The Hedonic Analysis : an Application to the Belgian Housing Market

Jury :

Promoteur : Joseph Tharakan

Lecteurs : Lionel Artige
Bernard Lejeune

Mémoire présenté par **Sébastien Broos**

En vue de l'obtention du diplôme de Master en Sciences Economiques, orientation générale, à finalité approfondie

Année académique 2012/2013

Executive Summary

The goal of this master's thesis is to build a hedonic model to analyse the Belgian housing market and in particular, the market for apartments. We will start by describing what the hedonic theory is and why it is useful. We will also present traditional price indices and the specificities of the housing market. A hedonic price index will be built through ordinary least squares regressions. With its help, we will see that prices of apartments, between 2011 and 2012 and at constant quality, have increased in some regions in Belgium. We will compare our results with another hedonic study about Belgium and with the results from traditional price indices. This comparison will show some conclusions that can be drawn thanks to hedonic price indices. It will also show that they are absolutely needed if economists want to understand the real estate market rigorously.

Résumé

Le but de ce mémoire est de construire un modèle hédonique afin d'analyser le marché immobilier Belge et en particuler le marché des appartements. Nous commencerons par décrire ce qu'est la théorie hedonique et pourquoi elle est intéressante. Nous présenterons aussi des indices de prix traditionnels et les spécificités du marché immobilier. Nous construirons un indice de prix hédonique grâce à des régressions des moindres carrés ordinaires. Grâce à cet indice, nous verrons que les prix des appartements ont augmenté dans certaines régions belges et cela, à qualité constante et entre 2011 et 2012. Nous comparerons nos résultats avec ceux d'une autre étude hédonique sur la Belgique et avec les résultats des indices de prix traditionnels. Cette comparaison montrera certaines conclusions que les indices de prix hédoniques peuvent aider à formuler et à quel point ils sont nécessaires pour une analyse rigoureuse du marché immobilier.

Remerciements

Tout d'abord, je souhaiterais remercier Monsieur Tharakan pour le temps qu'il a passé à m'aiguiller dans les méandres de la pensée économique. Le processus fut parfois compliqué et frustrant, mais au final, formateur et fructueux. Je le remercie aussi pour son attention toujours prompte et constante.

Mes remerciements vont aussi à Monsieur Lejeune et Monsieur Artige. Le premier pour les nombreux conseils méthodologiques et économétriques qu'il m'a prodigués. Le second pour m'avoir orienté lors du choix du sujet de ce mémoire et pour ses conseils tout court.

Finalement, ce mémoire n'aurait pu être réalisé sans les données de la Fédération Royale du Notariat Belge. Nous les remercions, en particulier Madame Nathalie Carion, qui fut toujours prête à répondre à nos questions.

"The curious task of economics is to demonstrate to men how little they really know about what they imagine they can design."

Friedrich von Hayek

"Everything is related to everything else, but near things are more related than distant things."

Waldo Tobler's First Law of Geography

Contents

List of Figures

List of Tables

Part I

Introduction

In 1977, an Apple II computer cost around 1300 dollars. Today[1], a basic iMac also costs 1300 dollars. Yet, absolutely no one will conclude that the prices of computers have been constant. What happened ? There have been changes in quality. While this paper takes today 3 seconds to be compiled by LaTeX, it would have taken hours or days in 1977 (if possible at all). Most of what was barely imaginable thirty years ago can now be realized in a matter of seconds. Our master's thesis will focus on this problem, namely, how should price indices take quality into account ?

More precisely, we will focus on the sales of apartments in Belgium. Indeed, what we said about computers is also true for many other goods, especially housing. We will try, with the help of the hedonic analysis, to see the impact of quality changes on prices of apartments. The two main questions that we will try to answer are 1) Have prices of apartments changed between 2011 and 2012, at given quality ? 2) How can we use this knowledge to understand the statement of traditional price indices "prices of apartments have changed in Belgium during these years" ? In other words, we will try to calculate the price change of Belgian apartments sold between 2011 and 2012 at constant quality and compare the results with traditional price indices.

We will begin our analysis with theory. We will first explain what the hedonic hypothesis is and the theoretical model behind it. One of the main uses of hedonic analysis, demand identification, will then be rapidly covered. Before entering the realm of hedonic indices, we will examine traditional price indices, their shortcomings and the reasons why better ones are needed for the real estate market. A few empirical applications will also be reviewed. We will continue the theoretical part with a section about the housing market and its specificities. We will conclude with a short review of spatial econometrics, a new approach to take into account location in economic models.

[1] We obviously omit one important element : inflation.

We will then turn to the core of our master's thesis and to our two empirical questions. The available data, as well as its advantages and its shortcomings, will be examined. Finally, we will try to estimate a hedonic equation and, through it, a hedonic price index for different geographical areas of Belgium. Last but not least, we will attempt to compare our hedonic price index with traditional Belgian price indices for the real estate market.

Part II

Theoretical Analysis

1 The Hedonic Theory

The goal of this this section is to explain what the hedonic theory is. The explanation will be divided amongst three parts : the origins of hedonic analysis, its uses and the formal model of Rosen.

1.1 The Origins of Hedonic Theory

For a long time, the first contribution to the hedonic literature was thought to be in 1939 by Court[2] until, in 1999, Colwell and Dilmore discovered an earlier paper by Haas (1922). The historical controversy does not interest us but it is interesting to note that while Court coined the term hedonic and was interested in the car industry, Haas was the first to apply the method to the housing market. Griliches (1961) introduced it into mainstream economics with his study of the automotive industry.

Hedonic theory is an answer to the heterogeneity problem. Economists have known for a long time what the difference between butter and margarine is but they could not explain why a castle and an apartment were very different and still were both called "housing goods". Lancaster's (1966) approach to the problem was that agents do not derive their utility from the goods themselves but from the characteristics contained in the goods. In Lancaster's words :

> "Utility or preference orderings are assumed to rank collections of characteristics and only to rank collections of goods indirectly through the characteristics that they possess."

For example, when someone buys a house, he buys it because he derives utility from the view, from the location, from the wooden floor, etc. Before Lancaster's theory, a popular idea (Muth, 1969) was to model goods as having different quantities of a same intrinsic homogeneous non-observable quality. A house worth a million euros simply contained more "housing" than a house worth half that amount. With Lancaster's approach, we can on the other hand say that one house has more rooms than the other, that is, it contains more of the characteristic "room". Both houses are representing the same good, but quantities of

[2] See any pre-1999 study for example.

characteristics vary. Although it was not always explicitly mentioned, this approach was always implicitly present in earlier hedonic studies.

From this stems the hedonic approach. If goods are valued for their characteristics then it should be possible to decompose the price of a good in terms of the prices of its features. There is in fact an implicit market for characteristics and an explicit market for goods.[3] Hence, hedonic prices are the prices of the characteristics on the implicit markets (Rosen, 1974).

1.2 The Uses of the Hedonic Method

Through the recovery of the implicit prices, the hedonic method can be used for two main purposes. Each use will have its own section but let us explain them briefly here.

First, it allows for the construction of price indices which take into account quality. Indeed, if the price of a good increases, is it because it has improved, that is, quantities of its characteristics have also increased, or because of something else, an increase in demand for instance ? If we know the implicit prices and quantities of the characteristics contained in the good, we can answer that question by building price indices at constant quality.

Another useful feature of the hedonic method is the identification of the structure of the demand and the supply functions for characteristics or in other words, the identification of supply and demand for each feature of a good. If goods are valued only for the levels of characteristics embedded in them, it would be interesting to know why and how these levels change. A typical example is the impact of pollution, defined broadly, on the valuation of houses. Think of the payment of damages around airports for instance : what is the impact of noise pollution on the price of a house and how much should air companies compensate "victims" ? This has very wide policy applications.

[3] Sheppard (1999) defines an implicit market as "the process of production, exchange, and consumptions of commodities that are primarily (perhaps exclusively) traded in "bundles"" and the explicit market as "the market for bundles, with prices and transactions observed."

1.3 Rosen's Model

How can we formalize what we have just said about goods being nothing more that bundles of characteristics ? Rosen was the first to really delve into what hedonic prices are and to deduce a theoretical model from them. Let us follow the approach he developed in his seminal paper (Rosen, 1974).

Let $\mathbf{z} = (z_1, z_2, ..., z_n)$ be a vector[4] of characteristics where z_i is the quantity of the ith characteristic contained in a certain good. Goods are completely defined by the n characteristics of vector $\mathbf{z}$. The levels of the characteristics are objectively observed but subjectively valued: different consumers know exactly how much of a given feature a good contains but can value that level differently. Sellers offer all possible combinations of bundles of qualities possible. This is a very important assumption since it allows the use of differential calculus on $\mathbf{z}$ and hence of marginal analysis.

We consider a multidimensional space where coordinates are represented by the vector $\mathbf{z}$. A price $p(\mathbf{z}) = p(z_1, z_2, ..., z_n)$ is associated with each point. Hence, to each set of characteristics corresponds a price. We are in perfect competition and the price is a given for both buyers and sellers. They thus locate in the space using this information. The price is such that, at equilibrium, buyers and sellers locate at the same point. p is increasing in z and its second derivatives are continuous. However, there is no other restriction on the form of the price function. Importantly, this implies that p is not necessarily linear. $p(\mathbf{z})$ is called the hedonic price function because it links the price of a commodity with the quantities and the prices of its characteristics.

There is no second-hand market. This is a simplifying assumption to avoid having to take into account depreciation.

A final but very important hypothesis is that goods are indivisible, that is, they can not be divided in any small quantity. It seems straightforward for housing but is not always so, think for example of gas or water. The consequence of this is that marginal analysis does not apply to goods, only to characteristics.

[4]Boldness in formulae denotes matrix or vector form.

1.3.1 The Market Equilibrium

First, let us analyse the consumer's side. He only buys one good[5], containing n characteristics as specified before. The utility function of the consumer is

$$U(x, \mathbf{z}) = U(x, z_1, ..., z_n) \tag{1.1}$$

Where x is the quantity of all other goods consumed. Its price is normalized to 1. U is strictly concave in both its arguments. The budget of the consumer is

$$y = x + p(\mathbf{z}) \tag{1.2}$$

Since $p(\mathbf{z})$ is not necessarily linear, the budget constraint of the consumer is not necessarily linear either. Through the Lagrange multiplier method, we can find that the first order condition of the maximization of utility under the budget constraint is

$$p_i = \frac{\partial p}{\partial z_i} = \frac{\partial U}{\partial z_i}\frac{\partial x}{\partial U} \qquad i = 1, ..., n \tag{1.3}$$

The second-order condition is negative given the strict concavity assumption. Let us now define a bid function $\theta(\mathbf{z}; u, y)$ such that

$$U(y - \theta, \mathbf{z}) = u \tag{1.4}$$

θ represents the willingness to pay of a consumer for some $\mathbf{z}$ given income y and utility u. In other words, it describes an indifference surface since it shows the bundles of characteristics which, given that they provide utility u and that the consumer has income y, will lead to a willingness to pay that is the same, that is, bundles of characteristics between which the consumer is indifferent. Hence, it relates prices of characteristics and money[6] :

$$\frac{\partial \theta}{\partial z_i} = \frac{\partial U}{\partial z_i}\frac{\partial x}{\partial U} \tag{1.5}$$

[5]This is Rosen's way of solving the problem of not being able to use marginal analysis on goods.
[6]Since the price of x, which represents all other goods, is normalized to 1.

We recognize the marginal rate of substitution between characteristic i and money. We can also show that the second derivative of the bid function with regards to characteristic i is negative. Hence, the bid for a characteristic will increase with the quantity of the characteristic but at a decreasing rate.

On the one hand, the price is the minimal amount the consumer needs to pay to get the good while on the other hand, the bid function is the maximum amount he wants to pay. At the equilibrium we thus have the two conditions (where the second is nothing but equation 1.3)

$$p(\mathbf{z}) = \theta(\mathbf{z}; u, y) \tag{1.6}$$

$$p_i = \frac{\partial \theta}{\partial z_i} \tag{1.7}$$

In words, a consumer will buy a good whose price is equal to his willingness to pay and for which his marginal valuation of each characteristic is equal to the implicit price of it. Note that Rosen extends his model to fit consumers' tastes. The problem is very similar and the utility function becomes $U(x, \mathbf{z}; \alpha)$, where α is a consumer's tastes. There is a joint distribution $F(y, \alpha)$ in the population.

Let us now turn to the producer's side. The process is very similar. Let $M(\mathbf{z})$ be the quantity produced of a good containing characteristics $\mathbf{z}$. The cost function of the firm is

$$C(M, \mathbf{z}, \beta) \tag{1.8}$$

β is a parameter representing factor costs and parameters of the production function with a distribution $G(\beta)$ in the population and with

$$\frac{\partial C}{\partial M} > 0 \qquad \frac{\partial^2 C}{\partial M^2} > 0$$

$$\frac{\partial C}{\partial z_i} > 0 \qquad \frac{\partial^2 C}{\partial z_i^2} \geq 0$$

That is, total costs of production of goods are positive and increasing while those of production of characteristics are positive and non-decreasing. Each firm maximizes its

profit

$$\pi = Mp(\mathbf{z}) - C(M, \mathbf{z})$$

Which gives first-order conditions

$$p_i(\mathbf{z}) = \frac{\partial C}{\partial z_i}\frac{1}{M} \tag{1.9}$$

$$p(\mathbf{z}) = \frac{\partial C}{\partial M} \tag{1.10}$$

The second condition is the usual "price must be equal to marginal cost of production of the good". The first is simply an adaptation of it to our context : the implicit price of characteristic i must be equal to the marginal cost of production of characteristic i per unit of good.

Let us now define an offer function $\phi(\mathbf{z}, \pi, \beta)$ which shows the combinations of $\mathbf{z}$, π and β for which the producer is indifferent. The point is the same as for consumers : ϕ represents an indifference surface for producers. To find first order conditions, let us first replace p by ϕ in the profit and take first derivatives.

$$\pi = M\phi - C \quad \Leftrightarrow \quad \phi = \frac{C + \pi}{M} \quad \Leftrightarrow \quad C = M\phi - \pi \tag{1.11}$$

$$\frac{\partial \phi}{\partial z_i} = \frac{\partial C}{\partial z_i}\frac{1}{M} > 0 \tag{1.12}$$

$$\frac{\partial C}{\partial M} = \phi \tag{1.13}$$

And hence, from these and from 1.9, at equilibrium we must have

$$p_i(\mathbf{z}) = \frac{\partial \phi}{\partial z_i} \tag{1.14}$$

$$p(\mathbf{z}) = \phi(\mathbf{z}, \pi, \beta) \tag{1.15}$$

In words, goods will be produced until price is equal to marginal cost, for the good as well as for its characteristics. Equilibrium will happen when conditions are satisfied both for the consumer and the producer.

1.3.2 Graphical Interpretation

We know that a derivative can be interpreted geometrically as a tangent and we can therefore represent our equilibrium graphically. We also know that the bid function is concave since its first and second derivatives with respect to z_i are respectively positive and negative. On the other hand, the offer function could be linear or convex.[7] We will however follow Rosen's example and depict it as convex.

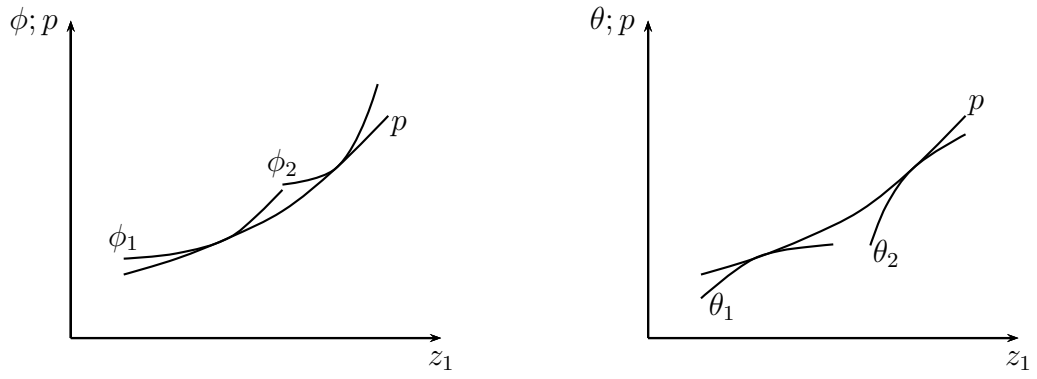
On figure 1.1 we have graphed offer and bid curves. Given that goods may have many characteristics, it might be difficult to draw a graph representing these curves for each feature. Here, Rosen's method is to draw the graph for z_1 and take quantities of the other characteristics as well as profit as being in equilibrium. For clarity, we have written ϕ_1 but it should be $\phi_1(z_1, z_2^*, ..., z_n^*; \pi_1^*; \beta)$ and the same for ϕ_2, p, θ_1 and θ_2 *mutatis mutandis*.

ϕ_1 is the offer curve for a firm that is best at producing a small amount of z_1 for a low cost while ϕ_2 is more efficient at producing a bigger amount.

For the other graph, θ_1 is the bid curve of a consumer who maximizes his utility for a small amount of characteristic 1 while consumer 2 would require a bigger amount to maximize his.

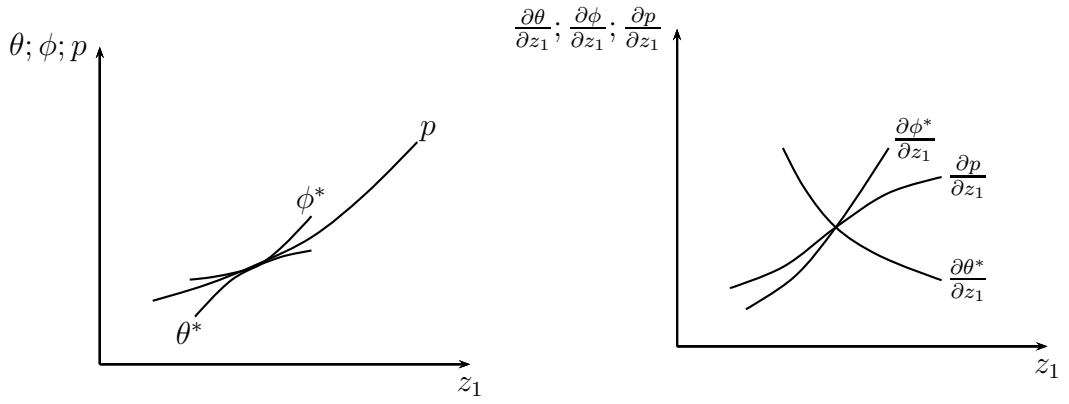
The equilibrium is depicted in two different ways on graphs 1.2a and 1.2b. On the right are the derivatives of each curve with respect to z_1. This shows that, at the equilibrium, the marginal price of characteristic 1 must be equal to the marginal willingness to pay of the consumer for z_1 and the marginal willingness to offer of the producer. The graph on the left represents the needed equalities (and tangencies) on the one hand, between the valuation of the consumer and the price of the market for the good and on the other hand, between the minimum price needed by the producer and the good's price on the market.

[7] The second derivative of the cost function is either positive or null.



(a) Offer Curves (b) Bid Curves

Figure 1.1: Offer and Bid Curves



(a) Equilibrium (b) Equilibrium viewed differently

Figure 1.2: Hedonic Equilibrium

Finally, let us note that the marginal bid and offer curves are nothing but inverse compensated demand and supply curves. We may thus be able to deduce marshallian demand curves for characteristics from them. However, the matter is quite complicated since the price function and hence the budget constraint are not linear. We now turn to this problem.

2 The Hedonic Method to Identify Demand and Supply

We have mentioned that one of the two main objectives of hedonic theory is the identification of the structures of demand and supply for characteristics. Since we do not have enough data to perform such an analysis, we will limit ourselves to a short theoretical review.

2.1 Rosen's Two-Step Method

The fact is, we only observe the prices of the goods exchanged on the market and not the prices of the underlying characteristics. Rosen (1974) proposed a two-step procedure, firstly to recover the implicit prices of attributes and secondly, to estimate the demand (and supply) functions for them. Suppose we have data about the parameters α and β and that we observe the quantities and prices of goods exchanged on the markets as well as their characteristics. First, regress the price of the goods on their characteristics (we ignore residuals for clarity):

$$p(\mathbf{z}) = f(z_1, z_2, ..., z_n) \tag{2.1}$$

This is a direct result of the hedonic hypothesis : prices are nothing but bundles of characteristics. Note that nothing is said about the econometric technique to be used. Nothing is said either on the form of the function f.

Then, compute the partial derivative (evaluated at exchanged quantities) of $p(\mathbf{z})$ with regards to characteristic i. This derivative is the implicit price of that characteristic. Using this as endogenous variable, we can then estimate two equations which will give us demand and supply functions for characteristic i :

$$p_i(\mathbf{z}) = \frac{\partial p(\mathbf{z})}{\partial z_i} = F_i(\mathbf{z}, \mathbf{x}_1) \qquad i = 1, ..., n \tag{2.2}$$

$$p_i(\mathbf{z}) = \frac{\partial p(\mathbf{z})}{\partial z_i} = G_i(\mathbf{z}, \mathbf{x}_2) \qquad i = 1, ..., n \tag{2.3}$$

Where $\mathbf{x}_1$ and $\mathbf{x}_2$ are respectively empirical counterparts of α and β, what Rosen calls "exogenous shift variables". F and G are the demand and supply functions (for character-

istic i) to be estimated. We thus have $2n$ equations to estimate the $2n$ variables.

This is all very well except that Rosen does not provide us with functional forms or econometric insight. Unfortunately, there are many problems to be reckoned with.

2.2 The Problems of the Hedonic Analysis of Demand and Supply.

There are various issues when one tries to infer demand and supply for characteristics in a hedonic model. We will briefly summarize some of them.

The first problem is that Rosen's second step simply does not provide any additional information. It can be proven (Brown and Rosen, 1982) that if both the hedonic price equation and the demand and supply equations are linear, it is generally not possible to recover the structural parameters (from supply and demand) from the two step method. Even if this problem is solved[8], they argue that we fall back on a classical simultaneity problem except that instruments are not available. Their solution is called the multi-market approach. Marginal prices should be estimated from different markets[9] with the same structural parameters. In that case, information would be created through the second-step of Rosen's method.

Unfortunately, Epple (1987) has shown that even with the multi-market approach, the characteristics of suppliers and consumers could not be used as instruments to solve the simultaneity problem. He shows mathematically that for a quadratic hedonic function estimated for k different markets, there will be correlation between the instruments and the errors. This stems from the probable relation between the characteristics of both consumers and suppliers. Indeed, it would not be surprising if high-income consumers bought high-end goods which must be produced by top-notch firms. He solves the problem by finding appropriate rank and order conditions but unfortunately they both depend on the functional form of the equations and on the forms of errors in the residuals. Bartik (1987) makes the same point but suggests using instruments from controlled experiments. For

[8]They suggest that an appropriate restriction solving the problem would be to have a hedonic equation of degree m and a demand (or supply) function of degree $m-1$ maximum. In that case, the information contained in the second step would be useful. Rosen made a similar suggestion in his paper.

[9]Different means geographically (e.g. Brussels and Wallonia) distinct or from different time periods.

example, studying one thousand households and giving a subsidy for rent to some of them would be their solution. Obviously, such data is rarely available in reality.

Ekeland, Heckman, and Nesheim (2004) argue that all the problems we have mentioned are misleading. They believe researchers do not extract all the information available in the hedonic function because they try to linearize hedonic functions, and that through non-parametric methods and non-linear instrumental variables, it is possible to identify demand and supply functions with very few restrictions. We do not go into further details as we do not have sufficient data to be able to apply any of the techniques mentioned.

3 Price Indices and the Real Estate Market

3.1 Traditional Price Indices

The second objective of the hedonic method is to produce a quality-constant price index. Indeed, traditional price indices (Eurostat, 2008) such as the Laspeyres, Paasche or Fischer indices only take into account prices and quantities, not qualities. The Laspeyres or "base-weighted" index (equation 3.1) shows the evolution of the price of the basket of goods bought at time 0. The Paasche (equation 3.2) or "current-weighted index" indicates the evolution of the price of the basket of goods bought at time t. Finally, the Fischer index is the geometric average of the two.

$$\text{Index}_{Laspeyres} = \frac{\sum_{i=1}^{n} p_{i,t} q_{i,0}}{\sum_{i=1}^{n} p_{i,0} q_{i,0}} \tag{3.1}$$

$$\text{Index}_{Paasche} = \frac{\sum_{i=1}^{n} p_{i,t} q_{i,t}}{\sum_{i=1}^{n} p_{i,0} q_{i,t}} \tag{3.2}$$

Computers are often mentioned as a typical example of the quality problem. Computer prices have not changed that much in recent years although they are faster, have more memory and so forth. Traditional price indices would simply state that prices have been more or less stable, which would hide the actual decrease of the price of the exact same good. Another example is cars. The main objective of one of the first hedonic studies Court (1939) was to show that prices of cars have gone down even though their nominal prices have gone up. Indeed Court argued that quality has risen much faster than nominal prices. Hence, if we compare the exact same car, "prices" would have gone down over time. This is something that traditional price indices can not take into account. This problem is not new and in 1942, Schumpeter (1942, p. 66) was already saying :

> "[...]improvements in quality almost completely fail to assert themselves although they constitute, in many lines, the core of the progress achieved - there is no way of expressing adequately the difference between a motorcar of 1940 and a motorcar of 1900 or the extent to which the price of motorcars per unit of utility has fallen."

Before analysing the hedonic price index in more details, let us make a detour to see why we need better price indices for the housing market and what are the main methods to improve them.

3.2 Why We Need Better Price Indices For The Real Estate Market

There are at least seven good reasons why we need better price indices for real estate markets (Eurostat, 2013).

1. Real estate markets are indicators of economic growth. Usually, a decline in prices in the housing market goes hand in hand with a decline in economic activity. We can distinguish three effects when prices increase. First, the income and employment of the construction sector and of the real estate sector go up. Secondly, the number of sales of current houses increases implying an increase in tax revenues. Thirdly, there is the "wealth effect" : as prices appreciate, the balance sheet of households also does and people consume more.

2. Because the real estate market is a large sector, variations in its prices have very important effects. Hence, more accurate information is needed for central banks and their inflation targeting.

3. For the same reason, better price indices are required to compute efficiently the Consumer Price Index and the deflator.

4. Since a house or an apartment is usually the main asset of a household, we need better price indices to be able to measure aggregate wealth.

5. Because financial and real estate markets are linked, the stability of financial markets could be monitored more easily.

6. It would provide better information for individuals and firms willing to invest in housing.

7. It would make international comparisons of housing markets much easier. At the moment, housing price indices have very different methodologies and many do not take into account quality.

3.3 Methods To Build Better Indices

There are usually four methods (Triplett (2006) and Eurostat (2013)) used to build constant quality price indices : the hedonic method, the stratification, the repeat sales and the assessment-based. Since it is the main focus of this master's thesis, the hedonic method will have its own section. Let us discuss the other three.

3.3.1 Stratification

The two easiest, most common and maybe most flawed measures of prices are the mean and the median. These however do not take into account changes in quality over time or samples. The (ad hoc) solution is to "stratify", that is, divide the sample into more homogeneous subsamples and give a weight to each of these subsamples. Over time, the comparisons are made between appropriate subsamples and weighted to give an aggregate index. For instance, a stratification that is often performed in Belgium is to divide goods based on whether they are located in Wallonia, Brussels or Flanders. The major issue is simple : the more we divide the sample, the more homogenous the subsample gets, but the more we divide, the less observations we have in each subsample. Hence, it is very problematic if we want to take into account a great number of characteristics.

3.3.2 Repeat-Sales

The repeat sales method is a clever way of using the hedonic hypothesis without needing data about characteristics. Suppose the hedonic price function for good n, with K characteristics, at time t is

$$\ln P_n^t = \beta_0^t + \sum_{k=1}^{K} \beta_k z_{nk} + u_n^t \tag{3.3}$$

As can be seen, implicit prices of characteristics are not supposed to vary across time. The repeat-sales method makes the assumption that we have data about housing goods which have been sold multiple times. Suppose we have hedonic equations for periods t and s with $0 < s < t < T$ in the form of 3.3 for good n. In that case :

$$\ln P_n^t - \ln P_n^s = (\beta_0^t - \beta_0^s) + (u_n^t - u_n^t) = \ln \frac{P_n^t}{P_n^s} \tag{3.4}$$

The characteristics terms disappear as they are time-constant. Moreover, the variation in price is the same for all goods since β_0 does not depend on n. We can now evaluate the following equation (by OLS) to determine price changes between the two periods at constant quality :

$$\ln\left(\frac{P_n^t}{P_n^s}\right) = \sum_{t=0}^{T} \gamma^t D_n^t + \epsilon_n^t \tag{3.5}$$

Where D_n^t is a dummy which takes value 1 at the time of resale and value -1 at the original period of sale.

The major positive aspect of this method is that it is quite easy to perform. Moreover, it does not require much data in terms of characteristics. There are however two major issues. First, if we have data about repeated sales, then...we have data about repeated sales. It means that we ignore a huge part of the market, for example new constructions. Secondly, it ignores changes in characteristics over time. Housing is known as a good where major changes can happen, for instance renovation.

3.3.3 Appraisal-Based Method

Suppose we have data about housing goods sold once during period t. We can not compare the prices of these dwelling with their prices at time 0 since we do not have data about that. Say however that we have appraisals for all goods at period 0. The idea is to use these as proxies for the values of the goods at the initial period.

4 Hedonic Price Indices

Even in the realm of hedonic price indices, there is not a unique way to proceed. There are mostly two popular methods : the direct (and adjacent) method and the characteristics price index.

4.1 The Direct and Adjacent Methods

The first method, often called the "direct" method, is to use time dummies (Eurostat (2013) and Triplett (2006)). Suppose that we have observations of house sales for different periods of times. The dwellings are characterized by K characteristics. For time period t and good n, from Rosen's first step, the model can be written as [10] :

$$\ln p_n^t = \beta_0^t + \sum_{k=1}^{K} \beta_k^t z_{nk}^t + u_n^t \tag{4.1}$$

Where β_0 is the intercept and z_{nk}^t is the quantity of characteristic k for good n at time t. In that case, the "time dummy variable hedonic model" is :

$$\ln p_n^t = \beta_0 + \sum_{\tau=1}^{T} \delta^\tau D_n^\tau + \sum_{k=1}^{K} \beta_k z_{nk}^t + u_n^t \tag{4.2}$$

D_n^τ is a dummy variable which takes value 1 if the good n was sold in period τ. We have as many dummies (minus one) as time periods. If we assume the model is correctly specified, the price index interpretation is straightforward : $\exp(\delta^\tau) - 1$ represents the percentage change in price from period 0 to τ, all else being equal. Or if we write it in a more traditional way :

$$\frac{P^\tau}{P^0} = \exp(\delta^\tau) \tag{4.3}$$

Note that the approximation that allows the interpretation of the coefficient directly as the percentage change in price, $\exp(\delta^\tau) - 1 \simeq \delta^\tau$, is only valid for small values of δ^τ. Slightly less useful but still theoretically important, we must also mention that $\exp(\hat{\delta})$[11] is

[10] More on the appropriate functional form in a few pages.

[11] We drop superscripts for clarity.

not an unbiased estimator of $\exp(\delta)$. Kennedy (1981) argues that an appropriate correction is $\exp(\hat{\delta} - \frac{1}{2}\hat{V}(\hat{\delta}))$, where $\hat{V}$ is the estimated variance. This is still not an unbiased estimator but is slightly better than the first one. In practice, the use of this correction rarely changes anything.

Also, notice that there is no time superscript on the coefficients to be estimated any more (except for the times dummies). It means that if we write the model like this, the coefficients are constrained to be the same in each time period. Whereas this may not be a problem if we study a short time interval, we can not be sure that coefficients will not change in the medium-run. Indeed, remember that coefficients in the hedonic framework are implicit prices and hence, restricting them would mean restricting demand and supply. It is easy to think of examples : if people start to have less children, it may be that the number of bedrooms becomes less important but that the overall quality of the dwelling counts for more. This is a major shortcoming of the direct method.

The adjacent method is a simple adaptation of the direct method. Instead of estimating an equation with $T-1$ time dummies for T time periods, we would estimate an equation for each set of two (or more) time periods. For example, if we have $T = 1, 2, 3$, we would estimate one model for $t = 1, 2$ and one for $t = 2, 3$. This gives us information about the price change between periods $1, 2$ and $2, 3$. The choice between the two methods depends on the objective of the researcher. This is useful if the number of periods is quite high and we are not sure that the time-stability condition of the implicit prices is met.

4.2 The Characteristics Price Index Method

The second method is the "characteristics price index method". Say we have two time periods. If we fit a regression to each of them, we obtain the implicit prices of characteristics in each period. We can then simply calculate (for the Laspeyres case) :

$$\text{Price index} = \frac{\sum_{i=1}^{n} p_{i,t} q_{i,0}}{\sum_{i=1}^{n} p_{i,0} q_{i,0}} \tag{4.4}$$

Where $p_{i,t}$ is the price of characteristic i in time period t and the same for quantities q. A major issue is to decide which quantities to use. We have described three types of weighting which could be used here (Laspeyeres, Paasche, Fischer) but the number is

virtually unlimited. Moreover, since nothing guarantees that the hedonic price function is linear, the problem is even worse. With non-linearity, different quantities would result in different implicit prices and different weights could then mean very different results. On the other hand, this way of proceeding does not require the time-stability condition.

4.3 The Functional Form

Economic theory does not seem to suggest a particular functional form to model hedonic equations. In applied work, most economists have chosen to use a log-lin form for simplicity but nothing requires us to do so and some have suggested the Box-Cox as more appropriate.

Cropper, Deck, and McConnell (1988) have conducted an experiment of hedonic models[12] of various functional forms using Monte-Carlo simulations. Their results show that in the case where some characteristics may be unobserved (which is probably true of any empirical study), the linear, log-log, semi-log and Box-Cox forms work better than other forms (quadratic and quadratic Box-Cox). The choice obviously depends on how we define "best". They use two indicators to do so : average percent bias and maximum percent bias. The Box-Cox specification seems to be the best as it mostly produces a lower average bias than others and usually does as well if we consider the maximum bias.

Let us analyse succinctly what the Box-Cox specification is. Usually (Davidson and MacKinnon, 1993), in linear regression models we define

$$y_t = \mathbf{x}_t^{'}\boldsymbol{\beta} + u_t = \mathbb{E}(y_t|\mathscr{G}) + u_t \qquad t = 1, ..., n \tag{4.5}$$

Where y_t is the the t^{th} observation of the dependent variable, $\mathbf{x}_t$ a vector of the t^{th} observations of the independent variables, $\boldsymbol{\beta}$ a vector of coefficients, u_t residuals and $\mathscr{G}$ the corresponding sigma-field.[13] However, we do not have to restrict ourselves to a linear

[12] They focused on hedonic functions, not on demand or supply for characteristics.

[13] Mikosch (1998) (p.62) defines a sigma-field as such : "A sigma-field $\mathscr{F}$ is a collection of subsets of Ω satisfying the following conditions :

1. $\emptyset \in \mathscr{F}$ and $\Omega \in \mathscr{F}$
2. If $A \in \mathscr{F}$ then $A^c \in \mathscr{F}$
3. If $A_1, A_2, ... \in \mathscr{F}$ then $\bigcup_{i=1}^{\infty} A_i \in \mathscr{F}$ and $\bigcap_{i=1}^{\infty} A_i \in \mathscr{F}$ "

relationship such as $\mathbb{E}(y_t|\Omega_t)$. We can also find the conditional mean of smooth, monotonic functions of y_t. That is called a transformation. The most common example is the logarithm. The Box-Cox Transformation (Box and Cox, 1964) is defined as follows :

$$B(x,\lambda) = \begin{cases} \frac{x^\lambda - 1}{\lambda} & \text{if } \lambda \neq 0 \\ log(x) & \text{if } \lambda = 0. \end{cases} \tag{4.6}$$

Where x must be positive because of the properties of the log. The Box-Cox transformation is attractive because if $\lambda = 1$, no transformation is made and if $\lambda = 0$, we get the usual log form. λ can be given a value if there are good theoretical reasons for it but a better way to proceed is to estimate it and see if it complies with the theory. More precisely, the transformation can be applied either to the dependent variable only or to both the dependent and the independent variables (or to some independent variables only).

The disadvantage of the Box-Cox transformation is that if there is no economic reason to fix λ it has to be estimated (Davidson and MacKinnon, 1993). Least squares techniques can not do that. Maximum likelihood techniques will not fully work either. They will not be able to estimate the covariance matrix consistently even though they will be consistent for λ and $\boldsymbol{\beta}$. Intuitively, that is because there is a sequential estimation that maximum likelihood does not take into account, that is, λ has to be estimated before $\boldsymbol{\beta}$. Other methods have to be used such as double-length artificial regressions (Davidson and MacKinnon, 2001) which are much more difficult to implement than OLS. For these reasons, the Box-Cox transformation is not commonly used in hedonic papers and we will not use it either.

5 Empirical Applications

Empirical applications of the hedonic method are virtually infinite. It has been applied to very different goods such as wines, computers, cars, houses, etc. We will focus here on some housing applications.

Maybe the most interesting study for us is that of Decoster and De Swerdt (2005). To our knowledge, it is the first paper that applies the hedonic analysis to the Belgian housing market. Unfortunately, they had to work with problematic data. They used data from the "Belgian Household Surveys" for several years. The "Belgian Household Surveys" record characteristics of the good at the moment of the survey and not at the moment of the sale. Given that housing goods are particularly prone to renovation, this may be a big issue. They try to control for it by eliminating some characteristics which are too often changed. Moreover, they do not really mention it but the number of observations they have for some years is very small. Finally, their data consist of averages at the levels of cities. Thus, they have no information about individual houses. Their conclusion is that, if their model is correct, there is an important discrepancy - sometimes positives, sometimes negative - between a hedonic price index and an average sales price index. It might thus be interesting to collect more data to investigate the real difference between the two.

Recently, De Bruyne and Van Hove (2013) have also tried to fit a hedonic regression to different regions of Belgium. Their study uses the same dataset as Decoster and De Swerdt (2005) and hence suffers from the same problems. Their major contribution resides in adding locational variables. For example, the distance by car to Brussels, the travel time by car to Brussels and the travel time by train to Brussels are all significant. The same is true for the travel time by car to the provincial capital. Unfortunately, they do not have many intrinsic variables, which renders any comparison with our model useless.

Another study of interest is that of Bilbao, Bilbao, and Labeaga (2010). They have been able to obtain data from real estate agents. They have then built a hedonic model for five different Spanish cities. They include two variables that are rarely present in hedonic models : distance to the city centre and SO_2 concentration at the location of the dwelling. Both are significant at the 1% level. Environmental quality and location thus seem very

important. Unfortunately, they only have data about one year and do not build a hedonic price index.

Closer to us and one of the most extensive application of the hedonic method is the work of Laferrère (2005) in France. In collaboration with the notaries, the INSEE produces a quarterly hedonic price index for many regions of France. A major shortcoming, as in Belgium, is that the information is provided by the notaries only if they are willing to. In some regions, information is thus very scarce. Regarding econometric aspects, there are three things to notice. First, they include environmental variables based on interviews with real estate agents. More precisely, they define three different kinds of neighbourhoods within cities. Secondly, the dependent variable is the price per square meter as opposed to the price in most studies. We do not know why they made that choice. Thirdly, they estimate their model for each zone separately. Surprisingly, to the best of our knowledge, no analysis has been performed with these hedonic prices.

Finally, we turn to the hedonic study about the real estate market of Hong-Kong performed by Chan, Cho, and Mok (1995). The data about housing goods are particularly easy to obtain there as the "Land Office" registers all sales and sells the information from time to time. The researchers are not interested in building a hedonic price index but just in deriving a hedonic price function. There are three original points in their paper. First, they consider a very narrow time-frame (one month). Secondly, they use very small geographic zones - we do not go into the the details but they select areas where the nearest train station is at most at a five minute walk - to control for different means of transportation. Thirdly, they include four interesting locational variables : distance from the business district, the presence of a sea view, the quality of the school district and the presence of entertainment facilities in the estate where where the good is located. All these variables are significant at the 5% level except for the school district which is insignificant at the 10% level.

A comparison of the findings of some of these studies with our own will be provided in the last section but we can still draw three important conclusions. The most obvious is the lack of data. Even in official studies such as in France, there are not too many variables. Secondly, "location, location, location" seems to be an appropriate motto. Many hedonic studies do not use any locational aspects while they appear to be very important. Finally,

there seems to be two sides regarding the modelling of regions. The hedonic model can be applied either to a whole country - Belgium or France - or to a very small region - the area around train stations in Hong-Kong . Data is one constraint but the more important aspect in our view is the objective of the researchers. In France, it was to establish a hedonic price index in a variety of zones, hence their approach is understandable. In Hong-Kong, they wanted to estimate a hedonic price function for certain attributes as precisely as possible, they thus tried to eliminate as many variables as possible. The only conclusion we can reach is that there are as many methods as there are hedonic studies.

6 The Housing Market

Before applying the hedonic techniques that we have seen to the housing market, it may be a good idea to know more about it and why we chose it. Housing is usually a very big part of an individual's wealth and consequently of a country. In 2009, housing represented 26.1% of a Belgian household's expenditures, 25.8% for the Flemish Region, 29.1% for the Brussels Capital Region and 25.7% for Wallonia (SPF Economie, 2011). This is thus a very important market and any improvement in its understanding would be welcome. Moreover, it is particularly suited for a hedonic analysis : under the name "housing" hides an incredible variety of goods. We will first focus on the properties of housing as a good and then on the characteristics of its demand and supply.

6.1 The Good "Housing"

Three properties[14] of the housing market make it particularly suited for a hedonic analysis: durability, heterogeneity and spatiality. Wheaton (1990) also highlight specific features of the housing market.

Durability stems from the extensive life span of a housing unit. For instance, according to numbers from the "Enquête socio-économique générale 2001" (Goossens, Thomas, and Vanneste, 2005), in 2001, 17% of Belgian houses had been built before 1919 and 35% before 1945. This is very interesting for hedonic price indices since it implies that the heterogeneity of the goods available on the market is quite strong.

Heterogeneity is a result of the complexity of the good "housing". Its characteristics are numerous : surface, number of rooms, number of bathrooms and bedrooms, the presence of a garden, of a garage, the insulation, the location, etc. The list could easily be extended *ad infinitum*. Again, this is very interesting in the hedonic context : no two dwellings are exactly the same and variations of the quality of the goods are considerable.

Thirdly, housing units are spatially fixed - one can not move a house easily - and its location can have a big influence on its price. The hedonic method may allow us to judge

[14]Studies on the properties of the housing market are numerous but we have mostly based ourselves on Arnott (1987) and Whitehead (1999).

exactly how much location influences the total price of a dwelling, all other characteristics fixed.

Wheaton (1990) also highlights four very specific and important traits of housing. First, real estate markets are characterized by a structural vacancy rate, that is, at any given time some proportion of housing goods are not occupied. Second, prices of housing goods seem to be quite sensitive to changes in the variation of the vacancy rate. Third, most transactions occur between agents already on the market because agents become mismatched with the housing good they own. Fourth, vacancy rates and market prices affect housing supply but at a slow pace. Wheaton is able to explain these four empirical facts by a matching model and a concept called "search"[15]. Households are buyers and sellers at the same time. Through Poisson processes, on the one hand, households become mismatched or matched with the house in which they live and, on the other hand, buyers and sellers match. For instance, when a young couple expects a child, they will want to move to a larger house. They will have to search for such a good. When they have found it - provided the Poisson process does not revert, that is, the child dies - they must now sell their old house or in other words, search for a household matched to their old dwelling. This model can explain the four facts aforementioned.

6.2 The Demand for Housing

Housing being a complex good, the demand for it is also complex and has many determinants. The first one is not surprising : income. Income obviously affects the housing goods one can afford. Prices of other goods and hence relative prices also play a role. Various studies have been made to determine the elasticities regarding income and prices. Results vary from from study to study but researchers usually agree that "Nowadays it is admitted that housing demand is inelastic with regard to income and prices" (Granelle, 1997, p.33) or that "on the demand side both income and price elasticities are generally less than one but vary between tenures, income and demographic groups as well as between different housing attributes" (Whitehead, 1999, p.1568).

[15]More precisely, "Economic search is the collection, transmission, and implementation of information required for the sequential decisions entailed in expected utility maximization" McCall and McCall (2008) p.5.

Another obvious determinant is demographic characteristics. Young professionals do not have the same needs as retired people. Arnott (1987) even argues that housing consumption is one of the three major decisions in an individual's life, with household composition and job choice.

Since housing is spatially fixed and is very durable, the location of the dwelling is also very important. It means choosing a regulatory environment - not all cities or countries have the same tax levels for example - and other things such as safety, proximity to the workplace and so on. This emphasizes the role governments can have on housing decisions, notably with regard to "voting with one's feet" (Tiebout, 1956).

Given that housing is a highly heterogeneous good, it is difficult to find a good suiting one's tastes perfectly. Because of this and of moving costs, households do not adjust very quickly to changes, they usually stay in a house or an apartment for many years, moving only when necessary. Housing policies trying to influence consumers' choices may thus take a long time to have any effect.

Another peculiar aspect is that demand is nearly equally divided between buying and renting. This is referred to as the tenure choice.

Finally, demand for housing is a consumption decision but it is also an investment decision. Indeed, it usually represents one of the most important transaction in one's life and is a very significant part of the capital of most individuals and hence of a country. Because of this, portfolio motives may be considered but it is rarely the case in hedonic models.

6.3 The Supply for Housing

A very specific trait of housing is that new supply represents a very small fraction of the existing stock of housing. It means that the reactions the supply could have to changes in market conditions will not have a huge impact. However, many studies have found that supply is actually very elastic to changes in demand, some even finding nearly-perfect elasticity (Butler, Gastler, Pitkin, and Rothenberg, 1991 as cited in Whitehead, 1999). Alas, researchers do not seem to agree on this point (see Wheaton, 1990 for example).

Secondly, housing supply can be divided into four main components : construction, maintenance, rehabilitation and conversion. Construction is the combination of land and capital to produce new housing units. Hence land markets also have to be taken into account. Maintenance is the act of maintaining the quality of a dwelling by simple tasks such as cleaning. This is very different in nature to construction because it requires a gradual - nearly constant - use of capital. Rehabilitation is an increase in quality by the application of capital. This application is not gradual but sudden. Finally, conversion is the act of changing the size of a good.

Last but not least - we already mentioned it - the housing market is thin, that is, it is very difficult to find a good fitting one's tastes perfectly. In consequence, housing goods are often modified by households to suit their needs, for instance by buying furniture. Routine maintenance is also generally provided by the household itself. Because of this, time and money are important inputs (Arnott, 1987). Moreover, it implies that some people will buy houses in a bad state to renovate it while others will want neat and ready homes.

7 Location, the housing market and hedonic functions

We have mentioned multiple times that, because of spatial location and immobility, location is especially important for housing goods. Can and Megbolugbe (1997) summarizes the issue especially well :

> "A major limitation of currently available [hedonic] indices is their insensitivity to the geographic location of dwellings within the metropolitan area. Indices are constructed on models that do not incorporate the underlying spatial structure in housing data sets."

Thus the question is : how can we incorporate locational aspects in hedonic models ? We will see that our sample does not allow us to do much but, if we had perfect data, say GPS location of housing goods, what should we do with it ?

A recent trend in hedonic analysis is interested in spatial econometrics. Suppose we are trying to evaluate the price of a house in some neighbourhood. Apart from obvious issues such as how to know which variables are relevant to measure a neighbourhood, there are two major problems (LeSage and Kelley Pace, 2009) with which a researcher trying to include locational variables is faced. The first is the definition of the boundaries of a neighbourhood. Say that we are considering an area where, in the west, there is a lot of crime and poverty while, in the east, the area is safe and rich. Fitting a single hedonic equation to the area may give us uninteresting results. For example, it is very probable that intrinsic characteristics of houses will be more valuable in the rich part than in the poor part. This problem is especially pervasive because researchers usually get data about some administratively defined regions and administrative divisions are rarely based on economic reasoning. This is called "spatial heterogeneity" : economic relations may be different in different areas.

The second problem is "spatial autocorrelation" (Dubin, 1998) : the residuals of regressions on the prices of nearby houses will probably be correlated. This can be thought of as an omitted variable problem. Say we have data about house prices in a small given region over one year. Imagine that, around the middle of that year, a local newspaper names one school of the region the best school of the country. Immediately, prices of houses around

this school will increase. Obviously, it is impossible to take such information into account in an econometric model. What will be the result ? Prices of houses located nearby the best school of the country are linked by a variable which is not taken into account in the model, hence, there is spatial autocorrelation in the residuals of the regression. Because of the omitted variable, OLS coefficients will be less precise and the covariance matrix biased.

Spatial econometricians have developed many models to take care of these two problems. Our goal here is not to explain spatial econometrics in detail and we will thus limit ourselves to a few basic principles. The basic process of spatial econometrics is the spatial autoregressive process :

$$\mathbf{y} = \rho \mathbf{w}\mathbf{y} + \mathbf{x}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{7.1}$$

$$\boldsymbol{\epsilon} \sim N(0, \sigma^2 \mathbf{I}_n)$$

Where $\mathbf{y}$ is an $n \times 1$ vector of observations of the dependent variable, $\mathbf{x}$ an $n \times n$ matrix of the observations of the independent variables, $\mathbf{w}$ an $n \times n$ spatial weight matrix and ρ a scalar which we call the "spatial autoregressive coefficient". This looks very much like an autoregressive process from time series analysis but the comparison is more complicated than it appears.[16] The spatial autoregressive process can usually go both ways : if the price of the house of neighbour A influences the price of the house of neighbour B, it is highly probable that the inverse is also true[17]. Actually, both ways is still too limited as interactions can be multilateral : a house often has more than one neighbour.

The weight-matrix is usually supposed to be known *a priori*. There are many ways to form it. The simplest one is for element $w_{i,j}$ to be 1 if the regions or neighbourhoods are adjacent and 0 otherwise. Note that usually a region is not considered to be a neighbour to itself. Another common way to define it is for $w_{i,j}$ to be 1 if region j is at most at some distance from region i and 0 otherwise. We can also use continuous values, for example $1/d_{i,j}$ with $d_{i,j}$ the distance between region i and region j. This is a major issue of the method since there is no consensus on which weight matrix to use (Dubin, 1998).

[16] See LeSage and Kelley Pace (2009) for a comparison of the processes. Note also that the process could be extended to include a time component.

[17] In time series, we usually include lags only, not leads.

The easiest way to understand the spatial autoregressive process is to decompose it in series expansion as is usually done with AR processes. In that case :

$$\mathbf{y} = (\mathbf{I}_n - \rho\mathbf{w})^{-1}\boldsymbol{\epsilon} = (\mathbf{I}_n + \rho\mathbf{w} + \rho^2\mathbf{w}^2 + ...)\boldsymbol{\epsilon} \tag{7.2}$$

For example, if we define $\mathbf{w}$ as $w_{ij} = 1$ if regions are contiguous and 0 otherwise, $\mathbf{w}^2$ will reflect second order neighbourhood effects, that is, the effects of neighbours of neighbours. If we suppose that $|\rho| < 1$ then the effects are disappearing as we get further away. While in time-series analysis, given appropriate conditions, a shock in the past will disappear with time, in spatially autoregressive processes, a shock in the price of a house 50 kilometres away will have less impact than the same shock in the price of a house located in the same street.

Is this useful in practice ? It seems so and some studies have shown that spatial autocorrelation was often found in data sets of hedonic studies and that, as a result, hedonic functions were better specified when taking it into account. For example, Anselin and Gallo (2006), while studying the effects of pollution on house prices find a ρ coefficient of 0.33 and the significance of some of their variables changes if they use an autoregressive model or not. Policy-wise this is very important. Indeed, when estimating the willingness to pay of two pollutants they find that in one case, OLS confidence intervals are much higher than that of a spatial autoregressive model while in another case, OLS confidence intervals are much smaller. Thus, if spatial autocorrelation is indeed present, not modelling it can have very serious effects which, moreover, can not be predicted to go one way or the other. Let us note however that the authors conclude that the bias of OLS may be due to endogeneity problems more than to the the presence of spatial autocorrelation. Yet, even after taking care of endogeneity, not modelling spatial autocorrelation leads, in their case, to confidence intervals which are too small and are thus misleading.

Another study is that of Montero, Fernández-Avilés, and Mínguez (2011). They evaluate the impact of noise on house prices. They try to see which type of model is the best and conclude that it is a Spatial Durbin Model[18]. Again, not taking into account spatial

[18]The Spatial Durbin Model is a type of model which represents the following DGP : $\mathbf{y} = \rho\mathbf{w}\mathbf{y} + \mathbf{x}\boldsymbol{\beta} + \gamma\mathbf{w}\mathbf{x} + \boldsymbol{\epsilon}$ with γ being a scalar and $\boldsymbol{\epsilon}$ defined as before. In other words, the autoregressive structure is applied both to the dependant and to the independent variables. (LeSage and Kelley Pace, 2009)

autocorrelation leads to biased coefficients which can lead to big differences when estimating willingness to pay. Andersson, Jonsson, and Ogren (2008) also conclude that spatial models work better than OLS for the modelling of the effect of noise on property prices.

Unfortunately, since the field is very young, it is quite difficult to find many serious empirical studies. Apart from those we have already named, let us cite, *inter alia*, Rouwendal and van der Straaten (2008) and Liao and Wang (2012).

Moreover, spatial econometrics is not without critiques. The fiercest criticisms concern the matrix $\mathbf{W}$ (Corrado and Fingleton, 2011). Firstly, the matrix is supposed to be known *a priori*, but this may be interpreted loosely and the researcher is usually free to choose it as long as he gives a reasonable justification. This is very problematic. In most applications it appears that all economic reasoning is evacuated. For example, many researchers use the geographic distance as weight whereas economic indicators such as GDP may be more suited. Unfortunately no solution exists for the economist who would like to "unfix" $\mathbf{W}$. Secondly, the $\mathbf{W}$ matrix is supposed to be constant. This is reasonable in cross-sections but may be far-fetched in panel data analysis. Even with these problems, nothing better than the matrix $\mathbf{W}$ has been found. Other methods have been proposed but they all have major problems (Corrado and Fingleton, 2011).

Another problem is what happens if $\rho = 1$, that is, if we have a spatial unit root. Unlike time series analysis where unit roots and cointegration are commonly treated, the field is very new in spatial econometrics. A first analysis was proposed by Fingleton (1999) but no definitive test for either unit root or cointegration seems to have appeared since (Beenstock and Felsenstein, 2008). This may be very problematic since, if the series are not stationary, the t-ratios are not correct and our regression is spurious.

Obviously spatial econometrics also depends on data availability which, at the moment, is not very good as our data set can attest. It seems however that often as methods get better, data also does.

Notwithstanding these problems, possibly arising from youth, we think spatial econometrics is the most promising technique for the improvement of hedonic analysis.

Part III

Empirical Analysis

Let us now turn to the empirical part of this master's thesis. We will look at the available data, the strategies for our study, the variables we obtained and some descriptive statistics. Finally, we will try to answer our two questions :

1. Has the price of apartments changed in Belgium between 2011 and 2012 at given quality ?

2. How can we use this knowledge to understand the statement of traditional price indices "prices of apartments have changed in Belgium during these years" ?

1 Analysis of the Data

1.1 Data Collection

Our sample comes from "La Fédération Royale du Notariat Belge". In Belgium, any sale of a housing good must be registered by a notary. The notaries are supposed to register some information concerning each sale going through their hands and to send it back to the federation. This scheme has been in place since 2010. Unfortunately, problems are numerous.

First of all, even though notaries enjoy a monopoly on housing sales registrations and should thus provide us with a lot of information, they do not have many obligations regarding data collection. In theory, data registration is compulsory but in practice there are no sanctions and most of the notaries do not write much. Worse, it seems that their sample is far from representative. For instance, in 2011, they register that Wallonia and Brussels only account for 6.28% and 14.5% of apartments' sales, respectively. This is very surprising given that according to Goossens, Thomas, and Vanneste (2005), Wallonia and Brussels represented 52.12% of the Belgian stock of apartments in 2001. Moreover, the notaries' dataset only contains $20,732$ observations for 2011 while the Stadim index (Stadim, 2013) reports $42,324$ observations for the same period.

Secondly, even when they do write something, there are many mistakes. For example, writing in the wrong column or writing an aberrant number is common. Fortunately, these mistakes are less frequent in 2012 than in 2011.

Another major problem is that for some variables, e.g. the variable "garage", a "0" means either "no garage" or "no data". It is difficult to see how these variables could be used with such a shortcoming.

We still think however that the dataset, because of its uniqueness, is quite valuable and that these problems will be solved through time and experience. Indeed, it has come to our knowledge that the database was going to be updated soon with new and better information. Moreover, this dataset is still a major improvement over preceding ones. We discussed it in the section about empirical applications but it is good to recall that the

only information available hitherto about the real estate market in Belgium consisted in characteristics aggregated at the communal level. In other words, previous studies had to be based on information such as "the average surface of an apartment in this city is x square meters" or "the average price in this village is y euros". There is no need to detail the weaknesses of such information.

1.2 The Data

We received two data sets, one for 2011 ($20,732$ observations) and one for 2012 ($15,082$ observations). Our data only concerns apartments. Data about houses were useless because of the number of errors and the number of missing observations.

In theory, this is a lot of information. Table A.1 in the appendix shows an apartment for which we have full data. In practice however, this dataset is barely usable. Indeed, as was mentioned, most notaries provide nearly no information. For instance, for the 2012 data set, the surface is known for only 213 goods. We thus had to decide which variables to keep. The problem was the following : more variables meant less observations but less variables meant useless information. This trade-off can only be solved if we know exactly what we want to do with the data we have.

Note also that we had to clean up the dataset from observations that seemed aberrant. For example, it is difficult to believe that a $500,000$ euro apartment can only have one bedroom and a surface of $40m^2$. We ended up with 442 observations.

1.3 Strategies

While reviewing empirical applications, we said that the modelling depended on the objective. It is also vital to select appropriate variables. There were three options for us : apply a single regression to Belgium, model each region separately or apply a single regression to Belgium but with dummy variables for different regions. Even though *a priori* arguments can be made for each possible choice, our modelling strategy was restrained by an empirical fact : as shown in table 1.1 our sample is clearly not representative. We do not think that this is a problem *per se*. This may however prove problematic for the coastal region as on the one hand we have very few observations for 2012 and on the other hand,

	2011		2012	
	Percentage	Observations	Percentage	Observations
Brussels	14.7%	35	12.26%	25
Flanders	47.9%	114	59.31%	121
Coast	23.95%	57	6.9%	14
Wallonia	13.45%	32	21.53%	44
Total	100%	238	100%	204

Note : Remember that Flanders does not correspond to the Flemish Region.
Sources : Fédération Royale du Notariat Belge and own calculations.

Table 1.1: Composition of the Sample

the number of observations outnumbers those for Wallonia in 2011 even though Wallonia has a much bigger population. Still, no data set is ever perfect and no better data can be found.

Still, this forces us to use dummy variables for different regions. If we applied a single regression for Belgium, we would not actually have results about the "real" Belgium. We would have biased results which would not take Wallonia enough into account for example. In a perfect world, we would have the exact proportions of each regions in the total number of apartments' sales in Belgium but we could not find such numbers. Even our full sample (the $35,614$ observations) seems biased : in 2011, Wallonia represents only 6.28% of the total number of observations.

In the end, we tried the first and the third methods. The first because even though it is difficult to interpret, it is still valuable to make comparisons, and the third because it is the method that makes the most sense to us. The second possibility of applying a regression to each different region does not make much sense since, on the one hand we do not have many observations and on the other hand we do not think that these regions are absolutely separate markets.

1.4 The Variables

Now that we have stated what our modelling strategies are, we can select the appropriate variables and solve the trade-off. Obviously, we could not keep dummies because of the ambivalent nature of 0.[19] Neither could we keep variables for which nearly no observations were available. We also decided not to use the "cadastral income" even though the information was available for nearly all goods. The cadastral income is defined as "the fictional revenue corresponding to the net yearly annual income that a housing good would provide to its owner" (Service Public Fédéral Finances, 2013). We do not go into the details but the definition is very vague and the cadastral incomes are calculated by controllers with some formula - probably full of subjectivity - not available to the public. We know it may be seen as a useful proxy to judge the general quality of the good and to approximate omitted variables but given the lack of information, we prefer not to include it.

After the selection, we ended up with the following variables : price, surface, age, bedrooms, postal code and year of sale. The first four variables are described in tables 1.2 and 1.3 [20]. We have 442 observations.

Price					Surface				
Mean	S.E.	Min.	Max.	Med.	Mean	S.E.	Min.	Max.	Med.
194.1	92.3	70	700	172.25	95.24	33.38	20	240	90

Note : The price is expressed in hundred thousand euros and the surface in square meters.
Sources : Fédération Royale du Notariat Belge and own calculations.

Table 1.2: Descriptive Statistics (1/2)

From the last two variables, we created dummies, a first set to represent the geographical location of the dwelling sold and a second set to take into account the year of sale. This last dummy (TimeDummy) takes the value 1 if the good was sold in 2012 and 0 if it was sold in 2011.

[19] Remember that for dummy variables such as "garage", a 0 means either "no garage" or "no information".
[20] More detailed tables can be found in the appendix.

Bedrooms					Age				
Mean	S.E.	Min.	Max.	Med.	Mean	S.E.	Min.	Max.	Med.
2	0.69	1	5	2	29.5	24.35	0	113	31

Note : Age is expressed in years.

Sources : Fédération Royale du Notariat Belge and own calculations.

Table 1.3: Descriptive Statistics (2/2)

The choice of geographical dummies was more complicated. We first wanted to include a dummy for each province but given the amount of data this was simply impossible. We then turned to formal "Regions" of Belgium. This worked quite well except that we discovered through the Stadim Index (Stadim, 2013) that the prices on the Belgian coast have been growing a lot faster than in the rest of Flanders. This discrepancy is confirmed by our model. In the end, we thus decided to include three dummies to represent four different regions :

1. The Region of Flanders without the coast is represented by the dummy "Flanders" which takes value 1 if the dwelling is located in the defined region and 0 otherwise. Note that except if we mention it explicitly, Flanders will always refer to this definition.

2. The Brussels region is represented by the dummy "Brussels" which takes value 1 if the dwelling is located in the Brussels Region.

3. The coastal region, that is, goods located in cities where the postal code starts with 83, 84 or 86[21]. The dummy "Coast" takes value 1 in this case and 0 otherwise.

4. If all these dummies are 0, the apartment is located in Wallonia.

Of course, the representativeness problem that we discussed earlier also applies here and some cities are overrepresented or underrepresented in the different regions. For example, the city of Anvers accounts for 28% of our sample in 2012. However, given that we do not have an infinity of observations, we have to limit the number of regional dummies.

[21] This is a definition of our own. It is very difficult to find what is meant by "the coast". No such definition is given by the Stadim index for instance. See the map in the appendix to understand what our definition corresponds to.

This fact has to be kept in mind when we will interpret our regression : it is very (almost certain) possible that we do not have a representative samples and our results could thus be biased.

Descriptive statistics for the regional and the year dummies are in table 1.4. Given that the variables are dummies, the means represent proportions. It can thus again be seen that our sample is not geographically representative. We have slightly more observations for 2012 (54%) than for 2011 (46%).

Flanders	Brussels	Coast	TimeDummy
0.53167	0.13575	0.16063	0.46154

Note : The proportion of observations from Wallonia is $1 - 0.532 - 0.136 - 0.161 = 0.171$.
Sources : Fédération Royale du Notariat Belge and own calculations.

Table 1.4: Means of the Dummies

1.5 Interesting Statistics

More interesting statistics can be produced from the decomposition of each variable into region and year of sale. The full tables are in the appendix for size reasons but the means are in table 1.5.

What can we say ?[22] First, regarding prices, it seems that the most expensive places to buy an apartment are Brussels and the coast. This is not really surprising. What is a bit more striking is that Walloon apartments appear to be more expensive than Flemish ones once the coast has been excluded. Our distinction of the coast and Flanders thus seems appropriate.

Secondly, more spacious apartments are located in Wallonia while the smallest ones are on the coast. But, excluding the coast, the differences between the three regions are not

[22]Note that these remarks are made about the data we kept for our regression. For a more general analysis of the whole dataset, see Fédération Royale du Notariat Belge (2013).

	All	Brussels	Flanders	Coast	Wallonia	2011	2012
Price	194.1	221.4	181.7	221.3	185.5	190.1	198.8
Surface	95.24	98.73	98.7	73.99	101.62	95.27	95.2
Nb. Bedrooms	2	1.917	2	2	2.08	2.05	1.96
Age	29.5	50.6	29.16	22.64	20.29	26.95	32.47

Notes : The columns All, 2011 and 2012 contain the four different regions. The price is expressed in hundred thousand euros.

Sources : Fédération Royale du Notariat Belge and own calculations.

Table 1.5: Means for Different Regions and Years

major. Regarding age, it seems that the apartment stock of the Brussels region is the oldest, the most recent being in Wallonia. This conclusion concurs with other studies' findings (Goossens, Thomas, and Vanneste, 2005).

We also have a possible foretaste of our model's conclusions. Indeed, the mean price seems to have increased between 2011 and 2012. On the other hand, all the characteristics have stayed put, except for the age : apartments sold were older in 2012 than in 2011. We can not say more from the descriptive statistics but a formal model might help.

2 The Regressions

2.1 The Regression with the Regions

Now that we have the variables, we can proceed with the first hedonic regression, where regions are included. We follow the general-to-specific approach also known as the LSE methodology (Hendry, 2002). We include the squares of each variable - those for which it makes sense at least - and the cross-products. Given that we have a limited number of variables and a relatively large sample, it is not problematic. Moreover, interaction variables are especially important in this context to check for the constancy of the parameters to be able to use the direct method.

Note also that the relevant independent variables are expressed in deviation from the mean. This will greatly facilitate the interpretation since we have interaction terms. The mean is calculated on the pooled sample, that is, on the 442 observations.

We choose a log-lin specification because a lin-lin model clearly displays heteroskedasticity.[23] Indeed, it is to be expected that the variance of the residuals will increase with the price of an apartment. Moreover, it is not a proof but most hedonic studies use a log-lin model. Concerning our model, we are not sure that the residuals are heteroskedastic. Looking at table 2.1[24], the tests are contradictory. The White tests (White, 1980) show heteroskedasticity while the Koenker test (Koenker, 1981) indicates homoskedasticity. Moreover, after using a robust variance-covariance matrix,[25] "Bedrooms" became significant at the 10% level, which it should *a priori* be. Other more informal tests were performed and can be found in the appendix (table and figure E.1).

We eliminated irrelevant variables by checking the p-values and the R^2 and by minimizing the information criteria. We also took into account theoretical *a priori* knowledge. In the end, we ended up with the regression in table 2.2 (see the appendix for more detail about our methodology). All variables are significant except Bedrooms^2, TimeDummy

[23] White tests and Koenker tests both have a 0 p-value.

[24] The tests were obviously performed on the model without the robust variance-covariance matrix.

[25] We use the HC1 variant which takes into account the number of variables and the number of observations : $var(\hat{\beta}) = \frac{N}{N-k}(X'X)^{-1}X'\hat{\Omega}X(X'X)^{-1}$ where N is the number of observations and k the number of regressors. Other variants were tried but did not change the results much.

Test	Statistic	P-value
White (cross products and squares)	141.11	0.0209
White (squares only)	43.62	0.0516
Koenker	24.68	0.214

Note : The null hypothesis is "residuals are homoskedastic" for all the tests.
Source : Own calculations and Fédération Royale du Notariat Belge.

Table 2.1: Heteroskedasticity Tests for the Log-Lin model

and TimeDummy×Flanders. We kept them nonetheless : for Bedrooms^2 because there are *a priori* reasons to believe that it may be significant "in reality", for the other two because they are variables of interest. Moreover, adding one too many variable is better than omitting an important one.[26]

We summarize the effects we expect and the effects we actually get in table 2.3. All the signs are as expected. We do not report cross variables because most of them are not of interest. Still, we must remark that strange results appear. For instance, we do not really understand why "Bedrooms×Coast" has such a high negative coefficient. Another thing to note is the different effect of the variable "Age" in the different regions. This can be seen from the values of the coefficients of the cross variables of the regions and "Age".

2.1.1 Interpretation of Some Variables

We will first interpret our regression mathematically and rigorously for the variables surface and age - the same thinking can be applied to the other variables - then simple

[26]Omitting a variable makes OLS biased while including a useless variable renders OLS less precise.

Variable	Coefficient	Standard Error	t-ratio
Constant	11.8101***	0.0431	273.76
Surface	0.0095***	0.0007	14.37
Bedrooms	0.0449*	0.0250	1.79
Age	-0.0093***	0.0014	-6.71
Surface^2	-0.00002***	8.03×10^{-6}	-2.62
Bedrooms^2	-0.0093	0.0159	-0.59
Age^2	0.0001***	1.87×10^{-5}	7.00
Brussels	0.3216***	0.0787	4.08
Flanders	0.1263***	0.0475	2.66
Coast	0.4784***	0.0713	6.72
TimeDummy	0.0250	0.0597	0.42
Surface$\times$Flanders	-0.0026***	0.0009	-3.01
Bedrooms$\times$Coast	-0.1825***	0.0561	-3.25
TimeDummy$\times$Brussels	0.1708*	0.0976	1.75
TimeDummy$\times$Flanders	0.0291	0.0672	0.43
TimeDummy$\times$Coast	0.3838***	0.1309	2.93
Brussels$\times$Age	0.0038**	0.0018	2.09
Flanders$\times$Age	0.0042***	0.0013	3.22
Coast$\times$Age	0.0066**	0.0033	2.00
TimeDummy$\times$Age	-0.0033**	0.0014	-2.38

Mean Dep. Var.	12.0818	S.D. Dep. Var.	0.4253
SSR	29.6368	S.E. of Reg.	0.2650
R^2	0.6285	Adjusted R^2	0.6117
Akaike Criterion	99.929	P-Value(F)	3.32×10^{-89}
Schwarz Criterion	181.755	Hannan-Quinn	132.203

Notes : The dependent variable is $\ln$(Price). We use the HC1 covariance matrix. ***, ** and * denote respectively significance at the 1%, 5% and 10% levels. N=442. Appropriate variables are demeaned.

Source : Own calculations and Fédération Royale du Notariat Belge.

Table 2.2: Hedonic Regression for Belgium With Regions

Variable	Expected	Actual
Surface	$+$	$+$
Age	$-$	$-$
Bedrooms	$+$	$+$
Surface^2	$-$	$-$
Bedrooms^2	$-$	$-$
Age^2	$+$	$+$
Brussels	$+$	$+$
Flanders	$+$	$+$
TimeDummy	$+$	$+$

Source : Own calculations and Fédération Royale du Notariat Belge.

Table 2.3: Signs of the Variables

examples will be given. Let us write our model explicitly :

$$\begin{aligned}\ln(\text{Price}_i) =&\beta_0 + \beta_1\text{Surface}_i + \beta_2\text{Bedrooms}_i + \beta_3\text{Age}_i + \beta_4\text{Surface}_i^2+\\ &\beta_5\text{Bedrooms}_i^2 + \beta_6\text{Age}_i^2 + \beta_7\text{Brussels}_i + \beta_8\text{Flanders}_i + \beta_9\text{TimeDummy}_i+\\ &\beta_{10}(\text{Surface}_i \times \text{Flanders}_i) + \beta_{11}(\text{Bedrooms}_i \times \text{Coast}_i)+\\ &\beta_{12}(\text{TimeDummy}_i \times \text{Brussels}_i) + \beta_{13}(\text{TimeDummy}_i \times \text{Flanders}_i)\\ &+ \beta_{14}(\text{TimeDummy}_i \times \text{Coast}_i) + \beta_{15}(\text{Brussels}_i \times \text{Age}_i) + \beta_{16}(\text{Flanders}_i\\ &\times \text{Age}_i) + \beta_{17}(\text{Coast}_i \times \text{Age}_i) + \beta_{18}(\text{TimeDummy}_i \times \text{Age}_i) + u_i\end{aligned}$$

The equation can also be interpreted as the "hedonic equation" defined in the theoretical part. Because of the interaction terms, the interpretation of the regression is a bit more complicated. The marginal effect of the variable "surface" on the logarithm of the price is

$$\frac{\partial \ln(\text{Price}_i)}{\partial \text{Surface}_i} = \beta_1 + 2\beta_4\text{Surface}_i + \beta_{10}\text{Flanders}_i$$

Since the variables are demeaned and we have a log-lin model, β_1 can be interpreted as the semi-elasticity of the price evaluated at a point where the surface is equal to the mean surface of the sample and the apartment is not located in Flanders. Note that this semi-elasticity is not constant. Since β_4 is negative, the bigger the apartment - with regards to the mean surface - the less the surface has an impact on the price. Because β_{10} is positive, the semi-elasticity of the price with regards to the surface is amplified if the dwelling is located in Flanders.

For the variable "age", we have :

$$\frac{\partial ln(\text{Price}_i)}{\partial \text{Age}_i} = \beta_3 + 2\beta_6 \text{Age}_i + \beta_{15}\text{Brussels}_i + \beta_{16}\text{Flanders}_i + \beta_{17}\text{Coast}_i + \beta_{18}\text{TimeDummy}_i$$

Hence, β_3 is the semi-elasticity of the price with regards to the age evaluated for a dwelling which is as old as the mean age of the sample, is located in Wallonia and which is sold in 2011. With an increase in age comes a decrease in price ($\beta_3 < 0$). The decrease becomes less and less important as we get further away from the mean age ($\beta_6 > 0$). If the good is located in another region the semi-elasticity becomes bigger. Finally, it is smaller for a good sold in 2012 than for a good sold in 2011.

Let us give examples in proper English. Say we are in 2011 and we are considering an apartment located in Wallonia which has mean mean characteristics, that is, an apartment with a surface of 95m^2, two bedrooms and which is 30 years old. We round up numbers and ignore cross variables for simplicity. The price varies like this :[27]

1. If we were to add 1m^2 of surface, the price should, on average, go up by 0.95%. The price of the mean good would jump from $176,635$ to $178,313$ euro, a rise of $1,678$ euros.

2. If a bedroom is added, the price goes up by 4.49%, an increase of $7,930$ euros.

3. If the apartment had been built one year earlier, it would sell for 0.93% less, a decrease of 1642 euros.

[27] We use the approximation to interpret directly the parameters. This is not a problem since they are quite small.

These prices are what we referred to in the theoretical review as implicit prices. Interestingly, we can see that Rosen was right in not restricting the hedonic equation to be linear : quadratic terms seem to matter.

2.1.2 The Regional Dummies

In a previous section, we said that "location, location, location" may be an appropriate motto for the real estate market. Unfortunately, due to the impreciseness of our data, the best we could do was to distinguish four different regions. In 2011, according to the "baromètre des notaires" (Fédération Royale du Notariat Belge, 2013), apartments in the Flemish Region and Brussels sold for 38% and 39.1% more than in Wallonia. Unfortunately, they include the coastal region in "Flanders" and distinctive figures are not available. Hence their numbers about the Flemish Region are biased upwards compared to our "Flanders". As has been explicitly shown in the theoretical review, this information is like comparing apples and oranges. What is more interesting is to consider what happens to the price if we compare apartments which are similar in quality and characteristics but which are located in different places, in other words, what is the price of being located somewhere.

First, we can note that all simple regional dummies are significant at the 1% level, which implies once again that location is crucial. The exact same good - with regards to the characteristics for which we control - will sell for a different price if located in a different region.

If we focus on a good with mean qualities, an apartment in Flanders should fetch a price which is 13.47%[28] higher than in Wallonia. For a given mean good, the increase in price is thus lower than the one computed by the notaries, which unfortunately does not really cover the same area. This discrepancy between our numbers and those of the notaries would imply that apartments in Flanders are of a better quality. Indeed, if we make the assumption that the prices reflect quantities of characteristics, the location would only explain 13.47% of the discrepancy and the rest would be due to differences in quantities of

[28]From here on, we will use the exact formula for dummy variables as discussed in the first part of the master's thesis. We will not however apply the bias correction proposed by Kennedy because it changes nearly nothing.

characteristics. Of course, it is very difficult to support any kind of definitive conclusion since on the one hand, the areas covered are different and on the other hand our confidence intervals are very big (see table F.1 in the annex). This last caveat is valid for all regions.

In Brussels, for the mean good, the price difference at given quality is 37.93%. This increase is nearly the same as that of the notaries. It means that a very big part of the increase in price between Wallonia and Brussels can be explained by location only and not by differences in quality.

The most impressive increase is for the coast with 61.35%. Is this too high ? According to the Stadim Index Stadim (2013), it would not be impossible. They calculate that between 1968 and 2011, prices on the coast have been multiplied by 12.8 while prices in Wallonia have been multiplied by 7.9 only. This is a bad approximation and we could not find exact numbers. Moreover, since Flanders for the notaries is the (weighted) sum of our "Flanders" and "Coast, the figure seems appropriate. Ideally, prices in all studies should include multiple separate regions but usually only Wallonia, Brussels and Flanders are mentioned. Location on the coastal region thus seems worth a lot of money even though our estimate must be taken with care.

2.1.3 The Hedonic Price Index

The main objective of this master's thesis is the application of the hedonic method to analyse the evolution of the prices of apartments between 2011 and 2012 in Belgium. The regression we have performed allows us to do that. We use the direct method as described in a previous section since we only have two time periods. Table 2.4 summarizes the hedonic price index for each region and the confidence intervals. Note that we use the exact formula for dummy variables and for the confidence intervals.

The first thing we notice is that the variables "TimeDummy" is not significant at the 10% level. The same is true for Flanders since the sum of "TimeDummy" and "TimeDummy $\times$ Flanders" is also not significant at 10%. Note however that the p-value is much lower than for Wallonia. Should we interpret that as "prices have not increased in Wallonia and Flanders at given quality between 2011 and 2012" ? On the one hand, we should because the coefficients are insignificant. On the other hand, confidence intervals are so big that

Region	Change 2011-2012 (%)	Lower Bound(%)	Upper Bound(%)	P-value
Wallonia	2.53	-10.499	15.394	0.68
Flanders	5.56	-0.373	13.100	0.12
Brussels	21.63	5.897	39.691	0
Coast	50.50	19.76	89.141	0

Notes : We use the exact formula for dummy variables for both coefficients and intervals. Figures for Wallonia and Flanders should be taken with care since the variables are insignificant at 10%.
Source : Own calculations and Fédération Royale du Notariat Belge.

Table 2.4: Hedonic Price Indices and 95% Confidence Intervals

we may have to admit that we do not know anything in the end. We choose to consider that the "true" value for Wallonia is 0% and 5.56% for Flanders, given that its p-value is nearly at 0.10 and that its confidence interval is nearly all above 0%.

For the Brussels region the increase from 2011 to 2012 is indicated by the sum of the coefficient of the variable "Brussels$\times$TimeDummy" and "TimeDummy" (with the appropriate correction). Prices, at given mean qualities have increased by 21.63%. Is this surprising ? Yes and no. Yes because the number seems too high. It is difficult to believe that prices can vary so much in one year. No because the confidence interval is very big. Actually, at the 95% level, the increase could be anywhere between 5.9 and 39.7%. We can not say much from that unfortunately.

The same is true for the coast. The increase is 50.5% which again seems unbelievable. The confidence interval is also very large : 19.76% and 89.14%. This is worse since even the lower bound is improbable. The reason may be that in 2012 we only have 14 observations, which implies very big standard errors and a probable bias in the sample.

Can we really say anything from this ? The first problem seems to be that confidence intervals are too big, hence whatever we say is bound to be extremely general. Secondly, there seems to be a problem with the coastal region, probably too few observations in 2012, which renders the value of the coefficient highly unlikely. Still, if we had to make

a conclusion it would be that, at given quality, prices have probably increased in Flanders, the coast and Brussels and have stayed constant in Wallonia.

2.2 The Regression for Belgium

Let us now try to fit a regression for Belgium as a whole to be able to compare our results with other studies. The preliminary steps are the same as for the precedent regression and we do not go over them again. The regression results are in table 2.5. No heteroskedasticity was found in the residuals of the log-lin model[29] and all the signs of the coefficients are as expected.

All variables are significant at least at the 10% level or less except for the square of bedrooms and, more interestingly, for TimeDummy. Hence, from that regression only, our conclusion should be that there has been no price increase, at given quality, for apartments in Belgium between 2011 and 2012. Is it consistent with our previous regression ? We concluded that it was probable that, at given quality, prices had increased in all regions except Wallonia. Still, we do not think the models are necessarily contradictory. Firstly, once again, the confidence interval[30] is quite big : it ranges between -0.03 and 0.094. Secondly, our previous conclusion was very precarious, especially for Flanders. Since Flanders represents the bulk of our sample, if the conclusion is erroneous there, it probably also is for the whole of Belgium. Hence, given that the model did not formally detect a price change in Wallonia and Flanders, it is normal that it will not detect one for the whole of Belgium since Wallonia and Flanders represent a big part of the observations.

What do we conclude from this regression in the end ? We have to concede that, if we look at this model and at this model only, no price change can be detected, at given quality, in Belgium between 2011 and 2012 for apartments.

On a slightly more positive note, this reinforces our thinking that fitting a model to Belgium with regional dummies is the best solution. Not including the dummies aggregates the price index in such a way that some local price changes become invisible.

[29] P-values are 0.06 (White Test) 0.56 (White test with squares only) and 0.45 (Koenker). P-values for the lin-lin model were all below 1%.

[30] See the appendix for the full table with all the confidence intervals of the regression.

Variable	Coefficient	Standard Error	t-ratio
Constant	12.1251***	0.02758	435.9
Surface	0.00625***	0.00067	9.35
Bedrooms	0.05008*	0.02815	1.78
Age	-0.00496***	0.00101	-4.92
Surface^2	-0.00002**	9.22×10^{-6}	-2.55
Bedrooms^2	-0.02001	0.02632	-9.76
Age^2	0.00011***	2.17×10^{-5}	5.14
TimeDummy	0.03192	0.0317	1.01
TimeDummy$\times$Age	-0.00276**	0.0014	-2.03
Surface$\times$Age	0.000003*	0.000002	1.75
Surface$\times$Bedrooms	0.00167**	0.0008	1.98

Mean Dep. Var.	12.0818	S.D. Dep. Var.	0.4253
SSR	44.1688	S.E. of Reg.	0.3201
R^2	0.4463	Adjusted R^2	0.4334
Akaike Criterion	258.2866	P-Value(F)	1.83×10^{-49}
Schwarz Criterion	303.2910	Hannan-Quinn	276.0376

Notes : The dependent variable is $\ln$(Price). $N = 442$. ***, ** and * denote significance at the $1, 5$ and 10% levels. Appropriate variables are demeaned.
Source : Own calculations and Fédération Royale du Notariat Belge.

Table 2.5: Hedonic Regression for Belgium without Regions

Another striking feature of the regression is that some new cross variables ("Surface $\times$ Age" and "Surface $\times$ Bedrooms") are now significant. It is difficult to interpret this result but it may be that the "true" effect actually comes from the regions since when we add them, these cross variables become insignificant.

The interpretation of the other coefficients is nearly the same as before[31]. At given mean characteristics, an apartment with an additional bedroom is expected to sell for 5% more. An additional square meter of surface, *ceteris paribus* and with mean characteristics, will increase the expected price by 0.6%. Other parameters can be interpreted identically.

[31]We use the approximation to interpret the coefficients directly. Again, this is not a problem since coefficients are small.

We do not go further in the interpretation of this model since we only performed it as a comparison basis.

2.3 Comparison of the Two Regressions and the Answer to the First Question

We summarize our two regressions in table 2.6. All the figures - except for the price indices - are for 2011. Some results are quite surprising. For instance, we might have expected the implicit price of the surface to be higher in Brussels than in other areas. Moreover - as we mentioned before - we can not explain the negative impact of an additional bedroom for an apartment on the coast. The coefficients for Belgium are very close to those of Flanders. It is not surprising since Flanders represents the lion's share of the observations.

Given the different results of the hedonic price indices, the answer to our first question "have prices of Belgian apartments changed between 2011 and 2012 at given quality ?" seems elusive. At least one thing is sure : prices, at given quality, have increased in Brussels and on the coast. The usual caveats apply : the sample is not representative, the confidence intervals are big, etc. But even with these problems, we will take this conclusion for granted for the rest of our research.

Variable	Wallonia	Flanders	Brussels	Coast	Belgium
Surface	0.00951	0.0069	0.00951	0.00951	0.00625
Bedrooms	0.04481	0.04481	0.04481	-0.1376	0.05008
Age	-0.00932	-0.00512	-0.00552	-0.00272	-0.00496
Price Index	2.5%	5.6%	21.63%	50.5%	3.22%

Notes : Price Index is calculated with the exact formula for dummies, it is not a variable *per se*. Remember that only the coefficients of Brussels and the Coast are significant at 10%.
Source : Own calculations and Fédération Royale du Notariat Belge.

Table 2.6: Comparison of the Two Regressions

Even though we are not sure, we will also assume that there may have been a slight increase in the prices of apartments in the region we call "Flanders", at given quality. Finally, given the huge confidence interval and the insignificance of the variable we will consider that we can not say anything about Wallonia. If we make these conclusions we also have to conclude that Belgian prices of apartments have probably increased between 2011 and 2012 at given quality.

We think that given the problems of the data set and the model, we can only give these qualitative conclusions. In our comparison with other price indices we will use the values of the coefficients, but it has to be borne in mind that we use them as default values and that we put more trust in our qualitative conclusions than in the quantitative ones.

3 Comparison

Now that we have hedonic price indices, we can attempt to answer the second question : how should we understand the statement "prices of apartments have changed in Belgium between 2011 and 2012" ? Our comparison will be twofold. First, we will compare our figures about implicit prices with those of Decoster and De Swerdt (2005). Then, we will try to answer the second question by looking at how our hedonic price index and the price index of the notaries fare.

3.1 Comparison With Other Studies

Of course, comparing our results with those of other studies is not an easy task to perform : the variety of hedonic studies is impressive and researchers use different functional forms, different characteristics, different methods, etc. For example, we can not compare our results with the hedonic index of France (Laferrère, 2005) because their dependent variable is price per square meter. We can not use numbers from the Spanish paper (Bilbao, Bilbao, and Labeaga, 2010) either as they use a linear specification. We thus focus on the previous hedonic study of Belgium (Decoster and De Swerdt, 2005). We only use the results of our regression for Belgium since they do not have results by region.

Table 3.1 summarizes the main coefficients of the two studies. The first column refers to our regression without the dummies and the second column refers to the study of Decoster and De Swerdt (2005). Unfortunately, the figures of Decoster and De Swerdt are for 1997 while ours are for 2011. Moreover, they consider all types of housing goods while we only consider apartments.

It is difficult to say much from this limited amount of information. However, three facts stand out. First, the signs of the different coefficients are the same. Secondly, the values do not seem to diverge too much. This is quite surprising given the very different variables used. Indeed, we include regional variables and not too many intrinsic characteristics while Decoster and De Swerdt use a lot of the latter such as the presence of a garden, of a garage, the type of building, the number of toilets, etc. Moreover - as we said - their study focuses on real estate in general while we only focus on apartments. Finally, we do not report in the table but the R^2 of our regression with the dummies is 0.6285, which is very

Variable	Belgium (2011)	Belgium(DC)
Surface	0.00625	0.0022
Bedrooms	0.05008	0.0631
Age	-0.00496	-0.0134
Surface^2	-2.34$\times 10^{-5}$	-9.04$\times 10^{-7}$
Age^2	0.00011	7.49$\times 10^{-5}$
R^2	0.4463	0.6391

Note : Price Index is calculated with the exact formula for dummies, it is not a variable *per se*. DC stands for Decoster.
Sources : Own calculations, Fédération Royale du Notariat Belge and Decoster and De Swerdt (2005)

Table 3.1: Comparison With the Results of Decoster and De Swerdt

close to the R^2 of Decoster and De Swerdt. It is striking that even though they have much more information than we do about intrinsic characteristics, we are nearly able to reach the same R^2 by adding regional dummies. Location thus seems once again crucial.

3.2 The Answer to the Second Question

Let us now turn to the main result of this master's thesis, the comparison of our hedonic price index with the two most cited price indices : the index of the notaries[32] and the Stadim Index. The main figures are in table 3.2.

It is immediately apparent that the indices do not agree. Notwithstanding the hedonic index, even traditional price indices have very different numbers. We have no idea why this is but we do know that the Stadim index has twice as many observations as the notaries. However, Stadim does not reveal its methodology. For example, we were not able to find what they mean exactly by "Coast". Still, it is very troubling that the two indices have such differences.

[32]The price index for the notaries has been computed by us with Fédération Royale du Notariat Belge (2011) and Fédération Royale du Notariat Belge (2013)

	Wallonia	Flanders	Brussels	Coast	Belgium
Hedonic Price Index	2.5%	5.6%	21.63%	50.05%	3.22%
Notaries' Index	1.3%	-2.54%	4.4%	/	-3.12%
Stadim Index	6.6%	5.55%	6.76%	0%	4.7%

Notes : Recall that Flanders does not refer to the same region for the two indices. The Coast also possibly refers to different zones.
Sources : Stadim, Fédération Royale du Notariat Belge, own calculations.

Table 3.2: Price Change Between 2011 and 2012

For Belgium, Flanders, and Wallonia, it seems that our model is not too improbable. Given the big differences in traditional price indices, we have to choose with which we want to compare. We have chosen the index of the notaries because their methodology is clearer but we certainly do not want to imply that their index is better than Stadim's. Both have major problems.

If we make the assumption that our model is well specified, we have to conclude that changes of quality account for a big part of the changes in prices of the notaries. We do not mean to say that the quality of Belgian apartments changes rapidly but it is possible that the stock considered each year is of different quality.

What sort of situation are we talking about ? In 2011, the growth of GDP per capita in Belgium was +0.9% while in 2012 it was -0.9% (Eurostat, 2013). A rapid and candid conclusion would be that people were poorer in 2012 that 2011. Hence, they would have bought less expensive goods in 2012 or in our case, less expensive apartments. The price index of the notaries should have recorded that as a diminishing mean price of apartment prices, which it did. Should we conclude that the price of apartments has diminished ? At given quality, the notaries can not give an answer. We have however concluded earlier that the probability that prices of apartments had slightly increased between 2011 and 2012, at given quality, was high. Hence, the diminishing price registered by the notaries' index stems from people buying apartments of a lesser quality, not from diminishing prices of characteristics. In other words, if we considered the same stock of apartment in 2011 and 2012, the mean price should be a bit *higher*.

The same is true for Flanders. The notaries find a decrease in price while we find an increase.[33]

We have said that we can not draw any conclusion for Wallonia. It is also very difficult to comment for Brussels since our figure seems off the track.

If our conclusions for Belgium and Flanders are true, they have important ramifications. Indeed, they imply that the prices on the real estate market for apartments not only have decreased on average, but that they have done so even though, at constant quality, prices have increased. Hence, the "true" state of the real estate market for apartments was even worse than it looked. It also adds more weight to the argument that the evolution of the prices of the real estate market can not be understood without hedonic price indices.

[33] It is not easy to make a weighted average of Flanders and the coast because the choice of the weight is quite difficult. However, an increase will be obtained for any weighting.

Part IV

Conclusion

The goal of this study was to provide a hedonic price index for apartments in Belgium and to compare it with traditional price indices. Three conclusions can be drawn.

Firstly, if a hedonic price index has to be built, economists need better data. We not only need better data about intrinsic characteristics of goods but also - more importantly - about locational aspects. For instance, we were able to achieve nearly the same R^2 as other studies, which have much more characteristics, simply by including some locational variables. We believe that adding data such as the distance from the capital may be beneficial. This is even truer if this data is incorporated in the form of spatial econometric models.

Secondly, we hope to have proven that a hedonic price index for Belgium is indeed needed. It is simply not possible to say anything meaningful from traditional price indices if they are not supplemented by hedonic ones. Indeed, our conclusion clearly shows that studying the evolution of the real estate market without taking into account quality is nearly useless.

Finally, even though our model and our dataset are not the most robust, we still have qualitative and quantitative conclusions. At given quality, prices have increased in the coastal region, in Brussels and - with less certainty - in what we call Flanders. We can not conclude anything for Wallonia unfortunately. Since Flanders constitutes the bulk of apartments transactions, it is probable that prices of apartments, in Belgium and between 2011 and 2012, have increased at given quality. We do not give much credit to our quantitative results, but if we had to summarize them we would say that the increase was quite strong on the coast and in Brussels and was mild in Flanders and Belgium. These results are important because after comparing them with other price indices, we were able to provide some evidence about the state of the real estate market : it is in a much worse state than traditional price indices indicate.

We believe these conclusions show the indisputable need for a hedonic price index for Belgium and, as a result, the need for better data concerning the housing market.

Part V

Bibliography

1 Books

ARNOTT, R. (1987): "Economic theory and housing," in *Handbook of Regional and Urban Economics*, ed. by E. S. Mills, vol. 2, chap. 24, pp. 959–988. Elsevier.

BUTLER, R., G. GASTLER, J. PITKIN, AND J. ROTHENBERG (1991): *The Maze of Urban Housing Markets*. University of Chicago Press, Chicago.

COURT, A. (1939): "Hedonic Price Indexes with Automotive Examples," in *The Dynamics of Automotive Demand*, ed. by C. Roos, pp. 99–117. General Motors, New-York.

DAVIDSON, R., AND J. G. MACKINNON (1993): *Estimation and Inference in Econometrics*, OUP Catalogue. Oxford University Press.

EUROSTAT (2008): *European Price Statistics*. Office for Official Publications of the European Communities.

——— (2013): *Handbook on Residential Property Prices Indices*. Office for Official Publications of the European Communities.

GOOSSENS, L., I. THOMAS, AND D. VANNESTE (2005): *Monographie sur les caractéristiques spatiales et sociologiques du Logement*. I.N.S, Bruxelles.

GRANELLE, J.-J. (1997): "Le logement comme bien de consommation [Housing as a Consumption Good]," in *Comprendre les marchés du logement*, ed. by B. Coloos, F. Calcoen, J.-C. Driant, and B. Filippi, pp. 25–59. L'Harmattan, Paris.

GRILICHES, Z. (1961): "Hedonic Price Indexes for Automobiles: An Econometric of Quality Change," in *The Price Statistics of the Federal Goverment*, ed. by T. P. S. R. Comittee, pp. 173–196. National Bureau of Economic Research.

LAFERRÈRE, A. (2005): "Hedonic housing price indexes: the French experience," in *Real estate indicators and financial stability*, ed. by B. for International Settlements, vol. 21 of *BIS Papers Chapters*, pp. 271–287. Bank for International Settlements.

LeSage, J., and R. Kelley Pace (2009): *Introduction to Spatial Econometrics*. CRC Press.

McCall, B., and J. McCall (2008): *The Economics of Search*. Routledge, London and New-York.

Mikosch, T. (1998): *Elementary Stochastic Calculus with Finance in View*. World Scientific Publishing, Singapore.

Muth, R. (1969): *Cities and Housing : the spatial pattern of urban residential land use*. University of Chicago Press, Chicago.

Schumpeter, J. (1942): *Capitalism, Socialism and Democracy*. Routledge.

Sheppard, S. (1999): "Hedonic analysis of housing markets," in *Handbook of Regional and Urban Economics*, ed. by P. C. Cheshire, and E. S. Mills, vol. 3 of *Handbook of Regional and Urban Economics*, chap. 41, pp. 1595–1635. Elsevier.

Triplett, J. (2006): *Handbook on Hedonic Indexes and Quality Adjustments in Price Indexes*. OECD Publishing.

Whitehead, C. M. (1999): "Urban housing markets: Theory and policy," in *Handbook of Regional and Urban Economics*, ed. by P. C. Cheshire, and E. S. Mills, vol. 3, chap. 40, pp. 1559–1594. Elsevier.

2 Scientific Articles

Andersson, H., L. Jonsson, and M. Ogren (2008): "Property Prices and Exposure to Multiple Noise Sources: Hedonic Regression with Road and Railway Noise," Lerna working papers, LERNA, University of Toulouse.

Anselin, L., and J. L. Gallo (2006): "Interpolation of Air Quality Measures in Hedonic House Price Models: Spatial Aspects," *Spatial Economic Analysis*, 1(1), 31–52.

Bartik, T. (1987): "The Estimation of Demand Parameters in Hedonic Price Models," *Journal of Political Economy*, 95(1), 81–88.

BEENSTOCK, M., AND D. FELSENSTEIN (2008): "Testing Spatial Stationarity and Spatial Cointegration," Discussion paper, Hebrew University of Jerusalem, Technical Report.

BILBAO, C., A. BILBAO, AND J. LABEAGA (2010): "The Welfare Loss Associated to Characteristics of the Goods : Application to Housing Policy," *Empirical Economics*, 38, 305–323.

BOX, G., AND D. COX (1964): "An Analysis of Transformations," *Journal of the Royal Statistical Society*, 26(2), 211–252.

BROWN, J. N., AND H. S. ROSEN (1982): "On the Estimation of Structural Hedonic Price Models," *Econometrica*, 50(3), 765–68.

CAN, A., AND I. MEGBOLUGBE (1997): "Spatial Dependence and House Price Index Construction," *Journal of Real Estate Finance and Economics*, 14, 203–222.

CHAN, P., Y.-S. CHO, AND H. MOK (1995): "A Hedonic Price Model for Private Properties in Hong Kong," *The Journal of Real Estate Finance and Economics*, 10(1), 37–48.

COLWELL, P. F., AND G. DILMORE (1999): "Who Was First? An Examination of an Early Hedonic Study," *Land Economics*, 75(4), 620–626.

CORRADO, L., AND B. FINGLETON (2011): "Where is the Economics in Spatial Econometrics?," Working Papers 1101, University of Strathclyde Business School, Department of Economics.

CROPPER, M. L., L. B. DECK, AND K. E. MCCONNELL (1988): "On the Choice of Functional Form for Hedonic Price Functions," *The Review of Economics and Statistics*, 70(4), 668–675.

DAVIDSON, R., AND J. G. MACKINNON (2001): "Artificial Regressions," Working Papers 1038, Queen's University, Department of Economics.

DE BRUYNE, K., AND J. VAN HOVE (2013): "Explaining the Spatial Variation in Housing Prices : an Economic Geography Approach," *Applied Economics*, 45(13), 1673–1689.

DECOSTER, A., AND C. DE SWERDT (2005): "Why and How to Construct a Genuine Belgian Price Index of House Sales," *Center for Economic Studies*, (05.15), Discussion Paper.

DUBIN, R. (1998): "Spatial Autocorrelation: A Primer," *Journal of Housing economics*, 7(4), 304–327.

EKELAND, I., J. J. HECKMAN, AND L. NESHEIM (2004): "Identification and Estimation of Hedonic Models," *Journal of Political Economy*, 112(1), S60–109, Part 2 Supplement.

EPPLE, D. (1987): "Hedonic Prices and Implicit Markets : Estimating Demand and Supply Functions for Differentiated Products," *Journal of Political Economy*, 95(1), 59–80.

FINGLETON, B. (1999): "Spurious Spatial Regression : Some Monte Carlo Results with a Spatial Unit Root and Spatial Cointegration," *Journal of Regional Science*, 39(1), 1–19.

HENDRY, D. F. (2002): "Applied Econometrics without Sinning," *Journal of Economic Surveys*, 16(4), 591–604.

KENNEDY, P. E. (1981): "Estimation with Correctly Interpreted Dummy Variables in Semilogarithmic Equations," *The American Economic Review*, 71(4), 801.

KOENKER, R. (1981): "A Note on Studentizing a Test for Heteroskedasticity," *Journal of Econometrics*, 17, 107–12.

LANCASTER, K. J. (1966): "A New Approach to Consumer Theory," *Journal of Political Economy*, 74, 132–157.

LIAO, W.-C., AND X. WANG (2012): "Hedonic house prices and spatial quantile regression," *Journal of Housing Economics*, 21(1), 16–27.

MONTERO, J.-M., G. FERNÁNDEZ-AVILÉS, AND R. MÍNGUEZ (2011): "Spatial Hedonic Pricing Models for Testing the Adequacy of Acoustic Areas in Madrid, Spain," *Investigaciones Regionales*, (21), 157–181.

ROSEN, S. (1974): "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition," *Journal of Political Economy*, 82(1), 34–55.

ROUWENDAL, J., AND J. W. VAN DER STRAATEN (2008): "The Costs and Benefits of Providing Open Space in Cities," Tinbergen Institute Discussion Papers 08-001/3, Tinbergen Institute.

TIEBOUT, C. M. (1956): "A Pure Theory of Local Expenditures," *Journal of Political Economy*, 64, 416–424.

WHEATON, W. (1990): "Vacancy, Search and Prices in a Housing Market Matching Model," *Journal of Political Economy*, 98, 1270–1292.

WHITE, H. (1980): "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity," *Econometrica*, 48(4), 817–38.

3 Non-Scientific Articles

FÉDÉRATION ROYALE DU NOTARIAT BELGE (2011): "Baromètre des Notaires," (11).

——— (2013): "Baromètre des Notaires," (15).

HAAS, G. (1922): "A Statistical Analysis of Farm Sales in Blue Earth County, Minnesota, as a Basis for Farm Land Appraisal," Master's thesis, The University of Minnesota.

SPF ECONOMIE (2011): "Budget des Ménages 2000-2009," .

4 Web Pages

EUROSTAT (2013): "Real GDP per capita, growth rate and totals," `http://epp.eurostat.ec.europa.eu/tgm/table.do?tab=table&plugin=1&language=en&pcode=tsdec100`.

SERVICE PUBLIC FÉDÉRAL FINANCES (2013): "Revenu Cadastral," http://minfin.fgov.be/portail2/fr/themes/dwelling/cadastral-income/.

STADIM (2013): http://www.stadim.be.

Part VI

Appendix

A Full Information

We give here the example of a good for which we have full information, that is, a good for which notaries have filled all the characteristics. We only have a handful of these.

We immediately see some of the problems of the dataset. First, there is no such thing as a "Brussels" province. Secondly, it is not really clear here but the columns "hectares, ares, centiares" were sometimes just different ways of saying the same things and some other times complements of information. For example, here it just repeats that the land is 5 centiares.

Date	**Trimester**	**Price**	**Surface**	**Type**	**Category**
8/23/12	3	140,000	67	Apartment	Apartment
New	**Hectares**	**Ares**	**Centiares**	**Nature**	**Year**
0	0.05	5	0	Private Sale	1961
Garages	**Storey**	**State**	**Garden**	**Cave**	**Attic**
1	1	To Renovate	0	1	0
PostalCode	**City**	**Region**	**Province**	**Arrondissement**	**CI**
1190	Forest	Brussels	Brussels	Brussels-Capital	808
Number	**Elevator**	**Bedrooms**			
24	0	1			

Note : "Year" refers to the year the good was built and CI to the cadastral income.
Source : Fédération Royale du Notariat Belge.

Table A.1: Full Data for One Apartment

B Descriptive Statistics

	Wallonia 2011					Wallonia 2012				
	Mean	S.E.	Min.	Max.	Med.	Mean	S.E.	Min.	Max.	Med.
Price	176	86.3	75	546	152.5	192.4	69.9	70	300	213.9
Surface	107.3	40.8	58	225	100	97.5	22.1	50	162	100
Bedrooms	2.1	0.64	1	3	2	2.07	0.55	1	3	2
Age	23.7	18.9	0	61	23	17.8	24.26	0	112	8

Note : The price is expressed in hundred thousand euros.
Sources : Fédération Royale du Notariat Belge and own calculations.

Table B.1: Descriptive Statistics for Wallonia

	Flanders 2011					Flanders 2012				
	Mean	S.E.	Min.	Max.	Med.	Mean	S.E.	Min.	Max.	Med.
Price	179.2	78.2	72.5	70	165	184.1	85.5	79	555	164
Surface	100.2	34.8	20	220	93.5	97.3	33	44	240	90
Bedrooms	2.1	0.68	1	5	2	1.93	0.7	1	3	2
Age	23.7	19.6	0	80	21	34.3	26.8	0	113	37

Notes : The price is expressed in hundred thousand euros. Remember that Flanders is not the Flemish Region.
Sources : Fédération Royale du Notariat Belge and own calculations.

Table B.2: Descriptive Statistics for Flanders

	Brussels 2011					Brussels 2012				
	Mean	S.E.	Min.	Max.	Med.	Mean	S.E.	Min.	Max.	Med.
Price	220.6	122.8	100	535	170	222.7	95.6	105	405	197
Surface	104.5	36.5	50	181	91	90.7	31.4	46	163	90
Bedrooms	1.9	0.74	1	3	2	1.9	0.81	1	3	2
Age	51.4	21.6	6	101	50	49.5	27.4	4	112	44

Note : The price is expressed in hundred thousand euros.
Sources : Fédération Royale du Notariat Belge and own calculations.

Table B.3: Descriptive Statistics for Brussels

	Coast 2011					Coast 2012				
	Mean	S.E.	Min.	Max.	Med.	Mean	S.E.	Min.	Max.	Med.
Price	200.1	93.4	70	500	180	304	142	150	635	260
Surface	73	29.9	23	158	69	78	15.2	54	111	78
Bedrooms	2	0.78	1	3	2	2	0.55	1	3	2
Age	20.9	15	0	48	17	32.2	15.7	3	61	32.2

Note : The price is expressed in hundred thousand euros.
Sources : Fédération Royale du Notariat Belge and own calculations.

Table B.4: Descriptive Statistics for the Coast

	Belgium 2011					Belgium 2012				
	Mean	S.E.	Min.	Max.	Med.	Mean	S.E.	Min.	Max.	Med.
Price	190.1	91.5	70	700	105	198.8	93.3	70	635	185
Surface	95.3	36.9	20	225	90	95.2	30.1	44	240	90
Bedrooms	2.05	0.7	1	5	2	1.96	0.67	1	3	2
Age	26.95	21.34	0	101	28	32.47	27.2	0	113	32.5

Note : The price is expressed in hundred thousand euros.
Sources : Fédération Royale du Notariat Belge and own calculations.

Table B.5: Descriptive Statistics for Belgium

C Map of Postal Codes

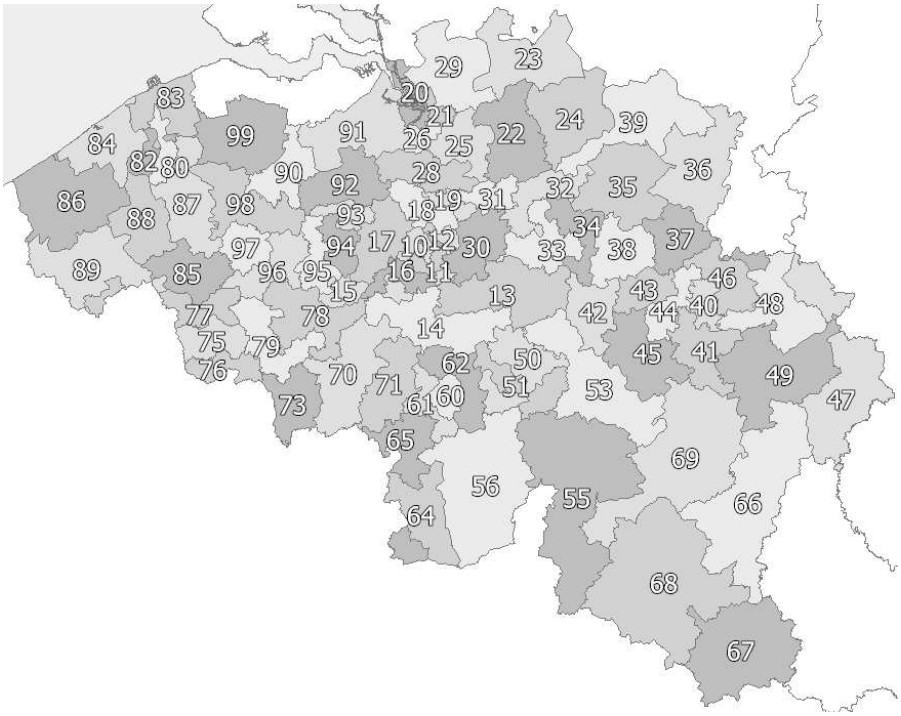


Table C.1: Map of Belgian Postal Codes. Source : GfK GeoMarketing

D Methodology For Our Choice Of Variables In The Regression

As mentioned before, we followed the LSE methodology in choosing our variables, that is, we estimated the most general model and tried to reduce its size with various criteria. The general regression is in tables D.1 and D.2.

Firstly, we kept all variables of interest (the time dummy, regional dummies and the interactions of both) and variables which *a priori* determine housing prices (intrinsic characteristics) even though some (bedrooms) were not significant at the 10% level. We also kept squared variables.

Secondly, since the interaction of regions and age seemed significant for all regions, we left them in the model. We also left Surface×Flanders since it was significant at the 10% level.

How to proceed after that ? We were left with the following variables to include or not : Surface×Bedrooms, Surface×Brussels, Surface×Coast, Surface×TimeDummy, Bedrooms×Brussels, Bedrooms×Flanders, Bedrooms×Coast, Bedrooms × TimeDummy, Surface×Age, Bedrooms×Age. We first eliminated three variables which had particularly high p-values : Bedrooms×Age (0.88), Surface×Brussels (0.9) and Bedrooms × TimeDummy. After this elimination, p-values were lowered for all variables except for Surface×Coast for which it became bigger. Since it was also the highest p-value, we eliminated it. Surface×Age was left out after this because its p-value was the highest and we could not think of a theoretical reason to keep it.

After that, the choice was more difficult. We simply proceeded by sequential elimination, starting with the highest p-values, until all variables were significant. How do we know that we may have succeeded ? Firstly, after elimination, all variables have the expected sign. Secondly, the coefficients for variables of interest did not change much after elimination, except for Bedrooms, but we gained in efficiency since standard-errors became smaller. The exception for Bedrooms may be strange and we do not really know how to interpret it. Indeed, the positive sign expected *a priori* for Bedrooms only appears after elimination. We tried to see if a mistake was made by adding the variables Bedrooms×Flanders and Bedrooms×Brussels back into our final model but they were highly insignificant.

Our final regression seems much better regarding information criteria. They are all smaller than in the general regression and the adjusted R^2 is also a bit bigger.

Variable	Coefficient	Standard Error	t-ratio
Constant	11.8117***	0.0421	280.5
Surface	0.0087***	0.0013	6.79
Bedrooms	-0.1997	0.0773	-0.26
Age	-0.0098***	0.0016	-6.25
Surface^2	-0.00002*	1.2×10^{-5}	-1.819
Bedrooms^2	-0.0147	0.0201	-0.73
Age^2	0.0001***	2.04×10^{-5}	6.68
Brussels	0.3214***	0.0774	4.15
Flanders	0.1234***	0.0469	2.63
Coast	0.4911***	0.0777	6.32
TimeDummy	0.0178	0.0555	0.2991
Surface×Bedrooms	0.0006	0.0008	0.7572
Surface×Brussels	0.0002	0.0019	0.1326
Surface×Flanders	-0.0028**	0.0014	-1.99
Surface×Coast	0.0014	0.0019	0.74
Surface×TimeDummy	0.0012	0.0013	0.89
Bedrooms×Brussels	0.704	0.0978	0.72
Bedrooms×Flanders	0.0656	0.0792	0.83
Bedrooms×Coast	-0.1034	0.1026	-1.01
Bedrooms×TimeDummy	0.0151	0.0562	0.27
TimeDummy×Brussels	0.1725*	0.0971	1.78
TimeDummy×Flanders	0.0307	0.0665	0.46
TimeDummy×Coast	0.4046***	0.1330	3.04
Surface×Age	1.9×10^{-5}	2.6×10^{-5}	0.73
Bedrooms×Age	-0.0002	0.0013	-0.13
Brussels×Age	0.0039*	0.0020	1.96
Flanders×Age	0.0043***	0.0015	2.97
Coast×Age	0.0076**	0.0036	2.08
TimeDummy×Age	-0.0030**	0.0015	-2.06

Notes : The dependent variable is $\ln$(Price). We use the HC1 covariance matrix. ***, ** and * denote respectively significance at the 1%, 5% and 10% levels.
Sources : Fédération Royale du Notariat Belge and own calculations.

Table D.1: General Regression (1/2)

Mean Dep. Var.	12.0818	S.D. Dep. Var.	0.4253
SSR	29.1208	S.E. of Reg.	0.2655
R^2	0.6349	Adjusted R^2	0.6102
Akaike Criterion	110.165	P-Value(F)	7.61×10^{-89}
Schwarz Criterion	228.813	Hannan-Quinn	156.963

Sources : Fédération Royale du Notariat Belge and own calculations.

Table D.2: General Regression (2/2)

E Informal Heteroskedasticty Tests

We perform two additional tests for heteroskedasticity for the regression with the regions. First, we look at the residuals. Observations are ordered by increasing price. From figure E.1 it is difficult to conclude, even though it looks like the variance may be a bit bigger for the last observations with the biggest prices. Secondly, we estimate the same model with and without a robust covariance matrix (table E.1). We do not report the standard errors of squared variables since they are either very small or the variable is insignificant. Resulting standard errors are slightly different which, again, leads us to conclude to the presence of heteroskedasticity.

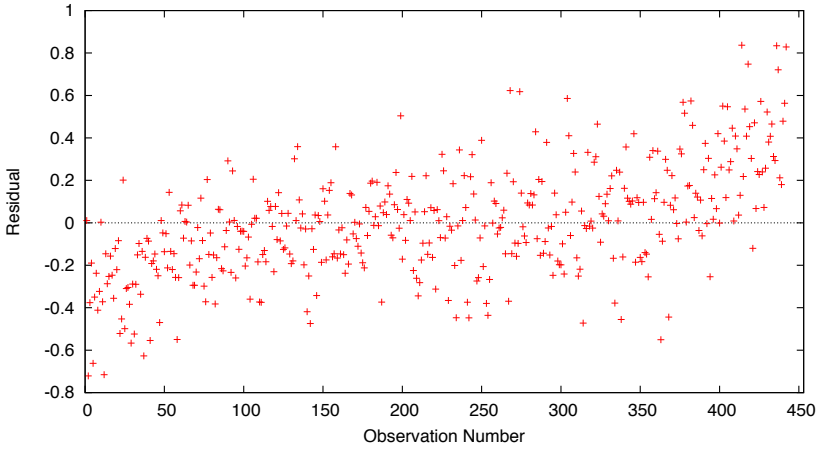


Figure E.1: Residuals from the OLS regression with the regions.
Source : Own Calculations.

Variable	Standard Errors	S.E. with HC1
Surface	0.000743	0.000661
Bedrooms	0.026339	0.025032
Age	0.001634	0.001388
Brussels	0.075186	0.078727
Flanders	0.054357	0.047506
Coast	0.067919	0.071285
TimeDummy	0.064413	0.059755
Surface×Flanders	0.000872	0.000880
Bedrooms×Coast	0.049801	0.056109
TimeDummy×Brussels	0.100669	0.097583
TimeDummy×Flanders	0.072700	0.067233
TimeDummy×Coast	0.107335	0.130919
Brussels×Age	0.002126	0.001827
Flanders×Age	0.001587	0.001311
Coast×Age	0.002639	0.003298
TimeDummy×Age	0.001310	0.001407

Sources : Fédération Royale du Notariat Belge and own calculations.

Table E.1: Comparison of Standard Errors

F Confidence Intervals

Variable	Coefficient	Lower Bound	Upper Bound
Surface	0.00951	0.00821	0.01081
Bedrooms	0.04481	-0.00440	0.09401
Age	-0.00932	-0.01205	-0.00659
Surface^2	-2.1×10^{-5}	-3.7×10^{-5}	-5.3×10^{-6}
Bedrooms^2	-0.00934	-0.04056	0.02188
Age^2	0.00013	9.4×10^{-5}	0.00017
Brussels	0.3216	0.16686	0.47635
Flanders	0.12636	0.03298	0.21973
Coast	0.47844	0.33832	0.61856
TimeDummy	0.02500	-0.09245	0.14246
Surface×Flanders	-0.00265	-0.00438	-0.00092
Bedrooms×Coast	-0.18247	-0.29276	-0.07218
TimeDummy×Brussels	0.17077	-0.02103	0.36258
TimeDummy×Flanders	0.02907	-0.10308	0.16122
TimeDummy×Coast	0.38382	0.12649	0.64116
Brussels×Age	0.00382	0.00023	0.00741
Flanders×Age	0.00423	0.00165	0.00681
Coast×Age	0.00660	0.00012	0.01309
TimeDummy×Age	-0.00334	-0.00611	-0.00058

Sources : Fédération Royale du Notariat Belge and own calculations.

Table F.1: 95% Confidence Intervals For The Regression With The Regions

Variable	Coefficient	Lower Bound	Upper Bound
Surface	0.00625	0.00494	0.00757
Bedrooms	0.05008	-0.00525	0.10541
Age	-0.00496	-0.00695	-0.00298
Surface^2	-2.5×10^{-5}	-4.2×10^{-5}	-5.3×10^{-6}
Bedrooms^2	-0.02001	-0.07175	0.03173
Age^2	0.00011	6.9×10^{-5}	0.00015
TimeDummy	0.03192	-0.03039	0.09422
TimeDummy$\times$Age	-0.00276	-0.00543	-8.3×10^{-5}
Surface$\times$Age	3.2×10^{-5}	-3.8×10^{-5}	-6.7×10^{-5}
Surface$\times$Bedrooms	0.00167	1.3×10^{-5}	0.00332

Sources : Fédération Royale du Notariat Belge and own calculations.

Table F.2: 95% Confidence Intervals For The Regression Without the Regions