of residuals and the number of rooms.



(h) Scatterplot of residuals and the House value index.

Figure 11: This figure represents several scatterplots of the residuals generated in the base model and one of the variables used in this base model.

Table 4: Overview of regression results using the 'News articles' variables of interest.

Modeled variable	Value parameter	t-value	p-value variable	R-squared model	Observations
Base Model				0.8242	28,892
Positive articles	0.0000479	104.74	0.000	0.8728	28,892
Total articles	0.0000441	114.95	0.000	0.8796	28,892

The results in Table 4 show that when, for example, a house which is 300 meters away from the closest metro station is, at a time 50% of the positive articles are published, is expected to

be worth about $0.50 * 0.0000479 * \big((13300 - 300) - (13300 - 1300)\big) * 100 = 2.395$ percent more than exactly the same house 1300 meters away from the metro station (Ceteris Paribus). The same calculation can be made for the second model given in Table 4 with only a change in the parameter. In both cases, our hypothesis concerning the effect of the distance to the closest metro station turns out to be true. We see that, under the ceteris paribus condition, the house price is expected to rise when the distance to the closest metro station of the Noord-Zuidlijn decreases or when more news on the Noord-Zuidlijn has been written after a period. This is in line with our first hypothesis of Section 2.

5.3 Trend Monitored by Google Trends

When doing a similar analysis for the Google Trends variables, we notice that the R-squared values are higher than in the models of the previous subsection. We also notice that, due to Google Trends only providing data for a limited amount of time, the number of observations is lower. This can potentially be an explanation for the better R-squared, as this gives less variables which have to be fit in the model.

In all three models, we see the variable of interest has the right sign and is significant at a significance level of 0.1%. This once again shows that our hypothesis cannot be rejected. This might potentially mean that the progress in the construction is related to the popularity both by news and internet searches.

Table 5: Overview of regression results using the 'Google Trend' variables of interest.

Modeled variable	Value parameter	t-value	p-value variable	R-squared model	Observations
Base Model				0.8274	23,422
'Noord-Zuidlijn' trend	0.0000506	113.26	0.000	0.8845	23,422
'Metrolijn 52' trend	0.0000395	77.01	0.000	0.8572	23,422
Base Model				0.8344	22,225
'Noord-Zuidlijn Amsterdam' trend	0.0000487	105.08	0.000	0.8886	22,225

Table 4 shows that the model which uses Google Trends in 'Noord-Zuidlijn' has the most significant variable of interest, while the R-squared statistic of the model using Google Trends in 'Noord-Zuidlijn Amsterdam' is slightly higher. Therefore, considering our interest, we have no clear way to say which of the two models performs better.

If we compare both these models with the best model of the previous section for the same data. Table 6 shows the 'Total articles' estimator is most significant and creates a model which explains the biggest part of the variance. Therefore, we keep the 'Total articles' as the model used for further analysis.

Table 6: Overview of the comparison of the 'News articles' and 'Google Trends' variables of interest.

Modeled variable	Value parameter	t-value	p-value variable	R-squared model	Observations
Total articles	0.0000584	121.70	0.000	0.8960	22,225
'Noord-Zuidlijn' trend	0.0000514	116.61	0.000	0.8925	22,225
'Noord-Zuidlijn Amsterdam' trend	0.0000503	111.03	0.000	0.8886	22,225

5.4 Isolation Components Variable of Interest

We start our analysis of isolated performance of the individual components by focussing on our interest in the effect of the distance component of our variable. Table 7 shows the situation create when we individually use the distance variable in the three situations described. We see that in every case the model which uses the 'Total articles' model outperforms the other instances. This confirms that, when information provided acts as proxy for the progress of the Noord-Zuidlijn, the progress of the noord-Zuidlijn has a significant influence on the added value of the distance component on house prices. When comparing with 'Distance, opening', which is used in previous research. We notice that the significance of the effect in our model is much higher. Besides, the effect when distance is combined with news in 'Total articles' is significantly higher than the effect in 'Distance, opening'. This is also the case for the other proposed variables, leading to the conclusion that the variable using the progress proxy outperforms the other distance variables.

Table 7: Overview of the individual effect of distance to the Noord-Zuidlijn compared to this variable combined with an information component.

Modeled variable	Value parameter	t-value	p-value variable	R-squared model	Observations
Total articles	0.0000441	114.95	0.000	0.8796	28,892
Distance, always	0.00000598	2.89	0.004	0.8243	28,892
Distance, construction	0.000000874	1.78	0.081	0.8243	28,892
Distance, opening	0.0000195	32.22	0.000	0.8304	28,892

When isolating the news variable, Table 8 shows that the news has a highly significant influence on the price as a whole. Note, that because of the way our variables of interest are constructed, it is not posible to compare the value parameters. However, the model in which the variable which also takes the distance to the Noord-Zuidlijn into consideration creates a higher explanatory power and significance. We are aware that the news itself apparently covers some sort of effect too, but this effect is not as strong as it is when combining it with the distance. Therefore, we conclude that this isolation gives worse values and the distance does add a lot of extra value.

Table 8: Overview of the individual effect of information about the Noord-Zuidlijn compared to this variable combined with a distance component.

Modeled variable	Value parameter	t-value	p-value variable	R-squared model	Observations
Total articles	0.0000441	114.95	0.000	0.8796	28,892
News	0.4588245	113.60	0.000	0.8787	28,892

5.5 Close Stations

We continue our research with analysis of the effect of houses located close to metro stations. Table 9 shows the effect of variables that compensate for houses close to the Noord-Zuidlijn. This table shows that taking the closer distance does not change the effect of the variable of interest much, as all its values and statistics remain roughly the same.

When we investigate the extra added variables, we notice a small decrease in price for all houses closer than 200 meters from the metro station. However, the value given to houses closer than 100 meter gives strong reasons to doubt whether this is a direct effect of the Noord-Zuidlijn. Besides, the maximum value of the parameter (-0.0727117) is not as high as the lowest value added to the house by 'total articles' in this scenario ($0.0000441 * (13300 - 200) * 1 = 0.57771$. This means that, although the effect of the Noord-Zuidlijn on house prices might not be as big compared to places a bit further away, there is still a positive effect. Therefore, we have reason to doubt whether nuisance has a big influence on the added value of the Noord-Zuidlijn being close. This may be because, as a significant part of the metro is located under the ground, the only nuisance complaints can be due to large amounts of travelers passing by, which is hard to compare with the trains and trucks mentioned in the papers used as

inspiration in the literature section. As we can not be sure whether the negative effect of houses closer than 200 meters is because of the metro station, we decide not to confirm the second hypothesis set in the literature section.

Table 9: Test for effect of houses very close to metro stations.

Modeled variable	Value parameter	t-value	p-value variable	R-squared model	Observations
Total articles	0.0000441	114.95	0.000	0.8796	28,892
Total articles	0.0000441	114.94	0.000	0.8796	28,892
$Dist < 100m$	0.0533366	0.71	0.477		
Total articles	0.0000441	111.34	0.000	0.8796	28,892
$Dist < 200m$	-0.0727117	-2.22	0.026		

5.6 Further Model Analysis

Now that we have a final model, we want to evaluate the overall performance of the model we have. When applying a Breusch-Pagan test to our model, it results in some heteroscedasticity being present in our model. However, the effects of all variables combined explains less than 8% of the variance (7,33%), and therefore we are not concerned for the values to differ heavily when using OLS. Figure 12 shows that the truncation used in our dependent variable, against our expectations did not have a big influence on the standard error. On the contrary, the errors turn out to have a bigger tail than the normal distribution. However, the trend does not differ for many variables, since the biggest share of residuals is on the $45°$-line. Additionally the tails do not differ very much, so we do not think that this non-normality affects the significance of or variables very much.

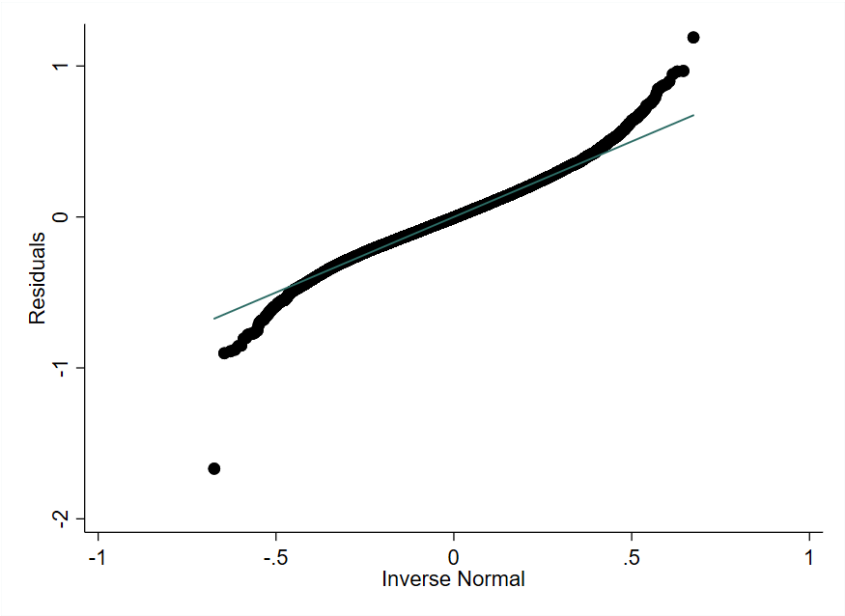


Figure 12: QQ-plot of residuals final model, fitting on a normal distribution.

We already have seen that our model is roughly linear in all variables used for the basis model. However, we did not check for the variable of interest. Figure 13 shows that the errors of the final model, appear to be linear around the x-axis, and therefore we conclude that the estimator is linear.

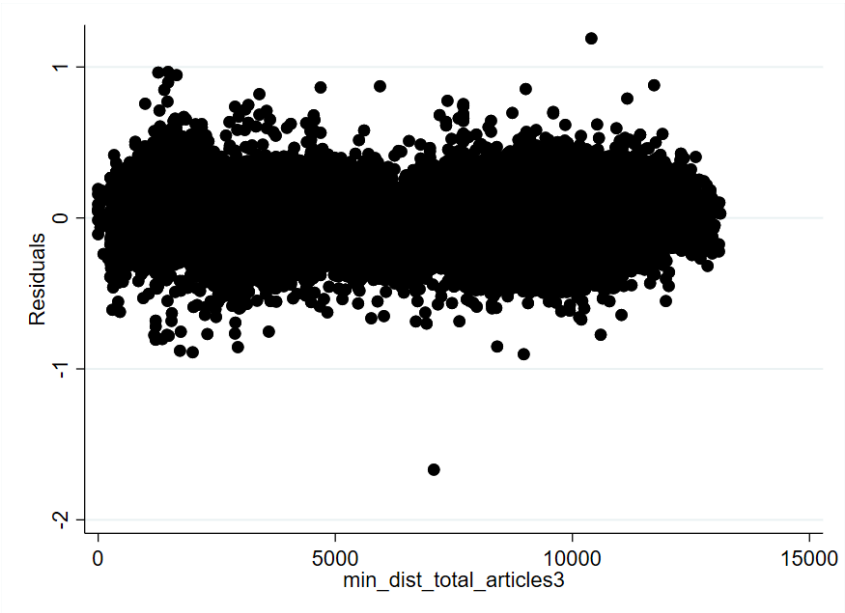


Figure 13: representation of a scatterplots of the residuals generated in the final modal and the variable of interest used in this regression.

Figure 13 also shows us that we only have two outliers with an absolute value higher than 1. When we check both outliers, we notice that their estimated house values are closer to the estimated value than to the value for which it was sold. The listed price for both these buildings does not differ much with the estimation we have, so these two sales might have been typing errors or a special situation under which the sale is made.

6 Conclusion

In this paper, we examine the effect of the presence of the Noord-Zuidlijn in Amsterdam on house prices. Based on the literature we hypothesize that the Noord-Zuidlijn raises the accesibility of regions, leading to higher house prices. This rise is not sudden but grows with the construction of the Noord-Zuidlijn due to the future value being taken into account. We decide to use the amount of information provided as a proxy for the construction progress. Furthermore, following the literature, we expect that externalities in the nearby area of metro stations, such as nuisance, have a negative influence on house prices. We find a significant result confirming our theoretical hypothesis both in case studies found in the literature. However, we notice that none of these case studies take a construction progress into account.

In our empirical study of Amsterdam, we decide to use a variable that incorporates both the progress of the Noord-Zuidlijn and the distance from a house to the closest metro station in a hedonic pricing model. This model, which is used to estimate house prices based on sales in Amsterdam, generates a significant negative effect on distance to the Noord-Zuidlijn. This effect grew gradually as the construction of the Noord-Zuidlijn progressed. We find that this effect is strongest when we use the total number of relevant news articles to embody the amount of information provided. This result holds when comparing with an individual effect of news or distance to the metro stations. In addition, we show that, although the houses located close to the metro station might have a lower price due to nuisance issues, we find this effect not to outweigh the effect of the price increase due to proximity to the Noord-Zuidlijn.

The price rises around the Noord-Zuidlijn are an enormous opportunity in the Netherlands, as people lobbying for new transportation projects get the confirmation that metro lines add value to the houses in the region which is an extra argument for construction.

The significance of a combined effect of information and distance in our study provide room to further analyse and optimise the contribution of both distance and news articles to this combined effect. We are aware that these results are case-specific and cannot set the standard for a similar effect in all other cities. However, our results create a promising basis for future study of similar transportation projects.

References

Adkins, W. G. (1957). *Effects of the Dallas Central Expressway on Land Values and Land Use.* Texas Transportation Inst.

Alonso, W. et al. (1964). Location and land use. toward a general theory of land rent. *Location and land use. Toward a general theory of land rent.*

Armstrong Jr, R. J. (1994). Impacts of commuter rail service as reflected in single-family residential property values. *Transportation Research Record*, (1466).

Bollinger, C. R., Ihlanfeldt, K. R., and Bowes, D. R. (1998). Spatial variation in office rents within the atlanta region. *Urban Studies*, 35(7):1097–1118.

Breusch, T. S. and Pagan, A. R. (1979). A simple test for heteroscedasticity and random coefficient variation. *Econometrica: Journal of the Econometric Society*, pages 1287–1294.

Brueckner, J. K. et al. (1987). The structure of urban equilibria: A unified treatment of the muth-mills model. *Handbook of regional and urban economics*, 2(20):821–845.

CBS (2020). House value indices. data retrieved from the CBS database.

Cervero, R. (2010). Effects of light and commuter rail transit on land prices: Experiences in san diego county. In *Journal of the Transportation Research Forum*, volume 43.

Cervero, R. and Duncan, M. (2002). Land value impacts of rail transit services in los angeles county. *Report prepared for National Association of Realtors Urban Land Institute.*

Chen, H., Rufolo, A. M., and Dueker, K. (1997). Measuring the impact of light rail systems on single family home values: A hedonic approach with gis application.

Chowdhury, S., Ceder, A. A., and Schwalger, B. (2015). The effects of travel time and cost savings on commuters' decision to travel on public transport routes involving transfers. *Journal of Transport Geography*, 43:151–159.

Cragg, J. G. (1971). Some statistical models for limited dependent variables with application to the demand for durable goods. *Econometrica: Journal of the Econometric Society*, pages 829–844.

Damm, D., Lerman, S. R., Lerner-Lam, E., and Young, J. (1980). Response of urban real estate values in anticipation of the washington metro. *Journal of Transport Economics and Policy*, pages 315–336.

Daniels, R. and Mulley, C. (2013). Explaining walking distance to public transport: The dominance of public transport supply. *Journal of Transport and Land Use*, 6(2):5–20.

Debrezion, G., Pels, E., and Rietveld, P. (2007). The impact of railway stations on residential and commercial property value: a meta-analysis. *The Journal of Real Estate Finance and Economics*, 35(2):161–180.

Debrezion, G., Pels, E., and Rietveld, P. (2011). The impact of rail transport on real estate prices: an empirical analysis of the dutch housing market. *Urban Studies*, 48(5):997–1015.

Dorantes, L. M., Paez, A., and Vassallo, J. M. (2011). Analysis of house prices to assess economic impacts of new public transport infrastructure: Madrid metro line 12. *Transportation Research Record*, 2245(1):131–139.

Du, H. and Mulley, C. (2006). Relationship between transport accessibility and land value: Local model approach with geographically weighted regression. *Transportation Research Record*, 1977(1):197–205.

Du, H. and Mulley, C. (2007). The short-term land value impacts of urban rail transit: Quantitative evidence from sunderland, uk. *Land Use Policy*, 24(1):223–233.

El-Geneidy, A., Grimsrud, M., Wasfi, R., Tétreault, P., and Surprenant-Legault, J. (2014). New evidence on walking distances to transit stops: Identifying redundancies and gaps using variable service areas. *Transportation*, 41(1):193–210.

Fejarang, R. A. (1993). Impact on property values: A study of the los angeles metro rail. In *Public transport planning and operations. Proceedings of seminar H held at the European transport, highways and planning 21st summer annual meeting, (September 13-17, 1993), Umist. Volume P:370.*

Fisher, I. (1930). *Theory of interest: as determined by impatience to spend income and opportunity to invest it.* Augustusm Kelly Publishers, Clifton.

Forrest, D., Glen, J., and Ward, R. (1996). The impact of a light rail system on the structure of house prices: a hedonic longitudinal study. *Journal of Transport Economics and Policy*, pages 15–29.

Gatzlaff, D. H. and Smith, M. T. (1993). The impact of the miami metrorail on the value of residences near station locations. *Land Economics*, pages 54–66.

Google Maps (2020). Distance to closest metro stations. data retrieved from various metro-stations, https://www.google.nl/maps/preview.

Google Trends (2020). Various 'Noord-Zuidlijn' trends. data retrieved from Dutch search trends, https://trends.google.com/trends/?geo=NL.

Heij, C., Heij, C., de Boer, P., Franses, P. H., Kloek, T., van Dijk, H. K., et al. (2004). *Econometric methods with applications in business and economics*. Oxford University Press.

Herwartz, H. (2010). A note on model selection in (time series) regression models–general-to-specific or specific-to-general? *Applied Economics Letters*, 17(12):1157–1160.

Hess, D. B. and Almeida, T. M. (2007). Impact of proximity to light rail rapid transit on station-area property values in buffalo, new york. *Urban studies*, 44(5-6):1041–1068.

Kiel, K. A. and Zabel, J. E. (2008). Location, location, location: The 3l approach to house price determination. *Journal of Housing Economics*, 17(2):175–190.

Kristoufek, L. (2013). Bitcoin meets google trends and wikipedia: Quantifying the relationship between phenomena of the internet era. *Scientific reports*, 3:3415.

Lancaster, K. J. (1966). A new approach to consumer theory. *Journal of political economy*, 74(2):132–157.

Landis, J., Guhathakurta, S., and Zhang, M. (1994). Capitalization of transit investments into single-family home prices: A comparative analysis of five california rail transit systems. *UC Berkeley: University of California Transportation Center*. Retrieved from https://escholarship.org/uc/item/80f3p5n1.

Lawless, P. and Gore, T. (1999). Urban regeneration and transport investment: a case study of sheffield 1992-96. *Urban Studies*, 36(3):527–545.

Li, M. M. and Brown, H. J. (1980). Micro-neighborhood externalities and hedonic housing prices. *Land economics*, 56(2):125–141.

Li, S., Chen, L., and Zhao, P. (2019). The impact of metro services on housing prices: a case study from beijing. *Transportation*, 46(4):1291–1317.

Li, X., Xie, H., Chen, L., Wang, J., and Deng, X. (2014). News impact on stock price return via sentiment analysis. *Knowledge-Based Systems*, 69:14–23.

Luttik, J. (2000). The value of trees, water and open space as reflected by house prices in the netherlands. *Landscape and urban planning*, 48(3-4):161–167.

Mills, E. S. (1967). An aggregative model of resource allocation in a metropolitan area. *The American Economic Review*, 57(2):197–210.

Monson, M. (2009). Valuation using hedonic pricing models. *Cornell Real Estate Review*, 7(1):10.

Muth, R. F. (1969). Cities and housing; the spatial pattern of urban residential land use.

Nelson, A. C. (1992). Effects of elevated heavy-rail transit stations on house prices with respect to neighborhood income. *Transportation Research Record*, (1359).

Nexis Uni (2020). Noord-zuidlijn articles. data retrieved from Dutch articles, `https://advance-lexis-com.eur.idm.oclc.org/`.

Ottensmann, J. R., Payton, S., and Man, J. (2008). Urban location and housing prices within a hedonic model. *Journal of Regional Analysis and Policy*, 38(1100-2016-89822).

Palos-Sanchez, P. R. and Correia, M. B. (2018). The collaborative economy based analysis of demand: Study of airbnb case in spain and portugal. *Journal of theoretical and applied electronic commerce research*, 13(3):85–98.

Rosen, S. (1974). Hedonic prices and implicit markets: product differentiation in pure competition. *Journal of political economy*, 82(1):34–55.

Sands, B. (1993). The development effects of high-speed rail stations and implications for california. *Built Environment (1978-)*, pages 257–284.

Schumaker, R. P. and Chen, H. (2009). Textual analysis of stock market prediction using breaking financial news: The azfin text system. *ACM Transactions on Information Systems (TOIS)*, 27(2):1–19.

Simons, R. A. and El Jaouhari, A. (2004). The effect of freight railroad tracks and train activity on residential property values. *Appraisal Journal*, 72(3):223–233.

Sirmans, G. S., MacDonald, L., Macpherson, D. A., and Zietz, E. N. (2006). The value of housing characteristics: a meta analysis. *The Journal of Real Estate Finance and Economics*, 33(3):215–240.

Sirmans, S., Macpherson, D., and Zietz, E. (2005). The composition of hedonic pricing models. *Journal of real estate literature*, 13(1):1–44.

Spengler, E. H. (1930). *Land values in New York in relation to transit facilities*. Faculty of Political Science Columbia university, New York.

Von Thünen, J. H. (1875). *Der isolirte staat in beziehung auf landwirtschaft und nationalökonomie*, volume 1. Wiegant, Hempel & Parey.

Weinstein, B. L., Clower, T. L., Means, F., Gage, L., Pharr, M., Pettibon, G., and Gillis, S. (2002). An assessment of the dart lrt on taxable property valuations and transit oriented development.

Yang, L., Zhou, J., Shyr, O. F., et al. (2019). Does bus accessibility affect property prices? *Cities*, 84:56–65.

A Hedonic Pricing Model

This table shows the 20 most used variables in hedonic pricing models (Sirmans et al., 2005). The numbers are compared with the Research specific papers. Note, these papers have a bias towards using significant variables only.

Table 10: Variables used by Sirmans et al. (2005).

Variable	Appearances	% Times significant	# Times present in Papers	% Times significant
Lot Size	52	86.53846154	13	84.61538462
Ln Lot Size	12	75		
Square Feet	69	89.85507246	10	100
ln Square Feet	12	100	1	100
Brick	13	69.23076923		
Age	78	89.74358974	10	100
# stories	13	84.61538462	3	100
#Bathrooms	40	87.5	7	100
#Rooms	14	78.57142857	1	100
Bedrooms	40	75	7	100
Full Baths	37	86.48648649		
Fireplace	57	80.70175439	4	50
Air-Conditioning	37	94.59459459		
Basement	21	76.19047619	3	66.66666667
Garage Spaces	61	78.68852459	3	66.66666667
Deck	12	83.33333333		
Pool	31	87.09677419	1	0
Distance	15	66.66666667	17	82.35294118
Time on Market	18	50		
Time trend	13	38.46153846	2	100

B Data analysis

Table 11: Information on all variables provided by the NVM.

(a)

Variable	original	value	categories	meaning	Unit	usage
Zipcode 6 digits	yes	string		Zip code of sold house		last two letters are removed to create 'zipcode 4 digits'
Zipcode 4 digits	no	Categorical		Zip code of sold house		Potential regression variable
House number	yes	string		House number		Not used in regression
House number Addition	yes	string		Addition to house number		Not used in regression
Building type	yes	Categorical	1: House 2: Apartment	Type of building		Used to create 'Apartment house dummy'
Apartment house dummy	no	binary	0: House 1: Apartment	Type of Building		Unused as 'Dwelling type' is more complete
Building Period	yes	Categorical	0: Unknown or before 1500 1: 1500-1905 2: 1906-1930 3: 1931-1944 4: 1945-1960 5: 1961-1970 6: 1971-1980 7: 1981-1990 8: 1991-2000 9: 2001 and later	Building Period		Values zero are dropped Potential regression variable
Plot size	yes	Continuous	0: Upstairs apartment/unknown	Plot size	m^2	Potential regression variable
Usable floor area	yes	Continuous	99999: Unknown	Usable floor area	m^2	Used in total floor area
Floor area	yes	Continuous	0: Unknown	Floor area *usable floor area in case floor area is unknown/unlikely	m^2	Used to create 'Ceiling Height' Unknown values are dropped Potential regression variable
Volume	yes	Continuous		Volume	m^3	Used to create 'Ceiling Height' Potential regression variable
LN Volume	no	Continuous		Natural logarithm of Volume		Potential regression variable
Ceiling height	no	Continuous	Smaller than 2 and bigger than 4: unrealistic	Ceiling height	m	Unrealistic values are dropped Potential regression variable
LN Ceiling height	no	Continuous		Natural logarithm of Ceiling height		Potential regression variable

Table 11: Information on all variables provided by the NVM, continued.

(b)

Variable	original	value	categories	meaning	Unit	usage
House class	yes	Categorical	-1: Apartment 0: House Type unknown 1: Mid-terrace house 2: Linked semi-detached house 3: Corner house 4: Semi-detached House 5: Detached House	Type of house		Not used
House class 2	yes	Categorical	-1: Apartment 0: Other sort of house 1: Simple house 2: Single family house, House boat or Recreational house 3: Mansion or Canal house 4: Living farm or Bungalow 5: Villa	House class		Unused as 'House type' is more detailed
House Type	yes	Categorical	-1: Apartment 0: Other sort of house 2: Simple house 3: Boat house 4: Recreational house 5: Single family house 6: Canal house 7: Mansion 8: Living farm 9: Bungalow 10: villa 11: Countryhouse	House Type		Unused as 'Dwelling type' is more complete
House Feature	yes	Categorical	-1: Unknown 0: No feature 2: Drive-in house 3: Dyke house 4: Semi-bungalow 5: Patio-bungalow	House Type		Unused as too many datapoints are unknown

Table 11: Information on all variables provided by the NVM, continued.

(c)

Variable	original	value	categories	meaning	Unit	usage
Apartment type	yes	Categorical	-1: House 0: Other 1: Ground floor apartment 2: Upstairs apartment 3: Maisonette 4: Porch apartment 5: Gallery flat 6: Welfare flat 7: Ground-floor apartment with an upstairs	Apartment type		Unused as 'Dwelling type' is more complete
Dwelling type	yes	Categorical	0: Other house 2: Simple house 3: Boat house 4: Recreational house 5: Single family house 6: Canal house 7: Mansion 8: Living farm 9: Bungalow 10: villa 11: Countryhouse 20: Other apartment 21: Ground floor apartment 22: Upstairs apartment 23: Maisonette 24: Porch apartment 25: Gallery flat 26: Welfare flat 27: Ground-floor apartment with an upstairs	Apartment type		Unused as 'Dwelling type' is more complete

Table 11: Information on all variables provided by the NVM, continued.

(d)

Variable	original	value	categories	meaning	Unit	usage
NVM grade	yes	Categorical	1: Unknown 2: Mid-terrace house 3: Linked semi-detached house 4: Corner house 5: Semi-detached house 6: Detached house 7: Apartment, building period unknown 8: Apartment, built before 1945 9: Apartment, built between 1945 and 1970 10: Apartment, built after 1970	Apartment type		Unused as 'Dwelling type' is more complete
Original asking price	yes	Continuous		Original asking price		Unused (contains many unrealistic values)
Last asking price	yes	Continuous		Last asking price		Unused (contains many unrealistic values)
Transaction price	yes	Continuous		Transaction price	€	Values lower than 100k and higher than 1M are dropped used to create 'LN Transaction price'
LN Transaction price	no	Continuous		Natural log of 'Transaction price'	€	Dependent variable of regression
Percentage difference	yes	Continuous		Difference between asking and transaction price	%	Unused due to simultaneity bias
Sale Condition	yes	Categorical	1: Purchasing costs payable by the purchaser 2: No additional costs payable by the purchaser 2: Auctioned or sold by public tender	Sale condition		Potential regression variable
Offering Date	yes	date		Date the house was offered		Used to create 'time for sale'
Selling Date	yes	date		Date the house was sold		Used to create 'time for sale'
Time for sale	no	Continuous		Period the house was for sale	days	Values lower than 0 and higher than 500 are dropped Potential regression variable
Selling year	no	ordinal		Year in which the house was sold		Potential regression variable

Table 11: Information on all variables provided by the NVM, continued.

(e)

Variable	original	value	categories	meaning	Unit	usage
Open porch	yes	Binary	-1: Not applicable 0: No open porch 1: Open porch	Open or closed porch		Potential regression variable
Elevator	yes	Binary	-1: Not applicable 0: No elevator 1: Elevator	Presence of an elevator		Potential regression variable
Apartment Quality	yes	Categorical	-1: No apartment 0: Simple 1: Normal/not filled in 2: Luxurious	Apartment quality		Dropped as it is unrealistic
Num floors	yes	Integer		Number of floors		Potential regression variable
Num rooms	yes	Integer		Number of rooms		Potential regression variable
Attic	yes	Binary		Presence of an attic		Potential regression variable
Attic stairs	yes	Binary		Presence of an attic stairs		Potential regression variable
Loft	yes	Binary		presence of loft		Potential regression variable
Living room shape	yes	Binary	0: Other sort of living room 1: L-room 2: T-room 3: Z-room or U-room 4: Open room 5: Room en suite	Type of living room		
Num balconies	yes	Integer		Number of balconies		Used to create balcony Unused as only 1.8% of the values has more than 1 balcony
Balcony	yes	Binary		Presence of a balcony		Potential regression variable
Num dormers	yes	Integer		Number of dormers		Used to create dormer, due to lack in number diversity
Dormer	yes	Binary		Presence of a dormer		Potential regression variable
Num roof terraces	yes	Integer		Number of roof terraces		Used to create Roof terrace, due to lack in number diversity
Roof terrace	yes	Binary		Presence of a roof terrace		Potential regression variable
Num kitchens	yes	Integer		Number of kitchens		Potential regression variable
Num sculleries	yes	Integer		Number of sculleries		Potential regression variable
Num toilets	yes	Integer		Number of toilets		Unrealistic variable
Num bathrooms	yes	Integer		Number of bathrooms		Potential regression variable

Table 11: Information on all variables provided by the NVM, continued.

(f)

Variable	original	value	categories	meaning	Unit	usage
Practice room	yes	Binary		Presence of a practice room		Potential regression variable
Practice room area	yes	Binary		Size of the practice room	m^2	Potential regression variable
Parking type	yes	Categorical	0: no parking space 1: Parking space 2: Carport and no garage 3: Garage and no carport 4: Garage and carport 5: Garage for multiple cars	Type of parking space availible		Used to create 'parking'
Parking	no	Binary		Presence of a parking space		Potential regression variable
Indoor parking	yes	Binary		Presence of an indoor parking space		Potential regression variable
Garden position	yes	Categorical	0: Unknown or no garden 1: North 2: North-East 3: East 4: South-East 5: South 6: South-West 7: West 8: North-West	Position of the garden relative to the house		Used to create 'Garden'
Garden	no	Binary		Presence of a Garden		Potential regression variable
Garden Quality	yes	Categorical	0: No garden 1: Neglected 2: Unknown or normal 3: Well-maintained 4: Very good finishing	Quality of the garden		Unused as it might be subjective
Indoor maintenance level	yes	Categorical	2: Mediocre to bad 3: Mediocre 4: Mediocre to reasonable 5: Reasonable 6: Reasonable to good or unknown 7: Good 8: Good to excellent 9: Excellent	Indoor maintenance level		Unused as it might be subjective

Table 11: Information on all variables provided by the NVM, continued.

(g)

Variable	original	value	categories	meaning	Unit	usage
Outdoor maintenance level	yes	Categorical	2: Mediocre to bad 3: Mediocre 4: Mediocre to reasonable 5: Reasonable 6: Reasonable to good or unknown 7: Good 8: Good to excellent 9: Excellent	Outdoor maintenance level		Unused as it might be subjective
Num insulation types	yes	Ordinal	5: Five or more	Number of kinds of insulation		Potential regression variable
Heating type	yes	Categorical	0: No heating 1: Gas or coal stove 2: Central heating, hot air heating or city heating 3: Air conditioning or sun collectors	Type of heating		Potential regression variable
Central location	yes	Categorical	0: Outside built area 1: Unknown 2: In residential area 3: In centre	Location compared to the centre		Dropped as it is covered in the Geographical data
Location beautiful environment	yes	Categorical	0: Unknown/no 1: Near forest 2: Near water 3: Near park 4: Clear view	Location near some sort of beautiful environment		Potential regression variable
Busy road	yes	Categorical	0: On a quiet road 1: On a medium-busy road 2: On a busy road	Quietness of adjacent road		Potential regression variable
Ground lease construction	yes	binary	-1: Unknown 0: No ground lease 1: Ground lease	presence of ground lease construction		Potential regression variable
Permanently inhabited	yes	binary		Permanently inhabited		Potential regression variable
Partially rented out	yes	binary		Partially rented out		Potential regression variable

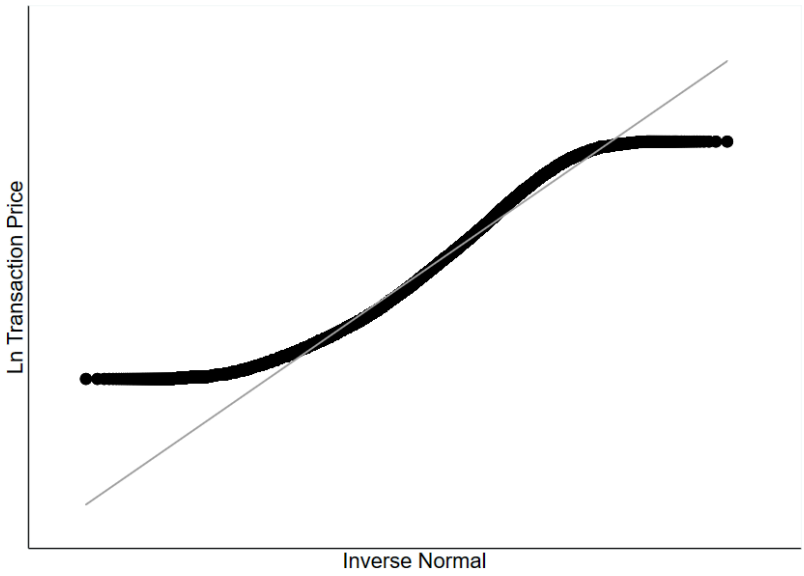


Figure 14: QQ-plot of the log of the transaction price and a normal distribution.

Table 12: overview of correlation between all independent (continuous) variables and the log of the transaction price.

Variable	correlation	Variable	correlation
LN Transaction price	1		
Attic	0.0047	Attic stairs	-0.0056
Balcony	-0.0723	Dormer	0.0717
Elevator	-0.0284	Garden	0.1558
Ground lease construction	-0.0542	Indoor parking space	0.1259
House value index	0.2990	Loft	0.0189
Num bathrooms	0.1396	Num floors	0.3468
Num insulation types	0.1222	Num kitchens	0.0754
Num rooms	0.4639	Num sculleries	0.0901
Open porch	-0.1244	parking	0.1591
Partially rented out	0.0083	Permanently inhabited	0.0253
Practicing room area	0.0231	roof terrace	0.2903
Time for sale	-0.0673	Year	0.3655
Usable floor area	0.5786	LN Usable floor area	0.6028
Volume	0.5979	LN Volume	0.6326
Ceiling height	0.2984	LN Ceiling height	0.2953

Table 13: Overview of correlations between the independent variables.

	Ceiling height	Elevator	Garden	House apartment dummy	House Value Index	Indoor parking space	LN Usable floor area	LN Volume
Elevator & apartment	-0.0824							
Garden	0.0828	-0.1772						
House apartment dummy	-0.1203	0.1710	**-0.4594**					
House value index	0.0046	-0.0248	-0.0495	-0.0020				
Indoor parking space	0.0321	0.1154	0.0560	-0.1190	-0.0239			
LN Usable floor area	0.1445	0.1084	0.2521	-0.3603	0.1196	0.1721		
LN Volume	0.3478	0.0855	0.2557	-0.3647	0.1134	0.1691	**0.9779**	
Num bathrooms	0.0212	0.0049	0.0629	-0.0820	-0.0138	0.0333	0.1759	0.1708
Num floors	0.1558	-0.1814	0.3322	**-0.6111**	0.0755	0.1032	**0.5425**	**0.5456**
Num insulation types	-0.0353	0.2182	0.0617	-0.1188	0.1174	0.1354	0.1118	0.0979
Num kitchens	0.1079	-0.0066	0.0458	-0.0495	-0.1999	0.0205	0.0898	0.1076
Num rooms	0.1570	-0.0531	0.2315	**-0.4107**	-0.0498	0.0878	**0.7337**	**0.7276**
Open porch	-0.1000	0.1336	**-0.4100**	**0.9114**	0.0038	-0.1149	-0.3237	-0.3259
Parking	0.0192	0.3313	0.0530	-0.1668	0.0064	**0.4536**	0.2635	0.2530
Roof terrace	0.0708	-0.0616	-0.0415	-0.0728	0.0442	0.0681	0.215	0.2151
Year	0.1822	-0.0336	-0.0640	0.0358	**0.6058**	-0.0165	-0.1371	-0.0928

Table 14: Overview of correlations between the independent variables, continued.

	Num bathrooms	Num floors	Num insulation types	Num kitchens	Num rooms	Open porch	Parking	Roof terrace
Ceiling height								
Elevator & apartment								
Garden								
House apartment dummy								
House value index								
Indoor parking space								
LN Usable floor area								
LN Volume								
Num bathrooms								
Num floors	0.2022							
Num insulation types	0.0453	0.0566						
Num kitchens	0.3393	0.1343	-0.0112					
Num rooms	0.2031	**0.6188**	0.0245	0.1246				
Open porch	-0.0691	**-0.5535**	-0.1206	-0.0405	-0.3615			
Parking	0.0346	0.0818	0.2915	0.0131	0.1194	-0.1666		
Roof terrace	0.1091	0.2709	0.1005	0.0533	0.1735	-0.0648	0.0418	
Year	-0.0613	-0.0866	0.1065	0.0219	-0.0384	0.0371	0.0147	0.0400

Table 15: Overview of correlations between independent variables and the variables of interest.

	Positive articles	**Total Articles**	**'Noord-Zuidlijn' trends**	**'Metrolijn 52' trends**	**'Noord-Zuidlijn Amsterdam' trends**
Apartment Elevator	-0.0534	-0.0476	-0.0460	-0.0186	-0.0396
Ceiling height	0.2682	0.2964	0.2863	0.2947	0.2917
House apartment Dummy	0.1298	0.1272	0.1203	0.0858	0.1130
House value index	-0.2551	-0.2198	-0.1908	-0.1386	-0.2479
Indoor parking	-0.0175	-0.0158	-0.0148	-0.0070	-0.0150
LN usable floor area	-0.0421	-0.0356	-0.0378	-0.0001	-0.0317
Num bathrooms	-0.0126	-0.0215	-0.0196	-0.0256	-0.0236
Num floors	-0.0630	-0.0590	-0.0590	-0.0226	-0.0532
Num insulation types	-0.0715	-0.0774	-0.0631	-0.0927	-0.0732
Num rooms	-0.0050	-0.0041	-0.0030	0.0097	-0.0008
Parking	-0.0556	-0.0520	-0.0463	-0.0317	-0.0440
Roof terrace	0.0557	0.0573	0.0536	0.0542	0.0520
Year	**0.7935**	**0.8216**	**0.8318**	**0.6209**	**0.8018**

Table 16: List of variables used for Base model construction.

Apartment elevator	Building period
Ceiling height	Dwelling type
Garden	Heating type
House value index	Indoor parking
Living room shape	LN usable floor area
Location beautiful environment	Num bathrooms
Num floors	Num insulation types
Num rooms	Parking
Roof terrace	Sale condition
Zipcode	

C Regression model

Table 17: Results OLS regression of Base model.

Variable	Coeficient	std. Error	t-value	p-value
Constant	7.474225	.0406881	183.70	0.000
Apartment elevator	.0218794	.0046197	4.74	0.000
Building period				
- 1500-1905	basis			
- 1906-1930	-.0719906	.0045311	-15.89	0.000
- 1931-1944	-.0884885	.0063838	-13.86	0.000
- 1945-1960	-.1368561	.0079397	-17.24	0.000
- 1961-1970	-.1675131	.0080069	-20.92	0.000
- 1971-1980	-.1366072	.0128227	-10.65	0.000
- 1981-1990	-.1090788	.0066836	-16.32	0.000
- 1991-2000	-.0620209	.0067273	-9.22	0.000
- 2001 and later	-.0171263	.0078036	-2.19	0.028
Ceiling height	.294217	.0049679	59.22	0.000
Dwelling type				
- Simple house	basis			
- Boat house	.067371	.0329088	2.05	0.041
- Single family house	.0829669	.0179253	4.63	0.000
- Canal house	.0181348	.0329313	0.55	0.582
- Mansion	.0584197	.019502	3.00	0.003
- Living farm	.1592044	.1035399	1.54	0.124
- Bungalow	.3480871	.0365697	9.41	0.000
- Villa	.3453397	.0313252	11.02	0.000
- Countryhouse	.1706164	.206821	0.82	0.409
- Ground floor apartment	.0566635	.017964	3.15	0.002
- Upstairs apartment	.0236756	.017956	1.32	0.187
- Maisonette	.0039482	.0187861	0.21	0.834
- Porch apartment	-.0032623	.0184129	-0.18	0.859
- Gallery flat	-.0261	.0185533	-1.41	0.160
- Welfare flat	-.2752741	.0451758	-6.09	0.000
- Ground floor apartment withan upstairs	.1687064	.0211448	7.98	0.000
Garden	.0343733	.0043507	7.90	0.000

Table 17: Results OLS regression of Base model,continued.

Variable	Coeficient	std. Error	t-value	p-value
Heating type				
- No heating	basis			
- Gas or coal stove	-.0976801	.0075269	-12.98	0.000
- Central heating, city heating or hot air heating	.074962	.0052014	14.41	0.000
- Air conditioning or sun collectors	.1862309	.084085	2.21	0.027
Living room shape				
- Another sort of living room	basis			
- L-room	-.033215	.005494	-6.05	0.000
- T-room	-.0022663	.0396799	0.06	0.961
- Z-room or U-room	-.0002774	.0221885	0.01	0.936
- Open room	-.0075852	.0050844	-1.49	0.149
- Room en suite	-.0530633	.0056122	-9.45	0.000
Parking	.0834482	.0056045	14.89	0.000
LN Usable floor area	.6627857	.0054017	122.70	0.000
Location beautiful environment				
- None	basis			
- Near a forest	.0860429	.0139382	6.17	0.000
- Near water	.070741	.0039147	18.07	0.000
- Near a park	.0210119	.0063787	3.29	0.001
- Clear view	.0107707	.0036719	2.93	0.022
Num bathrooms	-.0154395	.0030161	-5.12	0.000
Num floors	-.0094293	.0027688	-3.41	0.001
Num insulation types	.0091594	.0008567	10.69	0.000
Num rooms	.0252268	.0016745	15.07	0.000
Roof terrace	.0992101	.0041222	24.07	0.000
Sale condition				
- Purchasing costs, payable by purchaser	basis			
- No additional costs, payable by purchaser	-.0464857	.0062815	-7.40	0.000
- Auctioned or sold by public tender	.0928938	.0311794	2.98	0.003
House value index	.0126442	.0000882	143.34	0.000

Table 17: Results OLS regression of Base model, continued.

Variable	Coeficient	std. Error	t-value	p-value	Variable	Coeficient	std. Error	t-value	p-value
Zipcode4					Zipcode4				
1011	basis				1011	basis			
1012	-.0360973	.032744	-1.10	0.270	1013	.0238212	.0252689	0.94	0.346
1015	.0892178	.0254347	3.51	0.000	1016	.0763242	.0265364	2.88	0.004
1017	.0932748	.0266021	3.51	0.000	1018	-.0808661	.0253326	-3.19	0.001
1019	-.2239132	.0255061	-8.78	0.000	1021	-.3662403	.0276201	-13.26	0.000
1022	-.5001917	.0421547	-11.87	0.000	1023	-.3313562	.0449542	-7.37	0.000
1024	-.5155787	.0272543	-18.92	0.000	1025	-.4701811	.0266909	-17.62	0.000
1027	-.1310444	.060214	-2.18	0.030	1028	-.2209368	.1111221	-1.99	0.047
1031	-.3402248	.0368595	-9.23	0.000	1032	-.470286	.0299652	-15.69	0.000
1033	-.5043787	.0286563	-17.60	0.000	1034	-.4917668	.0266086	-18.48	0.000
1035	-.5999248	.02759	-21.74	0.000	1036	-.4944849	.0769792	-6.42	0.000
1051	-.1483905	.0258454	-5.74	0.000	1052	-.1034584	.025384	-4.08	0.000
1053	-.0939547	.024525	-3.83	0.000	1054	-.000521	.0251351	-0.02	0.983
1055	-.272653	.0247684	-11.01	0.000	1056	-.1710798	.0250597	-6.83	0.000
1057	-.1784902	.0252426	-7.07	0.000	1058	-.1214087	.0249538	-4.87	0.000
1059	-.1416018	.0256122	-5.53	0.000	1060	-.5560082	.0279834	-19.87	0.000
1061	-.3042958	.0287588	-10.58	0.000	1062	-.3778493	.02982	-12.67	0.000
1063	-.5463161	.0258091	-21.17	0.000	1064	-.500582	.0254054	-19.70	0.000
1065	-.4153209	.0278742	-14.90	0.000	1066	-.4186653	.0257934	-16.23	0.000
1067	-.5005371	.0290451	-17.23	0.000	1068	-.5306333	.0338413	-15.68	0.000
1069	-.5673576	.0266617	-21.28	0.000	1071	.0913748	.0255391	3.58	0.000
1072	-.0544319	.0252217	-2.16	0.031	1073	-.059616	.0262235	-2.27	0.023
1074	-.049467	.02653	-1.86	0.062	1075	.0187672	.0256079	0.73	0.464
1076	.0313245	.0259763	1.21	0.228	1077	.1369333	.0254443	5.38	0.000
1078	.0220966	.0249121	0.89	0.375	1079	-.0270344	.0273612	-0.99	0.323
1081	-.1510142	.0263601	-5.73	0.000	1082	-.2765779	.0273813	-10.10	0.000
1083	-.0608538	.0276477	-2.20	0.028	1086	-.4374404	.03674	-11.91	0.000
1087	-.3621538	.0293215	-12.35	0.000	1091	-.1323662	.0249075	-5.31	0.000
1092	-.107279	.0288142	-3.72	0.000	1093	-.1626977	.0274096	-5.94	0.000
1094	-.2651641	.0249921	-10.61	0.000	1095	-.2884292	.0272958	-10.57	0.000
1097	-.1558579	.0277717	-5.61	0.000	1098	-.1501831	.0257067	-5.84	0.000
1102	-.6791483	.0253992	-26.74	0.000	1103	-.776235	.028826	-26.93	0.000
1104	-.840655	.04693	-17.91	0.000	1106	-.6056338	.0335265	-18.06	0.000
1109	-.5462624	.0420226	-13.00	0.000					