Erasmus University Rotterdam
Erasmus School of Economics

Master Thesis Urban, Port and Transport Economics

Hedonic price indexes: A case study of the Dutch existing residential property market

Name student: Bart Groen

Student ID number: 369165

Supervisor: Jeroen van Haaren

Second assessor: Giuliano Mingardo

Date final version: 16-04-2018

Abstract

This study aims to investigate the possibility of a price index for the Dutch existing residential property market based on the hedonic method. Currently the official price index for the Dutch existing residential property is estimated using the "sales price appraisal ratio" method by Bourassa, Hoesli & Sun (2006). Partly due to a lack of reliable data on dwelling characteristics at the time (de Vries, de Haan, van der Wal, & Marién, 2009), a problem often encountered when making hedonic price indexes for residential property (Hill R. J., 2013). However, newly available data on dwelling and locational characteristics in the form of the "Basisregistratie adressen en gebouwen (BAG)" could make a hedonic price index feasible. To test this feasibility two regression models have been made. The first model only includes dwelling characteristics, whereas the second also includes locational characteristics in the form of factors estimated with a factor analysis. This factor analysis yields some interesting results resembling the "multi nuclei" theory of Harris & Ullman (1945). Ultimately, this study concludes that both models can be used to construct a decently reliable hedonic price index for the Dutch existing residential property market. However, the price index with the locational characteristics will most likely encounter problems with data availability if used to estimate up-to-date price changes.

Summary of contents

Abbreviations

CBS	Centraal Bureau voor de Statistiek (Central Bureau of Statistics)
SPAR	Sales Price Appraisal Ratio
PBK	Prijsindex Bestaande Koopwoningen (Price index existing residential property)
RPPI	Residential Property Price Index
OLS	Ordinary Least Squares
WLS	Weighted Least Squares
RESET	Ramsey Regression Equation Specification Error Test
KMO	Kaiser-Meyer-Olkin (-test)
MSA	Measured Sampling Adequacy
EFA	Explanatory Factor Analysis
VIF	Variance Inflation Factor
HTD	Hedonic Time Dummy (-method)
HDIL	Hedonic Double Imputation Laspeyres (-method)
HDIP	Hedonic Double Imputation Paasche (-method)
HDIF	Hedonic Double Imputation Fisher (-method)

1. Introduction

During the last decade there has been a strong push within the European Union to provide reliable price indexes for the residential property market, which can be used by European governments and the European Central Bank to monitor the owner-occupied property sector. The collapse of the residential property market in 2008 emphasized the importance of the reliability of these price indexes for residential property. Since 2008, the Central Bureau of Statistics (CBS) in the Netherlands publishes a monthly residential property price index based on the Sale Price Appraisal Ratio (SPAR) method. Before that, an index was determined by using a weighted version of the repeat sales approach (de Vries, de Haan, van der Wal, & Marién, 2009). This price index is calculated for the existing residential property market and is called the "Prijsindex Bestaande Koopwoningen" (PBK).

The SPAR method uses ratios of transaction prices and previous appraised values and, in contrast to repeat sales methods such as the Case-Shiller method that is widely used in the United States (Case & Shiller, 1987), utilizes almost all available data for the period under observation (Bourassa, Hoesli, & Sun, 2006). In general, there is a shortage of transaction prices for the base period, also known as the index reference period, because often the properties sold during the observation period were not sold during the base period. Therefore, these base period prices are estimated by using the appraisals of the properties. In the Netherlands these appraisals are gathered by the national government under the Valuation of Real Estate Law (Wet Waardering Onroerende Zaken) and can thus be used to estimate a SPAR index (de Vries et al., 2009). The study by de Vries et al. (2009) finds that the SPAR method performs well compared to repeat sales method for the owner-occupied residential property sector.

Eurostat states that an ideal Residential Property Price Index (RPPI) should represent changes in the prices of properties that are comparable in quality over time (Eurostat, 2013). However, in the past the development of reliable house price indexes was hampered by lack of suitable data sets and the extreme heterogeneity within the residential property market, meaning that every house is different both in terms of its physical characteristics and its location (Hill R. J., 2013). This study implies that hedonic methods, which express house price as a function of a vector of characteristics, might prove useful in solving the latter of these two issues. That leaves the issue concerning the unavailability of suitable data sets, which is one of the reasons the CBS chose to use SPAR in 2008 (De Vries et al., 2009). However recently additional data[1] on residential property, especially on the characteristics of individual dwellings, have become available that can be merged with existing data concerning residential property transactions.

This study aims to investigate whether it is possible to construct a price index for the existing residential property market in the Netherlands based on a hedonic price index model, with the use of the additional information from the BAG.

[1] ''Basisregistratie adressen en gebouwen'' (BAG)
https://www.kadaster.nl/documents/32706/37743/bag+grondslagen+catalogus/d6bb83b9-33e5-47bb-939c-fa7fde1f2b16

To this end the main research question is:

"To what extent is it possible to make a reliable hedonic price index for the Dutch existing residential property market by adding additional variables to the current CBS database 'Bestaande koopwoningen transacties'?"

To answer this research question comprehensively and as to discuss all aspects of it, the following sub-questions will be discussed:

Sub-question 1: *"Which variables from either the BAG or PBK hold explanatory value in respect to the transaction prices and what are the effects of these variables?"*

Sub-question 2: *"Which spatial variables could be added to enhance the model?"*

Sub-question 3: *"To what extent could a reliable hedonic price index be made with the addition of the BAG-variables?"*

Sub-question 4: *"To what extent could the hedonic price index be made more reliable using a factor analysis of the locational variables?"*

Sub-question 5: *"Which hedonic price index method has the highest reliability and/or is most suited for constructing a hedonic price index for the Dutch residential property market?"*

This study's usefulness firstly resides within the strive to create a house price index model that is as reliable and comprehensive as possible. In this context it is important to explore new opportunities when they arise. This study will shed some light on the possibility of a hedonic price index for the Dutch existing residential property market, either by showing its feasibility or by indicating what further developments will be needed in the future to realize a reliable hedonic price index.

Secondly, the results of this study will, most likely, give an insight into the reliability of the variables currently in the BAG and "Bestaande koopwoningen transacties" datasets regarding the transaction prices. Furthermore, the combining of these datasets, the addition of new spatial variables and cleansing of this data will hopefully create a more comprehensive dataset. This data could prove very useful for further research and for current analysis, e.g. news articles, of the existing residential property market.

The academic relevance of this study is twofold. On the one hand an extensive analysis of the advantages and drawbacks of specific hedonic price index models for the Dutch residential property market has, so far, not been made with the use of the official statistics in the Netherlands. On the other hand, the newly available data, hopefully, allows for better spatial variables to be constructed. This has already been done to capture the value of natural space for surrounding house prices (Daams, Sijtsma, & van der Vlist, 2016), but not with a large number of amenities which can be transformed into several overarching locational variables using a factor analysis. Furthermore, this type of locational variables is yet to be used in the Netherlands for creating a hedonic price index that also captures the value of a residential property's location.

Furthermore, a hedonic price index model would provide more detail on the variations in house prices and would make it possible to determine the impact of certain characteristics, including the aforementioned spatial characteristics, on house prices and on shifts in house prices over time.

Which is something most non-hedonic price indexes are unable to do (Hill R. J., 2013). Providing insight in the effects of these spatial characteristics appears especially relevant as there are sizable differences in residential property price index growth rates between different areas in the Netherlands. For instance, house prices in Amsterdam grew 13,5% in 2016 compared to the previous year, whereas house prices in Drenthe 'only' grew 2,3% (CBS, 2017). The proposed hedonic price index could thus be a first step in providing more insight in the characteristics, albeit locational or not, behind these differences in the Netherlands. Ultimately, providing more detail on the driving forces behind changes in house prices could help policymakers better assess the sustainability of these changes and might prove useful in preventing property market collapses such as in the one in 2008. Which is one of the main applications for national residential property market price indexes according to Eurostat (Eurostat, 2013).

The data used in this study mainly comprises of two datasets. The BAG "Basisregistratie adressen en gebouwen" from CBS and *Kadaster,* the Dutch land registry office, and the "Bestaande koopwoningen transacties" dataset from CBS and *Kadaster*. These two datasets will be combined using the "PHT-code", which combines the postal code, house number and house number addition of a specific residential property into a unique code.

The BAG database contains the details of addresses and buildings in the Netherlands. Details on existing residential property are therefore included in this database. Since 2012 *Kadaster* publishes several variables in the BAG that are relevant for the existing residential property including, but not limited to, size of the living area in square meters, age, address and geographic coordinates. The "Bestaande koopwoningen transacties" database contains details on individual transactions concerning existing residential property. The relevant information from this database for this study includes transaction prices, size of the lot in m^2, dwelling type, period in which the transaction took place and address. Furthermore, there will be an attempt to use the locational information on residential property from the BAG to construct one or more spatial variables that could be used in the hedonic price model.

The handbook from Eurostat on residential property price indexes (RPPI) will be the guideline for this study concerning the price index methods used (Eurostat, 2013). Since there are numerous hedonic price index methods. The three main hedonic approaches described in the RPPI handbook, the time dummy approach, the price characteristics approach and the imputations approach, will form the starting point in this study's attempt to estimate which hedonic price index model fits the data best.

To find the hedonic price index model that fits the data and its available variables best, the variables which should be considered in the hedonic price model will firstly be assessed and, secondly, an outlay will be given of the most common hedonic price index models. This outlay will describe the general theory behind these price index models and discuss the assumptions and limitations underlying these models.

Following this theoretic outlay, the databases will be combined as to create a dataset that contains transaction prices and several variables concerning the ''quality'' of the residential property. These variables will then be tested to assess their reliability and viability for the construction of a hedonic price index model. In addition to these existing variables in the two datasets described above, this study will try to add and test spatial variables by combining the BAG data on residential property location with spatial data. During this process there will also be an attempt to clean the data by

removing improbable, inaccurate or unreliable transaction records from the dataset. Furthermore, it will be assessed if it is preferable, or even necessary, to make a subsample within the dataset to perform the analyses described below. Preliminary research indicates that there are many locational variables available which can be added to the dataset described above. Therefore, it seems very likely that a factor analysis will be necessary to reduce the number of locational variables.

Finally, a choice must be made which regression model should be used to estimate the effects of the variables on the transaction prices. After a certain regression model has been chosen, it can be investigated what type of hedonic price index ought to be used taking into account: the chosen regression model, the available variables and its usefulness to the CBS. While doing this the various assumptions underlying the different models will be analysed to recognize whether assumptions have been violated.

2. Theoretical framework

In this chapter the theory behind hedonic price models and hedonic price indexes will be elaborated upon. Furthermore, it will address which characteristics might influence the price of a dwelling and how these characteristics should be included in the model. Finally, the theory behind the different hedonic price index methods and their strengths and weaknesses will also be discussed.

At the basis of a hedonic regression method lies the assumption that heterogeneous goods can be described by their attributes or characteristics (Lancaster, 1966; Rosen, 1974; Eurostat, 2013). In terms of price models this entails that the price of a good is determined by the values of its underlying characteristics, which are not independently observed however and are often called shadow prices (Hill R. J., 2013). In this context hedonic price functions are estimated for two primary reasons; firstly for use in construction of overall price indexes that account for changes in the quality of goods over time; and secondly as an input in the analysis of consumer demand for characteristics of heterogeneous goods, which is difficult or even impossible to observe separated from said heterogeneous good (Sheppard, 1999). An example of the second reason is the study into the effect of nature on housing by Daams et al. (2016). This study however is focussed on the first of these two reasons, which allows the hedonic regression method to be used for constructing quality-adjusted price indexes. This is especially useful for highly differentiated and heterogeneous products such as residential property. Because residential property might differ in quality from dwelling to dwelling, both in terms of physical characteristics and locational characteristics (Hill R. J., 2013). Eurostat (2013) notes that the most important characteristics of residential property are:

- the *area size of the structure*;
- the *area size of the land* that the structure sits on;
- the *age* of the structure;
- the *type of structure*;
- the *location* of the property;
- the *materials used*, and
- *other price determining characteristics*, e.g. number of rooms and the presence of, for instance, a garage and/or swimmingpool.

The size of a dwelling, or *living area size of the structure*, seems the most intuitive predictor of house prices. It is thus included in most hedonic house price models (Sheppard, 1999), whether it is included as a logarithmic transformation or a square root transformation or no transformation depends mostly on the model choice (Diewert W. E., 2003). Literature also shows that the lot size of a dwelling, or *area of the land* that the structure sits on, which is often closely tied to the size of the dwelling, is also a relevant predictor of house prices (Sheppard, 1999).

Literature seems to indicate that the age of the structure has an influence on the value of owner-occupied housing, this relationship is slightly ambigious however as it appears to be non-linear. This gives rise to the notion that including age as a continuous variable will not yield reliable results (Goodman & Thibodeau, 1995). More recent studies regarding the Dutch housing market underlined the conclusions of this older paper. It was found that the effect of age can only be partially explained by the deterioration of the physical condition of a structure over time. This causes depreciation of the value of a structure. However, for dwellings built in certain time periods this linear relationship between age and price is off-set by so-called “vintage-effects”, at least regarding the Dutch housing market (De Haan & Syed, 2016). Thus assuming that age has a strictly linear relationship to price might cause the age of a structure to have a positive effect on the price, ignoring the priviously noted depreciatory effect of age (Francke & van de Minne, 2016). It is therefore important, especially for the Dutch housing market, to weigh in both these effects when constructing a hedonic house price model.

Some hedonic models do not include type of dwelling as a predictor of housing prices, but it seems logical that, for instance, detached dwellings are more expensive than non-detached or even semi-detached dwellings. A study by Brounen & Kok (2011) appears to support the idea that dwelling type has an influence on housing price. The same study also indicates that the energy label attributed to a dwelling has a relation with its price. Either as an indicator of the quality of the dwelling or simply as an indicator of the energy efficiency of the dwelling (Brounen & Kok, 2011).

The location of a dwelling is important, since housing markets are intrinsically spatial. Not only do dwellings involve varying quantities of land, but also because every dwelling possesses a particular location (Sheppard, 1999). This particular location determines the locational characteristics of a dwelling, which in turn will have an effect on the price of a dwelling. This is indicated by (Koster & Rouwendal, 2012), but also by prices per squared meter that vary with location as shown by the study of (Cheshire & Sheppard, 1995), which rejects the hypothesis that land value is constant over locations. This might be caused by people willing to pay a higher price for a dwelling which provides more access to amenities (Glaeser, Kolko, & Saiz, 2001). However, it is important to note that people prefer not to have the negative externalities associated with living too close to certain amenities (Li & Brown, 1980).

The literature indicates that there are several ways to incorporate the location and geospatial data on dwellings in hedonic regression models. The simplest way of doing this is by including a neigborhood dummy for the neigborhood a property is situated in. However, when the adresses or coordinates of individual properties are available, more sophisticated geospatial techniques can be used. A relatively easy method of including geospatial data in a hedonic regression model is by measuring the distance of individual properties to amenities (Hill R. J., 2013). These estimated

distances can then be added to a hedonic regression model as additional variables. This can be done using distance nonparametrically (Martins-Filho & Bin, 2005) or parametrically (Hill & Melser, 2008).

Since it is likely that a factor analysis will be conducted to reduce the number of locational variables, it appears useful to analyse the consequences of such a factor analysis in the context of residential property. The factor analysis will likely group certain locational variables, or amenities, together (Krumm, 1980). These groups are locational in nature since the variables in our analysis are locational and therefore might reflect certain centres of amenities which are present in most cities, such as city centres. The concept that the structure of a city can be described as having multiple centres of amenities and businesses around which residential areas are situated best fits the multi neclei theory (Harris & Ullman, 1945). In "*The Nature of Cities"* they describe cities as having multiple discrete nuclei ranging from small to large and all fullfilling different purposes within the city, where major nuclei can be the central business district or the main retail area of a city. Multiple minor nuclei can also exist within the same city. These might be more culturally orientated such as parks and museums, but could also be lakes or smaller retail centres. Educational facilities can also be nuclei in this theoretical framework. They can be a major necleus in case of a large university, but also a minor necleus such as several elementary schools grouped together. According to this theorem all cities possess these internal nuclei, however these nuclei might be different for every city and most likely will be (Harris & Ullman, 1945).

2.1 Hedonic regression models

Converting these assumptions to a hedonic regression model leads to the starting point that the price P_i^t of residential property *i* in period *t* is a function of a fixed number, suppose K, characteristics measured by "quantities" X_{ik}^t and a random error term ε_i^t . With *T+1* time periods, going from base period 0 to period *T* this leads to the equation:

$$P_i^t = f(x_{i1}^t, \dots, x_{ik}^t, \varepsilon_i^t)$$

This equation can be specified as a parametric model. Two more well-known hedonic models are the fully linear model and the logarithmic model. These specifications both contain an intercept term β_0^t and term β_k^t , which shows the estimated characteristics parameters. Futhermore, both specifications allow characteristics to be transformations of continuous variables, e.g. logarithms. However, in hedonic models, many variables might be categorical rather than continuous (Eurostat, 2013). For instance a variable for dwelling type is likely to be categorical because it can only assume a limited amount of possible values, since dwellings are divided up in a fixed amount of categories. This type of variable will often be presented by a set of dummy variables, assuming value 1 if a observation belongs to the category and value 0 if it does not.

Fully linear model: $P_i^t = \beta_0^t + \sum_{k=1}^{K} \beta_k^t x_{ik}^t + \varepsilon_i^t$

Log-linear model: $\ln(P_i^t) = \beta_0^t + \sum_{k=1}^{K} \beta_k^t x_{ik}^t + \varepsilon_i^t$

The log-linear model is often used for products such as high-tech goods, because it reduces the problem of heteroskedasticity or non-constant variance of the errors. This because prices are usually log-nomally distributed (Diewert W. E., 2003). Another advantage of the log-linear model is that β_k^t reflects an estimation of the percentage change in price for one unit change in x_k (Griliches, 1971).

2.2 Hedonic price indexes

There are several ways in which hedonic regression models can be used to make quality-adjusted price indexes (Hill R. J., 2013). Given this information, it seems useful to give an overview of these indexes, concerning their underlying structure and how they relate to each other in the context of this study. As discussed earlier Eurostat (2013) considers two approaches in the creating of quality-adjusted price indexes.

The time dummy method is the original hedonic method of creating a quality-adjusted price index, typically using a semi-log functional form (Hill R. J., 2013). Its main advantage being its simplicity, as the price index follows directly from the estimated pooled time dummy regression equation (Eurostat, 2013). This method runs a single overall regression on the pooled data with the addition of a time dummy as an explanatory variable relating to periods $t = 0,\ldots,T$, where period 0 is the base period. The time dummy parameter can shift upwards or downwards and thus measures the effect of "time" on the logarithm of price, whilst the other explanatory variables in the regression model will control for changes in the quantities of the characteristics in the sample. In this sense the time dummy variable yields the quality-adjusted price change or index between the base period and each comparison period t (Eurostat, 2013; de Haan, 2004; Diewert, Heravi, & Silver, 2009).

The time dummy method has, of course, its own advantages and disadvantages. Its main advantage being its simplicity mentioned in the previous paragraph. However, its main disadvantage is the possible revision of previously calculated figures when addional periods are added, which is the case if one wants to extend the time series. If a new period, t + 1, is added to the model the characteristics coefficients of the periods will change and thus, subsequently, the price index numbers will also be different from the ones estimated without the new period, t + 1 (Eurostat, 2013; de Haan, 2004).

The second approach discussed in Eurostat (2013) concerns imputation methods. This approach to constructing hedonic price indexes runs a separate regression for each time period, after which an index can be composed using the predicted prices based on the regression coefficients (Eurostat, 2013).

Imputation methods utilize standard price index formulas. In this context the Laspeyres and Paasche formulas are the two best known, whereas also the Fisher method deserves a mention. The Laspeyres and Paasche formulas measure the movement of the price of a certain bundle of goods over time, generally estimating a hedonic model separately for each time period, whereas Fisher takes the geometric mean of Laspeyres and Paasche. However, these methods require the price of each of the goods in the bundle to be available in every relevant time period. This is problematic for house price indexes, as dwellings mostly sell at infrequent and irregular intervals. It is therefore difficult, if not impossible, to estimate a housing price index with the Laspeyres and Paasche formulas based on actual transaction prices (Hill R. J., 2013). This is where the imputation methods come in, because it is possible to construct a housing price index using the Laspeyres and Paasche formulas by substituting the actual transaction prices with imputed prices. The imputation method thus employs the estimated hedonic model to impute the prices of dwellings, whose actual prices are unknown in a certain period (Silver & Heravi, 2004). This ensures that the prices of all dwellings, included in the price index formula, are available in both the base period and the relevant measurement period (Hill R. J., 2013). Imputation methods can be further divided into single imputation and double imputation

methods. Single imputation means that the index formula only imputes the price of a property for one of the periods, either the period *t* or reference period *0*, and uses the actual price of the property for the other. Whereas the double imputation method imputes the prices of a property in period *t* as well as in the reference period (Hill R. J., 2013). There is some debate in the literature as to whether single or double imputation is better. However most studies agree that double imputation is preferred as it may reduce omitted variable bias, whilst single imputation does not (Hill & Melser, 2008).

Imputation methods have several strong points. Firstly, they permit shadow prices to evolve over time, meaning that the value of certain dwelling characteristics is allowed to change between time periods, which makes these methods very flexible. Secondly, the double imputation method can help reduce omitted variables bias, a problem that hedonic models usually struggle with. Finally, the imputation method is able to deal with missing characteristics for some properties in the dataset. A drawback of the imputation method is that the neccesity to estimate a new hedonic model for each period makes it difficult to exploit the interactions between functions. Furthermore, the imputation method does not automatically generate standard errors on the price indexes and therefore requires standard errors to be estimated indirectly (Hill R. J., 2013; Diewert et al., 2009).

Similar to the imputations method, the characteristics approach usually estimates a hedonic model for each period, after which a standard price index formula is used to calculate a price index. The critical difference between the characteristics method and imputation method is that the Characteristics price index is defined in characteristics space (Hill R. J., 2013; Diewert E. W., 2004). This means that the Characteristics method usually assembles a standardized property for each period, which can be viewed as the average dwelling in that specific time period. The standardized property of a period is strictly hypothetical and does not need to have realistic values (it could for instance have half a balcony). After this standardized property is assembled, its price is imputed. A price index can then be acquired by taking the ratio of the imputed price of the same standardized property in the base period and the measurement period (Eurostat, 2013). It is important to note that the standardized property may be an arithmetic mean as well as a median (Hill R. J., 2013).

Similar to the time dummy, the main advantage of the characteristics method is its relatively easy interpretation as it portrays the change in price of the standardized, or average, property over time. Furthermore, the characteristics method can also be used with a double imputation, which reduces the vulnerability of the model to omitted variables bias. The method also has its weaknesses, which are very similar to those of the imputation method, because the characteristics method also estimates a new model for every period, preventing it from exploiting the interactions between functions. The method also does not directly estimate standard errors, which is also similar to the imputation method. Finally, unlike the imputation method, the characteristics method can not handle missing characteristics for some properties in the dataset (Hill R. J., 2013; Diewert, 2003).

2.3 Price index assumptions

As discussed previously there is a large number of different formulas and approaches to estimating price indexes. These different indexes all have there strengths and weaknessess, which can be evaluated by looking at which assumptions they either do or do not violate. This axiomatic approach to judging the quality of price indexes can be found in the international CPI-manual (Diewert E. W.,

2004), Van der Grient & De Haan (2008) and more recently in Balk (2012). These studies indentify the following assumptions or axiomas underlying price index formulas:

- *Proportionality test*

  This axioma requires that a proportional change of all prices should cause an identical proportional change in the index.

- *Identity test*

  This axioma states that if the prices of all objects in period *t* are equal to the prices in reference period *0*, the index in period *t* should be equal to 1.

- *Commensurability test*

  This axioma says that the price index must not change if the unit of a good, of which the price is observed, changes. The homogeneity of the object is crucial for this.

- *Circularity test*

  This axioma requires that a direct comparison between periods *0* and *t* should result in the same price index as an indirect comparison with one or more periods between period *0* and *t*, which is the case with so-called "chain- indexes".

- *Time reversal test*

  This axioma states that if the prices of period *t* and the prices of reference period *0* are switched with each other, the ensueing index should be identical to the reciprical of the original index. This axioma is the result of combining the *identity* and *circularity* tests.

- *Consistency in aggregation*

  This axioma is satisfied if for every single index a weight exists which can be used to estimate the overall-index and the resulting overall-index is independent of the number of levels for which indexes are estimated.

Figure 1: *The table shows whether the different methods comply to the axiomas.*

Axioma	Laspeyres	Paasche	Fisher	Time dummy
Proportionality test	Yes	Yes	Yes	Yes
Identity test	Yes	Yes	Yes	Yes
Commensurability test	Yes	Yes	Yes	Yes
Circularity test	No	No	No	No
Time reversal test	No	No	Yes	Yes
Consistency in aggregation	Yes	Yes	No	Yes

Looking at figure 1 the first three assumptions are most important, whereas the circularity and time reversal tests are less crucial. This last assumption is mainly important for reasons of user-friendliness and clarity of the price index for its users. It is clear that all methods satisfy the first three important assumptions. Ultimately, using the axiomatic approach, the Fisher index is often deemed the best of the methods described above (Hill & Melser, 2008). However, even though this approach is a very useful guideline, small deviations can be acceptible and under specific circumstances other methods might therefore be better for constructing a hedonic price index (Van der Grient & De Haan, 2008).

3. Data

In this chapter the data used in this study will be discussed. Firstly, the sources from which the data originated will be elaborated. Secondly, the choices and underlying reasons regarding what data to use and how to use it will be discussed.

3.1 Data sources

To construct the hedonic regression model and ultimately create a house price index, data is used from two primary sources, namely "Kadaster", the Dutch land registry office, on the one hand and "Central Bureau of Statistics" on the other. This means that the base of the data comprises of two datasets. The BAG "Basisregistratie adressen en gebouwen" from both CBS and *Kadaster* and the "Bestaande koopwoningen transacties" dataset, also from CBS and *Kadaster*. These two datasets can be combined using the "PHT-code", which uses the postal code, house number and house number addition of a specific residential property to create a unique code.

The BAG database contains the details of addresses and buildings in the Netherlands. Details on existing residential property are therefore included in this database. Since 2012 the Dutch municipalities publish several variables in the BAG, which is managed by *Kadaster*, that can be used for constructing a hedonic price model. This includes the living area in square meters, the year a property was built, address and geographic coordinates of the property. The CBS dataset "Bestaande koopwoningen transacties" contains transactions of Dutch residential properties, some 780.000 in the period 2012 to 2016. It comprises of the following variables: transaction prices, type of structure, transaction date and the province in which a property is situated.

3.2 Proximity & other variables

Combining the variables mentioned in the previous paragraph creates a "base" dataset. This dataset can be enriched by adding a long list of variables related to the proximity of amenities, which are constructed by the CBS using the geographical coordinates of the properties and the amenities which are extracted from the BAG database. These amenities comprise of 95 variables divided among six groups: 'culture', 'retail', 'hotel and catering', 'education', 'accessibility', 'child day-care' and 'healthcare'. Since the data concerning the proximity of amenities is not yet available for 2016, it was decided to omit this year from the dataset. Furthermore, the data for museums, general medical practices and physiotherapy practices are not consistently available in the period 2012-2015. The 10

variables concerning these amenities are thus also omitted from the dataset, which leaves 85 variables that are related to the proximity of amenities. This group of 85 variables can be categorized in two groups: the first group of 28 variables gives the distance in meters from the property to the nearest amenity of a certain type. Whereas the second group is comprised of variables that count the number of times a certain amenity is present within a certain distance. This second group of variables has three buffer distances for each amenity. For instance, the number of primary schools within 3, 5 and 10 kilometers of a certain dwelling, the size of these buffer distances varies between amenities, but the concept of a small, medium and larger buffer distance is the same for all amenities. Energy labels were ultimately excluded from the dataset as adding these variables would result in the loss of approximately 300.000 observation, which seemed too much, even though the inclusion of the energy labels does explain about 8% of the variance in transaction prices. The resulting dataset consists of 425.000 observations between 2012-2015 of 99 variables.

3.3 Unusable values

The 425.000 observations contain values for all variables, however not all these values appear logical. Therefore, a quality cleansing of the data seems necessary. As a starting point, using the same parameters as the current existing residential property price index of the CBS, only observations with a transaction price between €10.000 and €5.000.000 are included in the dataset. The variable for living area also contained several highly unlikely values. It appears that 99.8% of the values are below $1000m^2$, whereas the largest value is above $108000m^2$. Therefore, it is decided to, at least, exclude all values above $1000m^2$. Upon further inspection it appears that many observations close to this $1000m^2$ threshold are also very unlikely, e.g. observations with a price below €100.000 but a size of over $750m^2$. On the other side of the spectrum there are also several observations with an unrealistically small living area, $5m^2$ for instance. This study considers all observations under $25m^2$ unrealistic to the extent that they have to be omitted. Since these outliers would heavily influence the relationship between price and living area, only property transactions with a stated living area between $25m^2$ and $750m^2$ are included in the models. Properties smaller than $25m^2$ and larger than $750m^2$ appear to be mostly administrative errors.

One more important note regarding the omission of data concerns the variable for structure type. This variable can assume 6 different values, “apartment”, “detached”, “semi-detached”, “mid-terrace-house”, “end-terrace-house” and “unknown”. The value “unknown” means that it was not possible to assign a structure to one of the five categories. The observations with the value “unknown” for structure type will also be omitted from the dataset, conform with the current PBK.

4. Methodology

This chapter elobarates on the methods used to analyze the data. Firstly, the way in which the regression models are constructed will be discussed and the underlying assumptions will also be tested. Secondly, this chapter explains how the hedonic price indexes are estimated using the aforementioned regression models.

From the dataset, described in the previous chapter, two main hedonic regression models will be estimated with the intention to make a constant quality price index using these regressions. The first hedonic price model will contain four “base” variables, namely the size of the living area, the age of the structure, the type of structure and the province the structure is located in. In the second hedonic price model the variables related to the proximity of amenities will be included. In order to do this a factor analysis is made to reduce the number of variables. After the two different hedonic regression models are made, it is possible to use these models in an attempt to create several hedonic price indexes for both models.

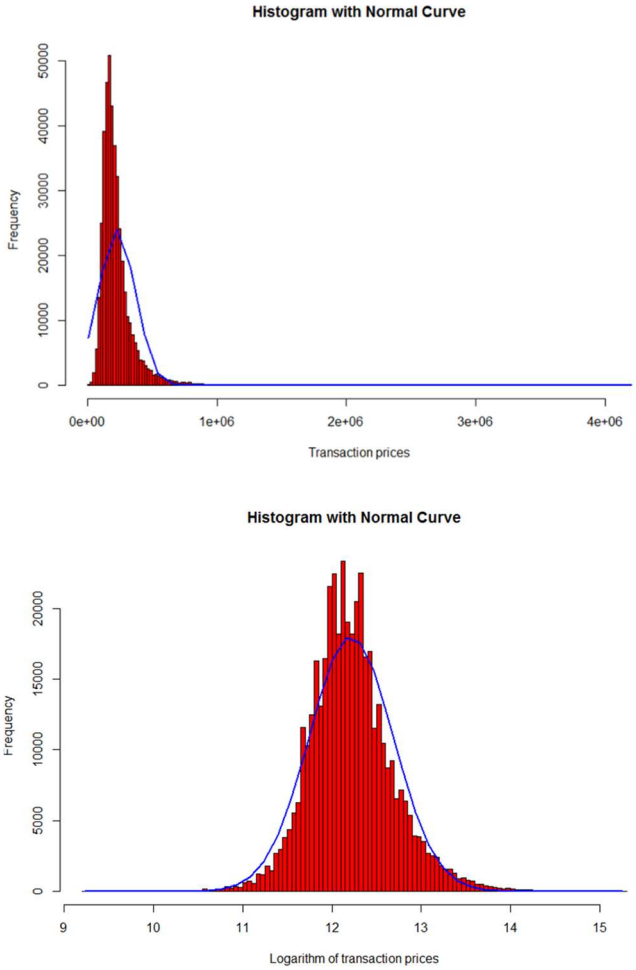
Estimating these two different hedonic regression models has two primary reasons. The first reason has to do with the academic relevancy of this study. One of the main objectives of this study is to find out whether adding variables related to the proximity of amenities could be an improvement compared to a hedonic model which does not control for the effects amenities can have on the price of a residential property. To establish this, it is necessary to first estimate a model without these variables in order to compare it with the model that does incorporate these variables.

The second reason is related to the practical relevancy of this study to the CBS. The data for the first model is readily available, whereas the data for the proximity variables, necessary for the second model, is not. Therefore, the first model could be repeated in future using the data of additional periods without major alterations to the methodology. Whereas the factor analysis in the second model would, most likely, require several choices, made in the proces of said factor analysis, to be re-evaluated with the addition of new data. For these reasons it is more practical to also preserve a model that can be ‘easily’ reproduced in the future, using data of future periods, rather than only having the more complicated second model. From this point on these models will be referred to as ‘hedonic price model 1’ for the model without the proximity variables and ‘hedonic price model 2’ for the model with these proximity variables.

4.1 Hedonic price model 1

In hedonic price model 1 several data transformations are made to create a better fit. In appendix 1 and figure 2 the spread of the variable “transaction price” can be seen. Most of the observations are concentrated on the bottom part of the graph, thus indicating that this variable is positively skewed. Appendix 2 shows the spread of the logarithm of the transaction price, in this graph the positive skew seems to have disappeared and the spread of the variable appears to closer resemble a normal distribution.

Figure 2: spread of variable "transaction price" (top), spread of the logarithm of "transaction price" (below)



Therefore a logarithm of transaction price is used in the model as the dependent variable. Furthermore, the first independent variable, living area in square meters, is transformed with a square root. This was done because the size of the living area of a property appears to have a marginally diminishing effect on the transaction price, meaning that every extra square meter adds less value to a property than the one before. Appendix 3 and 4 show the relationship between "transaction price" and "living area", without and with this transformation respectively. From these graphs it appears that performing data transformations creates a more linear relationship between the variables "transaction price" and "living area". The second independent variable is the type of structure, which assumes one of the five different values described previously and is thus categorical. The third independent variable is the province in which the dwelling is located. This variable is thus also categorical and assumes the name of the province as value. The province variable is added to model 1 in order to have some form of locational variable in the model which enables it to control for possible differences in the values of properties related to their general location. Province is chosen over municipality for this purpose, because there are a large number of municipalities, around 390, of which some have very few property transactions in certain periods. Furthermore, during the timespan of this study some new municipallities have emerged and some have ceased to exist. This would make the model much more unstable, which is very undesireable in the price index setting of this study. The fourth and final independent variable of this first model concerns the age of the

dwelling. Adding this variable as a numerical variable to the model would assume a linear relationship between the dependent and independent variable. However, literature indicates that this may not be the case (Goodman & Thibodeau, 1995). To account for this the variable is divided in several age categories, meaning that the variable "age" is thus added as a categorical variable. Using the above mentioned variables and transformation a hedonic linear regression model is estimated by means of Ordinary Least Squares (OLS), which can be found in appendix 5.

4.2 Outlier detection and assumptions

After the model is estimated a more quantitative outlier detection test can be used, namely Cooks' distance. This test estimates the influence that each data point or observation has on the estimations of the regression model (Cook, 1977). These distances are then compared to a cut-off point. If the Cooks's distance value of a data point is larger than this cut-off value a data point is considered to have an influential effect on the regression model and can thus be deemed an outlier. The consensus in the literature appears to be that the cut-off value should be 4 divided by the number of observations (Bollen & Jackman, 1990). A total of 23.370 observations have a Cooks' distance greater than the above-mentioned cut-off value, this amounts to 5,6% of the observations. A new model is thus estimated, excluding the observations which were deemed outliers by the Cooks' distance test (appendix 5). A residual scatterplot is made to assess if there are any obvious problems with this model. It appears that the residuals are not randomly distributed, as a pattern can be observed. The residuals seem to be left-skewed (appendix 6). This indicates that there might be heteroskedasticity in the model. Therefore, a Breusch-Pagan test is conducted on the new regression model in order to detect the presence of heteroskedasticity. The null-hypothesis of the Breusch-Pagan test assumes homoscedasticity, meaning that if this null-hypothesis is rejected the presence of heteroskedasticity is presumed (Breusch & Pagan, 1979). Figure 2 (appendix 7) shows that the P-value of the Breusch-Pagan test is smaller than 0,05.

Figure 2: *Breusch pagan test for hedonic price model 1*

Breusch-Pagan test	BP	Degrees of freedom	P-value
	31812	24	0.000

This means that the null-hypothesis is rejected, which indicates the presence of heteroskedasticity in the regression model. This means that OLS, although still unbiased, is inefficient because it underestimates the true variance and covariance (Johnston, 1972). Therefore, a different linear regression estimator, such as Weighted Least Squares (WLS), might prove better or more efficient than OLS. However, since the weights of WLS are based on assumptions regarding the structure of the heteroskedasticity, they can be rather arbitrary. Therefore, instead of WLS, White-Huber standard errors is used to estimate the regression model with heteroskedasticity-consistent standard errors (White, 1980) (Appendix 8). As hedonic price model 1 only has one linear predictor, 'living area', multicollinearity is not a concern in this model.

Finally, as it appears that the independent variable "living area" does not have a strictly linear relation with the dependent variable "transaction price", the model is checked with a Ramsey Regression Equation Specification Error Test (RESET) test. This test checks if non-linear combinations of the fitted values explain part of the dependent variable. Meaning that if the null-hypothesis of the

RESET-test is rejected there is some form of misspecification present in the model (Ramsey, 1969). This test shows that the hedonic price model 1 does suffer from some form of misspecification (p-value < 0.000, appendix 9), which is most likely caused by the, partly, non-linear relationship between "living area" and "transaction price". In an attempt to solve this problem two alternative models have been estimated, using different approaches to estimate the non-linear relationship between "living area" and "transaction price". The first of these alternative models uses a combination of a linear term and a quadratic term for "living area" to predict the dependent variable, which will be referred to as the "quadratic model". The second alternative model also uses this quadratic approach, but adds an additional variable which consists of the interaction effect between "living area" and "structure type", which will be called the "interaction model" henceforth. The interaction model is made because the effect of "living area" on "transaction price" could vary between different structure types, thus causing the observed misspecification.

For both alternative models the RESET-test null-hypothesis of correct specification is rejected ($p < 0.000$, appendix 9). It appears that the problem of misspecification persists through attempts to make the relationship between the dependent and independent variables more linear. Log-likelihood tests show that the alternative models, especially the interaction model, are a better fit than the original hedonic price model ($p < 0.000$, appendix 9). However, looking at the R^2 values, it seems that the quadratic and interaction models only provide an additional 0.05% and 0.13% explained variance. The fact that they are a better fit according to the log-likelihood-test probably has to do with the large amount of observations and the, subsequently, large number of degrees of freedom, which causes a slight increase in explained variance to be a better fit according to the log-likelihood test. Therefore, it is decided to continue using the original model. Firstly, because the alternative models would add more complexity to the model whilst only providing a marginal increase in explained variance. Secondly, because neither of the above-mentioned models satisfy the null-hypothesis of correct specification of the RESET-test. The results of the Ramsey RESET test can however be fickly with large datasets and must be interpreted carefully in this context (Kempf, 2015). As the dataset consists of around 400.000 observations, the graphical indication of linearity between the dependent and independent variables is deemed satisfactory (appendix 4).

4.3 Factor analyses

The hedonic price model 2 can be made created by adding the locational variables to the aforementioned hedonic price model 1. However, since the primary aim of this study is to create a constant quality price index , it is deemed that including 85 locational variables is excessive. Factor analysis is therefore used to reduce the number of variables by using the underlying factors, without disregarding large quantities of data, which would be the case if only a certain selection of these variables would be included. To conduct the multiple factor analysis, the 85 variables related to the proximity of amenities are categorized into the two groups previously described. Otherwise the factors, found by the factor analysis, would group variables based on their data structure, as the first group of 28 variables gives the distance in meters from the property to the nearest amenity of a certain type. Whereas the second group is comprised of variables that count the number of times a certain amenity is present within a certain distance. After having re-evaluated these two groups it was decided to split group one up into the three sub-groups of 19 variables regarding the buffer distances 'small', 'medium' and 'large', as described previously in the data chapter, in order to prevent the variables to load on factors based buffer distance rather than based on similarities in the

type of amenity the variable measures. It is deemed that performing a factor analysis for all of these three sub-groups with the intention to use the factor scores in the regression would not provide more useful results than just only including one of these groups, as they all measure a very similar phenomenon, namely the density of amenities surrounding a property. As one would expect that the effect of the density of amenities on house prices would be strongest in the smallest buffer distance, this sub-group is used to measure the density of amenities. From this point on group one of 28 variables will be referred to as 'closeness (of amenities)' and sub-group one of the second group of 19 variables will be called 'density (of amenities)'.

Before performing the actual factor analysis, it is important to check if it is appropriate to perform a factor analysis on the dataset in the first place. The Kaiser-Meyer-Olkin (KMO) test is therefore used to assess if the data is suited for factor analysis and the Barlett's test of sphericity is carried out to check for intercorrelation between the items. The KMO test returns a value between 0 and 1 and it is generally accepted that values larger than 0,6 are acceptable and the closer the value is to 1 the better the data is suited for factor analysis (Kaiser & Cerny, 1977). The overall measured sampling adequacy (MSA) for 'closeness' is 0.94, whereas the MSA for 'density' is 0.93 (appendix 10). These MSA scores are very high and reflect that the data is suited for factor analysis. Furthermore, the Bartlett's test of sphericity showed no signs of intercorrelation between the items as both the correlation matrices of 'density' and 'closeness' had a p-value of 0 (appendix 11). A factor analysis using the oblique "promax" rotation is made to assess which rotation should be used for the factor analysis of the variables for 'closeness' and 'density' (respectively, appendix 12 & 13). Both factor analyses show that the factors are correlated with one another, this indicates that an oblique rotation is better suited for this factor analysis than an orthogonal rotation such as the often used "varimax" rotation. Due to the oblique nature of the factors, it was thus decided to make the factor analysis for both groups using a "promax" rotation.

The number of factors is determined based on the eigenvalues of the factors, portrayed in appendix 14. This means that a cut-off point is chosen and only factors with an eigenvalue above this point are included in the factor analysis. The cut-off point is 1, because factors with eigenvalues below 1 are believed to be unstable. They explain less variability than a single variable and are therefore excluded from the analysis. This way one ends up with fewer factors than original variables (Girden, 2001), which is, as stated earlier, the aim of the factor analysis in this study. Using this method to establish the number of factors results in six factors for the 'closeness (of amenities)' and five factors regarding the 'density (of amenities)'. If we add one more factor to either of these factor analyses, the eigenvalue of the last factor is smaller than 1. Factor seven of 'closeness' has an eigenvalue of 0.941 and factor six regarding 'density' has an eigenvalue of 0.796 (appendix 14). Therefore, it is determined to make the factor analysis for 'closeness' with six factors and the factor analysis for 'density' with five factors.

To use the factor analysis in the regression model the factor scores have to be extracted from the analysis. This can be done in several different ways, using non-refined and refined methods. Non-refined methods are generally easier to use, but, as the word 'non-refined' already illustrates, are less accurate in estimating factor scores (Di Stefano, Zhu, & Mîndrilă, 2009). Refined methods can be used when both principal components and common factor extraction methods are used with explanatory factor analysis (EFA)(Di Stefano et al., 2009). The factor scores that stem from these methods are linear combinations of the observed variables. These not only take the shared variance

between the item and the factor into consideration, but also the uniqueness which indicates the part of the variance that is not measured (Gorsuch, 1983). In order to construct factor scores the more often used refined methods use standardized information, which creates standardized scores resembling a Z-score metric, with values ranging from about -3,0 to 3,0 (Di Stefano et al., 2009). This type of refined method attempts to maximize validity and acquire unbiased estimates of the true factor scores by creating factor scores that are highly correlated with a given factor (Di Stefano et al., 2009). These methods aim to preserve the relationships between factors. The factor scores should thus be uncorrelated with other factors if the EFA is orthogonal and the correlation among factor scores should be the same as the correlation among factors if the solution is oblique (Gorsuch, 1983; Di Stefano et al., 2009).

The three most well-known refined methods are 'Thurstone scores', 'Bartlett scores' and 'Anderson-Rubin scores'. The 'Thurstone scores' or 'regression scores' method uses a least squares regression approach to estimate factor scores. The scores estimate the position of each observation on the factor. The 'Bartlett scores' method considers only the common factors to have an influence on factor scores. The factor scores are estimated by minimizing the sum of squared components for the unique factors across the set of variables (Thurstone, 1935; Bartlett, 1937; Di Stefano et al., 2009). The advantage the 'Bartlett scores' method has over the other two is that the factor scores are created by using maximum likelihood estimates, which generates estimates that are the most likely to portray the "true" factor scores. 'Bartlett scores' is thus the most likely to generate unbiased estimates of the true factor scores (Hershberger, 2005). The 'Anderson-Rubin scores' method is a variation of the 'Bartlett scores' method. It alters the least squares formula, used in 'Bartlett scores', to create factor scores that are uncorrelated with each other as well as other factors. The resulting factor scores are always orthogonal with a mean of 0 and a standard deviation of 1 (Anderson & Rubin, 1956; Di Stefano et al., 2009). Ultimately, the 'Bartlett scores' method of estimating the factor scores was chosen over the other two methods to optimize validity and still have the highest likelihood of unbiased estimates of factor score parameters (Di Stefano et al., 2009).

4.4 Hedonic price model 2

Having added the scores of the six 'closeness' factors and five 'density' factors to the dataset, it is possible to estimate hedonic price model 2. Starting again with the exclusion of outlier observations by looking at the Cooks' distances using the same method as for hedonic price model 1. The resulting OLS regression model is found in appendix 15. As heteroskedasticity was detected in the previous model and the residual scatterplot of model 2 shows a similar pattern (appendix 16), it seems advisable to use another Breusch-Pagan test to check for heteroskedasticity in this model. The Breusch-Pagan test shows that heteroskedasticity is also present in hedonic price model 2 (appendix 17), therefore another regression model is created using White-Huber standard errors to estimate the regression model with heteroskedasticity-consistent standard errors (appendix 18). In this second model it is important to check for multicollinearity as there are now twelve linear predictors, 'living area' and the eleven factors. The correlation matrix (appendix 21) of these linear predictors shows that there is indeed correlation between the predictors, which was to be expected due to the oblique nature of the factors. However, it is does warrant a statistical test for multicollinearity. To check for multicollinearity the variance inflation factor (VIF) values of the linear variables are estimated (appendix 22). Fortunately, none of the variables have a VIF value higher than 4.0, and

therefore it is possible to conclude that the assumption of multicollinearity has not been violated (Pan & Jackson, 2008).

Ultimately, it seems advisable to also test whether the factor scores estimated with the factor analysis fit the data better than a more rudimentary method of reducing the available locational variables. To do this the highest scoring variable of each of the eleven factors (six "closeness" & five "density") has been determined. In this context these variables can be referred to as proxies of the underlying factors. These proxy variables are respectively: 'distance to stores for daily groceries', 'distance to public transport transfer station', 'distance to preparatory vocational education school', 'distance to hospitals without advisory doctor's practice', 'distance to department stores', 'distance to out-of-school care centre', 'presence of cafeteria 1km', 'presence of hotel 5km', 'presence of elementary school 1km', 'presence of hospitals with advisory doctor's practice 5km', 'presence of preparatory vocational education school 3km'(appendix 19, see also: appendix 12 & 13). The same hedonic price model as described above is estimated again, except now the factors have been swapped out with the previously described proxy variables, this model will henceforth be called the "proxy model" (appendix 20).

4.5 Price indexes

This study will estimate four price-index models, for both hedonic regression models described previously, following the methods outlined by Eurostat (Eurostat, 2013). The resulting price indexes of hedonic price model 1 and 2 are, respectively, found in appendix 23 and 24. Firstly a time dummy price index is estimated and secondly a double imputation method is used to calculate three more price indexes. These three indexes are estimated using Laspeyres and Paasche methods after which it is possible to use these two indexes to calculate a third index with the Fisher method.

The time dummy price index can be estimated relatively easy by adding a categorical variable that consists of dummy variables of the time periods, being the quarters of each year, to the regression and using the regression estimates to formulate a price index. This price index can be estimated with a single regression model. The formula for the hedonic time dummy (HTD) model thus becomes:

$$\ln(p_i^t) = \beta_0 + \sum_{t=1}^{T} \delta^t x_k^t + \sum_{t=1}^{T} \delta^t x_{ik}^t + \varepsilon_i^t$$

And the model to extract the hedonic time dummy price index:

$$I_{HTD}(t,0) = \frac{\left(\prod_{i \in S_t} p_i^t\right)^{1/N_t}}{\left(\prod_{i \in S_0} p_i^0\right)^{1/N_0}} \exp\left(\sum_{k=1}^{K} \widehat{\beta}_k \left(\bar{x}_k^0 - \bar{x}_k^t\right)\right)$$

Where p_i^t denotes the transaction price of residential property i in period t and S_t is the number of residential properties in the sample of period t. Furthermore, $\hat{\beta}_k$ denotes the average value of the k-th coefficient in periods 0 and t. Finally, $\bar{x}_k^t$ is the mean of the property characteristics k of the N_t property in period t and K is the number of property characteristics used in the hedonic regression model.

The other three indexes are different variants of the double imputation method. The Laspeyres double imputation index imputes the prices for the residential properties of base period *0* and of period *t*. It does this by estimating a model for both period *0* and period *t*. It then multiplies the coefficients generated by these two models with the average property characteristics of period *0*, which yields an imputed price for both periods. A price index can thus be estimated by comparing the imputed price of the base period to the imputed price of period *t.* This means that the Laspeyres method evaluates price change at the average property characteristics of the base period to ensure that quality changes between periods are accounted for (Eurostat, 2013). The formula for extracting the Laspeyres hedonic double imputation price index (HDIL) thus reads:

$$I_{HDIL}(t,0) = \frac{\exp\left(\sum_{k=1}^{K} \widehat{\beta}_k^t \bar{x}_k^0\right)}{\exp\left(\sum_{k=1}^{K} \widehat{\beta}_k^0 \bar{x}_k^0\right)}$$

where $\hat{\beta}_k$denotes the average value of the k-th coefficient in periods 0 and t. Furthermore, $\bar{x}_k^0$ is the mean of the property characteristics k in period 0 and K is the number of property characteristics used in the hedonic regression model.

The Paasche double imputation index method is very similar to the Laspeyres double imputation method as it also imputes the prices for both base period *0* and period *t*. However, the difference lies within the method that is used to impute these prices. The Paasche method uses the average property characteristics of period *t*, instead of period *0* like the Laspeyres method, to calculate the imputed prices. This means that the Paasche method evaluates price change at period *t* average property characteristics to ensure quality changes between periods are accounted for (Eurostat, 2013). Thus, yielding the following formula for the extraction of the Paasche hedonic double imputation price index (HDIP):

$$I_{HDIP}(t,0) = \frac{\exp\left(\sum_{k=1}^{K} \widehat{\beta}_k^t \bar{x}_k^t\right)}{\exp\left(\sum_{k=1}^{K} \widehat{\beta}_k^0 \bar{x}_k^t\right)}$$

where $\hat{\beta}_k$ again denotes the average value of the k-th coefficient in periods 0 and t. Furthermore, $\bar{x}_k^t$ is the mean of the property characteristics k in period t and K is the number of property characteristics used in the hedonic regression model.

The Fisher double imputation index method is a combination of the results of the Laspeyres and Paasche methods. It estimates a price index based on the geometric mean of the Laspeyres and Paasche indexes. Therefore, the Fisher hedonic double imputation price index (HDIF) formula is:

$$I_{HDIF}(t,0) = \sqrt{(I_{LHDP}^{0t} I_{PHDP}^{0t})}$$

Where I_{HDIL}^{0t} and I_{HDIP}^{0t}, respectively, denote the Laspeyres double imputation and Paasche double imputation indexes.

To test the reliability of these hedonic price indexes the bootstrap-method is used to estimate the accuracy of the price indexes, usually in terms of confidence margins. The bootstrap-method or 'bootstrapping' is based on random sampling with replacement (Efron & Tibshirani, 1994). To assess the validity of the price indexes, a random sample is drawn from the total dataset for which a price index is calculated. This process is repeated a thousand times to assess what the variances of the price indexes are for each period. By taking the square root of the variance, the standard errors are obtained, which in turn are used to estimate the 95% confidence margins of the indexes for each period. However, the magnitude of these standard errors, and subsequently the confidence margins, is dependent on the level of the index. Therefore, it is advisable to also use a relative test of reliability, which is obtained by dividing the confidence margin by the index number and multiplying it times a hundred (de Vries et al., 2009). The aforementioned values are thus obtained with the following formulas:

$$Upperbound = I^t + 1{,}96 * \sqrt{V^t}$$

$$Lowerbound = I^t - 1{,}96 * \sqrt{V^t}$$

$$Confidencem\arg in(Wc_t) = (1{,}96 * \sqrt{V^t}) * 2$$

$$\Pr ecision = (Wc_t / PI_{(.),t}) * 100$$

Where I^t denotes the price index in period *t* and V^t is the variance, acquired by bootstrapping the price index, in period *t*. Furthermore, Wc_t, is the width of the confidence margin in period *t* and $PI_{(.),t}$ is the price index of method *(.)* in period *t*.

Both the confidence margins and the precision thus indicate the reliability of each of the price indexes described in this chapter. The wider the confidence margin the less reliable the price index, however to control for the effect of the index level a relative measure, the precision, is also determined. The higher the value of precision the less reliable the price index.

5. Results & Interpretation

In this chapter the final results of the two hedonic regression models, as well as the factor analyses and the price indexes described in the previous chapter 'methodology', will be portrayed and analysed. Firstly, hedonic price model 1 will be elaborated upon, followed by a discussion on the results of the factor analyses and an examination of the differences between model 1 and model 2. Lastly, the price indexes are analysed.

5.1 Hedonic price model 1

In appendix 8 the final version of hedonic price model 1 is located. This model is a hedonic regression model measuring the effects of the variables 'structure type', 'province', 'size of living area' and 'age of the structure' on the transaction price, while making use of White-Huber standard errors. The model has an adjusted-R^2 of 0.6651 and all independent variables have a significant effect on the transaction price as all p-values are smaller than 0.0001. The size of the living area has a positive effect on the transaction price, as was to be expected, meaning that an increase in the size leads to a higher transaction price. All structure types have a positive effect on the transaction price, compared to the reference structure type 'apartment'. The structure types 'detached' and 'semi-detached' have the biggest positive influence on transaction prices, compared to apartments, with 27.8% and 15.6% higher transaction prices respectively. 'end-terrace' and 'mid-terrace' properties have a smaller, but still significantly positive influence on transaction prices of 7.0% and 3.1% respectively, again compared to apartments. The effects of the province, in which a property is situated, is measured compared to reference category 'Drenthe'. The provinces 'Groningen' and 'Friesland' are the only provinces that have a negative effect, 1.5% and 4.9% respectively, while all other provinces have a positive effect on the transaction prices of properties, compared to 'Drenthe'. The provinces 'Noord-Holland' and 'Zuid-Holland' have the biggest positive effect with 47.4% and 40.0% respectively. The age of the structure is divided in time periods with the period '1991-2000' being the reference period. The time periods 'before 1905' and 'after 2001' are the only time periods with a positive effect on transaction prices of 9.5% and 2.8% respectively, while all time periods between 1905 and 1990 have a negative effect, compared to the '1991-2000' time period. However, it is interesting to note that the '1905-1930' and '1931-1944' time periods have a smaller negative effect on the transaction price, 3.2% and 3.7% respectively, than the other time periods between 1905 and 1990, compared to reference period '1991-2000'. These time periods: '1945-1960', '1961-1970', '1971-1980' and '1981-1990' have negative effects of 17.5%, 21.3%, 19.8% and 11.3% respectively.

5.2 Factor analyses

Appendix 12 and 13 hold the results for the factor analyses performed on the locational data concerning the proximity of amenities to residential properties. As described in the previous chapter, the variables are divided in two groups. The factor analysis of the first group, also known as the 'closeness (of amenities)', can be found in figure 3[2] and appendix 12. The factor analysis of the second group, known as the 'density (of amenities)', is located in appendix 13.

[2] A cut-off value of 0.3 is adopted, meaning that all factor loadings below this point are hidden. This is done to prevent small factor loadings from clogging up the results and making interpretation unnecessarily difficult.

Figure 3 (appendix 12): *Factor analysis 'closeness' with factor loadings and uniqueness of variables*

Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Uniqueness
Distance to library	0.335				0.399		0.619
Distance to swimming pool					0.546		0.511
Distance to artificial ice-skating track		0.800					0.513
Distance to pop stage		0.508		0.344			0.540
Distance to cinema		0.357		0.524			0.391
Distance to sauna		0.565					0.616
Distance to tanning salon				0.485			0.513
Distance to attraction		0.581					0.597
Distance to large supermarkets	0.885						0.315
Distance to stores for daily groceries	0.905						0.315
Distance to department store					0.751		0.316
Distance to cafe	0.537						0.670
Distance to cafeteria	0.685						0.411
Distance to restaurant	0.649						0.552
Distance to hotel							0.760
Distance to elementary school	0.487						0.666
Distance to secondary- or high school							0.014
Distance to preparatory vocational education school			1.007				0.060
Distance to higher general secondary education schools and preparatory academic schools			0.987				0.268
Distance to highway ramp			0.503				0.959
Distance to train station		0.738					0.446
Distance to public transport transfer station		0.882					0.230
Distance to kindergarten						0.551	0.418
Distance to out-of-school care centre						1.001	0.037
Distance to general doctor's practice	0.867						0.336
Distance to pharmacy	0.755						0.420
Distance to hospitals with advisory doctor's practice				0.660			0.346
Distance to hospitals without advisory doctor's practice				0.925			0.222

The factor analysis for the 'closeness' indicates that there are six factors with an eigenvalue of more than 1 underlying the data of the 28 variables (appendix 12). The closeness of the following amenities load on factor 1: 'libraries', 'large supermarkets', 'stores for daily groceries', 'cafes', 'cafeterias', 'restaurants', 'elementary schools', 'general doctor's practices' and 'pharmacies'. When looking at the similarities of these variables, it appears that these amenities could all be found in a district centre, which is likely to be the connecting factor of these amenities. Factor 1 of the factor analysis of 'closeness' thus seems to measure the closeness of a property to a district centre. The amenities that load on factor 2 are: 'artificial ice-skating tracks', 'pop culture stages', 'cinemas', 'saunas', 'attractions', 'train stations' and 'public transport transfer stations'. Factor 2, in this sense, appears to reflect the closeness of the city centre with many centralized amenities loading on this factor. The following amenities load on factor 3: 'secondary- or high schools', 'preparatory vocational education schools' and 'higher general secondary education schools and preparatory academic schools'. The interpretation of this factor is relatively straightforward as it strongly appears to measure the closeness of secondary education facilities. The amenities loading on factor 4 are: 'pop culture stages', 'cinemas', 'tanning salons', 'hospitals with advisory doctor's practice' and 'hospitals without advisory doctor's practice'. This factor is more difficult to interpret; though, it appears that these amenities are a mix of healthcare and cultural amenities. Factor 5 consists of these amenities:

'libraries', 'swimming pools' and 'department stores'. These can often be found in centres along the edge of cities. The final factor of the 'closeness' analysis is made up of the following two amenities: 'kindergartens' and 'out-of-school care centres'. This sixth and final factor seems to reflect child-care centres for children between the ages of 1 to 12.

The factor analysis for the 'density', shows five factors with an eigenvalue of more than 1 underlying the data of the 19 variables (appendix 13). The first factor consists of the following amenities: 'Large supermarkets', 'stores for daily groceries', 'cafes', 'cafeterias' and 'restaurants'. Factor 1 of the 'density' analysis thus looks very similar to the 'closeness' factor 1, because it appears to indicate the presence and size of a district centre near a property as it is based on the number of components (read amenities) of such a district centre present near the property. The amenities loading on factor 2 are: 'pop culture stages', 'cinemas', 'department stores' and 'hotels'. Factor 2 again, like in the 'closeness' analysis, appears to relate to city centres, as it measures the presence and size of such a city centre. Factor 3 consists of: 'Large supermarkets', 'stores for daily groceries', 'elementary schools', 'kindergartens', 'out-of-school care centres' and 'general doctor's practices'. The amenities loading on this factor are similar to the amenities loading on factor 1 of both analyses, which are both identified as appearing to reflect district centres. However, this factor is slightly different as it includes 'kindergartens' and 'out-of-school care centres', but excludes food and beverage amenities such as 'cafes' and 'restaurants', which might suggest that the factor indicates a neighbourhood centre rather than an often-larger district centre. The following amenities load on factor 4: 'Attractions', 'department stores', 'hospitals with advisory doctor's practice' and 'hospitals without advisory doctor's practice'. This factor appears to reflect centres of amenities on the edge of cities. The amenities loading on factor 5 are the following: 'secondary- or high schools', 'preparatory vocational education schools' and 'higher general secondary education schools and preparatory academic schools'. This strongly suggests that the final factor of 'density' reflects secondary education facilities, like factor 3 of the 'closeness' factor analysis.

Figure 4: *Overview of which factors seem to reflect which latent variables*

Factors	Latent variables
Factor closeness 1	District centre
Factor closeness 2	City centre
Factor closeness 3	Secondary education facilities
Factor closeness 4	Cultural & Healthcare centres
Factor closeness 5	City-edge centre
Factor closeness 6	Child-care centres
Factor density 1	District centre
Factor density 2	City centre
Factor density 3	Neighborhood centre
Factor density 4	City-edge centre
Factor density 5	Secondary education facilities

Looking at figure 4, it appears that the factor analysis groups certain amenities based on what type of centre they would belong in. It seems to divide the amenities up into several discrete centres, ranging from small neighbourhood centres with grocery stores to, often large, city centres with hotels, department stores and cultural amenities. This structure found by the factor analysis seems to line up with Harris' & Ullman's (1945) 'multi neclei' theory of city structure. Though of course the

amenities used in this study do not reflect all nuclei in a city mentioned by Harris & Ullman, as they also consider manufacturing centres and central business districts, something this study does not. However, it does seem that the amenities surrounding residential real estate can be brought back to certain discrete nuclei most cities possess.

5.3 Hedonic price model 2

In hedonic model 2 the above described 'closeness' and 'density' factors are added to hedonic model 1, yielding the results depicted in appendix 18. Hedonic model 2 has an adjusted-R^2 of 0.7074, thus explaining about 4.2% more variance than hedonic model 1. Not all independent variables have a significant effect in this model as dummy variable 'Limburg', which is a part of the categorical variable province name, is no longer significant. However, all other independent variables retain their significance and, furthermore, all factors added to this model have a significant effect on transaction prices ($p<0.0001$, appendix 18). The effect of the size of the living area on the transaction price of a property remains relatively the same in the model, compared to hedonic price model 1. All structure types still have a positive effect on transaction prices in hedonic price model 2, compared to reference category 'apartment'. However, the positive effect of all categories is substantially bigger than in hedonic price model 1. The positive effect of the structure types 'detached', 'end-terrace-house', 'mid-terrace-house' and 'semi-detached' are 39.9%, 16.1%, 11.9% and 26.6% respectively. This suggests that adding the factors controls for an effect that makes apartments more expensive compared to the other structure types, e.g. being in the vicinity amenities. Most provinces have a very similar effect on transaction prices in both models, compared to reference category Drenthe. Limburg is one that stands out as it no longer has a significant effect on transaction prices in hedonic model 2. However, it appears that this is the result of 'Limburg' not significantly differing from the reference category 'Drenthe' in regard to transaction prices when the factors are added to the model. The very small estimate of category 'Limburg' points in this direction. The categorical variable concerning the age of the structure also has some notable differences between the two hedonic models. In hedonic model 2 the time periods before 1945 have a more negative influence on transaction prices than in hedonic model 1, compared to reference period '1991-2000'. This to the extent that the period 'before 1905' has a negative effect on the transaction price in hedonic model 2, rather than a positive effect, which was the case in hedonic model 1. Periods '1905-1930' and '1931-1945' simply have a more negative effect on transaction prices, compared to the reference period, than they had in the previous model. This appears to indicate that part of the reason that houses, built in the aforementioned time periods, are viewed as favourable is due to their location.

Having discussed the other variables in hedonic price model 2, it is possible to look at the effects of the locational factors themselves. All of the 'closeness' and 'density' factors show a significant effect on the dependent variable 'transaction price'. Due to the manner in which the factor scores have been determined it is difficult to interpret the size of the coefficients. However, it is possible to compare the factors with one another and to examine the signs. The 'closeness' factors with a positive sign are 'district centre' and 'healthcare', whereas the factors 'city centre', 'secondary education', 'city-edge centre' and 'childcare' have a negative sign. In this context a positive sign means that the further a dwelling is from a 'closeness' factor the higher its price will be. A negative sign means that the price will be lower the further a dwelling is from a 'closeness' factor. So being closer to a city centre, secondary education facilities, a city-edge centre and childcare has a positive

effect on the price of a dwelling, whereas being closer to a district centre and healthcare affects prices negatively.

These results indicate that being close to certain amenities, even the ones which would seem beneficial at first glance, does not have to result in a positive change in the transaction price. This seems counterintuitive at first, but could be explained by assuming there are in fact two effects instead of just one. On the one hand people are willing to pay more for dwellings that are close to amenities (Glaeser, Kolko, & Saiz, 2001), whilst on the other hand people want to avoid being so close to the amenities that they would incur negative externalities of said amenities (Li & Brown, 1980). This creates a situation in which the distance to an amenity can have a positive and a negative effect on the price of a dwelling depending on the size of said distance. In this context, it would thus seem that the negative externalities of being 'too close' outweigh the positive effect for the 'district centre' and 'healthcare' factors. Comparing the different 'closeness' factors to each other it becomes clear that being close to the city centre has the strongest positive effect on transaction price and that being close to healthcare has the biggest negative effect on transaction price.

With the 'density' factors, a positive sign means there is a positive relationship between the number of nearby amenities and the price of a dwelling, while a negative sign suggests a negative relationship. The presence of district centres, city centres and education facilities have a positive effect on the price of a dwelling and, of these three, the presence of a city centre has the largest positive effect. However, the presence of neighbourhood centres and city-edge centres have a negative effect on the price of a dwelling. These negative effects could be attributed to a similar phenomenon as we have seen with the 'closeness' factors. People seem to want to live close to amenities, but one can imagine that too many amenities, such as schools or large supermarkets can give rise to negative externalities which have a negative effect on the price (Li & Brown, 1980).

Ultimately, it is possible to assess the effectiveness of the factor analysis by comparing it with the proxy model (appendix 20). The coefficients of the factors and proxies vary as the absolute values of the proxy coefficients are much smaller. This was to be expected as the factors combine the effects of multiple amenities, whereas the proxies only show the effect of a single amenity. Most of the proxies have the same sign as their corresponding factor, with one exception, because 'presence of a city-edge centre' has a negative sign, whereas the proxy 'presence of hospitals with advisory doctor's practice' has a positive sign. However, most striking is the fact that the proxy model has a higher R^2 than the factor model, 0.7112 and 0.7074 respectively. This would mean that the method of using proxies to reduce the number of variables explains the variance in transaction prices better than the more sophisticated method of using factor scores which allows for more data to be preserved. This is remarkable in and of itself, however, perhaps it is caused because the buyers of residential properties make decisions based on what they see rather than based on complicated econometric models such as a factor analysis. For example, people might use the closest grocery story to assess whether a dwelling has good access to groceries rather than calculating it using all amenities linked to retail in the proximity of the dwelling. This would explain why the more simplistic method fits the data better than the more sophisticated method.

5.4 Price indexes

Appendices 24 and 25 reflect the results of the price indexes calculations described in the methodology chapter, without and with the inclusion of the factors respectively. Figure 5 shows the price indexes of hedonic price model 1 and figure 6 reflects the price indexes of hedonic price model 2. Starting with the four price indexes for hedonic model 1, it is clear that the price indexes do not show large differences between one another. Although they do differ from the SPAR-index, especially around the fourth quarter of 2013 the hedonic indexes and the SPAR-index seem drift away from each other, after which they gradually converge again from the first quarter of 2015 up until the fourth quarter of 2015, which marks the end of the time series. The hedonic price indexes of hedonic model 2 do follow a similar pattern and, even though they naturally differ, they never differ more than 0.45 percentage points from their hedonic model 1 counterparts. Though it is worth noting that during the last period the hedonic model 1 indexes decline, whilst the hedonic model 2 indexes rise. However, it is difficult to pinpoint what causes this deviation, partly because it happens in the last time period and it is not possible to assess whether the deviation persists through in later time periods.

Figure 5: *hedonic price indexes for hedonic price model 1*

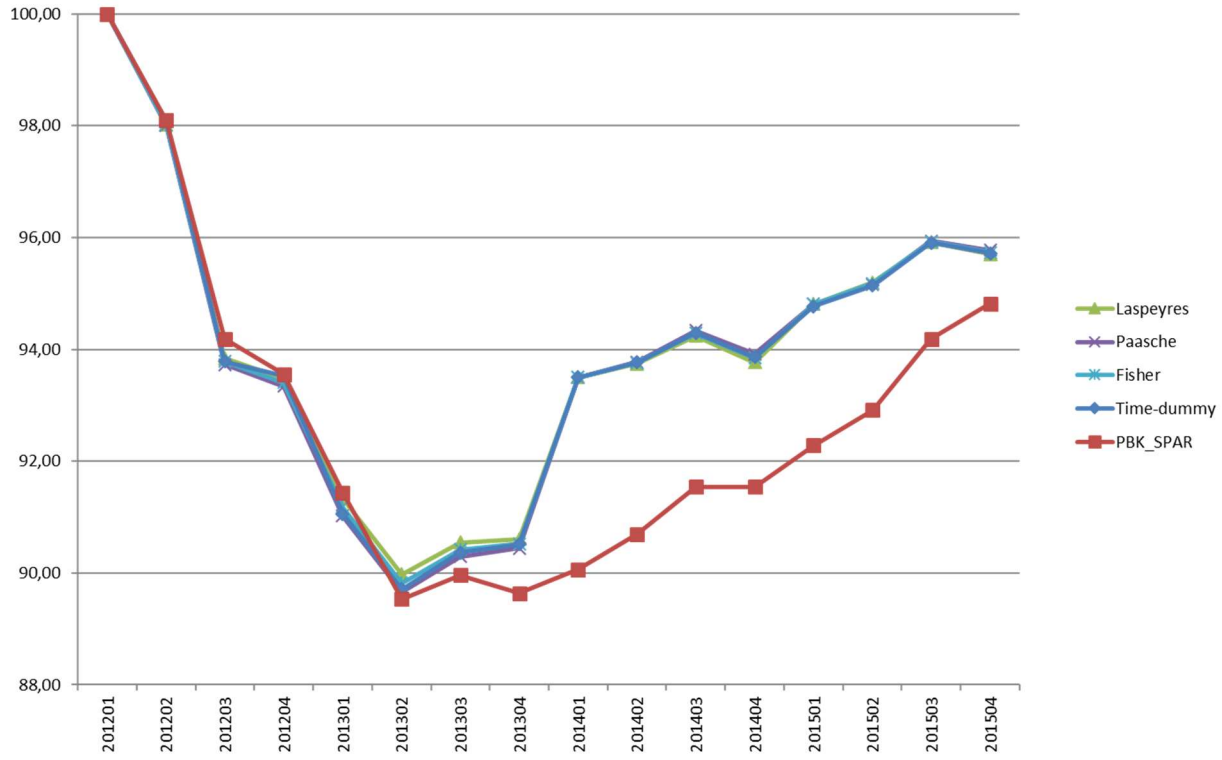
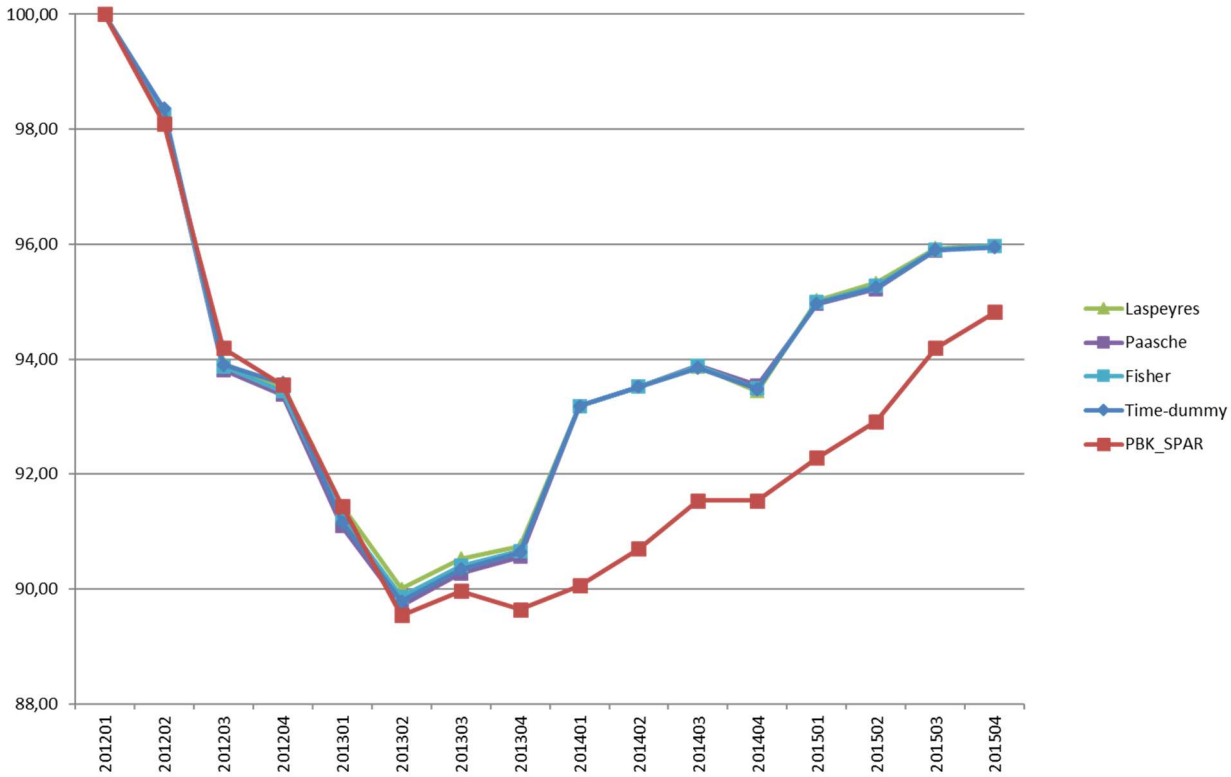


Figure 6: *hedonic price indexes for hedonic price model 2*



To analyse the quality of the price indexes the confidence margins are estimated. All hedonic price indexes estimated using model 1 show confidence margins between the 0.9 and 1.0 percentage point. The hedonic time dummy price index has an average margin of 0.946 percentage point, whilst the Laspeyres, Paasche and Fischer double imputation indexes have average margins of 0.959, 0.969, 0.951 percentage points respectively. Furthermore, it appears that the margins remain stable, as they rarely differ more than 0.05 percentage point from the mean. This indicates that the price index does not deteriorate over time as the margins do not widen with every period, something that usually happens with chained indexes.

The hedonic indexes of model 2 have lower average margins. The time dummy price index of model 2 has an average margin of 0.924 and the Laspeyres, Paasche and Fischer double imputation indexes of model 2 have an average margin of 0.939, 0.909 and 0.901 percentage point. These margins of model 2 also remain stable over the course of the time series and do not appear to deteriorate over time.

To control for the tendency of larger price index values to have wider margins a relative measure of precision was also estimated. The time dummy, HDIL, HDIP and HDIF price indexes of model 1 have a precision of 1.012, 1.025, 1.036 and 1.016 respectively (appendix 25), whereas for model 2 these methods have a precision of 0.988, 1.004, 0.973 and 0.963 respectively (appendix 26). The results of this measure are not too different from the absolute margin width because the index values of all methods are relatively close to 100%, rarely moving further than 10 percentage points away from this point. The absolute margins are corrected slightly upwards by the relative measure as most index

values are below 100%. However, the order from best to worst method, precision-wise, remains the same for both absolute and relative measures. The method with the best precision for model 1 is the time dummy model, followed by the Fisher, Laspeyres and Paasche methods respectively. Whereas the Fisher method has the best precision for model 2, followed by the Paasche, time dummy and Laspeyres methods respectively.

6. Conclusion & Discussion

At the base of this study lies the question if a hedonic price index could be made for the residential real estate market in the Netherlands. To this end it was investigated whether adding new variables to the current CBS residential real estate dataset would make a hedonic price index possible with the following research question:

"To what extent is it possible to make a hedonic price index for the Dutch existing residential property market by adding additional variables to the current CBS database 'Bestaande koopwoningen transacties'?"

However, in order to give a comprehensive answer to this main research question, several sub-questions were established. These sub-questions will be addressed first before answering the main research question.

The first sub-question is: *"Which variables from either the BAG or PBK hold explanatory value in respect to the transaction prices and what are the effects of these variables?".*

The results of the analysis show that the "structure type" and "province name" variables from the current PBK dataset and the "size of living area" and "structure age" variables from the BAG dataset hold explanatory value in respect to the transaction price. The effects of structure type how that a detached house is most valuable, and an apartment is least valuable relative to the other structure types. The effects of province name indicate that houses in North-Holland are valued the highest and houses in Friesland are valued lowest. Furthermore, the analysis shows that newer dwellings are often valued more than older dwellings, except for what appears to be a 'vintage' effect for dwellings built before 1944. Finally, the results indicate that the size of a dwelling has a positive effect on its price with diminishing returns.

The second sub-question is: *"Which spatial variables could be added to enhance the model?"*.

The results show that 37 locational variables from the CBS on the proximity of amenities can be added to the dataset. To enhance the model with these variables a factor analysis was made. The results of this factor analysis indicate that the factors do enhance the model and, furthermore, show a multi nuclei pattern like the one described by Harris & Ullman in "*The Nature of Cities*". However, ultimately, the addition of proxy variables instead of factor scores yields a better fit for the Dutch residential real estate market. This could be explained, because most people might make decisions based on what they perceive rather than based on complicated econometric models such as a factor analysis. Therefore, however, further research would be necessary to say with certainty if the aforementioned explanation is what truly causes this phenomenon.

The third sub-question is: *"To what extent could a reliable hedonic price index be made with the addition of the BAG-variables?"*

The estimated price indexes calculated with hedonic price model 1 indicate that it is possible to make a hedonic price index. The precisions estimated for these price indexes show that adding the two variables from the BAG results in a decently consistent and reliable price index for all price index methods.

Sub-question 4: *"To what extent could the hedonic price index be made more reliable using a factor analysis of the locational variables?"*

Hedonic price model 2 shows that adding the factor analysis of the locational variables yields a better fit, in terms of explained variance, than hedonic price model 1. Furthermore, this better fit translates itself into higher precisions for all price indexes calculated with hedonic price model 2, compared to their corresponding index from hedonic price model 1. It is therefore concluded that it is possible to make the hedonic price index more reliable using a factor analysis of the locational variables.

Sub-question 5: *"Which hedonic price index method has the highest reliability and/or is most suited for constructing the price index for the Dutch residential property market?"*

The results show that for the model without the locational variables the time dummy and Fisher methods provide the best precision and thus reliability, whereas, for the model with the locational variables the Fisher and Paasche methods yield the best precision. The Fisher method thus appears to provide a relatively good reliability for both models. Furthermore, the literature indicates that the Fisher method is one of the better, if not the best, method when taking an axiomatic approach to judging the quality of price index methods (Van der Grient & De Haan, 2008; Diewert E. W., 2004). Taking into consideration the revision effect drawback of the time dummy method, it seems that the Fisher double imputation method is the most suited and reliable method for both models.

Having discussed all the sub-questions, it is possible to give a comprehensive answer to the main research question. The results of this study strongly indicate that a reliable price index for the Dutch residential property market is achievable with the addition of two of the variables from the BAG to the current CBS database 'Bestaande koopwoningen transacties'. Furthermore, this price index can be made more reliable by adding factor scores based on locational variables. However, adding these factor scores comes with two main disadvantages. Firstly, it seems that for further research it might be advisable to use proxy variables instead of factor scores. Because in this study the proxies perform better than the factor scores whilst also being easier to model and interpret. Secondly, the data on locational variables is not readily available, whereas the variables from the BAG are. This is only a minor problem from an academic standpoint, however from a practical standpoint and looking at the usefulness to the CBS it is a major issue. This is the case since producing a monthly price index for the Dutch residential property market using the locational data is practically impossible considering the current availability of this locational data. This brings us to the conclusion that making a price index without the locational variables yields the most useful and consistent price index and, looking at the results and the literature, the Fisher double imputation appears the most suited method for making this index. Ultimately, this study provides a hedonic price index model that could help better understand how specific dwelling characteristics, locational and non-locational, influence house prices and shifts in house prices over time. As this hedonic price index shows which characteristics are the driving forces behind changes in house prices over time. Most non-hedonic price indexes are unable to do this. Furthermore, from a practical standpoint, the proposed regression models and price index model could be used as a benchmark for the construction of new existing residential or other types of property market price indexes. From an academic standpoint this study indicates that grouping locational characteristics of Dutch residential property yields a pattern resembling the multi nuclei pattern described by Harris & Ullman in "*The Nature of Cities*". It also provides a detailed overview of the available data and the challenges encountered when using this data for estimating

hedonic house price models. As such this study could be used as a starting point for further research into the effects of specific locational characteristics on house prices in the Netherlands. Which is something that is yet to be studied exhaustively for the Dutch residential property market.

7. Bibliography

Anderson, R. D., & Rubin, H. (1956). Statistical Inference in Factor Analysis. *Proceedings of the Third Berkeley Symposium of Mathematical Statistics and Probability*, (pp. 111-150). Berkeley, California.

Balk, B. M. (2012). *Prices and Quantities: Models for Measuring Aggregate Change and Difference.* Cambridge: Cambridge University Press.

Bartlett, M. S. (1937). The Statistical Conception of Mental Factors. *British Journal of Psychology*, 97-104.

Bollen, K., & Jackman, R. (1990). Regression Diagnostics: An Expository Treatment of Outliers and Influential Cases. *Modern Methods of Data Analysis*, 257-291.

Bourassa, S. C., Hoesli, M., & Sun, J. (2006). A Simple Alternative House Price Index Method. *Journal of Housing Economics*, 80-97.

Breusch, T. S., & Pagan, A. R. (1979). A Simple Test for Heteroscedasticity and Random Coefficient Variation. *Econometrica*, 1287-1294.

Brounen, D., & Kok, N. (2011). On the Economics of Energy Labels in the Housing Market. *Journal of Environmental Economics and Management*, 166-179.

Case, K. E., & Shiller, R. J. (1987). Prices of Single-Family Homes Since 1970: New Indexes for Four Cities. *New England Economic Review*, 45-56.

Chauncy, D. H., & Ullman, E. L. (1945). The Nature of Cities. *The Annals of the American Academy of Political and Social Science*, 7-17.

Cheshire, P., & Sheppard, S. (1995). On the Price of Land and the Value of Amenities. *Economica*, 247-267.

Cook, D. R. (1977). Detection of Influential Observations in Linear Regression. *Technometrics*, 15-18.

Daams, M. N., Sijtsma, F. J., & van der Vlist, A. J. (2016). The Effect of Natural Space on Nearby Property Prices: Accounting for Perceived Attractiveness. *Land Economics*, 389-410.

de Haan, J. (2004). Direct and indirect time dummy approaches to hedonic price measurement. *Journal of Economic and Social Measurement*, 427-443.

de Haan, J. (2004). Hedonic Regression: The Time Dummy Index As a Special Case of the Imputation Tornqvist Index. *8th Ottawa Group Meeting* (pp. 23-25). Ottawa: Statistics Netherlands.

De Haan, J., & Syed, I. A. (2016). Age, Time, Vintage, and Price Indexes: Measering the Deprecitation Pattern of Houses. *Economic Inquiry*, 580-600.

de Vries, P., de Haan, J., van der Wal, E., & Marién, G. (2009). A House Price Index Based on the SPAR Method. *Journal of Housing Economics*, 214-223.

Di Stefano, C., Zhu, M., & Mîndrilă, D. (2009). Understanding and Using Factor Scores: Considerations for the Applied Researcher. *Practical Assessment, Research & Evaluation*, 1-11.

Diewert, E. W. (2004). *Consumer Price Index Manual: Theory and Practice.* Genève: International Labour Office.

Diewert, E. W., Heravi, S., & Silver, M. (2009). Hedonic Imputation Versus Time Dummy Hedonic Indexes. In E. W. Diewert, J. S. Greenlees, & C. R. Hulten, *Price index concepts and measurement* (pp. 161-196). Chicago: University of Chicago Press.

Diewert, W. E. (2003). Hedonic Regressions: A Review of Some Unresolved Issues. Paris: In 7th Meeting of the Ottawa Group.

Efron, B., & Tibshirani, R. (1994). *An Introduction to the Bootstrap.* CRC Press.

Eurostat. (2013). *ec.europa.eu/eurostat.* Opgeroepen op Januari 2016, van http://ec.europa.eu/eurostat/documents/3859598/5925925/KS-RA-12-022-EN.PDF

Francke, M. K., & van de Minne, A. M. (2016). Land, Structure and Depreciation. *Real Estate Economics*, 415-451.

Girden, E. R. (2001). *Evaluating Research Articles from Start to Finish.* Thousand Oaks, California: Sage Publications.

Glaeser, E. L., Kolko, J., & Saiz, A. (2001). Consumer city. *Journal of Economic Geography*, 27-50.

Goodman, A. C., & Thibodeau, T. G. (1995). Age-Related Heteroskedasticity in Hedonic House Price Equations. *Journal of Housing Research*, 25-36.

Gorsuch, R. (1983). *Factor analysis.* Hillsdale: NJ: L. Erlbaum Associates.

Griliches, Z. (1971). Introduction: Hedonic Price Indexes Revisited. In Z. Griliches, *Price Indexes and Quality Change* (pp. 3-15). Cambridge MA: Harvard University Press.

Harris, C. D., & Ullman, E. L. (1945). The Nature of Cities. *The Annals of the American Academy of Political and Social Science*, 7-17.

Hershberger, S. L. (2005). Factor scores. In B. S. Everitt, & H. D. C., *Encyclopeida of Statistics in Behavioral Science* (pp. 636-644). New York: John Wiley.

Hill, R. J. (2013). Hedonic Price Indexes for Residential Housing: A Survey, Evaluation and Taxonomy. *Journal of Economic Surveys*, 879-914.

Hill, R., & Melser, D. (2008). Hedonic Imputation and the Price Index Problem: An Application to Housing. *Economic Inquiry*, 593-609.

Johnston, J. (1972). *Econometric Methods.* New York: McGraw-Hill.

Kaiser, H. F., & Cerny, C. A. (1977). A Study of a Measure of Sampling Adequacy for Factor-Analytic Correlation Matrices. *Multivariate Behavioral Research*, 43-47.

Kempf, S. (2015). *Development of Hedonic Office Rent Indices: Examples for German Metropolitan areas.* Springer.

Koster, H. R., & Rouwendal, J. (2012). The Impact of Mixed Land Use on . *Journal of Regional Science*, 733-761.

Krumm, R. J. (1980). Neighborhood amenities: An economic analysis. *Journal of Urban Economics*, 208-224.

Lancaster, K. J. (1966). A new approach to consumer theory. *Journal of Political Economy*, 132-157.

Li, M. M., & Brown, H. J. (1980). Micro-Neighborhood Externalities and Hedonic Housing Prices. *Land Economics*, 125-141.

Martins-Filho, C., & Bin, O. (2005). Estimation of hedonic price functions via additive nonparametric regression. *Empirical Economics*, 93-114.

Pan, Y., & Jackson, R. (2008). Ethnic Difference in the Relationship between Acute Inflammation and Serum Ferritin in US Adult Males. *Epidemiology & Infection*, 421-431.

Ramsey, J. B. (1969). Tests for Specification Errors in Classical Linear Least Squares Regression Analysis. *Journal of the Royal Statistical Society, Series B.*, 350-371.

Rosen, S. (1974). Hedonic prices and implicit markets: product differentiation in pure competition. *Journal of Political Economy*, 34-55.

Sheppard, S. (1999). Hedonic Analysis of Housing Markets. *Handbook of Regional and Urban Economics*, 1595-1635.

Silver, M., & Heravi, S. (2004). The Difference between Hedonic Imputation Indexes and time dummy hedonic indexes for desktop PCs. *CRIW Conference on Price Index Concepts and Measurement.* Vancouver.

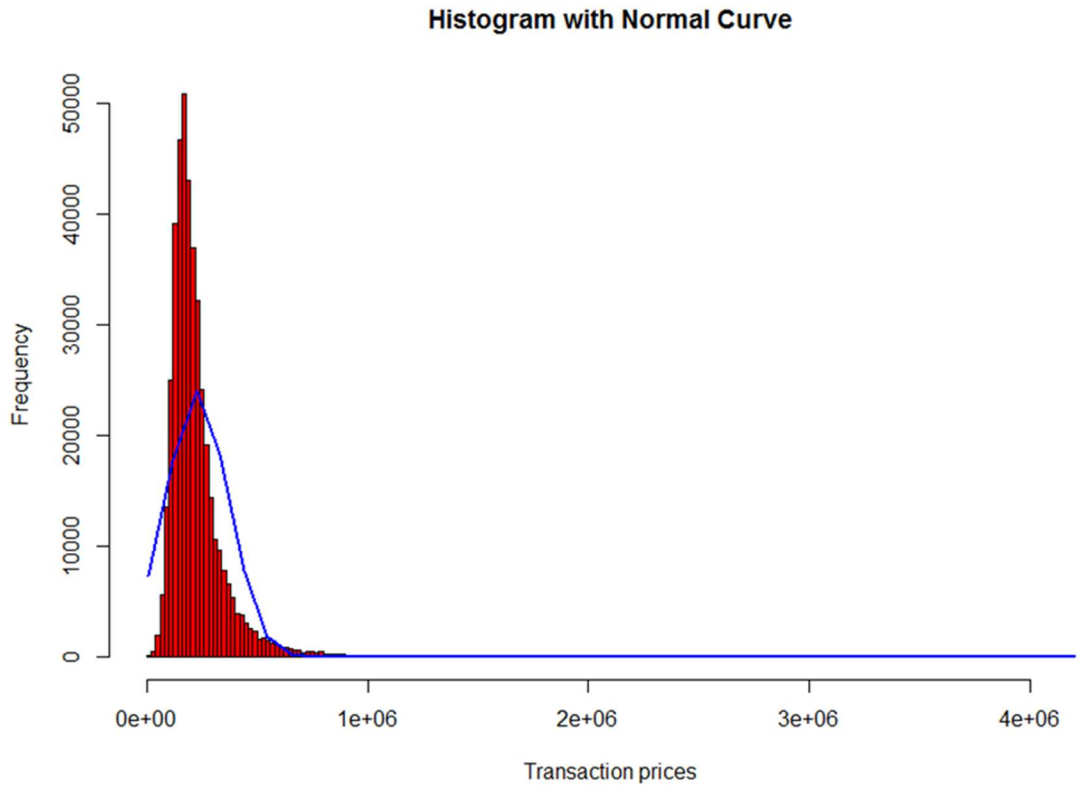
Thurstone, L. L. (1935). *The Vectors of Mind.* Chicago: University of Chicage Press.

Van der Grient, H., & De Haan, J. (2008). *Index Cijfers.* Den Haag/Heerlen: Centraal Bureau voor de Statistiek. Opgehaald van Centraal Bureau voor de Statistiek.

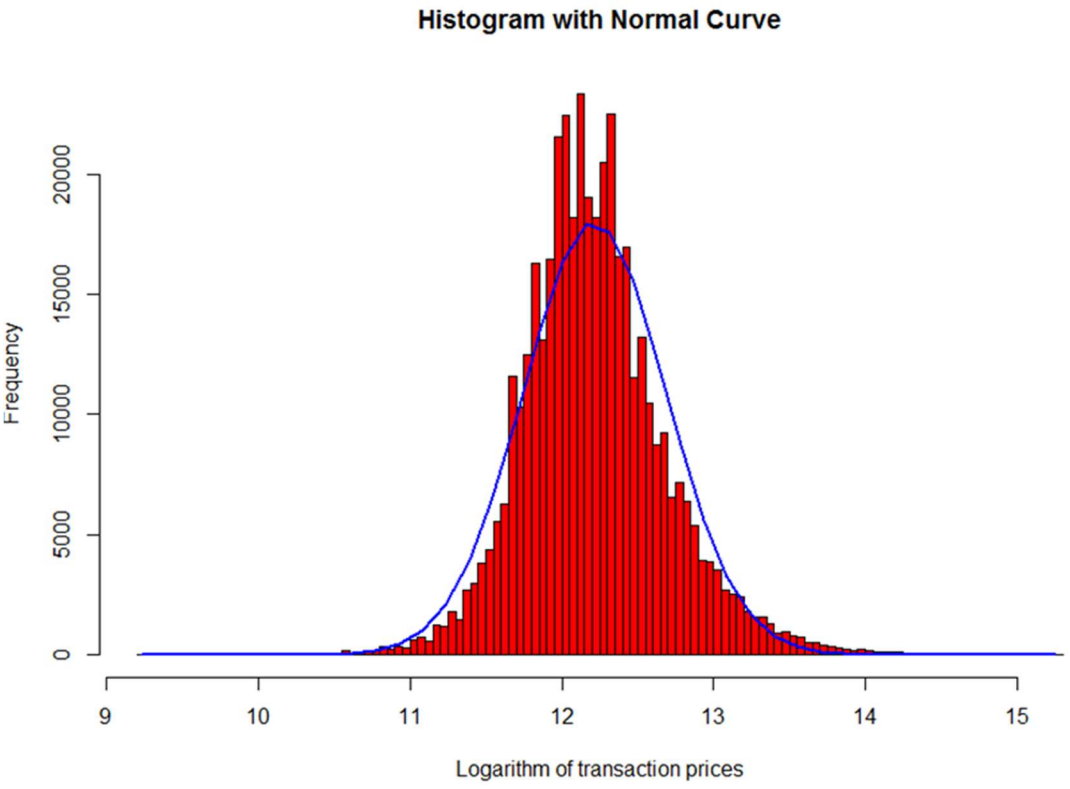
White, H. (1980). A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity. *Econometrica*, 817–838.

8. Appendix

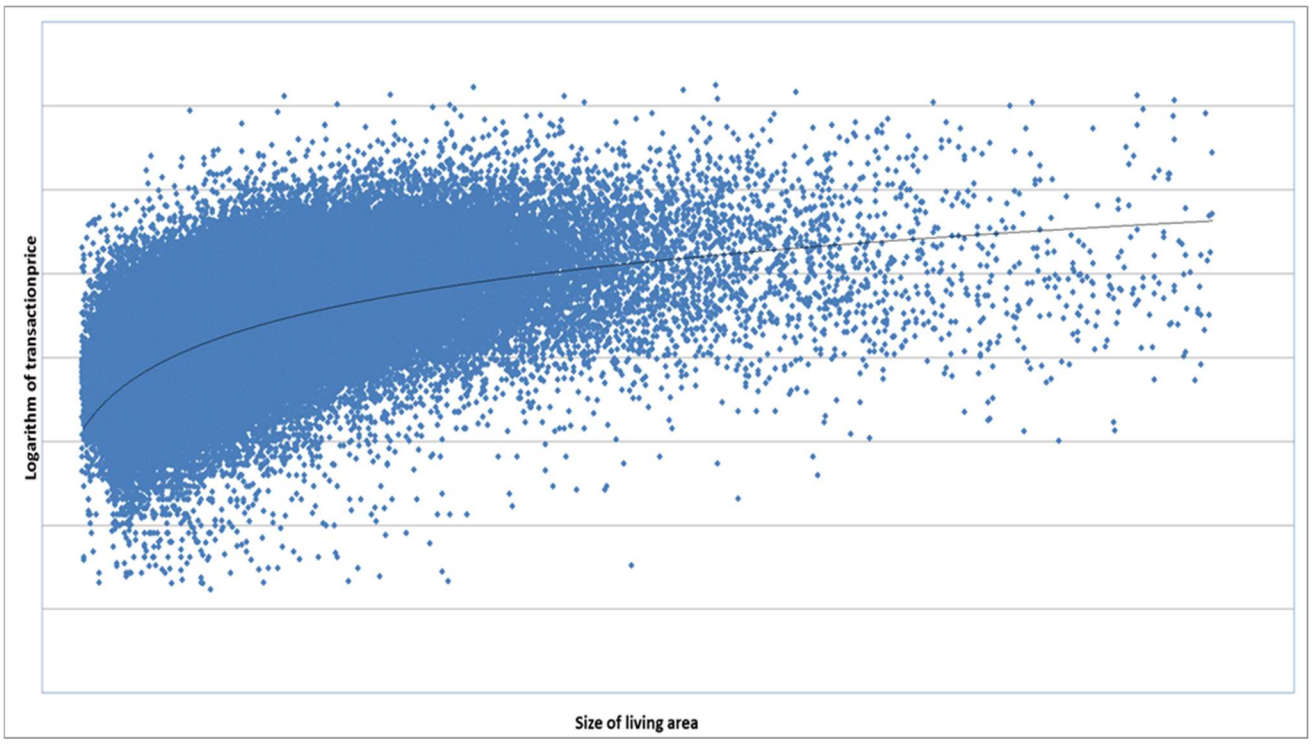
Appendix 1: The spread of transaction prices



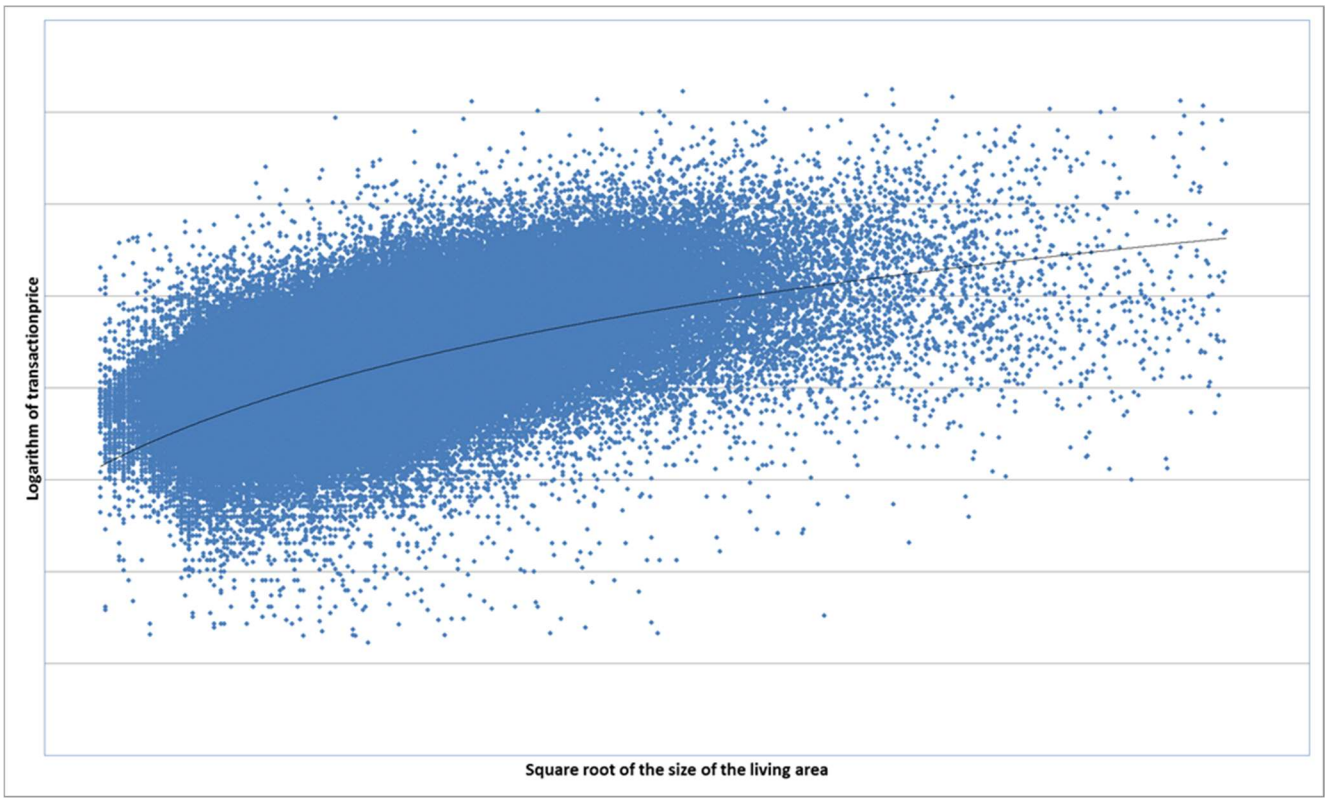
Appendix 2: The spread of the logarithm of transaction prices



Appendix 3: Scatterplot of logarithm of transaction prices and the size of the living area



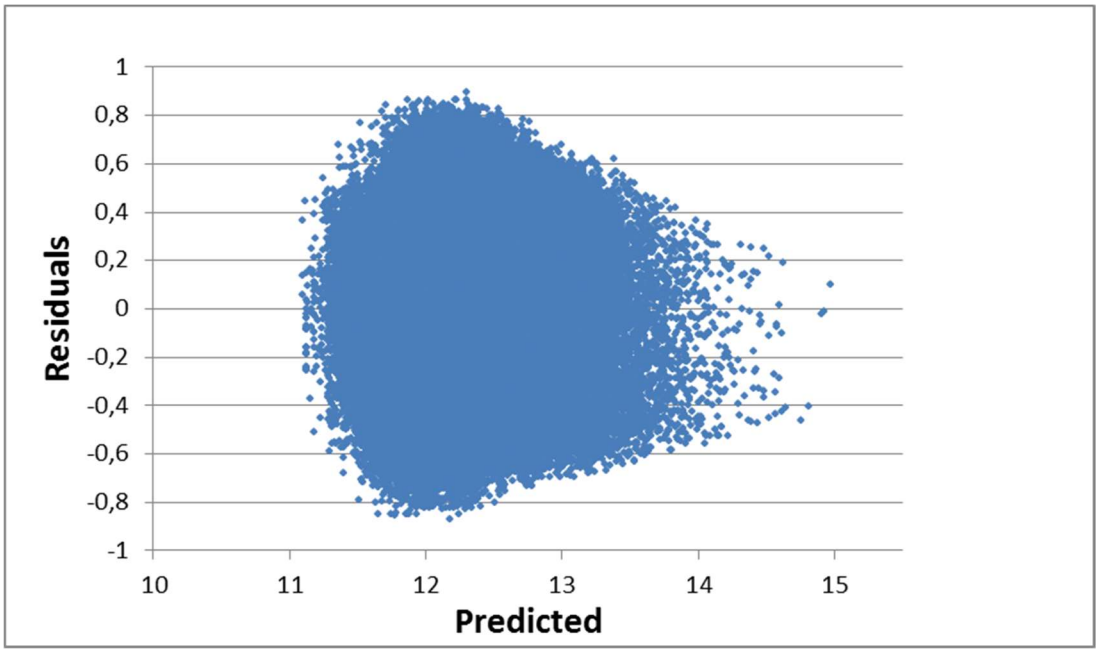
Appendix 4: scatterplot of logarithm of transaction prices and square root of the size of the living area



Appendix 5: OLS regression hedonic price model 1 with and without outliers

Variables	With outliers	Without outliers
√(living area)	YES***	YES***
Structure type	YES***	YES***
Province name	YES***	YES***
Building age cohorts	YES***	YES***
Adjusted-R^2	0.5550	0.6651
P-values: ~ < 0.05, * < 0.01, ** < 0.001, *** < 0.000		

Appendix 6: Residual scatterplot hedonic price model 1



Appendix 7: Breusch-Pagan test hedonic price model 1

BP	Degrees of freedom	P-value
31812	24	0.000

Appendix 8: Hedonic price model 1 with White-Huber standard errors

Variables	Coefficient estimates
√(living area)	0.1343***
Detached	0.2776***
End-terrace-house	0.0698***
Mid-terrace-house	0.0307***
Semi-detached	0.1690***
Flevoland	0.0638***
Friesland	-0.0487***
Gelderland	0.2284***
Groningen	-0.0148***
Limburg	0.0217***
Noord-Brabant	0.2664***
Noord-Holland	0.4742***
Overijsel	0.1085***
Utrecht	0.4429***
Zeeland	0.0826***
Zuid-Holland	0.3096***
Voor 1905	0.0950***
1905-1930	-0.0316***
1931-1944	-0.0368***
1945-1959	-0.1755***
1960-1970	-0.2128***
1971-1980	-0.1984***
1981-1990	-0.1134***
Na 2001	0.0277***
Adjusted-R^2	0.6651
Reference categories: Apartment, Drenthe and 1990-2001	
Heteroscedasticity-consistent standard errors	
P-values: ~ < 0.05, * < 0.01, ** < 0.001, *** < 0.000	

Appendix 9: Comparison of different types of model specification for hedonic price model 1

Variables	Square root model	quadratic model	quadratic model + interaction
√(living area)	0.1343***	NO	NO
Living area	NO	0.0086***	YES***
(Living area)^2	NO	-9.028^{e}-06***	YES***
Living area*structure type	NO	NO	YES***
Structure type	YES***	YES***	YES***
Province name	YES***	YES***	YES***
Building age cohorts	YES***	YES***	YES***
Adjusted-R^2	0.6651	0.6656	0.6664
Ramsey-reset test	0.000	0.000	0.000
Log-likelihood test	Reference model	0.000	0.000
P-values: ~ < 0.05, * < 0.01, ** < 0.001, *** < 0.000			

Appendix 10: Kaiser-Meyer-Olkin test

	Overall measured sampling adequacy
Closeness variables	0.94
Density variables	0.93

Appendix 11: Bartlett's test of sphericity

	Degrees of freedom	P-value
Closeness variables	378	0.000
Density variables	171	0.000

Appendix 12: Factor analysis 'closeness' with factor loadings and uniqueness of variables

cut-off value = 0.3

Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Uniqueness
Distance to library	0.335				0.399		0.619
Distance to swimming pool					0.546		0.511
Distance to artificial ice-skating track		0.800					0.513
Distance to pop stage		0.508		0.344			0.540
Distance to cinema		0.357		0.524			0.391
Distance to sauna		0.565					0.616
Distance to tanning salon				0.485			0.513
Distance to attraction		0.581					0.597
Distance to large supermarkets	0.885						0.315
Distance to stores for daily groceries	0.905						0.315
Distance to department store					0.751		0.316
Distance to cafe	0.537						0.670
Distance to cafeteria	0.685						0.411
Distance to restaurant	0.649						0.552
Distance to hotel							0.760
Distance to elementary school	0.487						0.666
Distance to secondary- or high school							0.014
Distance to preparatory vocational education school			1.007				0.060
Distance to higher general secondary education schools and preparatory academic schools			0.987				0.268
Distance to highway ramp			0.503				0.959
Distance to train station		0.738					0.446
Distance to public transport transfer station		0.882					0.230
Distance to kindergarten						0.551	0.418
Distance to out-of-school care centre						1.001	0.037
Distance to general doctor's practice	0.867						0.336
Distance to pharmacy	0.755						0.420
Distance to hospitals with advisory doctor's practice				0.660			0.346
Distance to hospitals without advisory doctor's practice				0.925			0.222

Appendix 13: Factor analysis 'density' with factor loadings and uniqueness of variables

cut-off value = 0.3

Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Uniqueness
Number of pop culture stages		0.950				0.005
Number of cinemas		0.761				0.156
Number of attractions				0.588		0.592
Number of large supermarkets	0.383		0.604			0.257
Number of stores for daily groceries	0.630		0.399			0.082
Number of department stores		0.462		0.441		0.129
Number of cafes	0.988					0.098
Number of cafeterias	1.008					0.037
Number of restaurants	0.974					0.076
Number of hotels		1.103				0.058
Number of elementary schools			0.970			0.417
Number of secondary/ or highschools					0.928	0.005
Number of preparatory vocational education schools					0.927	0.089
Number of higher general secondary education schools and preparatory academic school					0.846	0.108
Number of kindergartens			0.592			0.280
Number of out-of-school care centres			0.817			0.398
Number of general doctor's practices			0.677			0.333
hospitals with advisory doctor's practice				1.032		0.105
hospitals without advisory doctor's practice				0.927		0.103

Appendix 14: Eigenvalues and proportions of variance when using a certain number of factors

Closeness	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
Eigenvalue	4.626	3.076	2.331	2.192	1.548	1.463
Proportion variance	0.165	0.110	0.083	0.078	0.055	0.052

Closeness	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6	Factor 7
Eigenvalue	4.542	2.850	2.219	2.137	1.678	1.512	0.941
Proportion variance	0.162	0.102	0.079	0.076	0.060	0.054	0.034

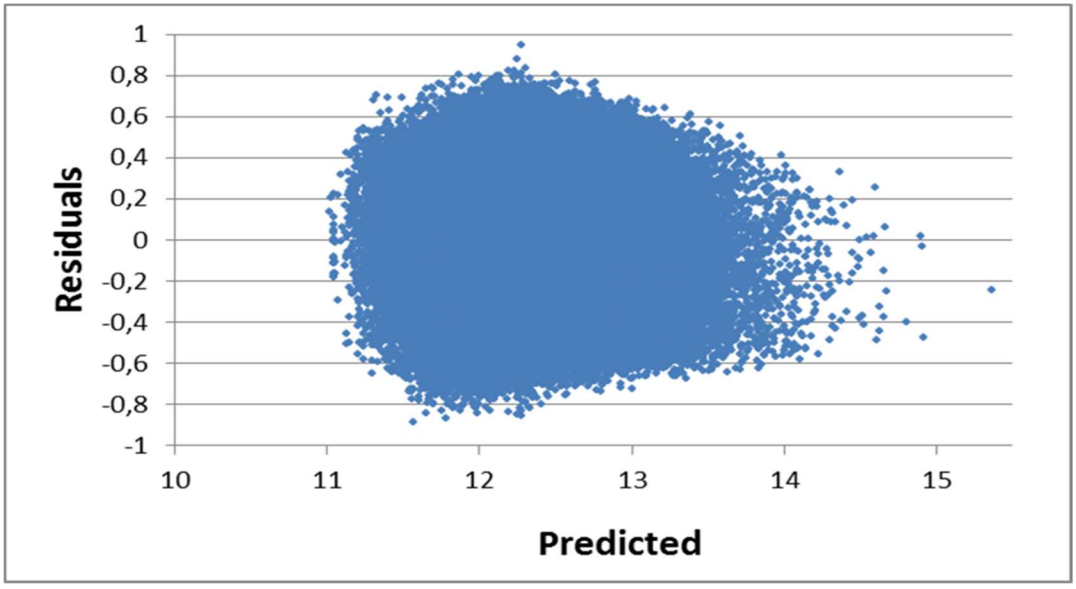
Density	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Eigenvalue	3.553	3.035	2.967	2.581	2.477
Proportion variance	0.187	0.160	0.156	0.136	0.130

Density	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
Eigenvalue	3.158	3.071	2.701	2.391	2.332	0.796
Proportion variance	0.166	0.162	0.142	0.126	0.123	0.042

Appendix 15: OLS regression for hedonic price model 2

Variables	Without outliers
√(living area)	YES***
Structure type	YES***
Province name	YES***
Building age cohorts	YES***
Factor scores	YES***
Adjusted-R^2	0.7074
P-values: ~ < 0.05, * < 0.01, ** < 0.001, *** < 0.000	

Appendix 16: Residual scatterplot hedonic price model 2



Appendix 17: Breusch-Pagan test hedonic price model 2

BP	Degrees of freedom	P-value
23944	35	0.000

Appendix 18: Hedonic price model 2 with White-Huber standard errors

Variable	Coefficient estimates
√ (living area)	0.1363***
Detached	0.3988***
End-terrace-house	0.1613***
Mid-terrace-house	0.1187***
Semi-detached	0.2664***
Flevoland	0.0656***
Friesland	-0.0433***
Gelderland	0.2160***
Groningen	-0.0183***
Limburg	-0.0013
Noord-Brabant	0.2468***
Noord-Holland	0.3823***
Overijsel	0.0849***
Utrecht	0.4307***
Zeeland	0.1861***
Zuid-Holland	0.3182***
Voor 1905	-0.0544***
1905-1930	-0.1110***
1931-1944	-0.0977***
1945-1959	-0.1821***
1960-1970	-0.1988***
1971-1980	-0.1918***
1981-1990	-0.1131***
Na 2001	0.0383***
Closeness to district centre	0.0030***
Closeness to city centre	-0.0289***
Closeness to secondary education centre	-0.0042***
Closeness to healthcare & cultural centres	0.0095***
Closeness to city-edge centre	-0.0136***
Closeness to child-care centre	-0.0163***
Presence of district centre	0.0212***
Presence of city centre	0.1151***
Presence of neighborhood centre	-0.0230***
Presence of city-edge centre	-0.0228***
Presence of secondary education centre	0.0117***
Adjusted-R^2	0.7074
Reference categories: Apartment, Drenthe and 1990-2001	
Heteroscedasticity-consistent standard errors	
P-values: ~ < 0.05, * < 0.01, ** < 0.001, *** < 0.000	

Appendix 19: Proxy variables and corresponding factors

Proxy variables	Factors
Distance to stores for daily groceries	Closeness to district centre
Distance to public transport transfer station	Closeness to city centre
Distance to preparatory vocational education school	Closeness to secondary education
Distance to hospitals without advisory doctor's practice	Closeness to healthcare
Distance to department store	Closeness to city-edge centre
Distance to out-of-school care centre	Closeness to childcare
Presence of cafeteria 1km	Presence of district centre
Presence of hotel 5km	Presence of city centre
Presence of elementary school 1km	Presence of neighborhood centre
Presence of hospitals with advisory doctor's practice 5km	Presence of city-edge centre
Presence of preparatory vocational education school 3km	Presence of education

Appendix 20: Comparison between hedonic price model 2 with factor scores and with proxy variables

Variables	Factor scores model	Proxy variables model
√ (living area)	YES***	YES***
Structure type	YES***	YES***
Province name	YES***	YES***
Building age cohorts	YES***	YES***
Closeness to district centre	0.0030***	-
Closeness to city centre	-0.0289***	-
Closeness to secondary education centre	-0.0042***	-
Closeness to healthcare & cultural centres	0.0095***	-
Closeness to city-edge centre	-0.0136***	-
Closeness to child-care	-0.0163***	-
Presence of district centre	0.0212***	-
Presence of city centre	0.1151***	-
Presence of neighborhood centre	-0.0230***	-
Presence of city-edge centre	-0.0228***	-
Presence of education secondary education centre	0.0117***	-
Distance to stores for daily groceries	-	2.820^{e}-06***
Distance to public transport transfer station	-	-2.969^{e}-06***
Distance to preparatory vocational education school	-	-1.395^{e}-06***
Distance to hospitals without advisory doctor's practice	-	1.763^{e}-06***
Distance to department store	-	-8.141^{e}-06***
Distance to out-of-school care centre	-	-2.038^{e}-05***
Presence of cafeteria 1km	-	8.812^{e}-04***
Presence of hotel 5km	-	1.675^{e}-03***
Presence of elementary school 1km	-	-1.260^{e}-02***
Presence of hospitals with advisory doctor's practice 5km	-	3.704^{e}-03***
Presence of preparatory vocational education school 3km	-	1.325^{e}-03***
Adjusted-R^2	0.7074	0.7112
P-values: ~ < 0.05, * < 0.01, ** < 0.001, *** < 0.000		

Appendix 21: Correlation matrix of continuous variables hedonic price model 2

	Size_living_area	Density_factor1	Density_factor2	Density_factor3	Density_factor4	Density_factor5	Closeness_factor1	Closeness_factor2	Closeness_factor3	Closeness_factor4	Closeness_factor5	Closeness_factor6
Size_living_area	1.0000000	-0.2293080	-0.2725369	-0.3101805	-0.2609479	-0.2728778	0.2805182	0.1763880	0.1721685	0.1993789	0.1920832	0.1635817
Density_factor1	-0.2293080	1.0000000	0.6190073	0.6532405	0.4754918	0.6045829	-0.3622341	-0.2939552	-0.2225589	-0.2956119	-0.3236780	-0.1664431
Density_factor2	-0.2725369	0.6190073	1.0000000	0.6715147	0.6492692	0.6774616	-0.2888184	-0.3573631	-0.2486728	-0.3497371	-0.2541064	-0.1694923
Density_factor3	-0.3101805	0.6532405	0.6715147	1.0000000	0.5995546	0.6507792	-0.5410917	-0.3679386	-0.3380531	-0.3760610	-0.4057113	-0.3567184
Density_factor4	-0.2609479	0.4754918	0.6492692	0.5995546	1.0000000	0.7302848	-0.3614459	-0.4866207	-0.4189231	-0.5944437	-0.3936916	-0.2502194
Density_factor5	-0.2728778	0.6045829	0.6774616	0.6507792	0.7302848	1.0000000	-0.4005490	-0.4523587	-0.4764962	-0.5189019	-0.4110456	-0.2446924
Closeness_factor1	0.2805182	-0.3622341	-0.2888184	-0.5410917	-0.3614459	-0.4005490	1.0000000	0.3589182	0.4144946	0.3883443	0.5595318	0.5819039
Closeness_factor2	0.1763880	-0.2939552	-0.3573631	-0.3679386	-0.4866207	-0.4523587	0.3589182	1.0000000	0.4527407	0.5664132	0.4728605	0.2972576
Closeness_factor3	0.1721685	-0.2225589	-0.2486728	-0.3380531	-0.4189231	-0.4764962	0.4144946	0.4527407	1.0000000	0.5692980	0.6668562	0.3913673
Closeness_factor4	0.1993789	-0.2956119	-0.3497371	-0.3760610	-0.5944437	-0.5189019	0.3883443	0.5664132	0.5692980	1.0000000	0.5363597	0.3493423
Closeness_factor5	0.1920832	-0.3236780	-0.2541064	-0.4057113	-0.3936916	-0.4110456	0.5595318	0.4728605	0.6668562	0.5363597	1.0000000	0.4783879
Closeness_factor6	0.1635817	-0.1664431	-0.1694923	-0.3567184	-0.2502194	-0.2446924	0.5819039	0.2972576	0.3913673	0.3493423	0.4783879	1.0000000

Appendix 22: Variance inflation factor test

Variables	GVIF	GVIF^(1/(2*df))	df
√ (living area)	1.82	1.35	1
Structure type	2.79	1.14	4
Province name	3.12	1.05	11
Building age cohorts	1.85	1.04	8
Closeness to district centre	2.28	1.53	1
Closeness to city centre	2.27	1.78	1
Closeness to secondary education centre	2.26	1.75	1
Closeness to healthcare & cultural centres	2.53	1.84	1
Closeness to city-edge centre	2.63	1.83	1
Closeness to child-care	1.71	1.51	1
Presence of district centre	2.35	1.51	1
Presence of city centre	3.17	1.50	1
Presence of neighborhood centre	3.07	1.59	1
Presence of city-edge centre	3.38	1.62	1
Presence of education secondary education centre	3.34	1.31	1

Appendix 23: Price index numbers and margins of price indexes produced with hedonic price model 1

Quarter	HTD_PI	TD_Margin	HDIL_PI	HDIL_Margin	HDIP_PI	HDIP_Margin	HDIF_PI	HDIF_Margin	PBK_SPAR
201201	100	-	100	-	100	-	100	-	100
201202	98,08	0,989	98,03	0,965	98,00	1,002	98,01	0,985	98,10
201203	93,78	1,039	93,84	1,035	93,73	1,029	93,79	1,078	94,19
201204	93,53	0,929	93,47	0,949	93,34	0,914	93,41	0,935	93,55
201301	91,06	0,964	91,33	1,001	91,02	0,982	91,17	0,958	91,44
201302	89,71	0,959	89,98	1,009	89,66	0,999	89,82	1,003	89,54
201303	90,37	0,912	90,54	0,956	90,29	0,934	90,42	0,913	89,96
201304	90,52	0,878	90,61	0,864	90,44	0,943	90,52	0,876	89,64
201401	93,50	1,032	93,50	1,032	93,50	1,032	93,50	1,014	90,06
201402	93,77	0,981	93,75	0,973	93,78	1,002	93,76	0,975	90,70
201403	94,30	0,928	94,24	0,934	94,33	0,972	94,29	0,927	91,54
201404	93,86	0,904	93,77	0,896	93,92	0,941	93,84	0,869	91,54
201501	94,77	0,951	94,81	1,006	94,81	0,991	94,81	0,984	92,28
201502	95,14	0,952	95,20	0,937	95,17	0,941	95,18	0,925	92,92
201503	95,91	0,890	95,92	0,917	95,94	0,927	95,93	0,911	94,19
201504	95,71	0,888	95,70	0,913	95,77	0,919	95,74	0,909	94,82

Appendix 24: Price index numbers and margins of price indexes produced with hedonic price model 2

Quarter	HTD_PI	TD_Margin	HDIL_PI	HDIL_Margin	HDIP_PI	HDIP_Margin	HDIF_PI	HDIF_Margin	PBK_SPAR
201201	100	0	100	0	100	0	100	0	100
201202	98,36	0,965	98,26	0,980	98,23	0,938	98,25	0,949	98,10
201203	93,90	0,995	93,93	1,012	93,81	1,008	93,87	0,988	94,19
201204	93,58	0,915	93,51	0,922	93,38	0,864	93,44	0,892	93,55
201301	91,16	0,940	91,44	0,975	91,09	0,929	91,26	0,890	91,44
201302	89,78	0,924	90,00	0,962	89,71	0,944	89,86	0,896	89,54
201303	90,34	0,899	90,53	0,913	90,27	0,902	90,40	0,883	89,96
201304	90,64	0,866	90,75	0,873	90,56	0,868	90,66	0,856	89,64
201401	93,18	0,986	93,18	0,975	93,18	0,955	93,18	0,959	90,06
201402	93,52	0,970	93,52	0,932	93,52	0,930	93,52	0,937	90,70
201403	93,85	0,921	93,87	0,922	93,89	0,858	93,88	0,895	91,54
201404	93,48	0,852	93,44	0,878	93,54	0,842	93,49	0,860	91,54
201501	94,97	0,943	95,02	0,988	94,96	0,945	94,99	0,928	92,28
201502	95,24	0,921	95,33	0,928	95,22	0,905	95,27	0,876	92,92
201503	95,90	0,862	95,93	0,914	95,89	0,872	95,91	0,853	94,19
201504	95,95	0,899	95,97	0,917	95,97	0,878	95,97	0,856	94,82

Appendix 25: Precision of price indexes produced with hedonic price model 1

Quarter	HTD	HDIL	HDIP	HDIF
201201	-	-	-	-
201202	1,008	0,984	1,022	1,005
201203	1,108	1,103	1,098	1,149
201204	0,993	1,015	0,979	1,001
201301	1,059	1,096	1,079	1,051
201302	1,069	1,121	1,114	1,117
201303	1,009	1,056	1,034	1,01
201304	0,97	0,954	1,043	0,968
201401	1,104	1,104	1,104	1,084
201402	1,046	1,038	1,068	1,04
201403	0,984	0,991	1,03	0,983
201404	0,963	0,956	1,002	0,926
201501	1,003	1,061	1,045	1,038
201502	1,001	0,984	0,989	0,972
201503	0,928	0,956	0,966	0,95
201504	0,928	0,954	0,96	0,949
Average	1,012	1,025	1,036	1,016

Appendix 26: Precision of price indexes produced with hedonic price model 1

Quarter	HTD_Precision	HDIL_Precision	HDIP_Precision	HDIF_Precision
201201	-	-	-	-
201202	0,981	0,997	0,955	0,966
201203	1,06	1,077	1,075	1,053
201204	0,978	0,986	0,925	0,955
201301	1,031	1,066	1,02	0,975
201302	1,029	1,069	1,052	0,997
201303	0,995	1,009	0,999	0,977
201304	0,955	0,962	0,958	0,944
201401	1,058	1,046	1,025	1,029
201402	1,037	0,997	0,994	1,002
201403	0,981	0,982	0,914	0,953
201404	0,911	0,94	0,9	0,92
201501	0,993	1,04	0,995	0,977
201502	0,967	0,973	0,95	0,919
201503	0,899	0,953	0,909	0,889
201504	0,937	0,956	0,915	0,892
Average	0,988	1,004	0,973	0,963