House price modelling of Denmark's municipalities using vector autoregression and gradient descent

Love Gillberg

Bachelor thesis supervised by:
Rolf Poulsen and Andrea Idini

Division of mathematical Physics
Department of Physics
Faculty of Natural Sciences

Half-time during 4 months

May 2019

LUND
UNIVERSITY

SWEDEN

Abstract

In this thesis, a house price evolution equation of Denmark's municipalities is proposed and solved using both an iterative optimization algorithm, and a closed form solution using Vector Autoregression. The quality of the solutions is investigated, analyzed and compared. Focus is put on the accuracy of the generated price predictions and to what degree the model parameters follow expected features from a well describing model, such as distance and population correlation with parameter values. Overfitting is a central problem using the closed form solution method due to the large number of parameters. It is found that the closed form solution performs badly in describing the system and that the iterative method generated much better models. Depending on the initial conditions, the iterative method more accurately captures the expected features and gives price predictions on four years within 5%. It is also found that the distance between the municipalities has a relatively large importance on the price correlations. The population size is not found to have any noticeable corresponding impact. It is clear that price inflation is a major factor which needs to be more accurately implemented in future work. For example, by exponential inflation adjustment or working with the logarithm of the prices.

Acknowledgements

I express my sincere thanks to my supervisors Andrea Idini (Associate senior lecturer at Mathematical Physics at Lund University) and Rolf Poulsen (Professor, Department of Mathematical Sciences, Copenhagen University) for providing excellent feedback and guidance during the work on this thesis. I learned a lot from their input and they provided valuable information for completing the thesis. Andrea gave the perfect feedback regarding physical connections to the project and about the report and Rolf gave his expertise in the subject of financial markets and perfect suggestions on what actions to take during the work with this thesis.

List of abbreviations

GD . . . . . Gradient Descent

VAR . . . . . Vector Autoregression

EP . . . . . Estimation Period

EP 1 step . . . . . First quarter in the Estimation Period

VP . . . . . Validation Period

VP 1 step . . . . . First quarter in the Validation Period

BS . . . . . Bootstrapping

VP BS . . . . . Bootstrapping during the Validation Period

Profit 1,2 . . . . . Predicted profit and actual profit using investment strategies 1 and 2 during a specified time period. For more details, see section 2.6.

Av/Cop profit . The average house price increase in all municipalities respectively only Copenhagen during a specified time period. For more details, see section 2.6.

MSE . . . . . Mean squared error

MAE . . . . . Mean absolute error

MAPE . . . . . Mean absolute percentage error

ME . . . . . Mean error

MPE . . . . . Mean percentage error

Contents

Introduction

This thesis proposes a model of house prices in Denmark's municipalities using tools and theories related to physics such as the return maps from Chaos Theory [1]. Each municipality has its own average house price, population density and location. Some different versions of the model are compared in relation to the physical properties of the system and in their ability to predict house prices.

The utilized set of data consists of quarterly house prices of all Denmark's municipalities (except five islands due to insufficient data) obtained from Finans Danmark [2]. It covers the time period from and including first quarter of 1991 until and including the last quarter of 2018. The prices are in danish kroner (dkk) per square meter and represent a total of 93 municipalities. For further details about the data, such as which the missing islands are, see Appendix. The municipalities will also be referred to as regions.

The price development for any goods over time is dependent on a lot of different factors, but an important one is price inflation, which is a general increase of prices with time. As the house prices used in this model span over 25 years, inflation has a huge effect on the price levels, see section 2.2. This is taken into account by adding parameters which represent inflation.

Let's briefly introduce the equation of price evolution. Let $x_{i,t}$ denote the average price in Danish kroner per square meter in region i at time t. The equation of price evolution, suggested by my supervisor, is

$$x_{j,t+1} = \alpha_i + \beta_i \cdot t + \sum_{j=1}^{93} \gamma_{i,j} \cdot x_{j,t} \tag{1.1}$$

for all regions $i \in [1, 93]$ and quarterly times $t \in [1, 108]$ where α_i is a price which represents the region, β_i is the inflation parameter and $\gamma_{i,j}$ represents the strength of price interaction between the regions. α_i, β_i and $\gamma_{i,j}$ are the model parameters to be found. The parameters are approximated as time-independent, which corresponds to the assumption that the structural and social conditions of the regions and the demand for them does not change too much during the time period. For further discussion of the equations, see section 2.4.

The purpose of this thesis is to extract as much information about the system as

possible from the model parameters. The model describes the system well if the model fits well to the data and also has predictive accuracy. In such a case, the parameters should be able to give information about how a country's spatial and population properties, affect the future price evolution in all regions. As we have data on the region's populations, population densities, and distances between the regions, it is possible to get estimates on how the influence on price levels changes with those quantities.

The set of equations is solved by minimizing the cost function, which is defined as the sum of the squared errors of the model's fitting to all region's data for a given time period. This is achieved using two different methods:

1. **Gradient descent method**, which is a first-order iterative optimization algorithm. The partial derivatives of the cost function with respect to all parameters are calculated. Starting values of the parameters are set. The value of every partial derivative is then be used to change the corresponding parameter an amount proportional to the partial derivative of the cost function with regards to that parameter to minimize the cost function. The proportionality constant is kept fixed during the process and chosen for optimal minimization behaviour of the cost function and around half its size for converging behaviour of the parameters with number of iterations. The process is iterated repeatedly to reduce the cost function by every iteration until a local minimum is reached closely enough.

2. **Vector Autoregressive method**. The VAR method gives an estimation in Ordinary Least Square sense to the matrix equation $Y = PZ + U$, containing all equations (1.1). Y and Z are matrices containing house prices, P is a matrix that contains all parameters in (1.1) and U are the forecast errors. For matrix definitions, see the Appendix. The ordinary least square estimation of $Y \approx PZ$ is given as $P = YZ^T(ZZ^T)^{-1}$.

We know from Chaos Theory that chaotic behaviour only appears in non-linear non-invertible maps [3]. Since the equations can be written as the matrix equation $Y = PZ$, the functional map is linear if P is invertible. Since the γ parameters are the only terms of non-linearity in (1.1), the functional map is linear if the coupling matrix (with $\gamma_{i,j}$ on row i and column j) is invertible.

Having the possibility to accurately predict the house price next year, could be extremely advantageous economically for the holders of the predictions. Having a better understanding of the underlying physics behind a country's price evolution (how every region's size, population, and location contributes to the other region's prices or what the effect the population densities has on the price development) contributes to the knowledge for possible interventions for a country to protect itself against future house crashes. Most people are aware of the housing crash in 2008, whereby house prices were at that time inflated due to the high loan approval rate by the banks.

When the market understood that the prices were too high, the prices fell drastically, and house owners experienced what happens when their house in reality is worth much less than they previously thought. Had the countries been better prepared for the crisis before it happened, there would have been a large chance that many of the problems could have been mitigated or avoided. Thus, this kind of research could be highly advantageous for governments and the society as a whole.

Method

2.1 First return maps and Chaos

Observing a dynamical system at discrete times generates a sequence of states of the system. In physics, a useful group of tools to study this data are known as iterated maps or just maps. They are used for example to analyze differential equations, to model natural phenomena and in the study of chaos [1]. One sort of those are the first return map, which is shown in one-dimensional form in equation (2.1).

$$x_{t+1} = f(x_t) \tag{2.1}$$

where x_t represents the state of a system at time $t \in N$ and f is called the logistic function [3]. Knowing the state at a specific time, the first return map thus describes the state of the system one unit of time later.

For simplicity, the state in equation 2.1 is represented by a single variable x making into a one-dimensional map. Some systems may be described with one-dimensional first return maps. For example, the time interval between two drops of a dripping faucet may approximately be described by the previous time interval. Another common example is the population of an animal species on an isolated island, where x represents the population fraction of a maximum population capacity of the island [3]. Assuming that environmental conditions remain constant, a simple suitable choice to describe this system is the logistic function

$$f(x) = rx(1 - x) \tag{2.2}$$

where the parameter $r > 1$. The logistic function should decrease for large populations since the food supply is limited. The logistic function for $r = 2$ is shown in the left diagram in figure 1 together with the line with the same derivative as the logistic function in the origo. Applying the logistic function on the first return map results in

the logistic map

$$x_{t+1} = rx_t(1 - x_t) \tag{2.3}$$

We can study this behaviour of the system after a long time. For low r ($r < 3$) the system will converge to a stable solution. For slightly higher r (up to around 3.4), this long-term value splits up in two values between which the logistic map is consecutively changing value with every time step. This is shown in the right diagram in figure 1. The splitting up of long-term value is called a bifurcation and at slightly larger r there occur again a new bifurcation for each long-term value. This process continues all the time as shown in the diagram, faster and faster with r, ensuing chaos, in which a deterministic system with aperiodic behaviour depends sensitively on the initial conditions. [3]

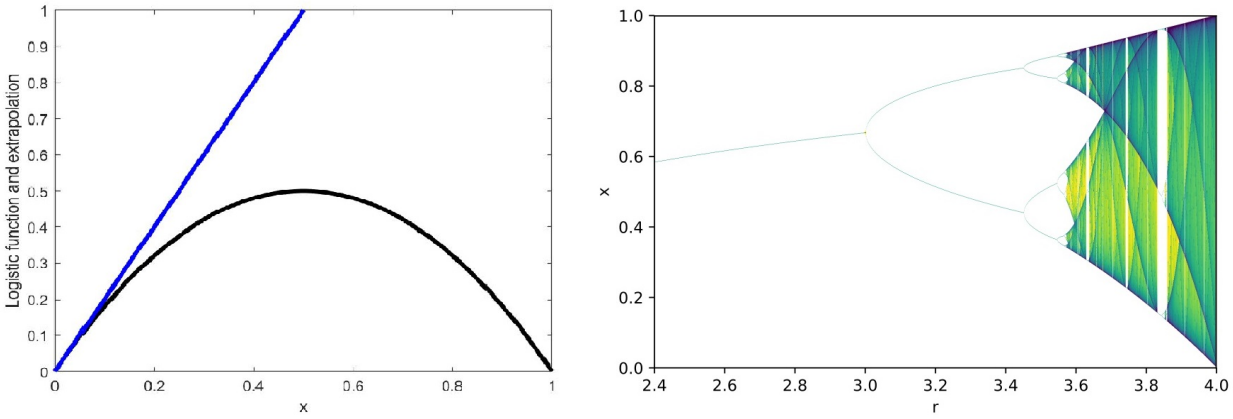

Figure 1: Left: The logistic function 2.2 for $r = 2$ where its derivative in the origo is represented by a blue line [3]. Right: Corresponding bifurcation diagram with r on the x-axis and long term value of the logistic map 2.3 on the y-axis [4].

For systems with higher dimension n, the first return map generalizes to [3]

$$\begin{cases} x_{1,t+1} = f_1(x_{1,t}, x_{2,t}, ...x_{n,t}) \\ x_{2,t+1} = f_2(x_{1,t}, x_{2,t}, ...x_{n,t}) \\ \vdots \\ x_{n,t+1} = f_n(x_{1,t}, x_{2,t}, ...x_{n,t}) \end{cases} \tag{2.4}$$

The number of variables used to describe many related systems may very well be higher. Two dimensional first return maps have been used to analyze the interaction between two species, which gave agreeable results with laboratory experiments [3]. In this thesis, we will use first return maps in many dimensions, such as the model we

introduced in equation (2.1), to study the price evolution of real estate. The number of municipalities gives the system 93 dimensions.

We know historically that markets have periods with unstable development. The earliest market models assumed that markets inherently contain aggregated fluctuations that may cause the market to be inherently unstable. Later on, from late 1950s to early 1990s market models instead mainly assumed that markets was inherently stable and that in the absence of an exogenous shock, the market would incline toward a firm growth path. Thus, the impact of exogenous events was investigated in those models. Examples of such events are wars, demographic events, natural disasters, and new technology. Some exogenous events relevant in time and space for this thesis are the establishment of the Øresund Bridge, 9/11, migration and BREXIT. From around 1990 the interest of models with the endogenous hypothesis again rose. This was partly due to the increased understanding that deterministic dynamical systems may generate chaotic dynamics and have properties that exactly mimic those of certain stable linear stochastic models. Among market models which show endogenous fluctuations, it has been shown that a class of perfect-foresight equilibrium models and another class of indeterminate equilibrium models both are compatible with optimizing behavior and competitive equilibrium.[5]

This thesis does not include modelling of exogenous shocks but they are briefly discussed in the conclusions section.

2.2 Inflation and money creation

Today and in recent history, the money is represented by fiat currency[6], which is not backed by any commodity. Fiat currency can be either cash or electronic money that one use while paying with payment cards or with bank transfers. In Denmark, less than 5 percent of the money supply consists of cash [7]. In countries going more and more cashless like Sweden for example, only around 2 percent of the money supply currently consists of cash, and the rest is electronical currency [8].

As the data used for modelling the house price spans over a long period of time, it is important to have an overview of the concept of inflation and money creation. Fiat currency or money, is created by either the central bank or by the private banks [8]. In the modern economy, the majority of the money is created by commercial banks making loans [9].

When the loan gets paid back, the money created by lending is destroyed [9]. It does never leave the banking system as a whole except if the underlying loan is paid back [10]. However, connected to the loan comes interest, inflating the money supply. At last, the amount of money created depends on the monetary policy of the central bank, usually by adjusting the interest rates [9]. The central bank may also affect the

amount of money directly through asset purchases, known as quantitative easing [9]. To get a feeling of how fast it may increase, one can consider the fact that during around 11 months, between March 2018 and February 2019, Denmark's money supply has increased more than the total amount of cash (bills and coins) in Denmark in February 2019 [7]. Most economists agree on the fact that if the money supply of a nation increases faster than the economical growth, inflation occurs, which results in higher prices in general. Many of the loans are issued for the purchase of a house, making the price increase on houses also directly depend on the size of the loans themselves. The diagram below shows Denmark's money supply M1 (solid orange line) together with the average house price for all regions (dashhot line in blue) during the same time period as the data in this thesis.

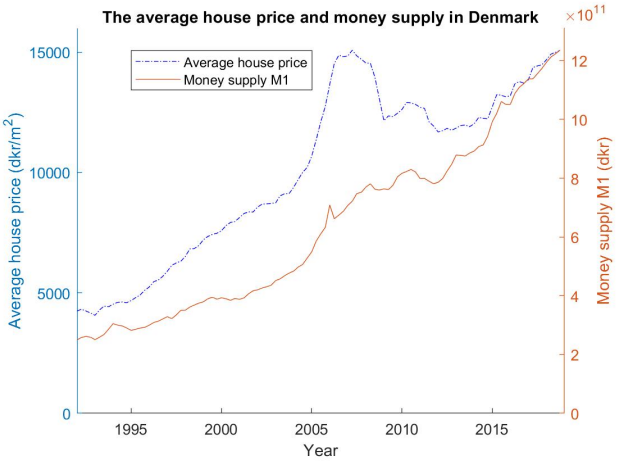

Figure 1: Dashdot in blue, left y-axis: The Average house price in Denmark for all 93 municipalities. Solid orange line, right y-axis: The Money Supply M1 of Denmark [11]. The x-axis represents time in years from first quarter 1992 to last quarter 2018. The data shows that the money supply has quintupled during the around 27 years.

2.3 General features of Denmark

Denmark consists of three large islands and other small islands. Figure 2 shows graphically the the prices last quarter of 2018 in Denmark by region on a map to the left, and the price development with time during the period 1992-2018. The price difference between different regions is huge and in general the prices are much lower the further away from the capitol the regions are located. From the right figure, it is noticeable that the price development for the different regions does not behave smoothly increasing.

It has a general increasing trend overall but otherwise it looks rather irregular, even though many different regions has similar up- and downtrends at similar times.

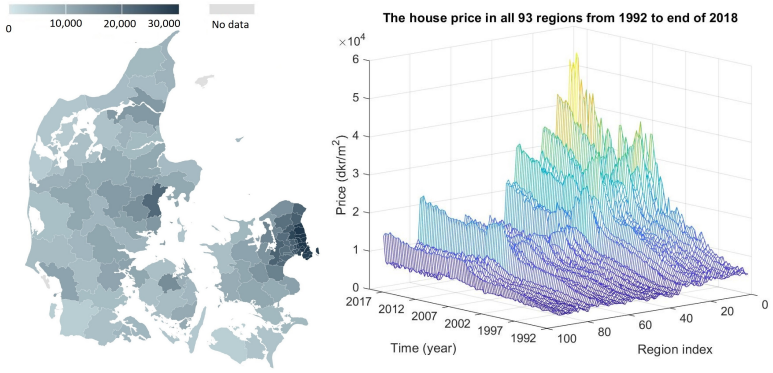

Figure 2: Left: Pricemap of the house price per square meter by region in Denmark [12]. Right: The house price for every region during the whole data period (1992-2018). Copenhagen has region index 1 and is in the diagram hidden behind the curve of Frederiksberg with region index 2, which reaches the highest price overall.

2.4 Equations of price evolution

Let $x_{i,t}$ denote the average price in Danish kroner per square meter in region i at time t. Then the task is to estimate the price in the same region the next quarter $x_{i,t+1}$, for all different regions. The first term introduced to represent this estimation is α_i that represents a time-independent kind of base price of the region. The next term in the estimation is $\beta_i \cdot t$, which defines a general temporal dependence on the price in region i that represents the inflation.

When the prices in Copenhagen increase, it is reasonable to expect that the house price in Århus, second largest city in Denmark, is affected. It would probably increase as well, as now more people would have an economical incentive to choose the second largest city to move to, which might be the most similar city to the largest city. By extending this reasoning, the estimation of $x_{i,t+1}$ should also consist of a fraction $\gamma_{i,j}$ of the price in any other region j, $x_{j,t}$. This will directly be taken into account to by adding the terms $\sum_{j=1}^{93} \gamma_{i,j} \cdot x_{j,t}, j \neq i$. As long as the model is describing the system well, the parameter $\gamma_{i,j}$ may be seen as the "price influence on region i by region j" and $\gamma_{i,j}$ and $\gamma_{j,i}$ both represent "price coupling" parameters between the regions. The assumption that the parameters are time-independent represents the assumption that the properties of the regions and the corresponding demand does not change very much during the time period. Examples of such properties of the system are houses,

roads, parks, shopping malls and pollution. This is an approximation done to limit the number of parameters.

The house price next year in, let's say Copenhagen, should depend to a large extent on the price there this year, probably more than the price last year. Therefore, the thesis should also take this into account. This is done by adding the missing term $\gamma_{i,i}$ in the expression above, i.e. to also allow that $j = i$. $\gamma_{i,i}$ will then represent the price coupling of region i's former price to its current price. Thus, the complete set of equations is

$$x_{i,t+1} = \alpha_i + \beta_i \cdot t + \sum_{j=1}^{93} \gamma_{i,j} \cdot x_{j,t} \tag{1.1}$$

for all regions $i \in [1, 93]$ and quarterly times $t \in [1, 108]$.

2.5 Methods for solving the equations

1. **Vector Autoregressive method**. The VAR method gives a closed form estimation in Ordinary Least Square sense to the matrix equation $Y = PZ + U$, where Y and Z are matricies containing house price data, P is a matrix that contains all the model's parameters in (1.1) and U is a matrix containing all the errors [13]. The ordinary least square estimation of $Y \approx PZ$ is given as $P = YZ^T(ZZ^T)^{-1}$ [13]. For definitions, see the Appendix. VAR methods are commonly used for forecasting variables within economics such as growth of Gross domestic product (GDP), inflation, other macroeconomic variables and oil prices [14]. This method adapts the parameters equally in the best possible least square sense to the training data. It will give a single model whose ability to describe and predict the system is compared to those of the gradient descent method.

2. **Gradient descent method**, which is a first-order iterative optimization algorithm. It minimizes the cost function, defined as

$$cost = \sum_{\forall i,t} (x_{i,t+1} - y_{i,t+1})^2 \tag{2.5}$$

where $i \in [1, 93]$ and $t \in [1, 108]$ and $y_{i,t+1}$ is the data. $x_{i,t+1}$ is in this case the one time iterated price (from the data points) according to equation (1.1), so that

$$x_{i,t+1} = \alpha_i + \beta_i \cdot t + \sum_{j=1}^{93} \gamma_{i,j} \cdot y_{j,t} \tag{2.6}$$

for all regions $i \in [1, 93]$ and quarterly times $t \in [1, 108]$.

The partial derivatives of the cost function with respect to all parameters is calculated separately. Starting values of the parameters is set. The value of every partial derivative is used to change the corresponding parameter value p to $p - L\frac{\partial cost}{\partial p}$ where $\frac{\partial cost}{\partial p}$ is the partial derivative of the cost function with regards to parameter p and L is the *Learning rate*. This step is called iteration or gradient descent iteration. In the model's parameter space, the parameter-vector is changed in exactly the direction that minimizes the cost function the fastest. The process will then be iterated repeatedly to reduce the cost function by every iteration until a local minimum is reached closely enough. The process is called the learning process. During this process, the learning rate, different for each set of parameters (see section 3.2.1): $L = L_\alpha, L_\beta, L_\gamma$, is kept fixed. The number of parameters is large, 8 835 $(93 + 93 + 93^2)$ and approximately as large as the data set, 8556 $(93 \cdot 92)$ or 10044 $(93 \cdot 108)$ (without or with out-of-sample data). Thus, many local minima is expected to exist. If not proper actions are taken, the model might be too specific to the sample data and miss out on general trends of the system or on important physical properties of the system. If the model is overfitted, it should not be accurate in making real predictions. To deal with this, two major actions are taken. The first one is out-of-sample validation, see section 2.5. The second one is to examine extra carefully the model's learning process by using parameter starting values that follow the expected general features in section 2.6, at the same time as monitoring the parameter evolution during the learning process.

2.6 Out-of-sample-validation and profit statistics

The data set, containing data for 108 quarters is divided into an *estimation period* (EP), containing the first 92 quarters of data, and a *validation period* (VP) containing the last 16 quarters of data. There have been suggested general recommendations on the length of VP in predictive models on 20% or more [15], however, it is desired to include data that cover enough of the time period in the backwash of the financial crash 2007-2008, which had huge effects on the housing market, see for example figure 1. The model parameters are estimated from the estimation data set and the model is then used to give predictions for the whole 16 quarter validation data set (VP) and for only the first quarter of the validation period (VP 1 step). The results of using bootstrapping[15] is also investigated, meaning that one prediction is made at a time, after which the model parameters are extracted again from the enlarged data set that now consists of the data in the EP plus those predicted prices. Then the process is repeated until all 16 quarters of prices are predicted and bootstrapping never use any data from the VP. The errors of the predictions with and without bootstrapping are analyzed and compared to the

errors of the model's fit to the data in the estimation data set.

It is important to keep in mind the difference between the model's fit to the estimation data from which the model parameters are estimated, and the model's prediction of the validation data, from which the model parameters are never estimated and lie in the "future" of the EP. When it comes to our own future, if new predictions of this kind would motivate new investments according to the predictions, there could be a risk of psychohistory, meaning that future price development would be altered due to the prediction itself. It corresponds to the measurement problem in quantum mechanics where the wave function is collapsed by measuring the system and the future development thus is altered compared to in the case of not having performed the measurement. This needs to be taken into account for investor if the use of these kind of models drastically changes, in the same way that traders have to understand the market behaviour and the group psychology of other investors. The values of the model predictions in the EP are called fitted prediction values and their errors are called residuals. The predictions in the VP are the prediction values with corresponding prediction errors. The accuracy of the model to predict the validation data is a measure of the model's ability of describing the system correctly.

The predictions made from the different models is also analyzed with regards to the hypothetically gained profit using two investment strategies.

1. Strategy 1. Corresponds to invest equally much in all regions that are predicted to increase in price. The investment is done at the beginning of the given time period. Thus, the predicted relative price change for all regions are calculated, and the average among those is calculated and compared to the actual average relative price change for the same regions during the same time.

2. Strategy 2. Corresponds to investing equally much in the five regions with the largest predicted relative price increase. The investment is done at the beginning of the given time period. Thus, the predicted relative price change for all regions are calculated, and the average among the predicted five largest positive relative price changes is calculated and compared to the actual average relative price change for those regions.

The profits for the strategies are correspondingly called "Profit 1" and "Profit 2" and are also compared to the average relative price increase of all regions ("Av profit") and of only Copenhagen ("Cop profit") during the time period.

2.7 Expected general features of the model

If the model describes the system accurately physically, the author expects the model to have the following general features

i. $\gamma_{j,i}$ for all $j, i \in [1, 93]$ has a general, but not strict or very strong trend towards being positive. The prices are generally governed by supply and demand, and if any house price increases in a region j, it is expected to push buyers or demand from this region into the rest of the regions, for example region i. Then, due to supply and demand, the price contribution on region i by region j is larger. From equation (1.1), we see that this price contribution on $x_{i,t+1}$ is $\gamma_{i,j} \cdot x_{j,t}$, meaning $\gamma_{i,j} > 0$. If the price in region j, $x_{j,t}$ instead decreases, the price contribution of region j, on region i must be smaller (than it would otherwise have been). This means that $\gamma_{i,j} \cdot x_{j,t}$ must still be positive and hence $\gamma_{i,j} > 0$ also in this case. There are reasons to why this inequality is not expected to be strong: 1) α, β alone by definition (but not to a high accuracy) is expected to tend to cover the general trend of the price evolution i.e. base price and inflation. In such a case, negative γ's are expected to exist to compensate for the positive ones. 2) Overfitting reasons: There are many times more γ parameters than α and β parameters that will give rise to the existence of many local minima of the cost function.

ii. the price coupling of the regions with themselves should in general be much higher than the price coupling with other regions i.e. $\gamma_{i,i} > \gamma_{j,i}, \gamma_{i,j}$ for $j \neq i$. Especially, the inequality $\gamma_{i,i} > \gamma_{j,i}$ for $j \neq i$ should hold stronger because $\gamma_{j,i}$ represents a measure on the price influence in region j by region i, which should be lower even if region i has a larger population.

iii. $\gamma_{i,j}$ and $\gamma_{j,i}$, $j \neq i$, are in general not the same unless the corresponding regions has similar population and other properties such as supply. If region j has the largest population, it is expected that $\gamma_{i,j} > \gamma_{j,i}$.

iv. the price coupling between regions that are closer to each other i.e. with a smaller Euclidean distance, are expected to be higher than if they were further away from each other. This means that $\gamma_{j,i}$ (that are expected to be mainly positive) should correlate negatively with the distance between regions j and i. This is under assumption that the relative populations of the compared pairs of regions remain roughly the same. This is expected because the attractive competing features of and in the different regions are located in or close to the region itself and further away from regions located further away.

v. the price in any given region is coupled stronger with the price in regions with larger populations i.e. $\gamma_{j,i}$, is larger the larger population region i has. This is under assumption that the Euclidean distance be-

tween the compared regions are roughly the same.

vi. the base price constants α_i remain positive as the prices are positive and the errors are reduced by fitting the graph closer to its average value.

vii. the inflation constants β_i remain positive and are larger for regions with overall higher prices as prices over time increase exponentially. In other words, $\beta_i > 0$ and $\vec{\beta}$ correlates positively with the average prices in each region $\vec{x}$.

It is important to observe the behaviour of these features during the supervised learning process of the GD-method. Thus, the parameter values is monitored and if needed limited carefully according to the expected general features. Focus is put on starting values that follows the features (i), (ii) and (vi), which are the most distinct features to align the model to. The behaviour and evolution of the model with the gradient descent iterations and different starting values is monitored.

Results and discussion

Equation (1.1) is solved using the different methods and obtained parameters are analyzed, discussed and presented method by method.

3.1 Vector Autoregressive method

3.1.1 Parameters α, β and γ

The price coupling matrix γ, extracted with data from the estimation period, is shown in image format in figure 3 together with the distributions of all price coupling parameters. The location of $\gamma_{i,j}$ in the image is on row i and column j.

As we can see from the image and diagram of figure 3, there are no observable diagonal element-patterns. This is not what is expected for a good model of this system according to (ii) in section 2.7. Table 3.1 shows the average value, average absolute value, minimum and maximum value of the price coupling parameters.

On average, diagonal γ are slightly higher than the rest, which is agreeing with the weakly expected feature (i) in section 2.7. There are some tendencies towards higher

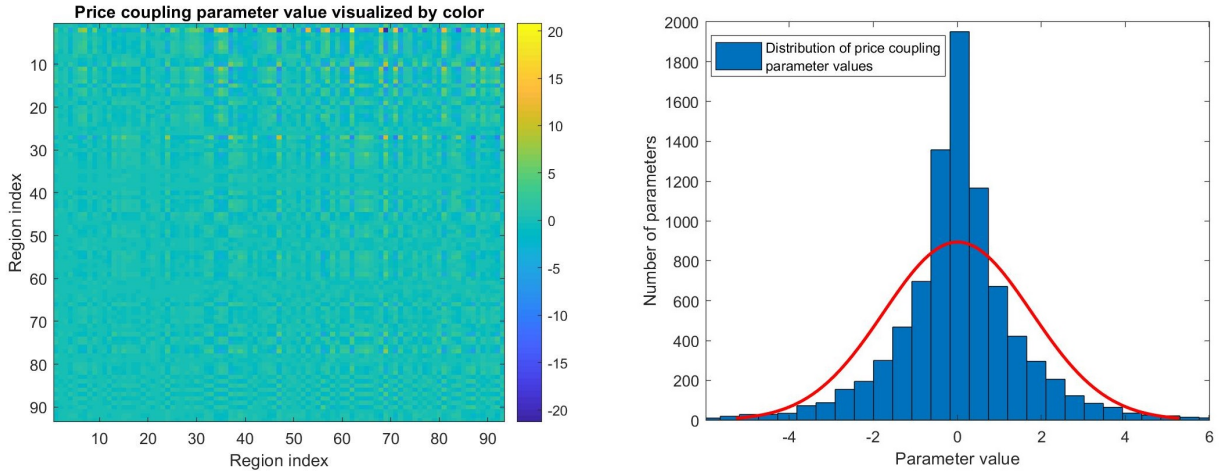

Figure 3: Left: The price coupling matrix γ in image presentation. The location of $\gamma_{i,j}$ in the image is on row i and column j. Horizontal and vertical patterns are visible, the diagonal elements does not constitute any visible pattern. Right: The distribution of the price coupling matrix parameters with a Gaussian distribution. The width if the bars are 0.46 and the standard deviation 2.82.

	$<\gamma>$	$<\lvert\gamma\rvert>$	$\sigma(\gamma)$	Min	Max
All γ	0.0108	1.10	1.76	-21.3	20.8
Diagonal γ	0.0136	0.8858	1.21	-2.80	3.29
Non-diagonal γ	0.0107	1.10	1.76	-21.3	20.8

Table 3.1: The average, average absolute, standard deviation, minimum and maximum values of the elements of the price coupling matrix γ, of its diagonal elements and of its non-diagonal elements.

diagonal parameter values. This tendency is too weak to draw any conclusions, but might indicate a very weak tendency towards expected feature (ii).

All region's centroid point is saved from Google Earth and a distance matrix containing all distances between each pair of region's centroids is created using the Haversine formula[16]. The correlation between the coupling parameters and the distances is found to be -0.0084, meaning a very weak negative correlation almost at the uncorrelated level, thus not significantly satisfying expected feature (iv) in section 2.7. The left scatter plot of Figure 4 shows the price coupling parameters as a function of the distance between each parameter's two corresponding regions. It is clear that the random behaviour is not what is expected.

For every region j in the coupling matrix, the correlation between $\gamma_{j,i}$ and the

population of region i is calculated. The average correlation is -0.0213. This does not agree with (vi) in section 2.7 which expects a positive correlation and this is another indication on that this model does not describe the system correctly. An example of the price coupling as a function of the array $\gamma_{1,i}$ is shown in the right scatterplot in figure 4. Note the different vertical scales between the diagrams, which are specified in order to display all the data points in the diagrams.

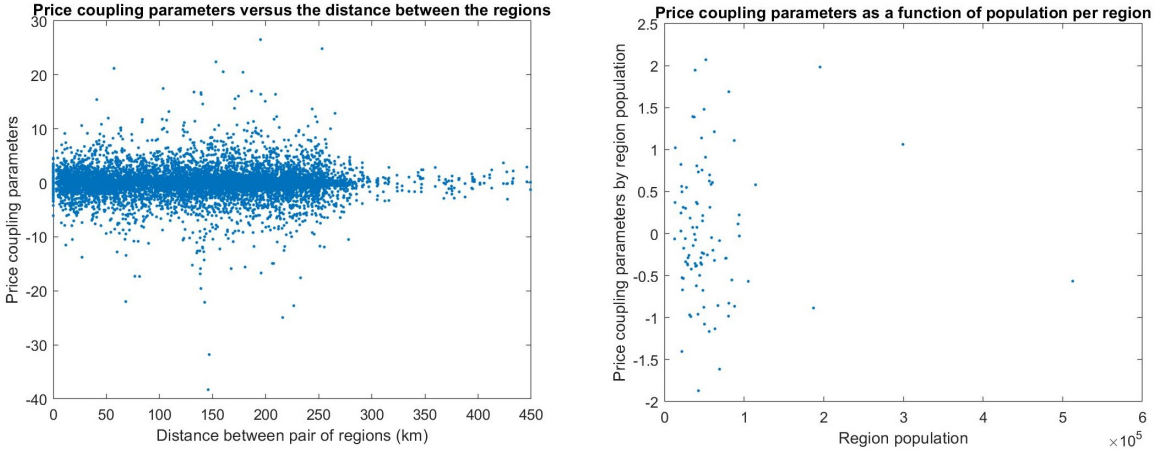

Figure 4: Left: Price coupling parameters as a function of the distance between the regions of the coupling parameter. A negative visible correlation is expected from a good model from (iv) in section 2.7. Right: Price coupling parameter as a function of region population. Both diagrams shows disagreement with expectations (iv) respectively (v). Note the different vertical scales.

The average value, average absolute value, standard deviation, minimum and maximum values of the elements of the α and β parameters are shown in table 3.2.

	$<\alpha,\beta>$	$<\lvert\alpha,\beta\rvert>$	$\sigma(\alpha,\beta)$	Min	Max
α	-5.25·10^{-5}	0.0084	0.0112	-0.0298	0.0474
β	-0.0344	0.896	1.20	-5.28	3.23

Table 3.2: The average, average absolute, standard deviation, minimum and maximum values of the elements of α and β.

The α and β parameters have a negative average value, which is contrary to (vi) respectively (vii) in section 2.7. The house prices are all over 2 thousands of Danish

kroner and α parameters are expected to be in this magnitude. The β parameters can be compared to the average increase in price for all regions and times in the estimation period which is close to 88 dkk/quarter. Some things are clearly wrong of this model representing the system physically and economically.

3.1.2 Combined parameter discussion

All parameters of the VAR method do in general not follow the expected features of a good model for the system. They seem to not pick up the behaviour of the system. The γ parameters had only very weak tendencies towards the expectations and will therefore most probably not be a good predictor for out-of-sample predictions.

Finally, we note that the coupling matrix in figure 3 has vague block-structures of several coupling parameters with the same value/color, meaning that it is common that regions that are close in their index space have similar coupling parameters. It is then interested to investigate the properties of the order of the regions. One way is to investigate the Euclidean distance between each pair of regions that represent the coupling parameters. It is defined as the distance between the centroids of the municipalities. Figure 6 shows this in a diagram with colour representation. The diagram is symmetric along the diagonal in the region index plane as there is only one distance for every two coupling parameters $\gamma_{i,j}$ and $\gamma_{j,i}$. Close to the diagonal, the distances are small, meaning that regions with similar indices are located closely. This might explain the block-structure of the coupling matrix. Then, this also means that the solution of the VAR method still picks up some physical behaviour of the system. The expected features in section 2.7. are partly based on arguments on supply and demand and the temporal description of the equation might not be the most accurate. It is most probably better to use an exponential time dependence as the price inflation is exponential. That has not been investigated in this thesis.

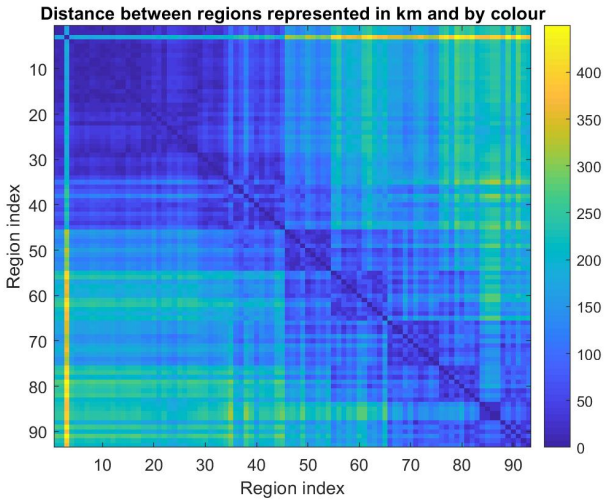

Figure 6: Euclidean distance between regions as a function of the region indices. It is visible that the indices are ordered such that close regions in index space is also close in Euclidean space.

3.1.3 Predictions

Table 3.3 presents statistics of the residuals of the fitted values in the estimation period together with statistics of the prediction errors.

From table 3.3 we see that the errors in the estimation period are very small, which means that the model has fitted the parameters very well to the data, so well that even fitted predictions are accurate after 4 years. But, these are not real predictions because the parameters are themselves extracted with data from these periods. The errors in the Validation period are extremely large, which is not surprising due to the discussion of the parameters and the conclusion that the model will not give accurate predictions.

What we further notice is that the bootstrapping method gives much larger errors than without bootstrapping. In figure 7 the prediction errors are shows both with and without bootstrapping.

	EP 1 step	EP	VP 1 step	VP	VP BS
MAE (dkk/m^2)	$2.81{\cdot}10^{-5}$	0.0199	$3.70{\cdot}10^{3}$	$4.50{\cdot}10^{5}$	$4.75{\cdot}10^{11}$
MAPE $(\%)$	$2.93{\cdot}10^{-7}$	$1.59{\cdot}10^{-4}$	31.1	$2.95{\cdot}10^{3}$	$3.13{\cdot}10^{9}$
MPE $(\%)$	$2.30{\cdot}10^{-7}$	$9.49{\cdot}10^{-5}$	-7.28	$-2.01{\cdot}10^{3}$	$-3.11{\cdot}10^{9}$
Profit 1 pre/real $(\%)$	5.41/5.41	5.61/5.61	35.4/3.84	$1.38{\cdot}10^{4}/22.2$	$5.83{\cdot}10^{10}/20.8$
Profit 2 pre/real $(\%)$	11.6/11.6	12.0/12.0	89.3/6.08	$3.05{\cdot}10^{4}/24.3$	$1.03{\cdot}10^{9}/21.7$
Av/Cop profit $(\%)$	-1.21/-4.74	-7.33/14.76	4.39/1.67	20.8/37.2	20.8/37.2

Table 3.3: Statistics of fitted predictions in the Estimation periods and real predictions in the validation periods. Columns from left: In-data one-time prediction of first quarter in the estimation period (EP 1 step), of all 16 quarters of the estimation period (EP), first quarter of validation period (VP 1 step), all quarters in the validation period (VP) and bootstrapped prediction statistics of all quarters in the validation period (VP BS). Rows from above: The mean absolute error, mean absolute percentage error, mean percentage error, predicted profit and real profit for investment strategy 1 and 2 as defined in section 2.6, lastly the average profit investing in all regions and in Copenhagen, as also defined in section 2.6. Notice that the first two columns correspond to in-data-predictions and are extracted from the same data so they do not represent any real prediction.

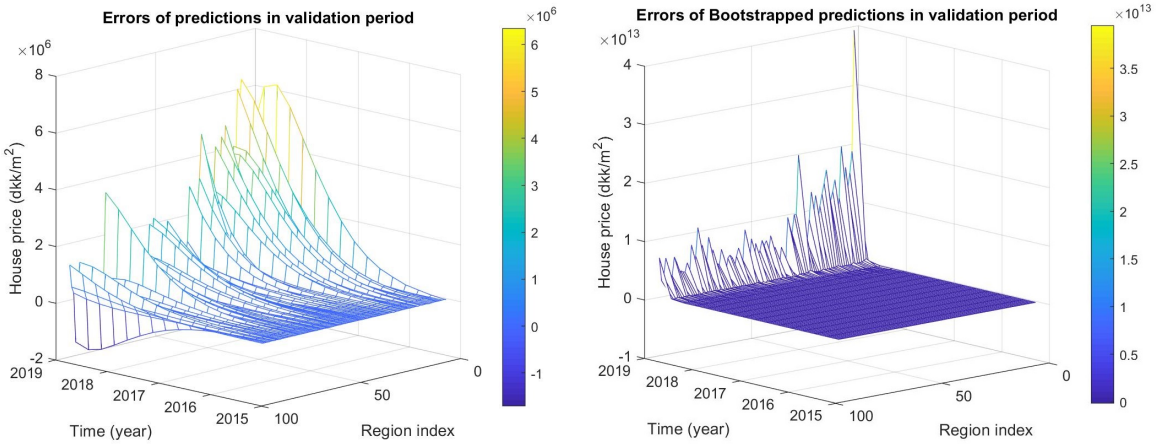

Figure 7: Errors of predictions in the validation period. Left: without bootstrapping. Right: with bootstrapping. Notice the scales, in million Danish kroner to the left and hundreds of billion Danish kroner to the right. The errors are very large in both cases compared to the prices, which shows that this model does not give any reliable predictions.

3.2 Gradient Descent method: Parameter starting values 0

The VAR method gives a model with global minima for the cost function which did not behave well. In this section, results from the GD method are presented. The starting parameters for the learning process for the GD method are very decisive for what local minima is reached, thus different starting parameters is used. First, the obtained results using modelling with using starting parameters values of 0 is presented and discussed in section 3.2. The corresponding procedure is done with starting parameter values according to the expected general features (in section 2.7) is presented and discussed in section 3.3.

3.2.1 Parameters α, β and γ, their evolution and learning rates

The size of the cost function for the estimation period as a function of number of GD-iterations is investigated. First, starting parameter values are set to 0 for all parameters. Different learning rates is investigated. Too high learning rates results in that the cost function instead increases and reaches too high values very fast. For example, it starts with values of the order 11 (dkk^2/m^4), and already after 13 iterations it reaches values of the order 183. Lowering the learning rate gives the desired results of a cost function that decreases with number of iterations.

It is found that grouping the learning rates into three different learning rates $L_\alpha, L_\beta, L_\gamma$ for the different parameters is better than using a single learning rate L. The explanation of this is that the different learning rates have different units (see them in next paragraph), and thus depend on the units of price, length and time. L_γ is the sensitive part making the cost function behave in the undesired above described manner and had to be kept much lower than the other.

There are also limitations of possible values of L_α and L_β and these values are 7 respectively 6 magnitudes higher. Keeping all 3 learning rates the same kept the α and β parameter very small, in the orders of -5 and -4, while the γ constants had developed well more. The learning rate is desired to be quite high to not miss out on the optimization of the cost function due to limitations of computer power, but not too high to miss out on reaching the best local minima. The three learning rates are thus chosen to around 40% of their maximum possible values; $L_\alpha = 1.7724 \cdot 10^{-5}$, $L_\beta = 8.2713 \cdot 10^{-7}$ (1/quarter^2) and $L_\gamma = 5.9081 \cdot 10^{-13}$ (m^4/dkr^2), which also gives as expected better results. These learning rates keeps the algorithm stable using a quite broad range of starting values in the sense of minimizing the cost function. In the diagram below is shown several different results regarding the parameter's expected general features in section 2.7: the fraction of negative parameters among its sort as

well as the root-mean-square-mean-error as a fraction of the average price of all times and all regions. The fraction of negative α and β parameters are expected to be close to 0 and the fraction of negative γ parameters are expected to be less than 0.5 for a well describing model. The root-mean-square-error is supposed to always decrease during the learning process, which it does.

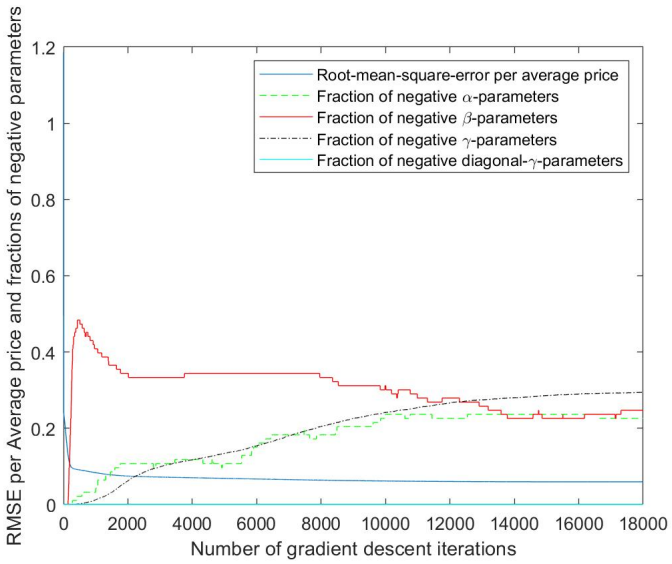

Figure 8: As a function of number of iterations; the root-mean-square-error as a fraction of the average price of all regions and times 9382 dkk/m^2 (solid blue line that approaches a value under 0.1), the fraction of negative α-parameters among the α-parameters (dashed in green), the fraction of negative β-parameters among the β-parameters (red line), the fraction of negative γ-parameters among the γ-parameters (dashdot in black), the fraction of negative diagonal-γ-parameters among the γ-parameters (cyan line at 0). Notice that RMSE per average price is proportional to the squared cost function. It reaches a minimum of 5.83 %.

It is noticed that the cost function is decreasing with the number of iterations. The parameter results are in general tending towards the expected general features with more positive than negative parameters especially for the α-parameters. The fraction of negative coupling parameters is around 30%, which is satisfying the non-strict expected feature (i). Trying different learning rates gives similar results and not anything significantly closer to the expected general features. We see that all diagonal γ-parameters are positive, which is a good sign regarding the discussion in section 2.7.

The average parameter values as a function of number of iteration is shown in the figure below.

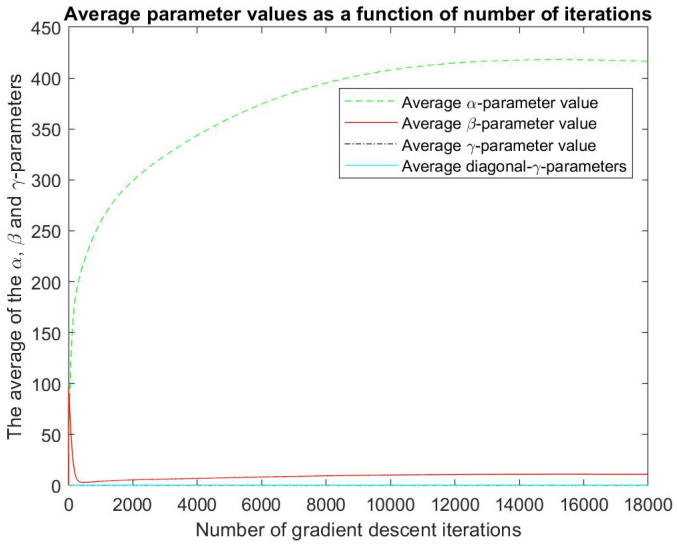

Figure 9: The average values of the α-parameters (dashed in green), β-parameters (solid line in red), γ-parameters (not visible because it is close to 0) and diagonal-γ-parameters (cyan-line close to 0). The final values are 416.4 dkk/m^2, 10.93 dkk/m^2/quarter, 0.0090 and 0.2348, respectively. A zoomed in verification shows that the two curves close to 0 stays almost flat except in the beginning of the learning process.

Figure 9 gives shows strong tendencies towards expected features; the average of base price constants α are in hundreds of danish kroner, which is relatively low compared to the average price of all regions and times; 9382 dkk/m^2, but it agrees much stronger with the expectations than the corresponding results for the VAR method does. The average β remains positive. It is also low compared to the average price increase of all times, 10.93 dkk/m^2/quarter compared to 87.78 dkk/m^2/quarter, but it is also in this case much closer towards expected general features than for the VAR method. Figure 8 & 9 shows thus much stronger tendencies towards expected general features (i), (ii), (vi) and (vii).

Figure 10 below shows the price coupling matrix γ in image format together with the distributions of all price coupling parameters, it is comparable to figure 4 of the VAR model.

We notice from figure 10 that expected general feature (ii) is well satisfied, and that the majority of the parameters are positive. The gathering of the diagonal elements in the distribution is visibly far to the right of the non-diagonal elements.

Table 3.4 shows the average value, average absolute value and minimum and maximum value of the price coupling parameters, and may be compared to table 3.1 of the

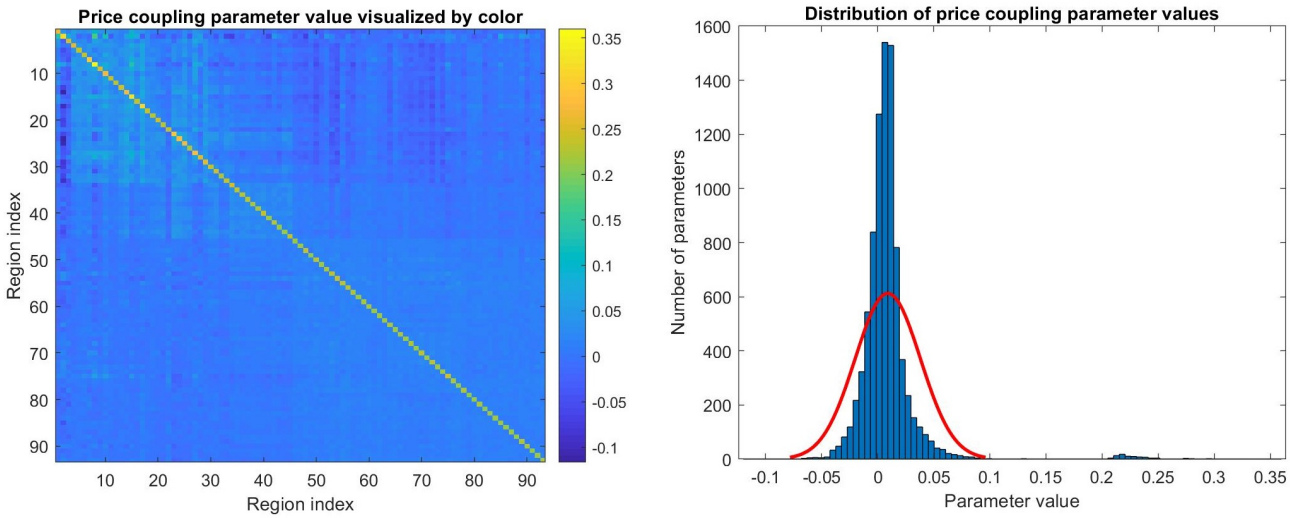

Figure 10: Left: The price coupling matrix γ in image representation. Right: The corresponding histogram distribution with a Gaussian distribution. The location of $\gamma_{i,j}$ in the image is on row i and column j. Notice the clear diagonal pattern and the parameters inclination to be positive.

VAR method. It is clear that there is a distinction between diagonal and non-diagonal coupling parameters, as is expected from general feature (ii). The average diagonal element is here around 18 times higher than the non-diagonal element, compared to in the VAR method the corresponding value is around 1.3.

	$< \gamma >$	$< \lvert\gamma\rvert >$	$\sigma(\gamma)$	Min	Max
All γ	0.0090	0.0153	0.0291	-0.117	0.360
Diagonal γ	0.235	0.235	0.0279	0.208	0.360
Non-diagonal γ	0.0065	0.0129	$2.86{\cdot}10^{-4}$	-0.117	0.168

Table 3.4: The average, average absolute, standard deviation, minimum and maximum values of the elements of the price coupling matrix γ, of its diagonal elements and of its non-diagonal elements.

The correlation between the coupling parameters and the distances is found to be -0.0897, which is more than ten times stronger negative correlation than with the VAR method and thus better satisfying expected feature (iv), but it is still not strong. The left diagram of Figure 11 shows the price coupling parameters as a function of

the distance between each parameter's two regions. Mainly at very short distances is the price coupling visibly higher. Since Denmark consist of islands, see figure 3, the distance is not well representation the human connections with many of the other regions, as one has to drive over certain bridges to come over to the other islands. This might partly explain the weak correlation.

The average correlation of the population and coupling parameter is -0.0497, opposing expected feature (v) in section 2.7 which expects a positive correlation and it is actually less than for the VAR method. An example of the price coupling as a function of the array $\gamma_{1,i}$ is shown in the right diagram in figure 11. Note the different vertical scales.

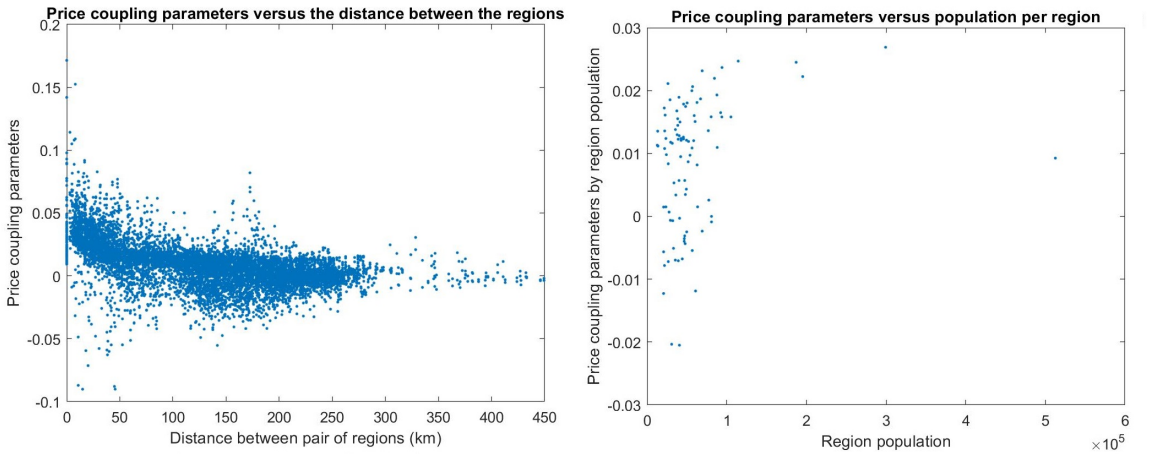

Figure 11: Left: Price coupling parameters as a function of the distance between the regions of the coupling parameter. A negative visible correlation is expected from a good model from (iv) in section 2.7. Right: Price coupling parameter as a function of region population. The left diagrams shows some visible agreement with expectation (iv) and the right does neither show agreement or not with expectation (v). Note the different vertical scales.

3.2.2 Combined parameter discussion

We notice that the coupling matrix has the expected diagonal feature. The fraction of negative γ parameters is around 30% and their average is positive, satisfying the non-strict expectation (i) in section 2.7. The base price parameters α has a large fraction positive values and an average of slightly over 400 dkk/m^2, which is as mentioned comparably low to the prices, but still well over 0. For the inflation parameters β, they tend to be more positive than negative, but still a quite large fraction is negative, which

is not well satisfying the expectations of a good model of this system, but at least they tend towards the right direction. Still in the coupling matrix in figure 10, we notice some block structures, which is most probably explained by the closeness of regions in index-space. This suggests that some closely located regions has a stronger positive price coupling to their neighbour than to region further away, which is a good sign with regards to (iv) in section 2.7.

3.2.3 Predictions

Table 3.5 presents statistics of the residuals of the fitted values in the estimation period together with statistics of the prediction errors.

	EP 1 step	EP	VP 1 step	VP	VP BS
MSE $(dkk/m^2)^2$	5.59·10^5	7.46e·10^5	5.14·10^5	1.31·10^6	1.35·10^6
MAE (dkk/m^2)	536	604	556	762	773
MAPE $(\%)$	4.21	5.23	4.93	5.58	5.64
ME (dkk/m^2)	2.59	-392	-353	186	202
MPE $(\%)$	-0.569	-3.12	-3.02	0.806	0.866
Profit 1 pre/real $(\%)$	2.71/3.28	6.31/1.10	3.10/6.42	28.4/21.1	28.6/21.1
Profit 2 pre/real $(\%)$	6.87/7.33	13.6/10.34	9.71/13.17	50.1/34.3	50.6/34.3
Av/Cop profit $(\%)$	-1.21/-4.74	-7.33/14.8	4.39/1.67	20.8/37.2	20.8/37.2

Table 3.5: Statistics of fitted predictions in the estimation periods and real predictions in the validation periods. Columns from left: In-data one-time prediction of first quarter in the estimation period (EP 1 step), of all 16 quarters of the estimation period (EP), first quarter of validation period (VP 1 step), all quarters in the validation period (VP) and bootstrapped prediction statistics of all quarters in the validation period (VP BS). Rows from above: The mean square error, mean absolute error, mean absolute percentage error, mean error, mean percentage error, predicted profit and real profit for investment strategy 1 and 2 as defined in section 2.6, lastly the average profit investing in all regions and in Copenhagen, as also defined in section 2.6. Notice that the first two columns correspond to in-data-predictions and are extracted from the same data so they do not represent any real prediction.

Comparing table 3.5 to the corresponding table of the VAR method, table 3.3 we notice that the predictions in the VP are at least 2 orders of magnitude smaller. The errors in the estimation period are almost the same as in the validation period, which

is a good sign of the model. It is thus not too overfitted. The bootstrapped predictions have larger errors, which is in general not what to expect from a good model and might be explained by that the MPE is quite large, close to one % which indicates that the model in general generates too high predictions instead of centered prediction errors. Thus, the prediction errors accumulate with every bootstrapped prediction.

The errors of the predictions in the validation period are shown in figure 12. They looked the same as with bootstrapping, thus only one of them is shown with the corresponding histogram.

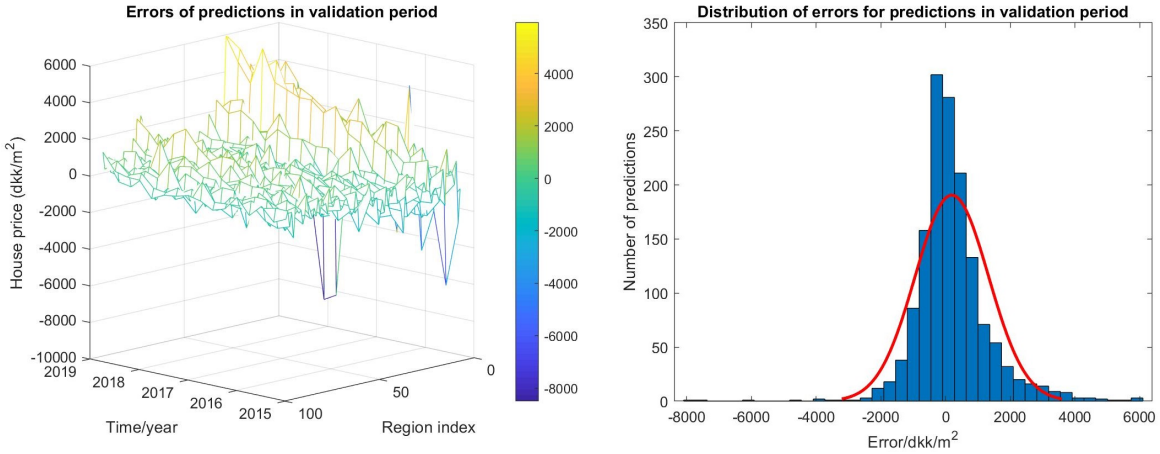

Figure 12: Left: Errors of predictions in the validation period. Right: Distribution of the errors with a fitted Gaussian distribution.

3.3 Gradient Descent method: Parameter starting values according to expected features

3.3.1 Parameters α, β and γ

We note that the development of the parameters with number of iterations, such as in figure 8, in some sense show the competition of the parameters staying positive which may represent the relative importance of inflation and price influence between regions. The starting values of the base price parameters α are set to half the initial price, the inflation parameters β to half the average inflation (see below) and the price coupling parameters such that their sums for any region adds up to 0.5. The motivation for this is normalization between the parameters. The learning rates are the same as in subsection 3.2.1.

1. Starting values of all base-prices parameters α_i are set to the first price in every region i.

2. Starting values of all inflation parameters β_i are set to $d(i)^{1/91}$ where $d(i)$ is the price difference between region i's first and last quarters and 91 is the number of quarters until the last quarter in the estimation period.

3. Starting values of non-diagonal γ are set to $0.5 \cdot 0.5/92 \approx 0.027$ and diagonal γ to $0.5 \cdot 0.5 = 0.25$. This is motivated by 1) letting the α and β parameters be half half of the starting value of what they would be assumed to be in a model without γ and 2) letting the γ parameters that represent $x_{j,t}$ in equation (1.1) be half of the starting value of a normalized (sum adds up to 1) value where each diagonal γ in equation (1.1) is as big as the non-diagonal γ.

Figure 13 shows the counterpart figure 8 and 9 i.e. the root-mean-square of the fitted errors, fractions of positive parameters, as well as the average parameter values, as a function of number of gradient descent-iterations.

We note from figure 13 that the fraction of negative α stays at 0, when for the case with starting values 0 it ended with over 20%. For β we end with around 7% negative parameters, while for the other case we had around 25%. Thus, (vi) and first part of (vii) is much better satisfied in this case than for starting values close to 0. Just as in the case with starting values at 0, all diagonal-γ is positive. Among the non-diagonal-γ, the fraction negative ones, 47%, is less than the fraction of positive ones, satisfying (i) in section 2.7. We also note that the α and β-parameters very fast tend positively compared to γ, which in some sense indicates that inflation is a stronger force than the price coupling between regions. As section 2.2 shows, the price inflation in the Danish house market is very large. Would the prices have been inflation adjusted, the system would have been much easier described by price coupling.

The coupling matrix in image format compared to the one in figure 10, looks very similar and is thus not shown. It has slightly less box structures. From table 3.6, we see that the average diagonal- and non-diagonal-γ have increased somewhat from the start of the learning process. The average diagonal parameter is 88 times higher than the average non-diagonal parameter, which is well satisfying (ii) in section 2.7, better than starting values of 0.

Overall, this set of parameters seem to have much more physical interpretations than the other sets, meaning that the predictions should also be better.

The correlation between the coupling parameters and the distances between the corresponding regions is found to be -0.405, meaning a quite strong negative correlation. It is well satisfying expected feature (iv) in section 2.7 and can be compared to

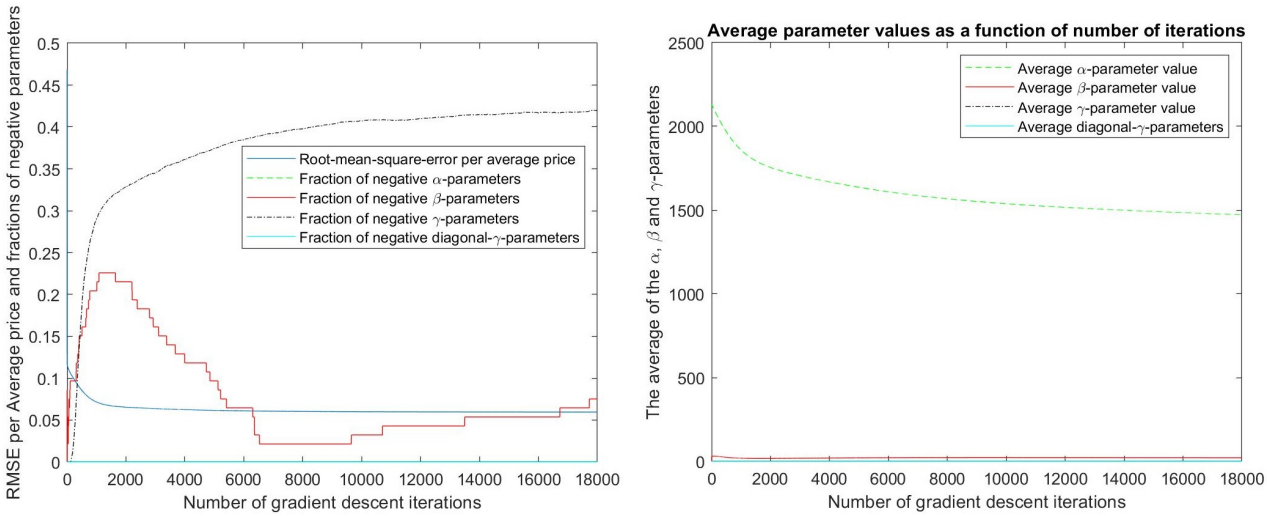

Figure 13: Left: As a function of number of iterations; the root-mean-square-error as a fraction of the average price of all regions and times 9382 dkk/m^2 (solid blue line that approaches a value under 0.1), the fraction of negative α-parameters among the α-parameters (stays at 0), the fraction of negative β-parameters among the β-parameters (red line), the fraction of negative γ-parameters among the γ-parameters (dashdot in black), the fraction of negative diagonal-γ-parameters among the γ-parameters (cyan line at 0). RMSE per average price reaches a minimum of 5.96 %. Right: The average values of the α-parameters (dashed in green), β-parameters (solid line in red), γ-parameters (dashdot in black close to 0) and diagonal-γ-parameters (cyan-line close to 0). The final values are 2097 dkk/m^2, 31.15 dkk/m^2/quarter, 0.0065 and 0.2551, respectively. A zoomed in verification shows that the two curves close to 0 stays almost flat except in the beginning of the learning process.

the same correlation for starting values at 0; 0.0897 (or the VAR method with -0.0084), which both validates this method and shows that the starting values are important. Moreover, as mentioned earlier, the island-structure of Denmark with a few bridges between them is moreover reducing the correlation between distances, and a better measure would have been an average time to travel between the bridges, preferably taking into account the travel cost as well. Sweden would be a good case to study because there are not too many relatively large populated islands.

The average correlation between the population arrays and the coupling parameters for each region is -0.0059 opposing expected feature (iii) in section 2.7 which expects a positive correlation. It is peculiar. Bigger regions should affect other regions price more than smaller regions. The same correlation for starting values at 0 had somewhat stronger negative value; -0.0497. This indicates that the population of the

	$<\gamma>$	$<\lvert\gamma\rvert>$	$\sigma(\gamma)$	Min	Max
All γ	0.0062	0.0166	0.0343	-0.102	0.422
Diagonal γ	0.282	0.235	0.0305	0.247	0.422
Non-diagonal γ	0.0032	0.0129	$3.41{\cdot}10^{-4}$	-0.102	0.179

Table 3.6: The average, average absolute, standard deviation, minimum and maximum values of the elements of the price coupling matrix γ, of its diagonal elements and of its non-diagonal elements.

regions seems not to have importance of the price evolution in different regions. This is counter-intuitive. As mentioned before, since the inflation is so large, it would have been very interesting to investigate the system with inflation-adjusted prices. Then more physical pattern would have been easier to detect. Unfortunately, the author did not have time to do this.

3.3.2 Predictions

Table 3.7 presents statistics of the residuals of the fitted values in the estimation period together with statistics of the prediction errors. By comparing with the corresponding table for parameter starting values 0, table 3.5, it is visible that the general trend is that the fitting errors are slightly higher, but the prediction errors are smaller. Choosing parameter starting values according to expected general features directs the learning process to find solutions that better describes the system. Moreover, the bootstrapping errors are consistently smaller than without bootstrapping, which is expected from a good model, as bootstrapping uses more available data for prediction. This is also s good sign.

	EP 1 step	EP	VP 1 step	VP	VP BS
MSE $(dkk/m^2)^2$	5.75·10^5	8.19·10^5	4.07·10^5	1.12·10^6	9.70·10^5
MAE (dkk/m^2)	543	641	495	720	670
MAPE $(\%)$	4.24	5.50	4.44	5.55	5.18
ME (dkk/m^2)	-55.7	428	-164	306	21.8
MPE $(\%)$	0.127	3.37	-1.32	2.55	0.458
Profit 1 pre/real $(\%)$	2.46/3.08	6.86/0.60	3.83/5.80	27.5/21.1	25.8/21.1
Profit 2 pre/real $(\%)$	5.95/7.42	14.6/10.3	10.8/11.0	45.1/35.6	42.2/35.6
Av/Cop profit $(\%)$	-1.21/-4.74	-7.33/14.76	4.39/1.67	20.8/37.2	20.8/37.2

Table 3.7: Statistics of fitted predictions in the Estimation periods (EP 1 step and EP) and real predictions in the validation periods (VP 1 step and VP). Rows from above: The mean square error, mean absolute error, mean absolute percentage error, mean error, mean percentage error, predicted profit and real profit for investment strategy 1 and 2 defined in section 2.6, lastly the average profit investing equally much in every region and the profit investing only in Copenhagen region. Columns from left: In-data one-time prediction of first quarter in the estimation period (EP 1 step), of all 16 quarters of the estimation period (EP), first quarter of validation period (VP 1 step), all quarters in the validation period (VP) and bootstrapped prediction statistics of all quarters in the validaton period (VP BS).

Conclusions and outlook

The author concludes that the parameters given from the VAR method do not follow the expected general features or correlation with distance and population. The prediction errors in the estimation period are extremely small and very large in the validation period, meaning that the model is very much overfitted. Thus, the model does not pick up on general physical behaviour of the system and is subjected to overfitting.

For the GD method with parameter starting values 0 the parameters overall picks up quite some features from the system. This is also verified by the fact that the errors in the estimation are similar to the errors in the validation period. The prediction errors are quite large but confined. The profits are not bad. However, the errors with the bootstrapped predictions are slightly larger, which it preferably should not be for a good model of the system. The conclusions so far is that the VAR model is highly overfitted and that the GD-model pick up on physical behaviour of the system. Since the correlation between price coupling parameters with population and distance is so weak, the features the model picks up is not strong enough to confidently draw conclusions. This is also verified by that the bootstrapped prediction errors are larger than without bootstrapping.

The GD method with parameter starting values according to expected general features in section 2.7 shows better agreement with expected general features than both other method. It better predicts future prices, and have much stronger correlation of its price coupling parameters with distance. Moreover, it also has better results with bootstrapping than without, validating the predictions and the model. We thus have confirmed that this method works in describing the system of house prices in Denmark and could work in other countries as well.

One important conclusion is that regions located close to each other affect each other's price positively and much more than regions located far away. The number of persons living in the region does however not seem to have an importance on the influence on other regions price levels.

We note (From figure 13) that the fraction of negative price coupling parameters is much larger than among other parameters, indicating that inflation is a strong force compared to the price influence between regions. We have already seen (from figure 1) that the correlation between the money supply and the average price (see figure 1) is noticeably very high and it is calculated to 0.90.

One of the most important changes for future work is to work with adjusted prices for inflation and/or money supply, or with the logarithm of the prices where the fast price increase gets reduced.

In this thesis, the impacts of exogenous shocks are not modelled. Equation (1.1) is limited to municipalities in Denmark and their interplay with time-independent parameters. It is understood that events like 9/11, the construction of the Øresund Bridge and BREXIT has a certain impact. We know that there exist endogenous equlibrium models that may explain unstable behaviour with aggregated fluctuations [5]. It has been suggested, that empirical validity of the exogenous and endogenous cycle hypotheses may be done by comparing predictions by the corresponding different models [5]. The time-dependence of parameters is thus very relevant to investigate in future work to reach a better understanding of the system.

It would be interesting to work with a price evolution equation that involves earlier prices than just one time step, for example such that $x_{t+3} = f(x_{t+2}, x_{t+1}, x_t)$.

It would also be interesting to work with a country without islands such as Germany. In that case, the distance between the regions may much better represent the travel time and thus the connections and price couplings between the regions. The distance correlations with the γ 's would probably be stronger, and it would be interesting to investigate the population correlations with the γ's.

Other enhancements of the model would be to include how long time the houses has been on the market before it gets sold or economical data and predictions such as GDP growth, interest rates, bond yield curve and average private debt. Another example is the property stock price index, consisting of shares of companies investing in properties and managing a portfolio of real estates. It has been found that there exist a long-term link between property stock price index, treasury bond interest and real estate price index [17]. It would also be interesting to work with fewer and larger price regions of the country while taking into account more factors.

Something else to study is possible chaotic behaviour for the predictions using especially the VAR model or from solution of the GD models, where some local minima solutions might have chaotic behavior and others might not.

The quality of the models also depends on how stable the real estate market is regarding the structural relationships, as it develops with time, as related studies has stated [18]. This is especially because the parameters are all time-independent, but in reality the houses age, gets modified, and new houses are built.

This kind of research will likely mainly be supported by either entities such as asset management firms and banks or by governments that want to protect the country against housing crashes or housing bubbles.

Bibliography

[1] Steven H. Strogatz, "NONLINEAR DYNAMICS AND CHAOS", *ADDISON WESLEY*, Seventh printing 1994,p 3,348. ISBN: 0-201-54344-3.

[2] Boligstatistik, http://rkr.statbank.dk/statbank5a/default.asp?w=1920. Accessed on 2-2-2019.

[3] Gunnar Ohlén, Sven Åberg, Per Östborn, "CHAOS", Division of Mathematical Physics LTH, Lund, 2007, Lund University, p 15-16, 37, 84.

[4] Bifurcation, https://en.wikipedia.org/wiki/Bifurcation_diagram#/media/File:Logistic_Map_Bifurcation_Diagram,_Matplotlib.svg. Accessed on 14-05-2019.

[5] M. Boldrin; M. Woodford "EQUILIBRIUM MODELS DISPLAYING ENDOGENOUS FLUCTUATIONS AND CHAOS", *Journal of Monetary Economics*, 25, 1990, 189-191, DOI: 10.1016/0304-3932(90)90013-T

[6] Boligstatistik, https://www.bankofengland.co.uk/knowledgebank/what-is-money. Accessed on 6-2-2019.

[7] BANK OG REALKREDIT. Choose "Bank og Realkredit" Then choose DNMNOGL: Nøgletal for MFI-sektoren efter nøgletal og sektor (2003M01-2019M02), http://nationalbanken.statistikbank.dk/statbank5a/default.asp?w=1600

[8] Vad är pengar?, Sveriges Riksbank https://www.riksbank.se/sv/betalningar--kontanter/riksbankens-uppdrag-inom-betalningar/vad-ar-pengar/ accessed on 15-03-2019.

[9] Michael McLeay, Amar Radia and Ryland Thomas "Money creation in the modern economy", *Bank of England Quarterly Bulletin*, 2014 Q1, ISSN: 0005-5166, Accession Number: 95261806, 2014, pp1. 2014

[10] Michael Kumhof and Zoltán Jakab "The TRUTH about BANKS", *Finance Development* Vol. 53 Issue 1, International Monetary Fund, ISSN: 0015-1947, Accession Number: 113458699, 2016, pp50-53.

[11] Main Economic Indicators - complete database, Main Economic Indicators (database), http://dx.doi.org/10.1787/data-00052-en (Accessed on 06-05-2019); Organization for Economic Co-operation and Development, M1 for Denmark [MANMM101DKM189S], retrieved from FRED, Federal Reserve Bank of St. Louis; https://fred.stlouisfed.org/series/MANMM101DKM189S

[12] PROPERTY PRICES IN HOUSING MARKET, https://rkr.statbank.dk/statbank5a/Graphics/mapanalyser.asp?maintable=BM010&lang=1. Accessed on 06-04-2019. Modified by enlarging the color bar and making the numbers bigger.

[13] Helmut Lütkepohl, New Introduction to Multiple Time Series Analysis, Springer, ISBN 3540401725, 2005.

[14] Clark, Todd E.; McCracken, Michael W. "Tests Of Predictive Ability For Vector Autoregressions Used For Conditional Forecasting", *Journal of Applied Econometrics*, Vol. 32 Issue 3, Apr2017, pp1. DOI: 10.1002/jae.2529.

[15] Simple forecasting models, https://people.duke.edu/~rnau/three.htm. Accessed on 04-04-2019.

[16] R.W. Sinnott "Virtues of the Haversine", Sky and Telescope, vol. 68, no. 2, 1984, p. 159.

[17] Seow Eng Ong. "Structural and Vector Autoregressive Approaches to Modelling Real Estate and Property Stock Prices in Singapore ", *Journal of Property Finance*, Vol. 5, Issue 4, 1994, pp17.

[18] Gloudemans, Robert J.; Miller, Dennis W. "MULTIPLE REGRESSION ANALYSIS APPLIED TO RESIDENTIAL PROPERTIES: A STUDY OF STRUCTURAL RELATIONSHIPS OVER TIME.", *Decision Sciences*, Vol. 7 Issue 2, p294-304. 1976, pp294. DOI: 10.1111/j.1540-5915.1976.tb00676.x.

Appendix

1. **House price data**

   The house prices were gained from http://rkr.statbank.dk/statbank5a/default.asp?w=1920 pressing on "BM010: Property prices in housing market by area, property category and prices of completed transactions". All municipalities were selected (meaning not selecting subregions and regions which contains groups of municipalities). The option "Detached/terraced house" and "Transaction price realised" were selected. All available quarters at the time was marked: from 1991Q1 to 2018Q4. It was found that a lot of data were missing for Bornholm, Læsø, Ærø, Fanø and Samsø and thus they were excluded. The house price in 2007Q4 for Frederiksberg was missing and we performed the arithmetic mean between it's adjacent two quarterly prices. It should also be noted that the house price data for 2018Q2-Q4 will be updated again, which was learned from contact with financedanmark.dk.

2. **Programming code**

   The programming code is divided into one part for the VAR method and another for the GD method and is available at https://github.com/swelov/MATLAB-code-for-thesis.git.

3. **Matrix definitions for the VAR method**

   The matrix definitions for the VAR matrix equation $Y = PZ + U$ in the introduction and in section 2.5 are as follows:

$$Y = \begin{bmatrix} y_{1,2} & y_{1,3} & \cdots & y_{1,t+1} \\ y_{2,2} & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ y_{93,2} & \cdots & \cdots & y_{93,t+1} \end{bmatrix} \tag{1.1}$$

$$P = \begin{bmatrix} \alpha_1 & \beta_1 & \gamma_{1,1} & \gamma_{1,2} & \cdots & \gamma_{1,93} \\ \alpha_2 & \beta_2 & \gamma_{2,1} & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ \alpha_{93} & \beta_{93} & \gamma_{93,1} & \cdots & \cdots & \gamma_{93,93} \end{bmatrix} \tag{1.2}$$

$$Z = \begin{bmatrix} 1 & 1 & ... & 1 \\ 1 & 2 & ... & t \\ y_{1,1} & y_{1,2} & \cdots & y_{1,t} \\ y_{2,1} & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \\ y_{93,1} & \cdots & \cdots & y_{93,t} \end{bmatrix} \tag{1.3}$$

Here $y_{i,t}$ is the price in region i at time t and α, β and γ are the parameters in (1.1). U is the matrix with the corresponding errors and 93 is the total number of regions.