UCLouvain

KU LEUVEN

LOUVAIN
School of Management

Louvain School of Management
and KULeuven

How do the demand and supply factors influence the number of transactions in the housing market, taking into account the disequilibrium of the market, the dependence between the demand and the supply, and the time dependency aspect?

Project/Research Master's Thesis submitted by
Lara Henne

With a view of getting the degrees:
Master in Business Engineering
Major in Data Science and Business Analytics (KULeuven)
Major in Supply Chain Management (LSM)

Supervisor :
Ingrid Van Keilegom

Academic Year [*2019-2020*]

Contents

Acknowledgments

I first wish to acknowledge the great support provided by my promotor Ingrid Van Keilegom. Her availability and deep involvement positively impacted the progress of this thesis and all her suggestions were incredibly helpful. Then, I also would like to express my great appreciation to the PhD student Negera Wakgari Deresa for helping me solve sporadic issues I encountered with the code. Finally, I would like to thank my father for his meaningful advice and my mother for the emotional support she brought me through the realisation of this thesis.

Chapter 1

Introduction

1.1 General Framework

We are all confronting the housing market at some time through the search of a home where to live, a profitable real estate investment, a buyer for our current housing and so many other reasons. However, few are those who actually understand its deep logic. Through time, it suffered from historic events like the world wars but also evolved with the Industrial Revolution, the emergence of the television, the digitalisation, etc. Therefore, studying the behaviour of the housing market quickly became necessary, or even vital, to adapt to the continuously changing societies.

Being aware of the complexity of this market, many researchers have focused on this part of the economic world for decades. An interesting observation is the global generalisation of the results obtained in limited regions of the world. Often, authors develop general model although it was based on a limited region. Similarly, we can find the same conclusions in two different scientific works focusing on distinct territories. The local cultures do not seem to be a crucial aspect for the behaviour of the housing market.

The two main elements of the housing market are the demand and the supply quantities. In this work, I will consider, on the first hand, the demand as the number of houses (excluding apartments) that households want to buy in a given semester in a given Flemish community. On the other hand, the supply will be defined as the number of houses (excluding apartments) that developers are willing to sell in a given semester in a given Flemish community.

Conversely, the approaches to the housing market can diverge significantly. First, the static point of view of the market, which considers the market in perfect equilibrium at any times, already brings new insights. By dividing the housing market into sub-markets, their states at the equilibrium can be revealed and providing conclusions about their mutual influences is feasible. In that respect, the relationship between the sub-market for real estate space (consumers) and the sub-market for real estate assets

(investors) (DiPasquale and William C. Wheaton, 1992) was discovered. Second, we can evaluate the dynamic characteristic of the market, which means considering the potential disequilibrium between the demand and the supply quantities. This results in an analysis of a more cyclical nature due to the short-term rigid supply (that thus cannot adjust itself easily to the demand that changes more rapidly). Third, this complex feature is often due not only to the housing market itself but also to the interaction of the housing market with other markets like the mortgage market (Ray C. Fair, 1972). Therefore, many investigations have been conducted around those interactions.

Nevertheless, the basis of a properly handled study concerning the housing market narrows down to adequately specifying the factors that influence the supply and/or the demand and providing a correct model using the convenient tools to avoid inconsistency of the results. Adequately specifying those factors implies selecting the relevant variables that have an impact on the demand and supply quantities. These variables are the price, the households' income, the interest rate, the inflation rate, the mortgage rate, the labour/construction costs. Adding other variables is admitted but removing one of them would lead to misspecifications. Then, the quality of the reports largely depends on the tools used like interactions schema, time series analyses, regression equations, etc.

1.2 Indicating the Gap

While the static point of view can already present interesting findings, it is not sufficient to give reliable solutions. Indeed, a dependency between the supply and the demand cannot be reflected if we consider them in equilibrium at any moment. The word "equilibrium" itself means that no adaptation is needed, this is a stable state. However, a strand of the literature only focus on this vision of the market regardless of its insufficiency.

Furthermore, if we consider the demand and supply quantities in disequilibrium, this means that the number of transactions observed is either one or the other. Indeed, if the supply is greater than the demand, the houses sold will only reflect the quantity demanded (the housing left being unsold and thus not reflected in the transactions). To be clear, this leads to right censoring.

Then, most papers that examine the disequilibrium of the housing market neglect other substantial aspects such as the interactions with other markets. Consequently, the results are again imperfect. In addition, certain models miss some of the essential variables mentioned above to formulate the demand and supply equation. Depending on the tool used, it can have a negative impact on the findings.

Moreover, across the various scientific works, I notice that mostly the same techniques are used to assess the impact of the housing factors on the demand and the supply quantities.

In short, the time dependency aspect of the demand and the supply is rarely taken into account. The authors Jang Sewoong, Lee Sanghyo, Kim Juhyung and Kimnise-Jaejun (2010) integrated the time dependency characteristic in their reasoning by using tools from time series analysis, but their work can be criticised at certain levels. For instance, the way they apply the impulse response function by putting an impulse on only one variable, assuming that this will not have an impact on the others, is quite naive.

1.3 This Work Contribution

First of all, I want to make sure that the integration of relevant markets is executed and that all variables needed are included in our model. The literature presents the mortgage market as the most influential and the key variables are already quoted above.

Then, I want to analyse the disequilibrium in the market itself to understand the dependency between the demand and the supply quantity, as this is missing in most of the literature. I analyse it as a right censoring which implies the use of a censoring indicator. This censoring indicator will be defined further in this work. I thus make reference to survival analysis techniques that I will try to apply in the context of the housing market. In order to take into account the time dependency factor, I will build an autoregressive model from the time series analysis techniques (this means including the dependent variable lagged of, for instance, one semester in the covariates). This would allow us to confirm our hypothesis that this market attribute is key for all the underlying aspects (rent levels, construction, labour costs, etc.).

At last, because of the lack of variability in the techniques used and their imperfection, I thought it was important to analyse the housing market in a novel way. Indeed, by multiplying the methods, we provide various points of view that, if converging to the same conclusions, will strengthen the theories and hypotheses already existing.

In the end, I will try to answer the following research question: *How do the demand and supply factors influence the number of transactions in the housing market, taking into account the disequilibrium of the market, the dependence between the demand and the supply, and the time dependency aspect?*

1.4 Outlining

To introduce the important aspects of the theoretical housing market context, a literature review is developed in the first place. Then, a section is devoted to the description of the models used and their adaptation to our housing market context. I also present the different variables and their sources. Finally, I reveal and discuss the results obtained.

Chapter 2

Literature Review

In this chapter, I will give an overview of the main theoretical findings related to the housing market context. Being one of the most substantial domains of a private and public economy, many studies have been conducted around the dynamic and static relationships within the housing market. Indeed, the latter is a very complex system, influenced by many interrelated parameters. The choice among those parameters to conduct an analysis of a market should be very meticulous. Actually, it depends on the objective of the analysis.

One of the main conducted research around the housing market is focused on the interactions between two sub-markets in a static state. For instance, Denise DiPasquale and William C. Wheaton (1992) centre their work around two aspects of the real estate market in order to show their relationship: the market for real estate space and the market for real estate assets. They illustrate the relation between those two markets through the rent levels (evaluated in function of the needs of tenants and the state of buildings). They argue that those rent levels (space market) influence the demand for real estate assets and the construction sector which, in turn, makes the price in the market for real estate assets vary and has an impact on the rent levels. They use a four-quadrant schema for long-run equilibrium to support their statements as shown in Figure 1. However, as underlined by the authors themselves, the entire study omits the disequilibrium aspect of the real estate market. This is, however, a substantial part that could lead to totally different conclusions.

Another strand of literature emphasises the disequilibrium part of the market though (but they are a minority). Addressing this characteristic, Jang Sewoong, Lee Sanghyo, Kim Juhyung and Kimnise Jaejun (2010) wrote a paper taking disequilibrium of the housing market in Korea into consideration. They make reference to Denise DiPasquale and William C. Wheaton (1992) by using their four-quadrant schema (cf. Figure 1) to study the interactions of the demand and the supply in the housing market on the first side, and the unsold new housing stock on the other side. Even if they inspired from their four-quadrant schema, the approach presented in this research diverges from the

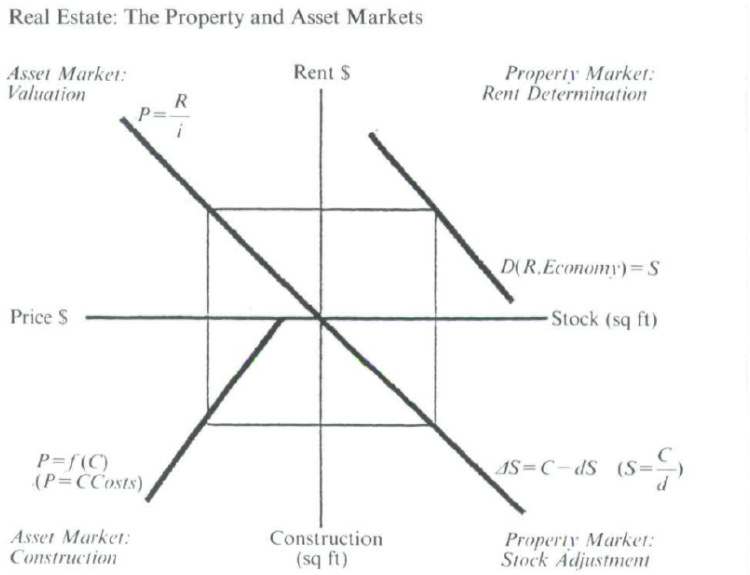

Figure 1: Four Quadrant Schema Used by DiPasquale and Wheaton (1992).

Source: DiPasquale , D., Wheaton, W. C. (1992). The Markets for Real Estate Assets and Space : A Conceptual Framework. *Journal of the American Real E.state and Urban Economics Association, 20(1)*, p.188.

approach of Denise DiPasquale and William C. by the techniques used and the disequilibrium aspect. Their main statement highlights that the unsold new housing stock is a direct consequence of the quantity difference between the supply and the demand. They thus decide to analyse the disequilibrium indirectly as the quantity difference is less observable which is, in our opinion, a very meaningful way to proceed. They follow an empirical method research using tools from advanced time series analysis and define the unsold new housing stock based on the four-quadrant schema in this way:

$$W_t = f(PS_t, I_t, M_t, L_t, Y_t)$$

Where W_t is the unsold new housing stock, I_t, M_t, L_t are the funds, the material, the manpower respectively and define altogether the construction costs, PS_t is the housing stock prices, which are assumed to influence the demand and supply of new houses and Y_t is the housing loans, which influence the demand for new houses.

Further to this gathering of data, the authors conduct time series diagrams to analyse the relationship between unsold new housing stock and the individual variables. Firstly, unsold housing stock and housing price time series (cf. Fig.2) help to analyse their correlated evolution. When the prices are low, the unsold housing stock increases because of the economic downturn, but when the prices increase, the new housing stock then

decreases thanks to business expansion. Then, the trend of unsold new housing stock increases again but this is mainly due to regulation policies and recent economic shocks not further developed here. Secondly, unsold new housing stock and construction costs time series (cf. Fig.3) do not show any related patterns regarding the material or labour supply. However, a small relationship between funds supply and unsold new housing stock is found. Thirdly, unsold new housing stock and housing loans time series (cf. Fig.4) indicate some negative correlations.

Figure 2: Trend of Housing Transaction Price Index[1]

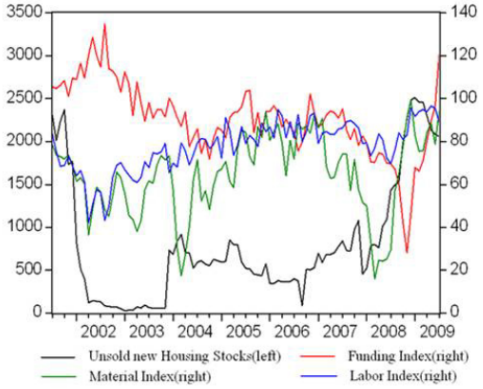

Figure 3: Trend of Construction Cost Variables[2]

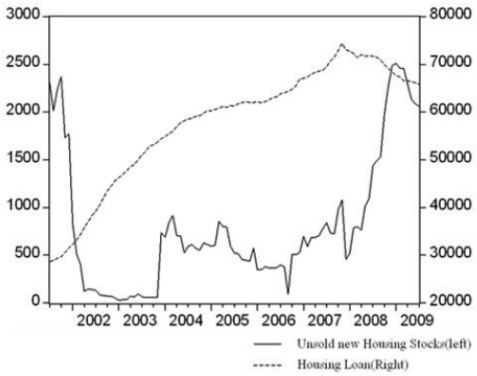

Figure 4: Trend of Housing Loan [3]

[1]Source: Sewoong, J., Sanghyo, L., Juhyung, K., Jaejun, K. (2010). Relationship Between Demand-supply in the Housing Market and Unsold New Housing Stocks. *Journal of Asian Architecture and Building Engineering*, 9, p.389

[2]Source: Sewoong, J., Sanghyo, L., Juhyung, K., Jaejun, K. (2010). Relationship Between Demand-supply in the Housing Market and Unsold New Housing Stocks. *Journal of Asian Architecture and Building Engineering*, 9, p.390

[3]Source: Sewoong, J., Sanghyo, L., Juhyung, K., Jaejun, K. (2010). Relationship Between Demand-supply in the Housing Market and Unsold New Housing Stocks. *Journal of Asian Architecture and Building Engineering*, 9, p.390

Through the implementation of a vector error correcting model, they realise that variations in unsold new housing stock are mostly provoked by selling price indexes and loans (cf. Table 1). On the other hand, production drivers have a hard impact on those variations. By applying an impulse response function, the conclusions are the same. However, the method they use can be criticised. For example, they make an impulse of one standard deviation at a time on one variable at a time. This is not particularly relevant as an impulse on one variable will certainly provoke an impulse on the others immediately. Therefore, outcomes could be biased. Also, as stated by Keating (1992) and McMillin (2001), short-run and long run impulses can lead to different outcomes.

Table 1: Variance Decomposition

Period	W_t	PS_t	I_t	M_t	L_t	Y_t
1	100.0000	0.000000	0.000000	0.000000	0.000000	0.000000
2	98.51575	0.565826	0.028310	0.074690	0.024779	0.790644
3	96.83944	1.548991	0.146758	0.053142	0.017514	1.394153
4	94.87509	2.726991	0.355214	0.043568	0.044533	1.954601
5	92.80579	3.913782	0.600461	0.047311	0.131785	2.500873
6	90.72648	5.047966	0.856922	0.058868	0.265924	3.043843
7	88.70435	6.104135	1.112071	0.074911	0.432850	3.571687
8	86.77450	7.075012	1.359266	0.093221	0.621595	4.076409
9	84.95485	7.961268	1.594654	0.112383	0.823057	4.553784
10	83.25241	8.767657	1.816328	0.131543	1.030067	5.001990

Source: Sewoong, J., Sanghyo, L., Juhyung, K., Jaejun, K. (2010). Relationship Between Demand-supply in the Housing Market and Unsold New Housing Stocks. *Journal of Asian Architecture and Building Engineering*, 9, p.391

However, understanding how the disequilibrium system in the housing market works is not an easy task. In general, in the literature regarding the disequilibrium in the real estate market, we observe that most studies are performed by separating the market into sub-markets (like DiPasquale and Wheaton (1992) do with the real estate market) or by comparing it to other markets and show that they are related which explains the disequilibrium. One relevant comparison that is made is the relation between the housing market and the mortgage market (Ray C. Fair, 1972). The findings put forward the constraining effect on the housing market when mortgage demand (by the builders) is too high. This excess demand provokes rationing in the mortgage commitments (builders receive fewer mortgage commitments than wanted and less financial support) and the constructors will build less "new" houses while the demand for new houses stays high. This results in an obvious disequilibrium. However, presented this way, the disequilibrium is asymmetric as an excess of supply in the mortgage market won't make the constructors build more (they adapt their constructions to the demand, producing in excess is not necessary). The author argues on this asymmetry by underlying other aspects that could influence the disequilibrium in the housing market in both ways. He

identifies the speed of adjustment of the selling prices (if too low, there will be too many demands compared to the supply) and an incorrect estimation of the quantity demanded by the suppliers.

Most methods used in the scientific research around the housing market do not include the disequilibrium aspect in their equations and models at all and the few that do, still do not manage to implement it properly. This can be due to incorrect specifications of the supply and the demand equations (by omitting some relevant variables), or to the inadequate integration of the interactions with another market. Addressing this issue, Fair and Jaffee (1972) and Augustyniak, Laszek, Olszewski and Waszczuk (2014), two groups originated from totally different parts of the world and eras, though, consider the disequilibrium correctly and at the centre of the housing market models. They assume the disequilibrium to be a consequence of, on the one hand, the relatively long construction operations when the demand is high and, on the other hand, the developers' reluctance to lower the prices faster when demand is low. The rapid demand changes compared to this fixed short-term supply provoke misalignment between the two. Indeed, the excess demand leads to an increase in house prices and the future need in houses supply is overestimated which will induce an excess of houses stock later on.

Augustyniak et al. (2014) construct a simple model from the Polish housing market data and focus on the primary market, the market for new constructions, putting aside the existing housing stock as they consider that both interact.

Before providing aggregated equations for the demand and the supply, the searchers analyse interesting individual behaviours of the actors and I think they are worth being described first as we can find similar analysis in other recent studies. On the demand side, households' demand will be influenced by the loan availability, consumer preferences for other goods and housing services. The households distribute their revenues between consumption of other goods and housing, housing being bought for consumption and investment. Therefore, the following households' utility function is built:

$$U(C, H) = (\theta C^{\mu} + (1 - \theta)A^{\gamma}(kpH)^{\mu})^{1/\mu}$$

Where kpH is the imputed rent (H for the house size, p for the price, k for the monetary value of the stream of housing service), θ is the consumption's share, μ is the elasticity substitution between the consumption of goods other than housing and the housing, $A = P_t/P_{t-1}$ is the appreciation which influence the household's forecasts on future house prices, often over- or underestimated by the households (Dunsky, R.M. Follain, J.R. (1997), Sommervoll, D. E., Borgensen, T.-A. Wennemo, T. (2010) or Lambertini, L.,Mendicino, C. Punzi, M. T.(2012)).

Due to restrictions, the income share allowed to repay the loan is limited to $x \in \{0, 1\}$ and is maximum b_H. Assuming that b represents the household's budget constraint:

$b = rpH + C$ (r is the constant loan, p is the price per square metre) we have:

$$b_H = xb \leq b$$

At last, the authors use the following formulas to illustrate the way households thus decide about the housing size (given that the income share to repay the loan is limited):

$$H = \begin{cases} H^*, & rpH^* \leq xb \\ \frac{xb}{rp}, & rpH^* > xb \end{cases}$$
$$C = \begin{cases} C^*, & rpH^* \leq xb \\ (1-x)b, & rpH^* > xb \end{cases}$$

With H^* and C^*, the optimum allocation between consumption and housing (derived from the utility and the budget constraint shown above. However, I won't present the mathematical details as this is not the most relevant part here).

Graphically, the demand will vary as follows:

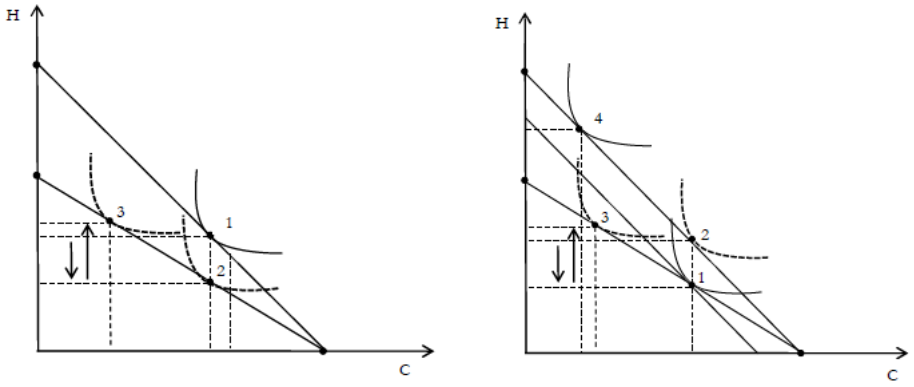

Figure 5: Demand Response

Source: Augustyniak, H., Laszek, J., Olszewski, K., Waszczuk, J. (2014). Housing market cycles – a disequilibrium model and its application to the primary housing market in Warsaw. *Ekonomia Journal*, 35, p.7

Concerning the left graph of Figure 5, from 1 to 2, there is an increase in prices and the housing is used in the form of consumption only while from 2 to 3, the consumer decides to consume less and invest in housing. As for the graph to the right, from 1 to 2, there is an income increase (with the same amount of consumption, we can buy more housing, this is represented by a translation of the budget line to the right), from 2 to 3

there is an increase in prices and from 3 to 4, there is a decrease in interest rates which makes the housing investment very attractive.

On the supply side, future production is predicted based on current prices taking into account the estimated future inflation. This induces an underestimation of future production costs if demand grows substantially. Indeed, if there is an important grow in supply in the midterm to adapt to this growing demand, a negative scale effect will occur: the materials, costs of transport, lands, etc. will be at a higher cost level. In particular, Augustyniak et al. (2014) present the real and virtual supply curves (cf. Figure 6).

The virtual supply curve (V on Figure 6) is built based on the supplier's calculation of future performance (subjective predictions). He relies on current prices and costs but do not take into account the dynamic of the market. This results in over- or underestimated costs/prices. If prices do not change but the profit decrease (because of increasing costs), the virtual curve will be higher and will considerably slow down the construction sector.

The real supply curve (F on Figure 6) is based on real modifications in investment profitability. This means that the dynamic of the market is here at the centre of the curve development (for instance, the increase of production factors if the production rises).

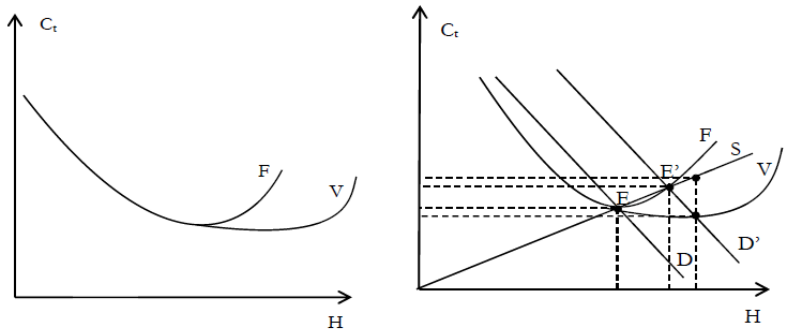

Figure 6: Virtual and Real Supply Curves

Source: Augustyniak, H., Laszek, J., Olszewski, K., Waszczuk, J. (2014). Housing market cycles – a disequilibrium model and its application to the primary housing market in Warsaw. *Ekonomia Journal*, 35, p.12

On Figure 6, we observe the difference in the real and virtual supply curves (left graph) and we can analyse the reaction to an increase in demand (from D to D' on the right graph). As I mentioned above, the production will increase considerably, raising the costs more quickly than what the virtual curve suggests. We can see that the point on the S curve (the price the suppliers are intended to ask for this number of houses) is

higher than the equilibrium point (E') and lower than the actual cost (curve F). Customers are thus not willing to pay this high price, especially if suppliers level up this price again to cover the costs.

Those individual behaviours being analysed, I can come back to the paper from Fair and Jaffee (1972) that provides a general model to estimate the supply and demand quantity in disequilibrium markets (applicable to the housing market but not only):

$$D_t = \alpha_0 X_t^D + \alpha_1 P_t + \mu_t^D \tag{2.1}$$
$$S_t = \beta_0 X_t^S + \beta_1 P_t + \mu_t^S \tag{2.2}$$

with t the time index, D_t the quantity demanded, S_t the quantity supplied, P_t the price, X_t the other covariates and μ_t the error terms.

Then, we can find an application for the housing market that is similar in the two papers. However, Fair and Jaffee (1972) focus on the primary and secondary market while Augustyniak et al. (2014) only take into account the primary market, but it does not have an impact on final outcomes. Therefore, I will only develop here the model proposed by Augustyniak et al. (2014). For the aggregate demand we have:

$$HD_t = \alpha 1 + \alpha 2 * P_t + \alpha 3 * D(P_t) + \alpha 4 * Intrate_t + \alpha 5 * Income_t + \epsilon_t$$

Where P_t is the price, $D(P_t)$ is the growth rate of the housing price, $Intrate_t$ is the interest rate, $Income_t$ is the household's income. By taking into account the interest rate in his demand equation, the writer takes into account the interaction with the mortgage market which is often wrongly omitted in previous studies.

For the aggregate supply we have:

$$HS_t = \beta 1 + \beta 2 * D(P_{t-4}) + \beta 3 * D(PC_{t-4}) + \beta 4 * Intrate_t + \epsilon_t$$

As here, only the construction market is considered, the suppliers react only one year after the price increase (building houses takes time). Therefore, we consider the growth rate in prices one year ago from now $D(P_{t-4})$. In this approach, we observe that the searcher explicitly presents the disequilibrium of the housing demand and supply as they are not determined by concomitant factors: there is always a late adjustment of the supply to the demand. For the price increase in construction costs, the builders react again one year later $D(PC_{t-4})$. Again, the interaction with the mortgage market

is present through the interest rate.

Then, the authors also provide an equation for the growth rate in prices $D(P_t)$, which depends on the disequilibrium between the demand and the supply $(HS_{t-1} - HD_{t-1})$ and on the growth rate in prices from the last period. An interesting behaviour mentioned by the author is the slowly lowering of the prices by the developers when there is an excess demand which maintains the imbalance of the market (their goal is to obtain the highest price as possible).

Finally, the technique used by Augustyniak et al. (2014) is the OLS regression (with a control for heteroscedasticity and autocorrelation) that confirms the direction of the influence of each factor on the housing demand and supply.

In the paper of Fair and Jaffee (1972), they use four different techniques and evaluate them in their ability to predict the demand and supply quantity. The first one is the maximisation of a likelihood function based on the equations (2.1) and (2.2). The others are called the directional method I, the directional method II and the quantitative method and are based on the assumption that the observed quantity equals the minimum of the demand and the supply quantity. The searchers obtain reliable results for the directional method I and the quantitative method which I will thus describe briefly.

In the directional method I, the graph on Figure 7 is analysed.

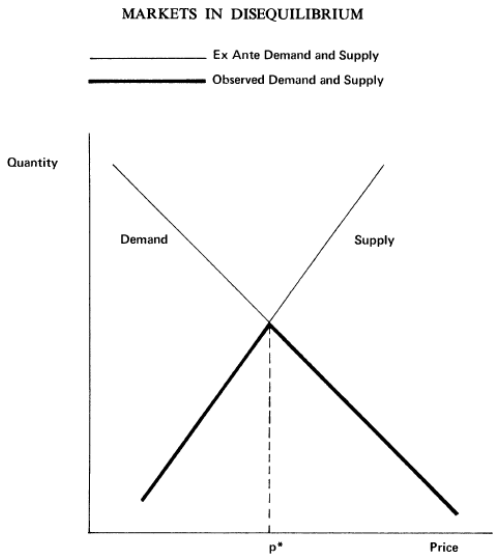

Figure 7 : Markets in disequilibrium

Source: Fair, R. C., Jaffee, D. M. (1972). Methods of Estimation for Markets in Disequilibrium. *Econometrica*, 40(3), p. 503.

If the price P is lower than the equilibrium price P*, this means there is an excess demand and the price will start to rise. Indeed, the price change is considered as an indicator of the demand state compared to the supply state and change in the following directions:

$$\Delta P_t \geq 0 \text{ while } D_t - S_t \geq 0$$
$$\Delta P_t \leq 0 \text{ while } D_t - S_t \leq 0$$

Therefore, when the price is rising, only the supply quantity is observed (given the assumption that only the minimum of the supply and the demand is observed). According to this price change, this method thus determines the times of excess demand, the times of excess supply and the times of equilibrium (when there is no price changes). The latter is included in both the demand and the supply samples.

In the quantitative method, they assume that the proportion (and not only the direction any more) of the price change is directly correlated with the proportion of excess demand which gives:

$$\Delta P_t = \gamma(D_t - S_t) \text{ with } 0 \leq \gamma \leq \infty$$
$$(D_t - S_t) = \tfrac{1}{\gamma}\Delta P_t \text{ with } 0 \leq \gamma \leq \infty$$

γ being the coefficient of proportion.

Then, during periods of rising prices, there is an excess demand and the quantity observed is the supply. Therefore, the quantity observed is determined as follows:

$$Q_t = S_t = D_t - \tfrac{1}{\gamma}\Delta P_t = \alpha_0 X_t^D + \alpha_1 P_t - \tfrac{1}{\gamma}\Delta P_t + \mu_t^D, \quad \Delta P_t \geq 0$$

During periods of falling prices, the principle for the excess supply is the same. Then, we can write a demand equation and a supply equation that can be estimated with the data:

$$Q_t = D_t - \frac{1}{\gamma}/\Delta P_t/ = \alpha_0 X_t^D + \alpha_1 P_t - \frac{1}{\gamma}/\Delta P_t/ + \mu_t^D$$

$$\text{where } /\Delta P_t/ = \begin{cases} \Delta P_t, & \Delta P_t \geq 0 \\ 0, & otherwise \end{cases}$$

and

$$Q_t = S_t - \frac{1}{\gamma}\ \Delta P_t\ = \alpha_0 X_t^S + \alpha_1 P_t - \frac{1}{\gamma}\ \Delta P_t \mu_t^S$$

$$\text{where } /\Delta P_t/ = \begin{cases} -\Delta P_t, & \Delta P_t \leq 0 \\ 0, & otherwise \end{cases}$$

As I mentioned above, the directional method I and the quantitative method give both reliable outcomes when applied to the housing market. They also prove that there is a disequilibrium which must be involved in any housing market analysis to give better results. I thus decided to inspire from one of those methods in this work to estimate if the quantity observed is either the supply quantity or the demand. As the authors encounter two small problems (not detailed here) with the quantitative method, I opted for the directional method I.

To sum up, I came across many different approaches to understand how the housing market works. The main literature sees it from a static angle which provides already interesting findings. Indeed, the equilibrium assumption still allowed searchers to divide the market into different segments and to analyse the behaviours in one segment relative to the other (like for instance the housing consumption and the housing investment described by DiPasquale and William C. Wheaton, 1992.). However, what is more of a matter of interest in this work is the cycle nature of the housing market as a consequence of the disequilibrium between the demand and the supply quantity. Only few searches have focused their analyses on this aspect of the real estate market and therefore, only few techniques have been applied. Furthermore, many weaknesses can be found in the construction of the models (omitting relevant variables for instance), or in the techniques used. All in all, I want to provide a novel way of analysing the housing market disequilibrium by taking into account the dependency between the supply and the demand quantity and by correcting previous erroneous models.

Chapter 3

Methodology

In this chapter, I present the theoretical models used, their adaptation to the housing market context and the data on which I apply them.

3.1 Survival Analysis Model

First of all, as presented in the Introduction section, I will analyse the dependency between the demand and the supply quantities through the use of a censoring indicator. This indicator makes even more sense through Fair and Jaffee's work (1972). Indeed, they argue that only the minimum of the supply or the demand quantity is observed (right censoring) according to the price variation ("directional method I"). This method consists in determining the equation as a supply equation if the price level in the previous period was lower than in the current period. A rising in price is a selection process by the suppliers among the demand in excess. The observed quantity, being the minimum of the supply and the demand quantity, is thus a supply quantity in this case because there is an excess demand (which is right censored).

Our main focus finds its origin in a paper from Ingrid Van Keilegom and Negera Wakgari Deresa (2019). This work builds a model for survival analysis. It considers that the survival time, which is the time before a precise event happens (like the death of a patient) and the censoring time, which is the time to something else before the event that is not observed (because the patient exits from the study before he dies, for instance) are dependent. This dependency aspect is not common in survival analysis that often assumes that those times are not related at all and thus independent. The statisticians build a joint regression model with an identifiability of the dependence between the survival time and the censoring time thanks to the bivariate normality of the errors. Also, the model uses an original parametric family of power transformations (Yeo-Johnson transformation described later) for the survival time and the event time, in order to improve the fit of the data. Largely applied in the health segment, this model has never been used in a housing market context as far as I know. Here is how it is presented:

To begin with the covariates, the ones for the demand (D) are gathered in $X = (1, \tilde{X}^T)^T$ and those for the supply (S) in $W = (1, \tilde{W}^T)^T$. In our case, they will be partially overlapping.

Furthermore, the joint regression model is as follows:

$$\begin{cases} \Lambda_\theta(D) = X^T\beta + \epsilon_D \\ \Lambda_\theta(S) = W^T\eta + \epsilon_S \end{cases} \tag{3.1}$$

with Λ_θ being a monotone increasing transformation function (same parametric transformation for both D and S) and β and η the vectors of coefficients. An important assumption in our model is the bivariate normal distribution of the error terms ϵ_T and ϵ_C (for the identifiability of the model):

$$\begin{pmatrix} \epsilon_D \\ \epsilon_S \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma = \begin{pmatrix} \sigma_D^2 & \rho\sigma_D\sigma_S \\ \rho\sigma_D\sigma_S & \sigma_S^2 \end{pmatrix}\right), \tag{3.2}$$

with Σ assumed to be a positive semi-definite matrix and with (ϵ_D, ϵ_S) and (X, W) independent. ρ is the association parameter: it indicates the direction and intensity of the correlation between the demand and the supply. Then, the authors use a parametric transformation based on the findings of Yeo and Johnson (2000) which is related to the Box-Cox transformations. The required conditions for this study are respected by the parametric transformation of Yeo and Johnson when $0 \leq \theta \leq 2$. This transformation takes the form:

$$\Lambda_\theta(t) = \begin{cases} \{(t+1)^\theta - 1\}/\theta & t \geq 0, \theta \neq 0 \\ log(t+1) & t \geq 0, \theta = 0 \\ -\{(-t+1)^{2-\theta} - 1\}/(2-\theta) & t < 0, \theta \neq 2 \\ -log(-t+1) & t < 0, \theta = 2 \end{cases} \tag{3.3}$$

The variables D and S are considered through the number of transactions $Z = min(D, S)$ (because of the censoring feature) and the censoring indicator δ that is defined by convention as $\delta = I(D \leq S)$ with I being the indicator function.

Moreover, the conditional distribution functions of D given $X = x$ and S given $W = w$ are:

$$F_{D|X}(d|x) = F_{\epsilon_D}(\Lambda_\theta(d) - x^T\beta) = \Phi\Big(\frac{\Lambda_\theta(d) - x^T\beta}{\sigma_D}\Big)$$
$$F_{S|W}(s|w) = F_{\epsilon_S}(\Lambda_\theta(s) - w^T\eta) = \Phi\Big(\frac{\Lambda_\theta(s) - w^T\eta}{\sigma_S}\Big) \tag{3.4}$$

with F_{ϵ_D} and F_{ϵ_S}, the distribution functions of ϵ_D and ϵ_S, and Φ the standard normal distribution. From there, the authors derive the conditional density of D given $X = x$ by:

$$f_{D|X}(d|x) = \sigma_D^{-1}\phi\Big(\frac{\Lambda_\theta(d) - x^T\beta}{\sigma_D}\Big)\Lambda'_\theta(d) \tag{3.5}$$

With ϕ the density function of a standard normal variable. Now the vector of parameters α is defined by $\alpha = (\theta, \beta, \eta, \sigma_D, \sigma_S, \rho) \in \mathbb{R}^{p+q+4}$. Then, integrating the number of transactions and the censoring indicator, we can obtain the sub-distribution $F_{Z,\delta|X,W}(\cdot,\cdot|x, w; \alpha)$ of (Z, δ) given $(X, W) = (x, w)$ for a given α. When we consider $Z = z$ and $\delta = 1$ we can define:

$$F_{Z,\delta|X,W}(z, 1|x, w; \alpha) = P(Z \leq z, \delta = 1|X = x, W = w)$$

If the minimum between D and S is D, then $Z = min(D, S) = D$. Then, applying the parametric transformation, we have:

$$\begin{aligned} F_{Z,\delta|X,W}(z, 1|x, w; \alpha) &= P(Z \leq z, \delta = 1|X = x, W = w) \\ &= P(\Lambda_\theta(D) \leq \Lambda_\theta(z), \Lambda_\theta(D) \leq \Lambda_\theta(S)|X = x, W = w) \end{aligned}$$

A probability is an expected value of an indicator, therefore:

$$\begin{aligned} F_{Z,\delta|X,W}(z, 1|x, w; \alpha) &= P(Z \leq z, \delta = 1|X = x, W = w) \\ &= P(\Lambda_\theta(D) \leq \Lambda_\theta(z), \Lambda_\theta(D) \leq \Lambda_\theta(S)|X = x, W = w) \end{aligned}$$

$$= E(I(\Lambda_\theta(D) \leq \Lambda_\theta(z), \Lambda_\theta(D) \leq \Lambda_\theta(S)|X = x, W = w))$$

By the law of iterated expectations:

$$= E\Bigg\{E\Big[I(\Lambda_\theta(D) \leq \Lambda_\theta(z), \Lambda_\theta(D) \leq \Lambda_\theta(S)|D, X = x, W = w)\Big]\Bigg\}$$

$$= E\Bigg\{I\Big[\Lambda_\theta(D) \leq \Lambda_\theta(z)\Big]E\Big[I(\Lambda_\theta(S) \geq \Lambda_\theta(D)|D, X = x, W = w)\Big]\Bigg\} \tag{3.6}$$

The second term of (3.6) can be rewritten as follows:

$$E\Big[I(\Lambda_\theta(S) \geq \Lambda_\theta(D)|D, X = x, W = w)\Big] = P(\Lambda_\theta(S) \geq \Lambda_\theta(D)|D, X = x, W = w)$$

Now, given the equations (3.1), if we express (3.6) with the error terms, we can rewrite the equation. For the second term, it gives:

$$P(\Lambda_\theta(S) \geq \Lambda_\theta(D)|D, X = x, W = w) = P(\epsilon_S \geq \epsilon_D + x^T\beta - w^T\eta|\epsilon_D, X, W)$$

As the error terms are bivariate normal, we can directly rewrite the probability as:

$$P(\epsilon_S \geq \epsilon_D + x^T\beta - w^T\eta|\epsilon_D, X, W) = 1 - \Phi\Big(\frac{\epsilon_D + x^T\beta - w^T\eta - \rho\frac{\sigma_S}{\sigma_D}\epsilon_D}{\sigma_S(1-\rho^2)^{1/2}}\Big)$$

For the first term of (3.6), we obtain:

$$I\Big[\Lambda_\theta(D) \leq \Lambda_\theta(z)\Big] = I\Big[\epsilon_D \leq \Lambda_\theta(z) - x^T\beta\Big]$$

Now, we can rewrite the equation (3.6) with the latest formulations and by taking the integral with respect to the density to represent the expected value:

$$F_{Z,\delta|X,W}(z,1|x,w;\alpha) = \frac{1}{\sigma_D}\int_{-\infty}^{\Lambda_\theta(z)-x^T\beta}\Big[1-\Phi\Big(\frac{e+x^T\beta-w^T\eta-\rho\frac{\sigma_S}{\sigma_D}e}{\sigma_S(1-\rho^2)^{1/2}}\Big)\Big]\phi\Big(\frac{e}{\sigma_D}\Big)de,$$

With $\epsilon_D = e$ and $(\epsilon_S|\epsilon_D = e) \sim N\Big(\rho\frac{\sigma_S}{\sigma_D}e, \sigma_S(1-\rho^2)\Big)$

The subdensity $f_{Z,\delta|X,W}(\cdot,\cdot|x,w;\alpha)$ of this function F is:

$$f_{Z,\delta|X,W}(z,1|x,w;\alpha) = \frac{1}{\sigma_D}\Big[1-\Phi\Big(\frac{\Lambda_\theta(z)-w^T\eta-\rho\frac{\sigma_S}{\sigma_D}(\Lambda_\theta(z)-x^T\beta)}{\sigma_S(1-\rho^2)^{1/2}}\Big)\Big] \times \phi\Big(\frac{\Lambda_\theta(z)-x^T\beta}{\sigma_D}\Big)\Lambda'_\theta(z), \quad (3.7)$$

When delta is equal to 0 instead of 1, we have:

$$f_{Z,\delta|X,W}(z,0|x,w;\alpha) = \frac{1}{\sigma_S}\Big[1-\Phi\Big(\frac{\Lambda_\theta(z)-x^T\beta-\rho\frac{\sigma_D}{\sigma_S}(\Lambda_\theta(z)-w^T\eta)}{\sigma_D(1-\rho^2)^{1/2}}\Big)\Big] \times \phi\Big(\frac{\Lambda_\theta(z)-w^T\eta}{\sigma_S}\Big)\Lambda'_\theta(z), \quad (3.8)$$

Now, we estimate the model parameters $\alpha = (\theta, \beta, \eta, \sigma_D, \sigma_S, \rho)$ with the likelihood function. Usually, the likelihood function is built with the joint density $f_{Z,\delta|X,W}f_{X,W}$ but we only consider here the conditional density part as the term $f_{X,W}$ does not depend on the model parameters. We multiply together the conditional density of the demand

(when the censoring indicator is worth 1) and the supply (when the censoring indicator is worth 0) which corresponds to equations (3.7) and (3.8):

$$
\begin{aligned}
L(\alpha) &= \prod_{i=1}^{n} f_{Z,\delta|X,W}(Z_i, \delta_i | X_i, W_i; \alpha) \\
&= \prod_{i=1}^{n} \Big\{ \frac{1}{\sigma_D} \Big[1 - \Phi\Big(\frac{\Lambda_\theta(Z_i) - W_i^T\eta - \rho\frac{\sigma_S}{\sigma_D}(\Lambda_\theta(Z_i) - X_i^T\beta)}{\sigma_S(1-\rho^2)^{1/2}} \Big) \Big] \times \phi\Big(\frac{\Lambda_\theta(Z_i) - X_i^T\beta}{\sigma_D} \Big) \Big\}^{\delta_i} \\
&\times \Big\{ \frac{1}{\sigma_S} \Big[1 - \Phi\Big(\frac{\Lambda_\theta(Z_i) - X_i^T\beta - \rho\frac{\sigma_D}{\sigma_S}(\Lambda_\theta(Z_i) - W_i^T\eta)}{\sigma_D(1-\rho^2)^{1/2}} \Big) \Big] \phi\Big(\frac{\Lambda_\theta(Z_i) - W_i^T\eta}{\sigma_S} \Big) \Big\}^{1-\delta_i} \times \Lambda'_\theta(Z_i)
\end{aligned}
\tag{3.9}
$$

3.2 Data

The data gathering was a meticulous part of the work. This one requires to be deeply informed about the housing market factors and about the statistical models used for those data. Then, the challenge was to find reliable sources to obtain those data.

3.2.1 Selected Variables

Before starting the description of the variables, Table 2 brings them together with their abbreviation used in the dataset to have a clear overview:

Table 2: Variables Used and Their Abbreviation

Abbreviation	Variable
Censor, Event	The censoring and event indicators
Trans, TransS_1	The number of transactions (lagged of 1 semester)
TransYJ, TransS_1YJ	The Yeo-Johnson transformed number of transactions (lagged of 1 semester)

Infl_rate, Infl_rateT_1	The inflation rate (lagged of 1 year)
DPrice, DPriceT_1	Deflated median house price (lagged of 1 year)
NH	The number of households
CC	The consumer confidence
IR, IRT_1	The mortgage interest rate (lagged of 1 year)
HI, HIT_1	The housing inflation (lagged of 1 year)
BankT_1	The number of bankruptcies in the construction sector
NC	The number of new housing constructions
DLCT_1	The labour cost index

First, I chose the **number of transactions** in the housing market as the observed quantity variable (the quantity sold on the market). As I made the choice to study the time dependency aspect of the market, I built an autoregressive model (from the time series analysis field) and put the **number of transactions lagged of one semester** in the demand and the supply equations. In addition, again for both equations, I took the **mortgage interest rate** variable to consider the interactions between the financial and housing market as they strongly impact each other. The **housing price** variable is also necessary as it allows to determine if we observe a demand or supply quantity through the **censoring indicator** described in the previous section using the "directional method I" of Fair and Jaffee (1972). The **Event indicator** is obviously the complementarity of the censoring indicator. Then, the general **inflation rate**, as well as the more specific **housing inflation rate**, were included as covariates in both the demand and supply formulas. I added the **number of households** as a household needs only one place to live, the **consumer confidence** and the **household income** for the demand equation. The consumer confidence is calculated according to a survey conducted every month to a representative sample of the Belgian population. Qualitative questions are asked about their perspective on the future economic and unemployment situation in Belgium, about their own financial position and about their intention to buy real estate. Based on the answers, a percentage of negative, positive and neutral responses are calculated: a "+10" means that there have been 10 percent more positive responses than negative ones. Moreover, on the supply side, I searched for the **number of bankruptcies** in the construction sector, the **number of new housing constructions** and the **labour cost index** (base 2012 = 100). The lagged form of certain variables (cf. Table 2) are the ones associated with the supply equation. Indeed, the few literature studying the

disequilibrium between the demand and the supply quantities emphasises the short-term rigid supply problem. The supply reacts to those variables only one year later (as the construction process takes a long time).

To produce a consistent aggregate dataset and ease the comparison between the variables, I decided to standardise them all. This brings more homogeneity to the dataset, without modifying the variations from one semester to the other. We can obtain more consistent results this way.

3.2.2 Time and Geographic Frame

As I wanted to study the time dependency of the demand and the supply quantities, I gathered data that was available for a certain interval of time, namely from the first semester of 2011 to the last semester of 2016. It was not possible to extend the time frame because one or more of the variables were not going beyond this period. However, four semesters in six years gives us already twenty-four periods of time. Then, as quoted in the second paragraph of the "Selected Variables" section, the demand and supply equations are built to form autoregressive models.

Concerning the geographic areas, I evaluated the demand and supply equations at the Flemish community level. Focusing on the community level is a way to analyse their particularities (households' number, wealth level, the number of new constructions every semester, etc.) and check if more aggregate data, like the country inflation rate, can be integrated in all of them without creating inconsistencies. Moreover, in several articles about the Belgian housing market, I noticed there were big differences between Wallonia and Flanders, especially in terms of price levels, economic growth, type of industries, unemployment rate (Degroof Petercam, web site, 2018 and L'Echo, 2019) and building permit regulations (Conseil francophone de la Fédération Royale du Notariat belge (Fednot), web site, 2020). Limiting our study to one of those regions strengthens the outcomes.

At last, I decided to focus on houses, excluding apartments because those are two different markets. People move from apartments more frequently and they tend to be rented more often, while houses purchase is the result of well-considered investments.

3.2.3 Limitations

During the data handling, I encountered some limitations. First, some data were only accessible for a more aggregate level (often for the whole country). This was the case for the inflation rate, the consumer confidence, the mortgage interest rate, the housing inflation and the labour cost. I thus assumed that the level of those variables was of the same value for all communities. Then, other data were well accessible for each community, but only for the whole year (and not for each semester). This restriction applied for the inflation rate (again), the number of households, the households' income and, only

for the year 2011, the number of new constructions. However, this is less disturbing as most of those data remain stable over the year and therefore, per semester. For instance, if we take the number of households, it should not vary extremely fast as families will not move every month and having a baby do not grow the number of households as it is still part of the same family. Similarly, the households' income per year can easily be divided per semester as this one is mostly composed of monthly wages and salaries. Finally, the number of new constructions is pretty regular through time and it should not be a problem to have split the year number through the different semesters (only for the year 2011).

For the mortgage interest rate lagged of one year variable, the semesters' values for the year 2010 was missing in the source used (cf. "Sources" section), except for the last semester. Through comparison with other sources, e.g. SPF Finances, that gave a similar value in January and in June, I concluded we could aggregate this last semester's value as the value for each semester.

Then, many data were containing missing values. At first, Filling those with the average value would have biased the outcomes for certain variables, like the price that determined the censoring indicator. Therefore, I decided to eliminate every line with missing values (to avoid misinterpretation by the computer software). The number of observations left is still high (4422) and will still provide significant results.

3.2.4 Sources

The choice of the source was not simple. After contacting many real estate firms without any concrete return of information (because of confidentiality reasons), I decided to use data available online. Obviously, I did not find a complete dataset available with all variables of interest. This was a real work of combining differently constructed datasets or documents that contained one or more variables that suited our search.

First, I found the mortgage interest rate from the documents *notaries' barometer* (Fednot, 2020) of first semester 2010 until last semester 2016 and I manually transcribed them into an excel file. This source is reliable because data are entered by notaries themselves.

The customer's confidence figures were coming both from this *notaries' barometer* and the National Bank of Belgium (2020) official web site as the two provided the same numbers.

The rest of the data was found on the *Statbel* web site which is the official Belgian Federal Government (2017) web site to provide Belgian figures and statistics. However, as I explained previously, every dataset was differently constructed and my part of the work was also to select the correct time and geographic frame and then, gather all data

in one dataset.

3.3 Theoretical Models Used

3.3.1 Wald Test

I come back now to the likelihood function (3.9) from which the Wald Test will be executed [4]. The maximum likelihood estimator (MLE) $\hat{\alpha}$ of $\alpha = (\theta, \beta, \eta, \sigma_D, \sigma_S, \rho)$ will be obtained by maximising the likelihood function with the given data over the parameter space $A = \{(\theta, \beta, \eta, \sigma_D, \sigma_S, \rho) : 0 \leq \theta \leq 2, \beta \in \mathbb{R}^p, \eta \in \mathbb{R}^q, \sigma_D > 0, \sigma_S > 0, -1 < \rho < 1\}$:

$$\hat{\alpha} = (\hat{\theta}, \hat{\beta}, \hat{\eta}, \hat{\sigma_D}, \hat{\sigma_S}, \hat{\rho}) = \text{argmax}_{\alpha \in A} \text{L}(\alpha)$$

Actually, $\hat{\alpha}$ is a consistent estimator of α if it has an asymptotic normal distribution with zero mean and the following estimated variance-covariance matrix V:

$$V = A(\hat{\alpha})^{-1} B(\hat{\alpha}) A(\hat{\alpha})^{-1}$$

Where $A(\hat{\alpha})$, the expectation of the second derivative of our likelihood function, and $B(\hat{\alpha})$ are worth:

$$A(\hat{\alpha}) = (E\big\{\frac{\partial^2}{\partial\hat{\alpha}_i \partial\hat{\alpha}_j} log f_{Z,\delta|X,W}(Z, \delta|X, W; \hat{\alpha})\big\})^{p+q+4_i, j=i}$$

$$B(\hat{\alpha}) = (E\big\{\frac{\partial}{\partial\hat{\alpha}_i} log f_{Z,\delta|X,W}(Z, \delta|X, W; \hat{\alpha}) \cdot \frac{\partial}{\partial\hat{\alpha}_j} log f_{Z,\delta|X,W}(Z, \delta|X, W; \hat{\alpha})\big\})^{p+q+4_i, j=i}$$

When we invert $A(\hat{\alpha})$, we obtained the well-known Fisher's information matrix. The Wald test will compare each coefficient parameter to the 0 value ($H_0 = 0$ for all parameters) and compute a p-value for each parameter based on this estimated variance-covariance matrix V.

The Wald test has been conducted through several approaches to test the stability of the outcomes. First, a survival regression and then, a linear regression have been applied to choose two sets of initial coefficients. Second, two algorithms (the "nloptr" and "nlminb" algorithms, both are optimisation algorithms for several parameters at a time) were conducted for each of those initial values to test if both were converging to

[4] the complete development can be found in the paper of Van Keilegom and Deresa (2019)

the same final coefficients. To avoid redundancy, only the coefficients results based on the initial values obtained from the survival regression were kept. The results obtained from the nlminb algorithm can be found in Appendix B while the results given by the nloptr algorithm are presented in the core text.

3.3.2 Survival Curve

In this study, I will also try to adapt the survival curve, often used in the survival analysis context, to our housing market context. Theoretically, the survival curve is drawn based on the survival function defined as follows:

$$S(t) = Pr(T > t)$$

with t, a point in the time and T the variable indicating the time at which the event occurs. Commonly, $S(t)$ indicates the probability that the time at which a patient dies occurs later than at time t. This function must be a decreasing function because, for example, the probability that a patient survives at least 3 days can never be lower than the probability that a patient survives at least 4 days (if the patient is not cured, the disease will worsen the patient's conditions through time and thus decrease its chances of survival through time). Similarly, the probability that the event time T is greater than $t = 0$ (the beginning of the study) is always 1, while when $t = 2$, it is lower than 1 as the patient could have died at $t = 1$. In this context, the survival curve asymptotically equals 0 because no human lives eternally. Here is a graph of a survival function extracted from the paper "Comment lire une courbe de survie?", translated as "How to read a survival curve?" (Silvy Laporte, 2005):

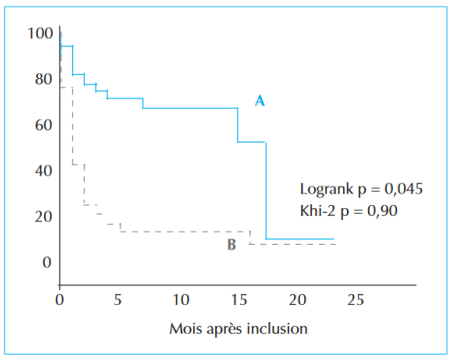

Figure 8: Two Represented Survival Curves for Two Different Treatments

Source: Laporte, S. (2005). Comment lire une courbe de survie? *Medecine thérapeutique*, 11(6), p.423

Actually, those curves look more like stairs. This is because it is an approximation of the survival curves by the Kaplan-meier method that assumes that the event time and the censor time are independent. This method is based on data and, in this context of diseased patients, measures the fraction of patients that are still living at time t with treatment A or B. the "flat" parts reflects no deaths in the provided data and thus estimate the fraction as constant until new deaths. This estimator is presented as:

$$\hat{S}(t) = \prod_{i:t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

With t_i the time at which at least one event occurs, d_i the number of events at time t_i and n_i the number of patients for which the event did not occur yet (and that is not censored yet either).

Now the challenge is to convert this survival curve properly into a demand and supply transaction curve. The t will now represent the number of transactions (instead of the time) and will thus start from the highest transaction found in our dataset instead of starting from zero to keep the aspect of a decreasing function. The survival curve in the housing market context will be interpreted as the probability that a number of demand transactions T (resp. supply transactions) will be lower than the number of transactions t. This will allow to analyse at which level of transactions we are more likely to see a demand quantity (resp. a supply quantity) as the quantity observed in the market.

3.3.3 Goodness of Fit Plot[5]

The goodness of fit plot is assessed through a PP plot (probability-probability plot or percent-percent plot) which compares two cumulative distribution functions plotted on top of each other's. Here, the estimated cumulative distribution function of the dependent variable is evaluated relative to the respective empirical quantities.

Those empirical quantities are defined by the $F_n(z)$ distribution function that assumes that every observation of the sample has a probability of $\frac{1}{n}$ to come up:

$$F_n(z) = n^{-1} \sum_{i=1}^{n} I(Z_i \leq z) \tag{3.10}$$

[5]The Goodness of fit plot and the Goodness of fit test are well-known models. However, their particular formulation based on model (3.1) are retrieved from the thesis written by Anaïs Monetaire (2019) that apply them in a survival analysis context.

I will compare it with the distribution function of the observed transactions number Z from model (3.1):

$$\begin{aligned} F_Z(z;\alpha) &= \int\int P(Z \leq z|X=x, W=w) f_{X,W}(x,w)dxdw \\ &= \int \Phi(\frac{\Lambda_\theta(z) - w^T\eta}{\sigma_S}) f_W(w)dw + \int \Phi(\frac{\Lambda_\theta(z) - x^T\beta}{\sigma_D}) f_X(x)dx \\ &- \int\int \Phi(\frac{\Lambda_\theta(z) - w^T\eta}{\sigma_S}, \frac{\Lambda_\theta(z) - x^T\beta}{\sigma_D}; \rho) f_{X,W}(x,w)dxdw \end{aligned} \tag{3.11}$$

In terms of the MLE , we can approximate the above formula as follows:

$$\begin{aligned} F_Z(z;\hat{\alpha}) = \frac{1}{n}\sum_{i=1}^{n} \Phi(\frac{\Lambda_{\hat{\theta}}(z) - W_i^T\hat{\eta}}{\hat{\sigma_S}}) + \frac{1}{n}\sum_{i=1}^{n} \Phi(\frac{\Lambda_{\hat{\theta}}(z) - X_i^T\hat{\beta}}{\hat{\sigma_D}}) \\ -\frac{1}{n}\sum_{i=1}^{n} \Phi(\frac{\Lambda_{\hat{\theta}}(z) - W_i^T\hat{\eta}}{\hat{\sigma_S}}, \frac{\Lambda_{\hat{\theta}}(z) - X_i^T\hat{\beta}}{\hat{\sigma_D}}; \hat{\rho}) \end{aligned} \tag{3.12}$$

To compare both distributions ($F_n(z)$ and $F_Z(z;\hat{\alpha})$), a diagonal line going from (0,0) to (1,1) is plotted and if the two distributions are equivalent, then the plots follow this line.

3.3.4 Goodness of Fit Test

This test is complementary to the P-P plot. Again, the purpose is to describe how well the distribution under the flexible parametric model fits the distribution of the observed number of transactions Z. First, I use the following null hypothesis:

$$H_0 : P(Z \leq z) = F_Z(z;\alpha), \text{ for some } \alpha$$

A difference with H_0 could be due to a model misspecification of either the event or censoring model parameters as both are part of the formula (3.11).

Again, I will compare $F_n(z)$ and $F_Z(z;\hat{\alpha})$ with the use of the Cramer-von Mises statistic this time, defined hereafter:

$$T_{CM} = \int_{\mathbb{R}} n\{F_n(z) - F_Z(z;\hat{\alpha})\}^2 dF_Z(z;\hat{\alpha}) \tag{3.13}$$

Therefore, a large value of this T_{CM} statistic means that the model is potentially wrong and the null hypothesis must be rejected. More specifically, if the observed statistic T_{CM}^{obs} is greater than the $(1-\alpha) \times 100\%$ quantile of the distribution of T_{CM}, H_0 is rejected at the $\alpha\%$ significance level.

Through a bootstrap method, I will characterise the T_{CM} distribution under the null hypothesis. First, I will create B bootstrap samples from the defined model in (3.1) by replacing α with the MLE estimator $\hat{\alpha}$. From each of those bootstrap samples, I can calculate the $T_{CM,b}^{*}$ statistic $(b = 1, ..., B)$ which will give me a T_{CM}^{*} distribution. If the T_{CM}^{obs} is greater than the $(1-\alpha) \times\ 100$ quantile of the distribution of T_{CM}^{*} just obtained from the bootstrap samples, I reject H_0 at the $\alpha\%$ significance level.

3.4 R Software

The software used to compute all those statistical models is R. The reason is that the original statistical models code has been developed by N.W. Deresa and I. Vand Keilegom (2019) in R. However, the latter had to be adapted to our more specific context: the housing market. The code for this work is provided in the appendix. Also, R is a powerful computer program and allows to construct our own functions and models which are not all preprogrammed yet.

Chapter 4

Results

4.1 Wald Test

The R code for this part is presented in Appendix A.3.1. Before going through the analysis, it is important to remember first that the coefficients of the standardised variables are obtained relative to the Yeo-Johnson transformed variable (3.3) and not to the original variable. I tried different values of θ (0.5,1,1.5), the parameter value for the Yeo-Johnson transformation (referred as "alp1" in the code), to test the stability of the results and they were almost constant from one value to another (with a difference of a tenth at most, cf. Appendix D). Furthermore, I tested the model only on three communities separately, then on all communities to see if there were any significant differences. It was not the case and I consider here only the aggregate outcomes to be general enough. Finally, I applied the Wald Test on a **dependent model**, which considers the dependency between the demand and the supply quantities, and an **independent model** (both considering the disequilibrium of the market). The direction of the coefficients were the same for both. However, some key variables like the price for the supply equation and the interest rate for the demand equation were not significant in the independent model. This does not seem plausible (cf. Literature Review and Introduction Chapters). Moreover, we obtained a significant coefficient for the dependency variable in the dependent model. Therefore, I will only present the dependent model coefficients here and provide the independent coefficients in the Appendix C. I will attempt to give an interpretation of those results in the Discussion Chapter.

First and foremost, the two algorithms converged to almost the same coefficients and I will thus only present one of them in this Chapter, the results of the other being provided in the Appendix section. Those analogous outcomes prove their stability.

The value of the parameters coefficients and their p-values using the survival regression and the "nloptr" algorithm are all presented in Table 3. The ρ value, which indicates the level of dependency between the supply and the demand quantity is negative and significant at the level of 1%.

Table 3: Wald Test (Dependent Model, Survival Regression, nloptr Algorithm)

	Estimate	Standard Error	P-value
beta_int	1.756	0.1056	0
beta_Infl_rate	-0.08003	0.01508	0
beta_DPrice	0.04411	0.008694	0
beta_TransQ-1	0.1451	0.02228	0
beta_NH	0.04724	0.01262	0.00018
beta_CC	-0.01646	0.005365	0.00215
beta_IR	-0.01264	0.00599	0.03482
beta_HI	0.06177	0.01293	0
beta_HINC	-0.03816	0.008694	1e-05
eta_int	1.516	0.06487	0
eta_HIT-1	0.0183	0.006125	0.00281
eta_Infl_rateT-1	-0.061	0.01116	0
eta_DPriceT-1	-0.0132	0.004161	0.00151
eta_TransQ-1	0.175	0.02715	0
eta_IRT-1	0.03987	0.00798	0
eta_BankT-1	0.01556	0.005908	0.00846
eta_NC	0.01103	0.004654	0.0178
eta_DLCT-1	0.01905	0.004922	0.00011
sigma1	0.255	0.03695	0
sigma2	0.2133	0.03087	0
rho	-0.6122	0.03722	0
alp1	0.5521	0.09489	0

Table: Wald test

>

Moreover, all coefficients are significant at the level of 5%. They thus all have a significant effect on their respective equation. At first sight, we come up with surprising

direction for certain coefficients. For instance, the price has a positive coefficient for the demand and a negative one for the supply (even if we start with a reverse direction in the initial coefficients). In an equilibrium model and without considering any dependency between demand and supply, we observe the opposite, but taking into account those characteristics, the situation looks different.

The interest rate keeps the usual negative direction despite the dependency and disequilibrium between the supply and the demand while the housing inflation rate and the consumer confidence coefficients have, again, surprising signs regarding the demand equation.

Looking more at the supply side of the test, other than for the price variable already examined before, we have a surprising direction for only one variable, the labour cost.

4.2 Survival Curves

The R code associated with the Survival Curves can be found in Appendix A.3.2. Recall that the quantity observed in the market is always the minimum of the supply and the demand quantities (Fair and Jaiffee, 1972). In Figure 9, the red curve represents the supply quantities observed through all communities and all time periods whereas the black curve stands for the demand quantities.

These curves show that at a higher level of transactions, supply is mainly observed (the red curve) while from about 20 to 600 transactions we observe both quantities and again from 0 to 20 mainly supply quantities (but they are less numerous, and thus less significant). Then, about 90% of the transactions numbers are below 100. Therefore, Figure 10 presents a second plot of the transaction levels, with transactions between 0 and 200 to have a clearer view within this interval.

4.3 Goodness of Fit Plot

The R code corresponding to the PP plot is situated in Appendix A.3.3. Thanks to the large number of observations, it seems pretty clear from the PP-plot in Figure 11 that the points are almost aligned on the straight line. This assumes there is probably no misspecification of the model causing deviation regarding the empirical distribution. A confirmation of this hypothesis is provided in the "Goodness of fit test" section.

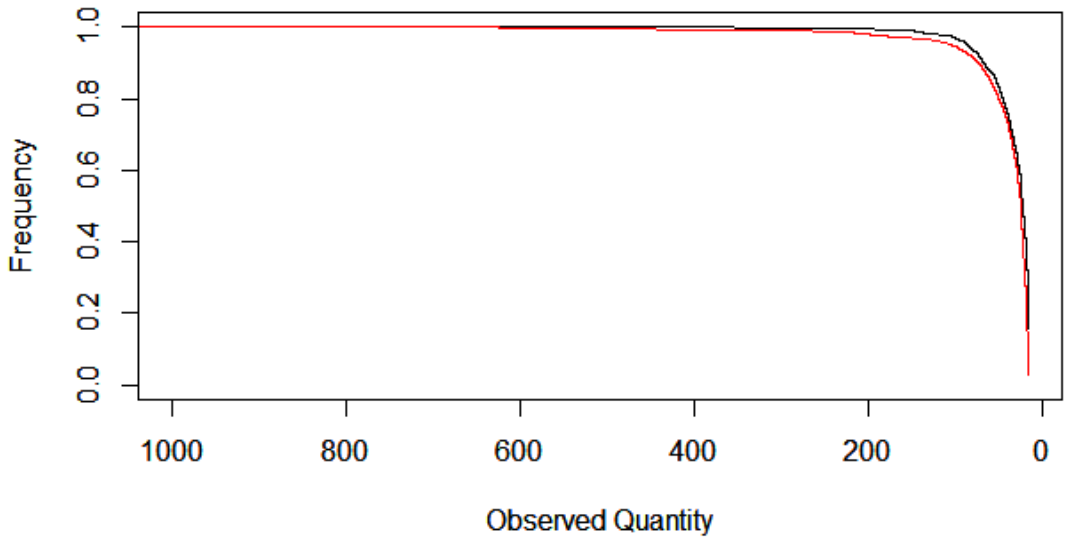

Figure 9: Demand and Supply Survival Curves

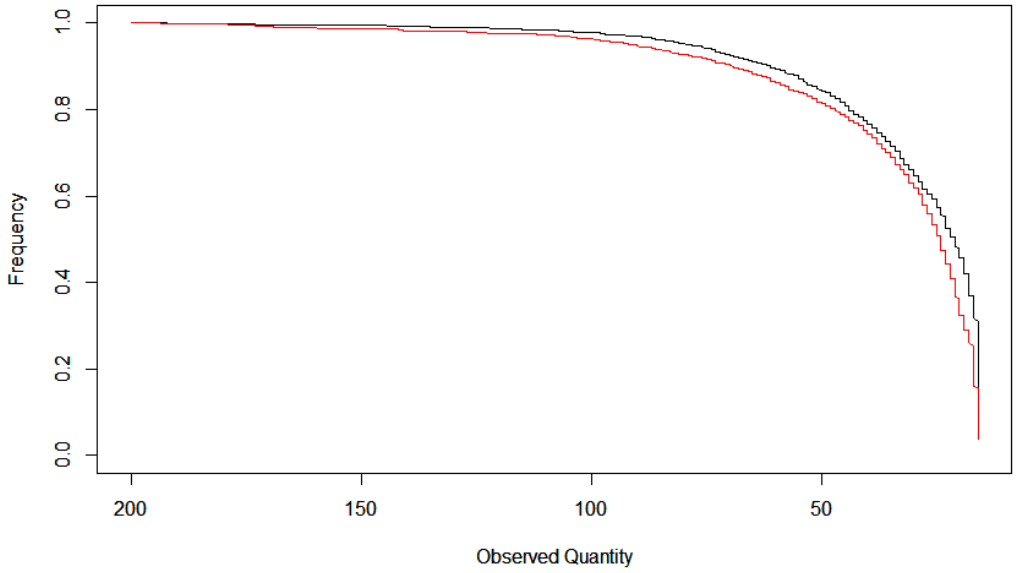

Figure 10: Demand and Supply Survival Curves from 0 to 200 Level of Transactions

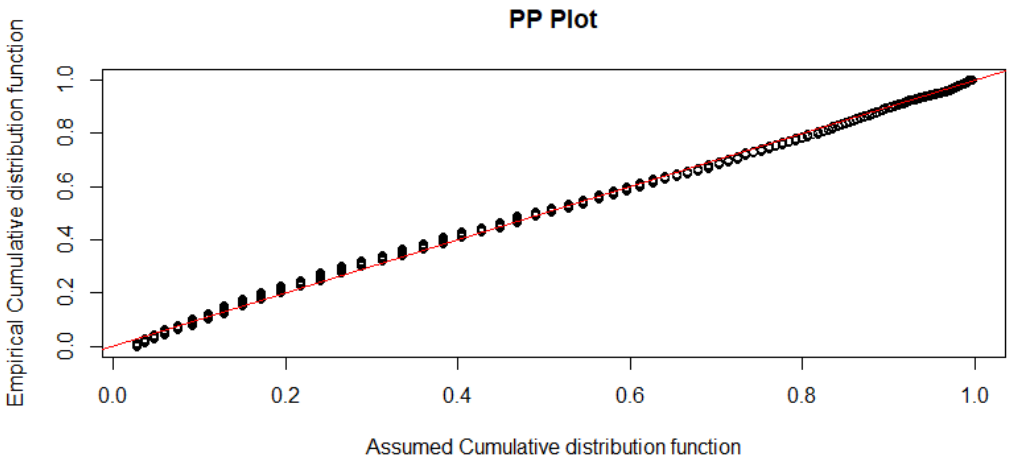

Figure 11: P-P Plot

4.4 Goodness of Fit Test

The p-value of the Cramer-von-Mise clearly indicates an acceptation of the null hypothesis. Through the 4422 observations, which gives us a large panel on which to run the algorithm, the latter even converged to a p-value of 1. Therefore, I cannot conclude that the proposed model is incorrect and there is no reason that it provokes deviation from the empirical distribution. The number of Bootstrap samples chosen to compute this Cramer-Von-Mise statistic was 200. The R code is provided in Appendix A.3.4.

Table 4: Cramer Von Mise Statistic and P-value

Q90	Q95	Observed Cramer-von-Mise	P_value
4.417	4.602	0.6505	1

Chapter 5

Discussion

Although the housing market has been studied by surely hundreds of searchers through time, it is still not completely understood in today's society. Being of a very high complexity, mixing economic mechanisms and inter-market relationships, its deep logic can only be revealed step by step thanks to technological, mathematical, statistical and economic innovations as they occur. In this Chapter, I will discuss the results in order to bring my own contribution as well.

5.1 Problematic and Results Interpretation

5.1.1 Wald Test

My main objective was to focus on the disequilibrium characteristic of the market and to take into account the potential dependency existing between the demand and the supply without omitting the time dependency aspect through the implementation of an autoregressive model. Filling this literature gap was a real challenge but with the use of a survival analysis model, I came up with unexpected outcomes that could change our view of basic theoretical frameworks.

First, a dependency between the supply and the demand quantities is confirmed by our model at a level of "-0,6". This indicates that there is well a high dependency between those two quantities that shouldn't be neglected when analysing their behaviour. A natural interpretation would be that when the demand is high the supply is low because of the inherent disequilibrium between the two due to the fixed short-term supply quantity (the construction process takes time). Therefore, the overestimated supply reaction often appears later in time and we observe the opposite situation.

Then, looking at the price variable and taking into account the disequilibrium between the demand and the supply quantities and their dependency level ρ, new interpretations about the direction of one given the other can be brought into light. Indeed,

if the price increases, the incentive to sell houses is higher and consequently induces a decreasing in the supply (because houses are sold faster and building new houses is not possible in the short-term), which makes, in turn, the demand increase because of their "-0,6" level of dependency (the ρ value). This demand increase could be due firstly to the fact that it can be less easily satisfied without new constructions immediately available, or secondly to anxiety caused by a further price increase anticipation (households want to buy faster to pay their home at the lowest price). Therefore, the price consequences on the supply quantities available have a more important effect on the demand quantities than the price itself. This negative direction of the price on the supply was argued by Jang Sewoong et al.(2010) also, the disequilibrium (evaluated through the unsold new housing stock in their work) being the main reason for this unexpected direction (cf. explanation of Figure 2, page 9).

Also, the choice of an autoregressive model was judicious because the number of transactions in the previous period has a significant effect on the number of transactions in the current period: the more we move forward in time, the more the number of transactions grows.

On the demand side, the interest rate variable appears as expected despite the disequilibrium and the dependency aspects of the market which, incidentally, justifies the policy of lowering key rates adopted by the European Central Bank to stimulate the economy.

Regarding families, The more households there are in a community, the more the demand will be high. Concerning the negative direction of the households' income, it can be due to the restriction I applied on the choice of the buildings: I only selected houses. Households buy houses when they want to start a family and are therefore maybe not at their highest work position: the more people earn, the less they buy houses, but maybe more apartments if they want to invest for instance.

On the supply side, we observed an unexpected positive direction for the labour cost. An explanation found in the broad literature easily justifies this correlation (cf. "literature review" section, Figure 6, Augustyniak et al. (2014)): building developers overestimate the demand and underestimate the future rising cost of an excessive production so they start too many constructions and have to assume the rising cost of materials and labour. This is thus more a correlation relationship than a causal relationship.

The variable number of bankruptcies, rarely included in supply equations, is worth giving an explanation. The positive coefficient regarding the number of bankruptcies along the period before can mean that those events lead to more monopoly from the firms left and they take advantages of this situation to build more.

To conclude, I notice that some covariates take surprising directions (the price, the

housing inflation rate, the consumer confidence, the labour costs), but this is probably due to the dependency and the disequilibrium between the supply and the demand quantities that become even more important than the initial effect of the variable on the demand or supply quantity. This dependency is validated by our ρ value that is significant. This ρ finding shows that analysing those quantities independently would lead to biased results which may explain that certain variables in the independent model are non-significant and overweight (even if the signs of the coefficients stay the same because this model still considers the disequilibrium in the market that also has a great influence).

5.1.2 Survival Curves

The survival curves mainly show that when the level of transaction is high, the quantity observed is mostly supply quantity (recall my consideration to always observe the minimum of the two on the housing market). Actually, having many supply quantities above 600 transactions is not surprising. When the transactions level is high, it becomes difficult for the supply to provide even more houses because of space limitations, building permits limitations, and so on.

Moreover, most transactions level we find on the Belgian market are between zero and 100 as we observed from Figure 9 and Figure 10 a drop in the survival curve from there. Within this interval, slightly more demand quantities are usually observed. However, the two curves look pretty similar and I do not see any particularities for one or the other.

5.1.3 Goodness of Fit Plot and Goodness of Fit Test

Finally, the goodness of fit plot and the goodness of fit test indicate no misspecification of the model. This comforts us to use the selected model.

5.2 Theoretical Framework Review

Furthermore, I assumed that the quantity observed on the market is the minimum of the supply and the demand quantities, as Fair and Jaiffee (1972) proposed. This may be called into questions. Muellbauer (1978) put forwards his doubts about this theory at the macro-level, suggesting instead: $Q \leq Min(D, S)$. Therefore, the minimum condition has to be validated and if not, it becomes really complicated to estimate the actual level of quantity demanded and supplied.

Moreover, the original sample separation is actually unknown but estimated through the technique proposed by Fair and Jaiffee (1972) presented in the literature review: a price increase means an excess demand and therefore, the quantity observed is the supply quantity (cf. "directional method I"). However, this method could not be 100% reliable

because the price mechanism still remains complex. The works of Goldfeld and Quandt (1975) and Kiefer (1979) even demonstrate loss of information when sample separation is unknown and its estimation is still subject to weaknesses.

Also, the survival model used assumed a bivariate normal distribution of the error terms but this might not necessarily be true.

Finally, an autoregressive model of order one was built to consider the time dependency. The lagged dependent value was significant and validated our choice. However, we could have built an autoregressive model of order two, three or even more and only retain the significant lagged values. Also, testing the stationarity of the time-series model would tell us if this time dependency evolve through time.

5.3 Data Limitations

Unfortunately, I encountered some limitations when gathering the data and made selection choices which could weaken my results. Firstly, the time frame I used could be even more extended to give more stability to the outcomes. Secondly, I focused myself on a small geographic scale that could not reflect entirely the mechanism at a more aggregate level (like a whole country, continent, etc.) owing to the excessive weight of particularities of each community. Thirdly, even if the study focuses on the communities at each semester period from 2011 to 2016, some variables were only available for one whole year or for the whole country. Those had to be adapted and, even though it was made in a meaningful way, it can produce some inaccuracies.

5.4 Contribution

In this thesis, in spite of obvious limitations in the data collected, in the validation of some theories and in the models used, I provide statistical evidence that the demand quantities in the housing market depend on the supply quantities through this significant ρ value while integrating the disequilibrium and time features in the model. This was revealed by using a novel model developed by Ingrid Van Keilegom and Negera Wakgari Deresa (2019). Thanks to the survival curves, I also show that when the level of transaction is high, we mainly observe supply quantities in the market.

5.5 What About Tomorrow?

From my point of view, there are so many analyses left to make around the housing market, especially through the late technology innovations. New insights to be brought are infinite. However, a first step could be to extend this study to other areas, at a different scale (land scale for instance) and/or within a different time frame.

List of Figures

List of Tables

Bibliography

[1] Augustyniak, H., Laszek, J., Olszewski, K., Waszczuk, J. (2014). Housing market cycles – a disequilibrium model and its application to the primary housing market in Warsaw. *Ekonomia Journal*, 35, 5-23.

[2] Belgian Federal Government. (2017), *Statbel, Belgium in figures*, Retrieved on the 15^{th} September 2019 from `https://statbel.fgov.be/en`

[3] Conseil francophone de la Fédération Royale du Notariat belge (Fednot). (2020). *Baromètre des notaires*. Retrieved on the 31^{th} July 2019 from `https://www.notaire.be/nouveautes/barometre-des-notaires`

[4] Conseil francophone de la Fédération Royale du Notariat belge (Fednot). (2020). *Permis d'urbanisme*. Retrieved on the 12^{th} march 2020 from `https://www.notaire.be/acheter-louer-emprunter/acheter-un-terrain-et-construire/permis-d-urbanisme`

[5] Degroof Petercam. 2018. *Les disparités économiques régionales sont-elles importantes dans notre pays ?*, Retrieved on 12^{th} march 2020 from `https://blog.degroofpetercam.com/fr/economie/1766/les-disparites-economiques-regionales-sont-elles-importantes-dans-notre-pays`

[6] DiPasquale , D., Wheaton, W. C. (1992). The Markets for Real Estate Assets and Space : A Conceptual Framework. *Journal of the American Real E.state and Urban Economics Association, 20(1)*, 181-197

[7] Dunsky, R.M. Follain, J.R. (1997), The demand for mortgage debt and the income tax, *Journal of Housing Research, 8*, 155-199.

[8] Fair, R. C. (1972). Disequilibrium in housing models. *The Journal of Finance*, 27, 207-221.

[9] Fair, R. C., Jaffee, D. M. (1972). Methods of Estimation for Markets in Disequilibrium. *Econometrica*, 40(3), 497-514.

[10] Goldfeld, S. M., Quandt, R. E. (1975). Estimation in a Disequilibrium Model and the Value of Information, *Journal of Econometrics, 3(3)*, 325-348.

[11] Keating, J. W. (1992). Structural Approaches to Vector Autoregressions. *Federal Reserve Bank of St. Louis Review*, 74(5), 35-57.

[12] Kiefer, N. (1979). On the Value of Sample Separation Information, *Econometrica*, 47(4), 997-1003.

[13] Lambertini, L.,Mendicino, C. Punzi, M. T. (2012), *Expectations-Driven Cycles in the Housing Market* [Discussion Paper] Bank of Finland Research, 2-2012.

[14] Laporte, S. (2005). Comment lire une courbe de survie ? *Medecine thérapeutique*, 11(6), 419-423.

[15] Loberto, M., Zollino, F. (2016), *Housing and Credit Markets in Italy in Times of Crisis*, No. 1087, Temi di Discussione [Working Paper], Bank of Italy, Economic Research and International Relations Area.

[16] Maddala, G. S. (1986). Chapter 28 : Disequilibrium, self-selection, and switching models [Book Chapter]. In *Handbook of Econometrics* (Vol. 3, pp. 1633-1688). University of Florida: Elsevier Science Publishers BV.

[17] Maddala, G. S., Nelson, F. D. (1974). Maximum Likelihood Methods for Models of Markets in Disequilibrium. *Econometrica, 42*(6), 1013-1030.

[18] McMillin, W. D. (2001). The Effects of Monetary Policy Shocks : Comparing Contemporaneous versus Long-Run Identifying Restrictions. *Southern Economic Journal*, 67(3), 618-636.

[19] Monetair, A. (2019). *Flexible parametric model for survival data subject to dependent censoring: Comparison, Extension Application of this model to a real dataset.* (Master). UCLouvain, Louvain-la-Neuve.

[20] Muellbauer , J., Portes, R. (1978). Macroeconomic Models with Quantity Rationing. *Economic Journal, 88*(352), 788-821

[21] National Bank of Belgium (NBB). (2020). *Consumer survey* [Online database], Retrieved on the 8^{th} september 2019 from `http://stat.nbb.be/?lang=en`

[22] Quandt, R. E. (1980). *Equilibrium and disequilibrium : transitional models.* Princeton, N.J.: Econometric Research Program, Princeton University.

[23] Sewoong, J., Sanghyo, L., Juhyung, K., Jaejun, K. (2010). Relationship Between Demand-supply in the Housing Market and Unsold New Housing Stocks. *Journal of Asian Architecture and Building Engineering*, 9, 387-394.

[24] Sommervoll, D. E., Borgensen, T.-A. Wennemo, T. (2010), Endogenous housing market cycles, *Journal of Banking Finance, 34*, 557 -567

[25] Van Keilegom, I., Wakgari Deresa, N. (2019). Flexible parametric model for survival data subject to dependent censoring. *Biometric Journal*, 62(1), 136-156.

[26] Wouter, V. (2019, 8 february). L'économie flamande croît toujours plus vite que celle de la Wallonie. *L'Echo*.

[27] Yeo, I. Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry, *Biometrika* 87, 954-959.

UNIVERSITÉ CATHOLIQUE DE LOUVAIN
Louvain School of Management
Place des Doyens, 1 bte L2.01.01, 1348 Louvain-la-Neuve, Belgique | **www.uclouvain.be/lsm**