Assessing housing market dynamics across a sample of European cities

A Random Forest Machine Learning Approach

Luca Begatti
108211

Corrado Mauceri
107318

A thesis presented for the degree of MSc. Advanced Economics and Finance

Thesis supervisor:
Chandler Lutz, Associate Professor of Economics, CBS

June 27, 2018

Copenhagen Business School

Contents

List of Figures

List of Tables

Assessing housing market dynamics across a sample of European cities

Luca Begatti, Corrado Mauceri

June 27, 2018

Abstract

This paper studies the housing market across European cities by combining regression analysis with a machine learning process. We identify that despite the limited data availability, there are time, national and local effects associated with real housing returns. According to our choice of explanatory variables, we document that mortgage rate, GDP per capita and unemployment rate are important determinants of housing returns. Generally speaking, we can infer that the European housing market does not show evidence of bubbles but there are some markets which deserve particular attention. We witness the presence of Shiller's Irrational Exuberance only with respect to the city of Nurnberg, Germany. We further witness in some cities a substantial adjustment rate suggesting market efficiency as well as strong mean reversion effects. Collectively, these results support the view that most of the European cities are not currently experiencing a bubble but in some instances the foundation for such inexplicable behaviour have been building up following the recent financial downturn.

[0]We want to thank the Copenhagen Business School for the opportunity to develop and improve our economic and financial skills in a challenging environment. We express our deep gratitude to the thesis supervisor, Chandler Lutz, for his feedback and support of the project. We also offer special thanks to all the National Statistical Institutes and private companies which provided their assistance with the collection of data: ECB, Eurostat, OECD, UK government National Statistical Office, Ministero de Fomiento, Die Regionaldatenbank Deutschland, Destatis.de, Statistics Denmark, Statistics Sweden, Statistics Finland, Statistics Norway, Statbel, Centraal Bureau voor de Statistiek, INSEE, Ministre de la Transition Ecologique et Solidaire, National Bank of Belgium, Statista, Bulwiengesa AG. Last but not the least, we wish to express our appreciation to our families and friends for their support and encouragement throughout our studies.

Introduction

What started as a random conversation on real estate prices in Copenhagen during a lecture break, turned into the spark for our thesis topic. The discussion started as an exchange of opinions on Copenhagen's house prices, rents and whether a house bubble explained such figures.

The question surfaced mainly because, as first time students in Denmark, we perceived everything, especially rents, as extremely high. Probably, this was a bias due to the fact that our home country economic situation was rather different than the Danish one. Although we were aware that house prices per se do not mean much, the pattern, still, was evident in our opinion, or at least worth investigating. The debate proceeded with the professor underlying the fact that actually only rents have been skyrocketing, meanwhile house prices have not been increasing that much. In other words, the price-to-rent ratio, a common measure used in the real estate literature, should be rather small in Copenhagen.
However, there is no point in denying the huge demand with respect to the supply of housing allowing landlords to set rents above what would be generally perceived as the fair price. Unfortunately, we never had the chance to properly continue the discussion but that short talk was enough for us to become curious about the topic and start thinking about it as a relevant Master Thesis topic.

After the lecture, we discussed what had been said and decided that keeping the discussion limited to Denmark would not be the best way to develop our thesis, for many reasons. Mainly the fact that the Danish population is condensed in a few cities, city level house prices statistical information are not always readily available and, ultimately, our goal remains to compare the housing market situation between and across European cities to determine where there is a surge in house prices which cannot be explained by taking into account fundamentals[1].

[1]Fundamentals are the explanatory variables intended as in Case and Shiller (2003), with the addition of two more variables

The idea of extending the discussion to the European level appeared an interesting shift from the usual focus on U.S. data. Indeed, despite the lack of research within this area, we truly believe that examining European house prices represents an interesting topic, especially when considering the current economic condition.
On one side, Nordic countries are experiencing a period of prosperity, whereas on the other side southern ones still have to fully recover from the 2008 downturn. In addition, the UK is facing the consequences of its post-Brexit decision.
Our quest is to grasp whether European cities are experiencing a housing bubble. In order to do so, our analysis will need to be carried through different complementary steps. The remainder of the paper is organized as follows. Section 1 presents the literature review. Section 2 contains a short introduction on house price methodologies. Data and all the variables of interest are introduced in Section 3. An overview of the local housing market is explained in Section 4.

In Section 5 all the methodology is carefully presented. Our dataset consists of panel data whose analysis may be developed in different ways depending on the type of task aiming to accomplish. Usually FE, RE, FD and (P)OLS are the most common estimators to analyze cross-sectional time series data. In order to select among the different models, we implement well-known specification tests allowing the user to select the appropriate method. This step will be based on a precedent work for the U.S. housing market. Indeed, the target is to replicate *Case and Shiller (2003)'s* Table 3 regressing house price returns on a mixture of national and city level variables. This will give us a preliminary idea of which factors affect our dependent variable.
After, the analysis will proceed by adding other two regressors not considered in *Case and Shiller (2003)* which, in our opinion, could have a statistically significant impact on house prices. Each of these two variables will be added to the model independently. Only in the final specification, these variables will be added together. The first variable is net migration and the second is urban sprawl.
The next step involves deciding which panel data specification best fits our data. Indeed, *Case and Shiller (2003)'s* Table 3 is replicated according to all specifications and with different R code functions to see if results are consistent across models. Once an estimator is chosen, then we would try to discern between increasing returns due to either national

or local effects. Indeed, we would need to include country and city effects in the selected model and observe their impact on the dependent variable and the other regressors.

Section 6 exhibits the econometric analysis together with the Random Forest machine learning procedure. The algorithm provides another measure, other than regression analysis, to understand the importance of each explanatory variable on housing returns. Indeed, a comparison between regression coefficients significance and results obtained via the random forest is presented at a later stage. According to *Kuhn and Johnson (2016)*, the procedure is an ensemble technique averaging the results of a set of decision trees and represents an improvement in predictive accuracy over the bagging procedure. The main distinctive feature added by the random forest is the inclusion of randomness in the number of variables considered to split the data.

Section 7 entails the last part of our analysis focusing on the estimation of a regression capturing local trends in the housing market via the inclusion of bubble persistence, mean reversion and adjustment coefficients. The idea was first presented by *Capozza et al. (2002)* on U.S. data. Our intention is to describe the situation in the European housing market by interpreting those coefficients and then establishing the eventuality of a bubble in some of the cities in the sample.

Ultimately, Section 8 presents the conclusions together with the limitations and suggestions for further research on the topic.

Problem Statement

The real estate market is an extremely broad topic, so we define questions tracing out the main milestones touched by our analysis. The principal question addressed is:

> Which are the housing price dynamics across the European market? Is there any evidence of frenzy behaviour causing actual prices to deviate considerably from their fundamentals? Alternatively, is there any evidence of Shiller's notorious "Irrational Exuberance"?

In order to fully answer that statement it is necessary to take into account other sub-questions:

- What are the most relevant determinants of real house price returns?
- Are there any time, national or local effects?
- What is the main predictor of house price returns according to the Random Forest machine learning procedure?
- Are housing markets predictable or is there somehow a degree of efficiency within the sector?

It is our priority to discuss possible explanations to these questions. In order to do so, we will consider previous research, our econometric knowledge and an innovative implementation of a machine learning algorithm to reconcile regression results.

1 Literature review

1.1 Approaches of housing economics

Housing economics distinguishes between four main different approaches:

1. No arbitrage condition between renting and owning a home
2. Financial no arbitrage condition, i.e. market efficiency
3. Spatial no arbitrage condition
4. New econometric approach

1.2 Main findings from the literature

The first approach, the no arbitrage condition between renting and owning a home, is adopted by *Poterba (1984)*. He presents an asset-market model of owner-occupied housing to analyze the impact of inflation and tax policy on relative home prices and on the size of the housing capital stock. In the paper, the analysis is grounded on the long-run relationship between house prices and construction costs. However, *Meen(2002)*, comparing the U.K. and U.S. market, argues that *Poterba (1984)'s* framework does not hold unless a full stock equilibrium occurs, which may take decades to manifest. Although, at a first glance, it may appear that the two housing markets behaved differently, *Meen(2002)* is able to show the existence of the same long-run transatlantic relationship between home prices and real income per household, real wealth, housing stock supply and real interest rate.

Muellbauer and Murphy (1997) take the cue from *Poterba (1984)'s* methodology and implement a model with an inverse demand and supply function of housing services. They analyze the volatility of the UK housing market between 1957 and 1994 and use the price-to-income ratio as an affordability measure to define housing booms and burst in a specific area. Their findings suggest that low initial debt levels, low house prices, growing wealth-to-income ratio and financial liberalization increased income growth expectations and represented key factors to the late 80s house price boom.

Poterba (1984) focuses on the no arbitrage condition between renting and owning a house, whereas *Glaeser and Gyourko (2007)* and *Case and Shiller (1987, 1989, 1990)* rely on

the second approach, the financial no arbitrage condition, where investors earn equal risk-adjusted returns by investing in housing or other assets. *Glaeser and Gyourko (2007)* argument considers one's ability to switch from being renter to owner in specific market conditions. They suggest the choice being severely limited by risk aversion and high volatility of housing prices. In the specific case of declining prices, it might be possible that households are not risk neutral.
After all, the house is perceived as the main investment for most individuals, so it is not possible to compare households decisions on single stocks and homes. House prices are highly volatile and risk aversion (when considering whether to buy or not) will entail one to delay purchase but on the other side when considering whether to sell or not, for the owner of the house, it is more convenient to rent it waiting for better economic conditions.

Case and Shiller (1989) demonstrates the inefficiency of the housing market in Atlanta, Chicago, Dallas and San Francisco. However, they also show that taking advantage of inefficiencies is not so simple and smooth, given the presence of transaction costs, carrying costs and tax considerations affecting real estate investment choices.
Additionally, by implementing time series cross-sectional regressions for those cities, *Case and Shiller (1990)* test for the forecastability of prices and excess returns. Specifically, they find out that the construction costs-to-prices ratio, adult population changes and real per capita income increases have a positive association with excess returns and price changes over the adjacent year. As a result, the single family homes market is inefficient.

Malpezzi (1999) develops an error correction model for house prices[2]. The analysis focuses on the long-run relationship of a dependent variable (house prices) on its lagged values and income. The key remark is a long run equilibrium price-to-income ratio denoted by k. This measure should vary depending on market conditions, time and some error term. The author thereby suggests two equivalent ways to estimate k. First option is to consider the change in prices. When, in two consecutive periods, there is no sign of variation, then the ratio is basically around its equilibrium level. Second alternative, instead, includes the choice of a market where k is relatively constant over the period of interest. Their analysis proceeds then by observing the movements of k if a shock occurs

[2]VEC: Vector Error Correction. For more information, please read Malpezzi (1999)

to the market. Results confirm previous work that housing price changes are not random walks and are at least partly forecastable. *Vyacheslav Mikhed and Petr Zemcik (2007)*, on the contrary, conclude that *Malpezzi (1999)'s* VEC model is not appropriate for modelling house prices and rents. They point out that it may take more than three decades for house prices to correct and return to fundamentals, i.e. converge to Malpezzi's k.

Case and Shiller (1989, 1990) point out that house prices dynamics present a positive serial correlation in the short run as well as a negative serial correlation in the long run. Conversely, *Poterba(1991)*, who also finds a positive serial correlation in yearly excess returns, does not find any evidence of house price reversion to the mean in the long run. *Dipasquale and Wheaton (1994)*, in accordance with *Case and Shiller(1989,1990)* and *Gyourko and Voith (1992)*, question the traditional stock-flow assumption that the housing market clears quickly and observe a tendency of house prices to follow a *"predictable cycle with a positive price serial correlation"*.

Gyourko and Voith (1992) analyze time series across 56 U.S. metropolitan areas finding evidence of positive serial correlation in some local appreciation series. This suggests that the housing market is predictable and foresighted investors could take advantage of some opportunities. Furthermore, they show unequal persistence in appreciation rates across MSAs suggesting local house price diverging from national trends.
Jud and Winkler (2002) obtain the same results with their analysis based on 130 U.S. metropolitan areas between 1984 and 1998. After identifying population growth, real changes in income, construction costs and interest rates as determinants of real house price appreciation, they observe that *"house price appreciation rates seem to vary because of location - specific fixed effect"*, whose magnitudes are positively correlated with restrictive growth management policies and limitations on land availability.

Capozza, Hendershott, Mack and Mayer (2002) analyze the dynamics of house prices of 62 U.S. metropolitan areas from 1979 to 1995. Their analysis not only provides additional evidence on serial correlation and mean reversion in house prices but also shows that considerable real construction costs and faster growth in both population and real income are determinants of greater autocorrelation.

Blanchard and Watson (1982) question the financial no arbitrage condition by arguing that the presence of arbitrage itself may not prevent the formation of a bubble in the market. Indeed, they argue that if there is a finite number of infinitely lived players and rationality, bubbles cannot emerge, while, instead they might arise until new participants enter in the market. Additionally, they point out that asset bubbles are more likely to emerge in markets where fundamentals are difficult to assess (gold market and real estate market) implying that participants have different perceptions about fundamentals.

According to *Glaeser and Gyourko (2007)*, *"it makes sense to conflate the rent-own no arbitrage relationship with the purely financial no arbitrage analysis of Case and Shiller (1989)"* because *"in both cases, the key prediction of the absence of arbitrage is that there will not be excess predictable returns for owning"*.

On the one hand, owing to inefficiencies, the financial no arbitrage condition as in *Case and Shiller (1989, 1990)* is based on the idea that an investor should not be indifferent between owning a house or investing in alternative assets.

On the other hand, the no arbitrage condition as in *Poterba (1984)* is based on the indifference between owning or renting a house. However, *Glaeser and Gyourko (2007)* show that the reservation price[3] of a prospective owner should be 40% higher than the one of a potential landlord willing to rent the same dwelling and facing the same costs. As a result, renting cannot be compared to owning the same property, thus those no arbitrage conditions are flawed and it does not make sense to use such common measures of housing economics as the price-to-rent ratio[4].

They underline structural differences between units available for sale or rent. Typically, most owned properties are single-family detached homes, whereas rentals belong to other categories, i.e. mainly apartments or semi-detached houses. Furthermore, they point out that owned properties have greater size (usually double) than rentals and they are located in more prestigious neighbourhoods. Last but not least in terms of importance, income distribution tends to be different among renters and owners, i.e. people choosing to buy and live in a specific house, when compared to renters, are richer and married

[3]Maximum price that a buyer is willing to pay for a product

[4]Price-to-rent ratio is seen as a sort of housing-market-replica of the well-known price-to-earnings ratio.

with children. Although there are some similarities between owner-occupied and rental housing, they are, at the same time, substantially different and belong to partially different categories experiencing different trends or market movements.
As it is pointed out in the introduction, in Copenhagen, house prices do not appear to have increased as much as rents, indeed the city shows a low ratio if compared to other Nordic agglomerations. In accordance with *Glaeser and Gyourko (2007)* but for diverse reasons *Mikhed and Zemcik (2007)* investigate the dynamics of the price-to-rent ratio and show that *"the non-stationarity of both house prices and price-to-rent imply that using standard regression techniques with price-to-rent ratio give invalid predictions about the growth of both house prices and rents"*.

Among many authors, *Himmelberg, Mayer and Sinai (2005)* consider the price-to-rent ratio in their paper. They make use of the OFHEO[5] price index which is a repeated sales price index controlling for minor quality improvements of the homes considered. Their main point is that when the price-to-rent ratio remains high for a while, it should be attributed to unrealistic expectations in house prices gains rather than to fundamental values. The authors also point out that house prices are more sensitive to variations in interest rates when rates are already below a certain threshold. In the time span investigated in the paper, real interest rates are low, which means that a decrease in interest rates is likely to foster demand contributing to a greater percentage increase in house prices than an equivalent downward change would bring if starting from a higher level.

The asymmetry between prices and rents pushes *Glaeser and Gyourko (2007)* towards a third approach, i.e. a spatial no arbitrage condition along with a no excess profits condition for builders, which is based on the idea that people should be indifferent to either buying or renting a similar house between different locations. In relation to this, differences in house prices may be explained by income levels and amenities.
The results from their model turn out to be somehow different from previous research. On the one hand, in contrast with *Case and Shiller (1989)*, there is no evidence of serial correlation in house prices, whereas on the other hand results suggest the presence of mean reversion in house prices over a 5 years horizon. *Glaeser and Gyourko (2007)*

[5]Office of Federal Housing Enterprise Oversight

explain that this pattern is mainly due to a mean reversion in economic shocks affecting the housing market along with the availability of new buildings determined by the response of the supply with housing construction. Although their model sheds light on some relevant points, it does present some drawbacks worth mentioning. Indeed, the authors themselves point out that *"Spatial equilibrium models clearly imply that housing should cost more in more pleasant climes but they cannot tell us whether people are overpaying for California sunshine. The heart of the model lies in spatial comparison, so it could never help us understand whether national housing prices are too high or too low"*.

The last approach entails the so-called new econometric models based on Vector Autoregressions, allowing to estimate and forecast the impact of the variable shock on the others. For instance, *McDonald and Stokes (2011, 2013)* analyze the impact of a shock to the federal funds interest rate on house prices.
The authors show that monetary policy is one of the main factors which *"contributed both to the housing price bubble and to the subsequent decline in housing prices"*.

Gupta and Miller (2009) analyze the housing market in Southern California by calculating out-of-sample forecasts[6] through many different Vector Autoregression and Vector Error Correction models, i.e. VAR[7], VEC [8], BVAR[9], SBVAR [10], SBVEC[11], CBVAR[12], CBVEC[13]. The main result of their paper is that *"housing market reflect, in large measure, run ups and then crashes in land values"*.

1.3 Theoretical framework

This research focuses mainly on econometric models trying to extrapolate the housing market behaviour. The starting point is *Case and Shiller (2003)* regression which looks at population, employment, mortgage rate, unemployment rate, housing starts and income

[6]Out-of-sample forecasts use all the available data to estimate observations which are not part of the sample, while In-sample forecasts are used to forecast observations which are part of the sample
[7]Vector Autoregression
[8]Vector Error Correction
[9]Bayesian Vector Autoregression, for further details see Litterman (1981), Doan et al., (1984), Todd (1984), Litterman (1986), and Spencer (1993)
[10]Spatial BVAR, for further details see LeSage and Pan(1995)
[11]Spatial BVEC
[12]Causality BVAR
[13]Causality BVEC

per capita to explain house price patterns. In their model, the change in employment and housing starts are proxies for housing demand and supply, respectively. The main results point out that when there is a significant correlation between house prices and income, adding other regressors does not augment the predictive power of the model. On the contrary, where income does not have such a strong effect on house prices, adding explanatory variables such as mortgage rate, housing starts, employment and unemployment contributes to the precision of the model.

In our analysis there is evidence of such strong relationship between income and house prices, despite the fact that we implement a different approach than *Case and Shiller (2003)*. They consider each state as a separate case and regress house price returns on income per capita for each state in order to obtain the R-squared measure. From here, they extrapolate such strong relationship between house price returns and income per capita. Their next step involves considering only the eight most volatile state-level housing markets in their sample and then adding other explanatory variables to try to improve the predictive power of the model.
In our case, instead, due to limited data availability we cannot exactly replicate their approach, so we opted for an alternative and more appropriate method according to the information at our disposal. Indeed, we first consider the determinants of house price returns on the European market and then include in another specification time, country and city effects. In other words, instead of running regression for each city, we include in a separate model city dummy variables (one for each city).

After a close replica of their model for the European market, the analysis starts including other factors and different specifications to augment the predictive power on house prices. Other authors, namely, *Poterba(1991)*, *Mayer and Sommerville (2000)*, *Capozza, Hendershott, Mack and Mayer (2002)* introduce a land supply index. Considering the relevance of such parameter, we decide to embrace their suggestion and insert it in our model through the urban sprawl index which should act as a land availability proxy.

Another variable not considered in *Case and Shiller (2003)* which could bring significant improvements to our model is net migration. The variable defines itself but what is most

important is that, especially in recent times, considering the inflows and outflows of population it should improve the predictive power of our model.
According to comments by David N. Weil on *Poterba (1991)'s* paper, the author underestimates the weight of migration at city level. This effect, according to Weil's critique, should alter changes in adult population due to natural increase, especially because migration flows are tightly linked to the destination country or city economic situation.
In other words, increases in housing demand should not depend solely on income per capita but also on income elasticity of migration. Indeed, if the responsiveness to variation in income is substantial, a tiny boost to households' wealth would attract a considerable amount of immigrants, leading to an increase in population, which in turn would result in higher house prices (through housing demand). These are the main reasons why we believe including the measure could significantly contribute to our cause.

Card (2007) also provides useful insights on the effects of immigration in U.S. cities. He points out that rents as well as earnings are larger in cities with substantial immigration and this parallel movement weakens the burden coming from rent increase. In other words, immigrants might be expected to have some effects on rents and house prices of existing residents. Shorty before *Card (2007)*, *Saiz (2006)* introduces a simple theoretical model in which an inflow of immigrants causes a short term increase in rents as families compete one another for a sticky supply of housing. In the long term, instead, assuming an elastic supply, the stock of housing will be able to accommodate those in needs and given that some people will move to other places, the initial increase in rents should smooth out.

According to *Abraham and Hendershott (1994)*, real price appreciation can be mainly attributed to two groups of factors. One explaining the change in equilibrium prices via real income, real construction costs and the after-tax interest rate, the other capturing the deviation from equilibrium prices by considering lagged real appreciation and the difference between actual and equilibrium real house prices. Indeed, one of the most interesting aspects of their model is the introduction of a proxy for the tendency of bubbles to burst (λ_2).
The term should capture the deviation between the fundamental price and the actual one. If the difference between the two measures is significant and negative, then the coefficient

indicates a higher chance of the bubble to burst as the deviation from fundamentals is allegedly substantial.

Blanchard and Watson (1982) also investigate the bubble phenomenon even though they do not make use of coefficients capturing a specific effect. However, they try to model a house bubble by providing two specifications, one is "deterministic", the other assigns probabilities measures for bubble persistence and crash. The former assumes that higher prices are justified by larger capital gains and that deviations grow exponentially. This is clearly not feasible in the long-run under rational expectations. The latter, including prospects, states that the likelihood of bubble to burst might be related to how much time has elapsed from the beginning of the bubble or how far the value is from fundamentals. Ultimately, *Vyacheslav Mikhed and Petr Zemcik (2007)* also introduce a bubble indicator which equals unity if house prices are non-stationary and rents are stationary, whereas it is zero if prices are stationary.

Upon considering evidence from various sources, our analysis focusing on the behaviour of house prices across European cities will adopt some "bubble persistence", "mean reversion" and "partial adjustment" factors. These coefficients should spot major trends and indicate whether some cities are experiencing a surge in real prices beyond fundamentals. Credits for the inclusion of these specific coefficients go to *Capozza et al. (2002)* who by adding the last coefficient elaborate on the work of *Abraham and Hendershott (1994)* published a few years before.

2 Comparison across House Prices methodologies

One of the main objectives of our analysis is to understand which factors drive house prices across different European cities and countries. Unfortunately, due to the greater data availability, most housing research and publications focus on the U.S market. Despite this inconvenient, that literature represents a great source of inspirations for our thesis. Skimming through numerous articles we came across different methodologies to compute house prices indices. This section is paramount to our analysis and we believe it is essential to provide an exhaustive explanation before focusing on the data description section.

2.1 Average House Prices

There are various methodologies to calculate average prices. Usually, the arithmetic, geometric mean and the median are the most widely used.

The first method, also known as simple mean, is a value derived by summing up prices and dividing them by the number of properties considered. Now, if the units are very similar in value, then this method is extremely straightforward and works quite well. On the contrary, if there is a lot of variance between house prices, the simple mean is not suitable as it will distort the true mean. In this case, the geometric mean is more appropriate as it gives less weight to expensive properties lowering the average (except when house prices are the same, in this case arithmetic and geometric mean coincide). Because of this weighting factor, the geometric mean value is closer to the median. Last but not the least, the median considers the value in the middle of the distribution.

In the housing market literature, the so-called median house price index has been extensively criticized. According to *Case and Shiller (1987) "Median house prices are not the best data to extrapolate appreciation*" because they are too volatile and excessively depending on the characteristics of sold dwellings. Indeed, if we assume a period in which only high-quality houses are sold, the index would imply an increase in the median price

even if there has not been a de-facto increase in home values.

McCarthy and Peach (2004) and *Poterba (1991)* follow *Case and Shiller (1987)* by labelling the median price index as "unpredictable", especially in the short run, and not being able to accurately reflect home values due to the fact that the measure mainly considers recent sales. Furthermore, the argument of *Case and Shiller (1987)* goes on considering the expectation of a better quality of life in the future (given an increase in income) implying that the quality of homes should rise as well. Their explanation follows a simple logic: new homes will soon become existing ones and in turn will affect the median price with their higher quality. Therefore, even without real appreciation, median prices will increase due to the higher quality of the new dwellings built in that period.

The one thing these three methodologies leave out is quality adjustments. Most statistical offices, indeed, decide to adjust their average/median values for quality.
There are a few ways to do so, as *Case and Shiller (1987)* suggest. The first one is the so-called hedonic regression method, the second one is the repeated-sales procedure.
Ultimately, there is another method, which *Case and Shiller (1987)* do not mention, the SPAR-method[14], which is used by some Statistical Offices in our sample.

2.2 Hedonic Adjustment

The hedonic regression considers all main house characteristics where each one has a weight on the price paid for real estate. For instance, the fact of having a garden and a private parking lot will each contribute to the final price but it is not possible to put a price on a specific characteristic via this regression. Indeed, as *Case and Shiller (1987)* point out, coefficients might be interpreted as attribute prices.
This method basically gives the user the possibility to obtain a value for the house by considering different combinations of features. Despite its benefits, this method is not without stains. *Meese and Wallace (1991)* criticize the hedonic strategy because it requires large and costly datasets including actual sales prices and property characteristics. Furthermore, the regression will become computationally demanding and cumbersome

[14]It stands for Sales Price Appraisal Ratio method

the more regressors are included, although the more the variables, the more precise and complete the model will be.

2.3 Repeat Sale Index

Case and Shiller (1987) also propose a second option to correct for quality, the repeated-sale index. They created such index for single family homes for 20 U.S. cities dating back to 1987. This method, however, does not adjust for quality, or perhaps it does but not in the same way as the previous one.
The index considers the differences in purchasing prices of the same unit over time, so tiny house improvements are considered but overall the value should not strongly depend on quality. The main drawback of this method is that it considers only properties sold more than once, i.e. all one-time purchases are not considered. Given the amount of data which is not taken into consideration, this method is useful when we have long time series.

2.4 SPAR method

The SPAR approach focuses on the relationship between the average sale price and the appraisal value, where the latter controls for quality.
van der Wal and Tamminga (2008) argue that exclusively comparing average sale prices in two consecutive months might lead to misleading conclusions. For instance, house prices may appear to increase when assuming that all the lower quality houses are sold in one month and all more expensive properties are sold in the following one.
The SPAR method, however, solves this problem by introducing the appraisal value, which takes into account differences in quality levels between houses sold in different periods. Therefore, if a house has a lower appraisal value, it means that it is of lower quality, i.e. it might be reasonable that its average price is lower than the one of another deluxe house. In other words, it depends on the value of the average price in relation to the appraisal value of the property. If the percentage change over two consecutive time periods between the average price-to-appraisal value ratio increases, then prices have increased even if looking at average house prices indicates a percentage decrease in the window of time considered.

In the next section where the data will be presented, it will be specified which method has been used to compute house price data. As a general trend, most statistical offices do not use median prices but rather averages which are then adjusted for quality.

3 Data

Tables 1-4 (Appendix) show the source and description of every variable in the dataset. It is an unbalanced panel which covers over 900 observations, about 50 cities and spans across 10 different countries. Since most of the research related to the housing market has been carried out for the U.S. market, finding data at city level for the European area revealed to be quite challenging. Indeed, the dataset is composed by hand-selected information extrapolated and pooled together from different sources.
Before focusing on the description of the dataset, a remark concerning the status of our data should be presented. Statistical offices tend to frequently upload figures and given that our thesis project covers a period of roughly six months, information might be revised by them within this time frame. To the best of our abilities, the most up-to-date figures are taken into consideration, however, in some instances, numbers might have been updated too close to our deadline, so that we had to maintain the previous version.

3.1 House prices

Throughout our analysis, we have encountered many different property definitions and since each statistical office reports the one they prefer, it is important for us to describe them to avoid any sort of confusion. Furthermore, our work is carried out according to our understanding and classification of these definitions.

Single-family homes are also known as detached houses and single dwellings (1 dwelling) referring to individual familial buildings where single-families live. This is different from semi-detached, terraced, row and multi-dwellings (2 dwellings) houses where instead more than one family shares the housing unit. The distinctive trait of the latter type of housing is that homes are usually identical and one next to the other, forming long rows of housing units.

The analysis is based on deflated house price data. When data are available in nominal terms, they are discounted by the National Consumer Price Index in order to generate a homogeneous house price index with 2005 (2005=100) as base year. Although real house

prices are not homogeneous across different cities, since the types of property (single, single-double or all residential family homes) may vary depending on data availability, these are homogeneous within the same city, i.e. the type of property is the same for each municipality across all different variables.

The UK government provides a full dataset on municipalities gathering average house prices for detached, semidetached, terraced houses and flats[15]. These are calculated as a three months moving average meaning that monthly figures are computed as the arithmetic average of the three previous periods. Such procedure, which is crucial for reducing the volatility of house prices, does not remove the bias related to differences in terms of quality reflected in average house prices. For this reason, data are also adjusted with a hedonic regression considering different amenities among houses sold during different time periods.

Regarding Germany, data have been extracted from two different sources: the online platform *Statista*[16] and the private database *Bulwiengesa AG*. The former collects and makes available prices per square meter for detached and semi-detached houses in Berlin, Hamburg, Munich, Koln, Stuttgart, Essen, Dresden, Dortmund, Dusseldorf and Bremen. These data are actually provided by the Empirica Institute database[17] and are adjusted for quality via hedonic regressions.
Bulwiengesa AG, instead, provides nominal house price indices for all residential dwellings of Bonn, Bochum, Mannheim, Freiburg and Nurnberg. The methodology of how Bulwiengesa AG computes these indices remains a bit unclear, however we have asked them directly to provide their methodology literature. In order to be as much truthful as possible we include a synthesis of their definition: *"In Germany, the past few years have seen a marked expansion in the availability of price indices for residential real estate ...The data of bulwiengesa AG are largely based on expert assessments... to determine the value of typical properties...The house price index is thus not confined to owner-occupied residential real estate but also incorporates the prices of rental property"*[18].

[15]For more information, visit: https://www.gov.uk
[16]For more information look up: https://de.statista.com
[17]Statista acts as an intermediary
[18]For the full definition, visit: https://www.bundesbank.de

Data about Spanish cities are provided by the *Ministero de Fomento*[19], whose database incorporates nominal house prices per square meter on *Vivienda Libre*, which, according to their description, include all the types of dwellings involved in private transactions in the market, namely not subjected to any public protection regime and can be transferred without any restriction among the counterparts of the transaction.
A main limitation of using average prices per square meter is that these data are not optimal indicators when differences in terms of both characteristics and locations are not taken into account. In order to fix this distortion, the *Ministero de Fomento* points out that houses are divided into subgroups on the basis of geographical areas and common characteristics. Every class receives a different weight, therefore the average price per square meter is given by the weighted arithmetic average of each class for each geographical location for each time period.

Data about Northern European cities, i.e. Denmark, Finland and Norway can be downloaded directly from their respective National Statistical Offices: *Statistics Denmark, Statistics Finland and Statistics Norway*[20] providing historical time series for single family homes. These are in terms of price per square meter for Helsinki and house price indices for Copenhagen and Oslo.

When it comes to describing the methodologies implemented, Copenhagen house price index accounts for quality differences through the SPAR-method, while both *Statistics Norway* and *Statistics Finland* adopt hedonic methods. *Statistics Sweden*[21] supplies prices per square meter for detached and semi-detached houses and data are based on a simple arithmetic mean and are not adjusted for any differences in terms of quality among houses sold. As a result, it is likely that these data are upward biased.

Lastly, with regard to the Western European countries, data at city level are available for Netherlands, Belgium and France. The *Dutch Statistical Office*[22] presents nominal

[19] http://www.fomento.gob.es
[20] Denmark: http://www.statbank.dk, Finland: https://www.stat.fi and Norway: https://www.ssb.no
[21] https://www.scb.se
[22] Central Bureau voor de Statistiek. For information visit: https://www.cbs.nl/en-gb/figures

house prices indices by city on existing owned homes. Data, also in this case, are adjusted for quality with the SPAR-method. *StatBel*[23] procures statistics on average house prices for all residential dwellings, whereas the *French National Statistical Institute*[24] supplies a house price index for logements. However, in both cases house prices are adjusted by the implementation of hedonic methods.
Most house prices data are available in yearly time series, even though some countries provide both yearly and quarterly series. Concerning other variables, data are mainly available year by year. This is why the analysis will focus on yearly data.

3.2 Housing Starts

Case and Shiller (2003) argue that housing starts may represent a proxy for supply restriction. The association between housing starts and house prices could be either positive or negative. On the one hand, low housing starts may boost prices due to specific factors such as land regulation and availability. On the other hand, a positive association may exists because of builders' construction activity in response to an increase in house prices[25].

In our dataset the variable housing starts consists on the number of building permits issued before construction of residential properties. However, since the UK government does not collect data about building permits, net additional dwellings are used as a substitute for all the British cities. It is likely, however, that these data are biased, since net additional dwellings include additional information, i.e. house demolition figures[26].

The *German Statistical Office*[27] provides data about building permits by building type and by municipality. Although data about Dresden and Leipzig are not available from the German Statistical Office, they have been collected via a formal request directly to the German real estate consulting company *Bulwiengesa AG*. It is important to mention

[23]For information refer to: https://data.gov.be/en
[24]INSEE: https://www.insee.fr/fr/accueil
[25]Case and Shiller, 2003
[26]Case and Shiller (2003), instead, use a time series based on the historical relationship between permits and starts and a proprietary data base on permits
[27]Die Regionaldatenbank Deutschland

that figures have been chosen accordingly to the type of property used for the house price variables, such that they refer to the same type of property for each city.
As described in the previous section, house prices are available in terms of detached and semidetached homes for some German cities and in terms of all residential dwellings for other cities, depending on the different source.

Regarding Spain, the Spanish *Ministero de Fomiento* publishes both monthly and yearly time series about building permits referring to *Vivienda Iniciada*, i.e. new residential dwellings. For the city of Bilbao, given the lack of data, housing starts are taken from the whole region Bizkaia. Data are likely to be upwardly biased in magnitude, although Bilbao is the largest city within the region.

For all the remaining countries, data at city level are taken from their respective National Institute of Statistics. However, Statistics Denmark does not adjust the data for delays in construction, implying that figures could include permits for delayed buildings. It may be the case that, for this reason, the number of permissions issued might be upwardly biased.

3.3 Population

Eurostat provides a detailed picture of the diverse EU territories by collecting disaggregated statistics at regional, metropolitan and city level. These data have been pooled and are accessible in the *Urban Audit* and the *Large City Audit project*, which incorporates 171 variables and 62 indicators. Statistics about Sweden, Finland, Netherlands and Belgium have been obtained from their respective National Statistic Institutes, due to either the unavailability of data or to the presence of shorter time series in the Eurostat database.

3.4 GDP per capita PPS

In its website section *Economic accounts by metropolitan regions*, Eurostat includes data at municipal level about GDP [28] per capita. The aforementioned variable is standardized per continuous PPS[29], i.e. it takes into account price level differences between countries

[28] Gross Domestic Product
[29] Purchasing Power Standard

but it cannot be used for time series comparisons. As a result, we manually compute constant PPP time series. First of all, we divide GDP (provided by Eurostat in millions of euro) by population in order to obtain a more accurate measure of nominal GDP per capita for each city. Second, we deflate GDP per capita by CPI[30]. The last step involves taking GDP per capita per continuous PPS (base year = 2011) and apply the growth rates of the real GDP per capita forwards and backwards in order to extend the series and to obtain a PPP constant measure for each city.

3.5 Unemployment and Employment Rate

OECD [31] collects data about unemployment and employment rate for 271 metropolitan areas all over the world. The unemployment rate stands for the number of unemployed people as a percentage of the labour force, i.e. the unemployed plus those in paid or self-employment. The employment rate, instead, is the ratio of the employed between 15 and 64 years to the working age population[32].

Given that OECD provides data with short time horizon, i.e. from 2000 to 2014, when longer time series are available we use data provided from the Eurostat database. Unemployment rate data about London, Leeds, Bradford, Sheffield, Bristol, Newcastle, Leicester, Portsmouth, Nottingham are collected from the Eurostat. Data about Liverpool, Birmingham and Manchester are collected from the OECD database. For the Employment rate, instead, data about London, Leeds, Bradford, Manchester, Sheffield, Bristol, Newcastle, Leicester, Portsmouth, Nottingham are collected from the Eurostat whereas data about Liverpool and Birmingham are collected from the OECD database.

3.6 Mortgage Rate

The mortgage rate refers to the long-term rate for house purchases. The variable is at national level and varies across time, so each city in a country presents the same rate for a specific year. The ECB[33] provides interest rates on loans over 5 years for house purchases across UK, Germany, Spain, Denmark, Sweden, Finland, the Netherlands and

[30] Consumer Price Index, index 2005=100

[31] Organization for Economic Cooperation and Development

[32] Definitions from OECD, for further details see https://data.oecd.org/emp/employment-rate.htm and https://data.oecd.org/unemp/unemployment-rate.htm

[33] European Central Bank

France. Unfortunately, this is the best mortgage rate proxy available. The main reason why this measure is chosen is given by its homogeneity across most observations and because it specifically refers to "home purchases", so we are certain that the values refer to mortgage rates and not to other types of rates.

The main drawback, however, comes in terms of loans length considered as "over 5 years" does not indicate a precise time span as "30 years mortgage" rates would. The ECB provides mortgage rates data also for Belgium for all house purchase loans, not just the ones over 5 years. However, the measures are identical and the more general data records longer time series, that is why it is preferred over the other.
Norwegian mortgage rate data are supplied by their national statistical office. Also in this case, however, data present only a general reference on the duration of the mortgage and refer to outstanding loans on secured dwellings.

3.7 Net Migration

Net migration is one of the two variables used to augment *Case and Shiller (2003)* paper's model. A generally shared definition of net migration is the difference between the number of immigrants and emigrants in a specific area at a specific time [34].

In our case, the variable is considered at city level and the time span depends on city specific data availability. Indeed, for instance, in the case of Spain, the Eurostat figures are preferred to those of National statistical offices because often the latter provides data at the province level, whereas the former provides more accurate city level data.

UK figures on net migration are provided by the Office of National Statistics, whereas for German cities data are collected from the Federal Statistical Office[35]. The Eurostat provides data on Spanish, Belgian and French Cities[36]. Data on Danish cities are provided by Statistics Denmark but net migration figures are not available, so they are manually computed by differencing immigration and emigration figures by city. Data for Swedish

[34]This definition is taken from the OECD website. For more information, consult : https://stats.oecd.org/glossary/detail.asp?ID=6639
[35]Statistisches Bundesamt. For more information, visit: www.destatis.de
[36]Net migration plus statistical adjustment. For more information, look up: http://ec.europa.eu/eurostat/data/database

cities, obtained from Statistics Sweden are computed in the same manner. Statistics Finland provides directly net migration figures[37]. Norway, as well, supplies information via Statistics Norway[38]. Lastly, the Netherlands supply data directly from their Statistical Office[39].

3.8 Urban Sprawl

Urban sprawl index is the second variable included which should augment the explanatory power with respect to house prices in the original *Case and Shiller (2003)* model.

The figures on the variable are provided at city level entirely by the OECD database[40]. The data do not present any time series but only cross sectional variation. The research on this variable sheds light on two periods covered by the index: from 2000-2006 and from 2000-2012.

In our case, the last specification is selected given the longer time span covered. The index considers how much a specific area progresses in terms of development with respect to its population growth. In more detail, it measures whether population growth is greater than the development in built-up areas. There is a specific benchmark for which, accordingly, an increase in population dictates a certain increase in built-up areas. If the two variables are stable over time, then the index is equal to zero.

Alternatively, if the "building" growth rate exceeds the predicted change by the benchmark measure, then the index will be greater than zero, otherwise it will be lower. In the former case, population density will diminish, whereas land availability will increase. For the supply and demand relationship, this should intuitively drive house prices downwards. On the contrary, if the index is negative, it means that the density has increased and land availability reduces inflating house prices. As previously mentioned, the urban sprawl index represents a proxy for land availability within a specific city.

Here, a remark must be made. We are not considering the fact that a city cannot keep increasing built-up areas indefinitely, as there are other cities which take control of the

[37]Data available at: http://pxnet2.stat.fi/PXWeb/pxweb/en/StatFin/ (choose 008 in the list)

[38]Data available at: https://www.ssb.no/en/statbank/table/05426?rxid=f81fe4f8-5e32-41ae-b8d7-c574e6d81b0b

[39]For more information, consider: https://data.overheid.nl/data/dataset/population-dynamics-birth-death-and-migration-per-region

[40]https://stats.oecd.org/Index.aspx?DataSetCode=CITIES

development process after a certain point. This is certainly a limitation but indeed the index constitutes only a proxy which by definition is not a perfect measure of the original variable.

3.9 CPI

The consumer price index is mainly used to deflate nominal house prices into real ones. In order to do so, statistical offices have been individually contacted on whether house prices extrapolated from their websites are in nominal or real terms, so that our data could be adjusted accordingly. The CPI figures are national and are taken from country statistical offices for Germany, Spain, Denmark, Sweden, Finland, the Netherlands, France and Norway. The UK data are provided by the government office, whereas those for Belgium are made available by the National Bank of Belgium[41].

[41]For more information, visit: https://stat.nbb.be/Index.aspx?DataSetCode=NICPHISTO

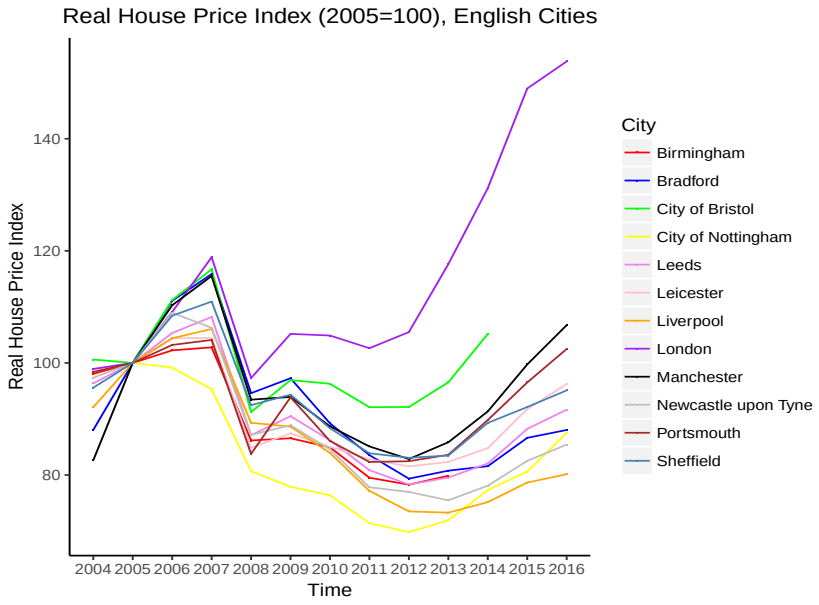
4 City and Time trends

The purpose of this section is to illustrate a preliminary analysis on house price levels among cities and across time. Due to data unavailability, the base year in all the graphs below is 2005. In other words, before getting into the phase where regression models are implemented to try to explain the variation in house prices, it is paramount to present and explain data as they are adjusted in real terms. The purpose of this section is to illustrate with graphs the data collected, assuming they will provide the reader with some insights before going through the other sections. The figures in this section will display real house prices indexes by city through time, providing some clues on the housing situation in specific cities.

4.1 English RHPI

Looking closely at the English Cities plot (Figure 1), it is immediate to notice the common trend among them.

Figure 1



Considering 2005 as base year, the graph shows an uptrend from 2005 to 2007, matching the run-up in prices anticipating the recent financial crisis. During the collapse, prices plummeted until the end of 2008, to then increase almost unanimously for approximately a year. However, this is a general picture, indeed there are some local exceptions. Some

cities, even after 2008, such as the City of Nottingham, Liverpool and Birmingham do not follow the same trend but rather show a decrease or no change in the index. Following the period where some differences are clearly visible, house prices started homogeneously decreasing until 2012. From 2012 to recent days, prices have been increasing at different rates across our subsample of 12 cities.

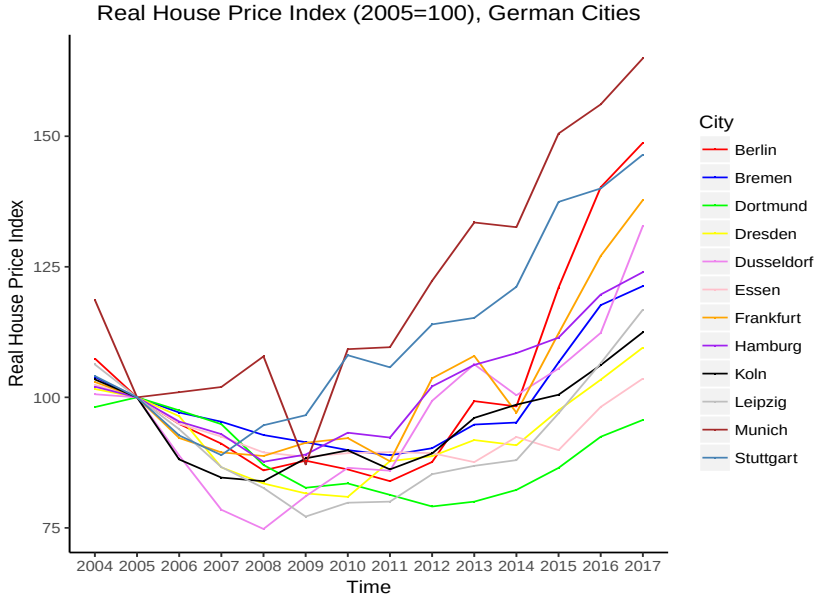
London, Liverpool and Bristol presented significantly different patterns than the other cities. London and Bristol from the end of 2008 display a stronger uptrend in prices, especially from 2012 onwards. This effect is clearly more evident for the former than the latter. Liverpool shows a continuous decrease in the price level (from 2005 to 2012) which is more marked than in all the other cities.

Focusing on the last years of the graph, it appears the other cities are following London with a lagged effect. Despite this, already in 2014 the difference between the group and London (excluding for a moment Bristol) is considerable, ca. 30 index points. Indeed, for all cities, except London, the pre-crisis real price level has not been reached yet.

4.2 German 1-2 Family homes RHPI

Analyzing the graph for 1-2 Family homes in German cities provides some interesting results (Figure 2).

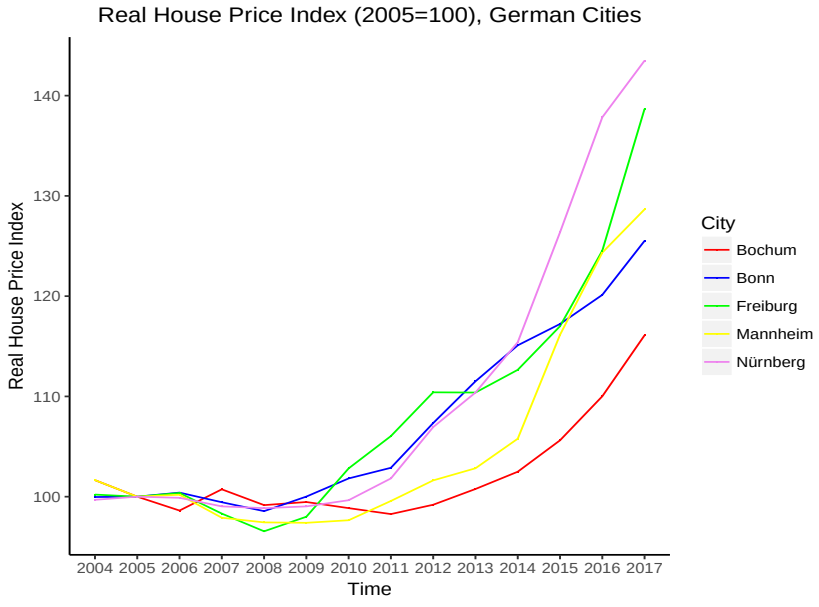
Figure 2



Although most German cities in the plot display a decrease in their respective house prices until the end of 2008, Stuttgart and Munich present different trends. The former depicts an increase starting at the end of 2007, which, considering some exceptions where prices have slightly declined (2010-2011) or remained stable (2008-2009), has not stopped yet. The latter shows a completely different pattern, similar to what happened across UK cities.

Indeed, in Munich house prices started increasing from 2005 onwards, especially from 2007 to 2008 and then plummeted until 2009 much more in relative terms than what happened in the other cities. Since the end of 2009, in Munich prices have been increasing, or rather alternating steep rises to flat periods. However, from 2014 property prices have been skyrocketing and the trend appears to be followed by other cities, especially Frankfurt, Berlin and Dusseldorf. Overall, real house prices are in all German cities but Dortmund above their respective pre-crisis level.

Figure 3



4.3 German Residential RHPI

This other fraction of german cities (Figure 3) shows similar patterns to the previous plot. The index appears to be stable from 2004 to 2008/2009, as if the financial recession does not impact house prices in this group of cities. Overall, in that time span data show

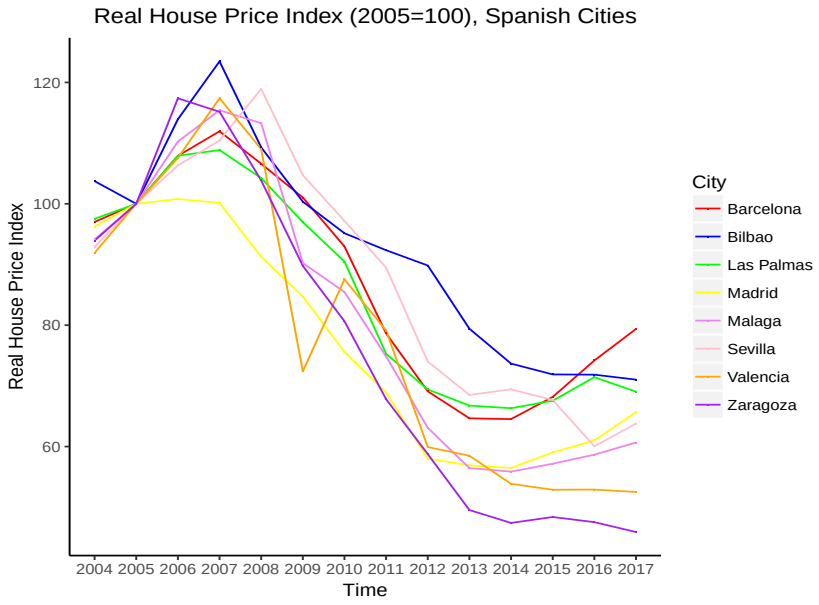
a shy decreasing trend.

Since then prices have been increasing with different growth rates. The surge is led by Nurnberg whose house prices behaviour from 2013 resembles the steep increase occurred in London, Stuttgart and Munich. Also Freiburg and Mannheim show significant increases in house prices which should be worth investigating further.

4.4 Spanish RHPI

As in the English city sample, in Spain, prices have increased just before the 2008 financial collapse (Figure 4). The city of Madrid presents the only exception where the index remains approximately flat from 2005 to 2007. During the recession, prices decrease in all Spanish cities homogeneously, especially in Valencia, where the index slump is steeper than in the other cities in relative terms but, at the same time, prices "recover" momentarily much more quickly.

Figure 4



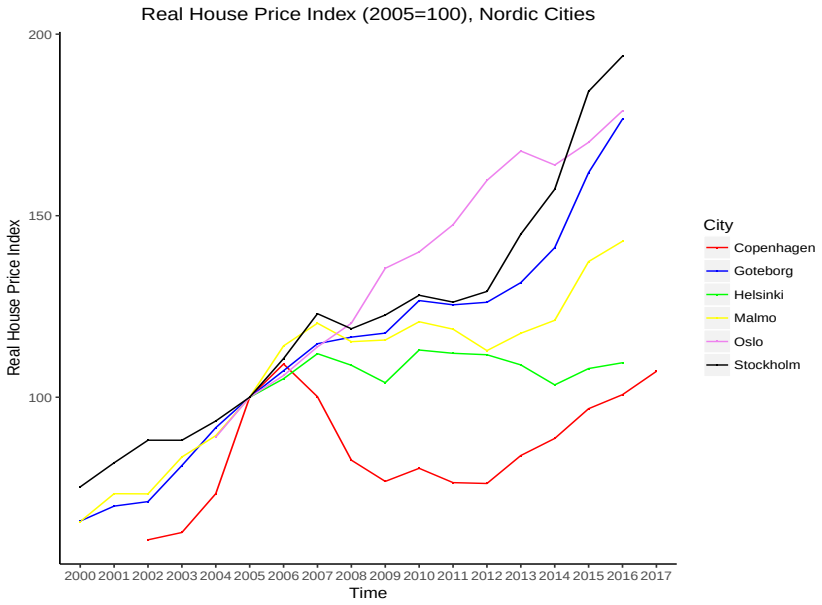
Indeed, Valencia house index shows an increase already at the end of 2009, even if it does not last more than a year, then prices begin falling again. A few remarks on the Spanish situation underline the fact that the country has not fully recovered from the disruptive consequences of the recent financial downturn. Indeed, starting in 2014, the prices show a shy recovery and in some cities such as Malaga the index is rather flat. Overall, all

house prices are according to the RHPI far below the pre-crisis level.

4.5 Northern European RHPI

Data about Northern Europe include Swedish, Danish, Finnish and Norwegian cities (Figure 5). Pooling together these data seems reasonable, since for some countries we only have time series relative to one city.

Figure 5



Concerning this clustering of cities, it is interesting to notice that the 2008 recession does not affect the housing market as much as those analyzed so far. Copenhagen appears to be the sole nordic city hit by the crisis in a similar way as the other European cities previously described. It appears that here prices start decreasing already at the end of 2006 (a year before the other cities just inspected).

Overall, prices seem to be constantly lower in Copenhagen, especially because they keep decreasing until 2012, whereas in Stockholm, Malmo, Goteborg prices drop until 2008-2009 and then alternate flat periods to steep increases, especially after 2012, leading to a price still above the pre-crisis level. Helsinki follows the initial pattern of the Swedish cities until 2009, but prices do not pick up from 2012 onwards. Indeed, as of 2016, the index for the Finnish capital reaches approximately the pre-crisis level. Ultimately, Oslo does not experience any significant decline due to the crisis as real prices have been

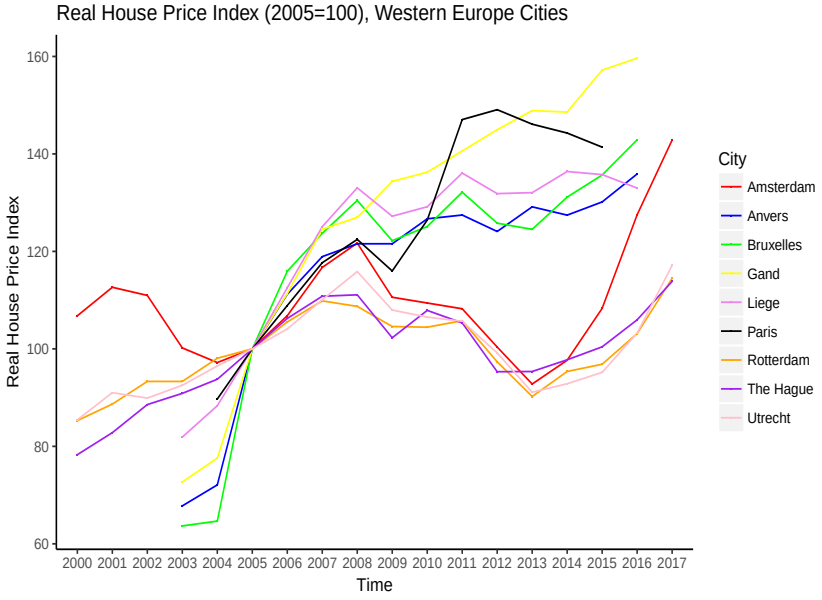
increasing throughout the time slot considered.

4.6 Western European RHPI

Zooming in the Western European graph (Figure 6), it is possible to observe the negative weight of the financial crisis (except for Gand). Immediately after the downturn period terminates, the graph parts ways.

On the lower group are displayed the Dutch cities. These appear to follow the same trend up until the end of 2013, then they all start rising. Above all, Amsterdam's prices skyrocket, creating a gap between its prices and the ones of the other Dutch cities. On the higher one, Liege, Bruxelles and Anvers follow a similar trend even after the recession but Paris and Gand indexes remain at a higher level (at least until 2012 for the French Capital). From this date onwards, Paris index start declining and shows convergence to the level of Liege, Bruxelles and Amsterdam. Gand house prices keep increasing during the time span, mimicking the behaviour of Oslo's index. As a final remark also for this cluster of cities (the higher group), pre-crisis real price levels have been abundantly overtaken.

Figure 6



5 Methodology

The first objective is to replicate Table 3 from *Case and Shiller (2003)*, i.e. regressing house price returns on a specific set of regressors to understand which factors are associated with the response variable. The main difference between our model and *Case and Shiller (2003)'s* is that they use quality adjusted median house price indices via the repeated sales methodology, whereas most of our data are average house prices corrected with either Hedonic or SPAR method. Additionally, longer time series are available for the U.S. housing market. Indeed, Case and Shiller use quarterly instead of yearly data.

Our dataset is composed by longitudinal data (also defined as panel data) disaggregated by city i and period t. In this section, different methodologies are compared, namely Pooling Ordinary Least Squares, Within or Fixed Effects, First Difference and Random Effects Estimator. Initially, estimates are obtained without making any correction on standard errors. All estimates are then corrected for serial correlation and heteroskedasticity, thus it is possible to compare our findings with those obtained by *Case and Shiller (2003)*.
The subsequent step involves the inclusion of two additional independent variables, i.e. net migration and urban sprawl.
Before starting with the analysis, an explanation about the model and the different estimators seems necessary. This theoretical part is extrapolated from *Wooldridge (2010, 2015)*.

5.1 Panel data

Panel data consist of time series for each cross sectional unit $i = 1, 2, 3...., N$ in the dataset. In other words, it is possible to track each cross sectional element (in our case each city) over time. The analysis is based on the long-run relationship between real house price returns and fundamentals including Δhousing starts, Δpopulation, ΔGDP per capita, unemployment rate, mortgage rate and Δemployment[42]. The panel spans from 2000 to 2017 but it is not balanced, i.e. time series are not available for the same time horizon for every city. For instance, data about Oslo are available only from 2011

[42] Δ stands for the log first difference of the respective variable

onwards. The baseline panel regression specification is:

$$y_{it} = \mathbf{x}_{it}\beta + u_{it}$$

where the dependent variable y_{it} represents house price returns for each city i at time t, where t indicates the year. Transposed vector $\mathbf{x}_{it}$ is a set of explanatory variables with dimension 1x K (K is the number of the explanatory variables) and u_{it} is the error term. In matrix notation, the extended form is:

$$\begin{bmatrix} y_{111} \\ y_{121} \\ \vdots \\ y_{1t1} \\ \vdots \\ y_{211} \\ y_{221} \\ \vdots \\ \vdots \\ y_{nt1} \end{bmatrix} = \begin{bmatrix} x_{111} & x_{112} & x_{113} & \dots & x_{11k} \\ x_{121} & x_{122} & x_{123} & \dots & x_{12k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{1t1} & x_{1t2} & x_{1t3} & \dots & x_{1tk} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{211} & x_{212} & x_{213} & \dots & x_{21k} \\ x_{221} & x_{222} & x_{223} & \dots & x_{22k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{nt1} & x_{nt2} & x_{nt3} & \dots & x_{ntk} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} u_{111} \\ u_{121} \\ \vdots \\ u_{1t1} \\ \vdots \\ u_{211} \\ u_{221} \\ \vdots \\ \vdots \\ u_{nt1} \end{bmatrix}$$

where x_{itk} is the observation for the city i at time t for the explanatory variable k. In stacked matrix notation:

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{u}$$

where $\mathbf{X}$ is NTxK , $\mathbf{y}$ is NTx1 and $\mathbf{u}$ is NTx1.

5.1.1 Pooled OLS Estimator

The simplest approach considers all data as cross-sections. The model can be written as:

$$y_{it} = \mathbf{x}_{it}\beta + u_{it}$$

where $\mathbf{x}_{it}$ is 1xK or in matrix notation:

$$\mathbf{Y}_i = \mathbf{X}_i\beta + \mathbf{u}_i$$

where $\mathbf{Y}_i$ is Tx1, $\mathbf{X}_i$ is TxK and $\mathbf{u}_i$ is Tx1. The Pooled OLS estimator is unbiased and consistent when the following assumptions are met:

1. The model is a static linear model, i.e. the model is linear in parameters and static because regressors contemporaneously determine the dependent variable
2. $E(X_i'X_i) = A$ is a nonsingular matrix with rank K , i.e. no perfect multicollinearity
3. $E(X_i'u_i) = 0$, i.e. zero conditional mean in the error term.

The derivation of the beta estimate is the following:

$$\hat{\beta}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (\sum_{i=1}^{N} \mathbf{X}_i'\mathbf{X}_i)^{-1}(\sum_{i=1}^{N} \mathbf{X}_i'\mathbf{y}_i) = (\sum_{i=1}^{N}\sum_{i=1}^{T} \mathbf{x}_{it}'\mathbf{x}_{it})^{-1}(\sum_{i=1}^{N}\sum_{i=1}^{T} \mathbf{x}_{it}'\mathbf{y}_{it})$$

A strong feature of the Pooled OLS is that no serial correlation and homoskedatisticity conditions are not required for the unbiasedness and consistency of the estimator. However, when serial correlation is in the error term, OLS standard errors are incorrect and Pooled OLS is inefficient.

5.1.2 First Differencing Estimator

The first differencing procedure, simply put, consists of subtracting, i.e. differencing, the lagged value of a variable from the variable itself. Then the (P)OLS are applied on these transformed data. Generally speaking, first differencing allows to relax exogeneity requirements on the regressors. Indeed, consider the base model:

$$y_{it} = \mathbf{x}_{it}\beta + u_{it} = \mathbf{x}_{it}\beta + a_i + v_{it}$$

with $i = 1, 2, .., N$ and $t = 1, 2, ..., T$.

In order to show the usefulness of the methodology, imagine the error term can be split into two components.

The u_{it} term composed by both the individual (unobserved) effect a_i (it is something we control for even if it is unknown) and an idiosyncratic component v_{it}. For example, when at the beginning of our analysis we shall replicate *Case and Shiller(2003)'s* regression without including the variable urban sprawl, a_i will reflect the impact of this variable, together with other individual effects. Land availability may indeed be fairly considered constant over time but varying depending on the city.

This abstract procedure of discerning the error term in two fragments can be arbitrarily done also in all specifications including the (P)OLS. However, in this case a problem may arise: the estimator would be biased and inconsistent if the unobserved error element a_i is correlated with at least one of the regressors, i.e. $Corr(a_i, X_i) \neq 0$. By taking first differences this problem is resolved by factoring out the unobserved fixed effect a_i such that the model becomes:

$$\Delta y_{it} = \Delta \mathbf{x}_{it}\beta + \Delta u_{it} = \Delta \mathbf{x}_{it}\beta + \Delta v_{it}$$

with $i = 1, 2, .., N$ and $t = 2, .., T$.
The FD estimator is unbiased and consistent under the following assumptions:

1. The model is a static linear model
2. $E(X_i' X_i) = A$ is a nonsingular matrix with rank K, i.e. no perfect multicollinearity
3. $E(v_{it}|X_i, a_i) = 0$, i.e. strict exogeneity of the regressors conditional on the unobserved component, which implies that $E(\Delta X_{it}' \Delta v_{it}) = 0$ and $E(\Delta v_{it}|\Delta X_{i2}, ..., \Delta X_{iT}) = 0$.

As the Pooled OLS , the FD estimator also allows the presence of serial correlation in the error term v_{it}. However, the latter is the most efficient estimator only when the previous three conditions along with

1. No serial correlation in the differenced errors Δv_{it}
2. Homoskedasticity $E(\Delta v_i' \Delta v_i|X) = \sigma^2_{\Delta v} I_{T-1}$

are satisfied. The derivation of the beta estimate is the following:

$$\hat{\beta}_{FD} = (\Delta \mathbf{X}' \Delta \mathbf{X})^{-1} \Delta \mathbf{X}' \Delta \mathbf{y} = (\sum_{i=1}^{N} \Delta \mathbf{X}_i' \Delta \mathbf{X}_i)^{-1} (\sum_{i=1}^{N} \Delta \mathbf{X}_i' \Delta \mathbf{y}_i) =$$

$$= (\sum_{i=1}^{N} \sum_{t=2}^{T} \Delta \mathbf{x}_{it}' \Delta \mathbf{x}_{it})^{-1} (\sum_{i=1}^{N} \sum_{t=2}^{T} \Delta \mathbf{x}_{it}' \Delta \mathbf{y}_{it})$$

5.1.3 Within or Fixed Effects Estimator

The fixed effects (FE) estimator represents an alternative to the first difference estimator. It consists in applying OLS to a transformed model whose data are demeaned within a

city in order to remove a part of the unobserved component. Consider the original model:

$$y_{it} = \mathbf{x}_{it}\beta + u_{it} = \mathbf{x}_{it}\beta + a_i + v_{it}$$

with $i = 1, ..., N$ and $t = 1, ...T$.
By demeaning the data, i.e. $\ddot{y}_{it} = y_{it} - \frac{1}{T}\sum_{t=1}^{T} y_{it}$, $\ddot{\mathbf{x}}_{it} = \mathbf{x}_{it} - \frac{1}{T}\sum_{t=1}^{T} \mathbf{x}_{it}$ and
$\ddot{u}_{it} = u_{it} - \frac{1}{T}\sum_{t=1}^{T} u_{it}$, the transformed model becomes:

$$\ddot{y}_{it} = \ddot{\mathbf{x}}_{it}\beta + \ddot{u}_{it} = \ddot{\mathbf{x}}_{it}\beta + \ddot{v}_{it}$$

with $i = 1, ..., N$ and $t = 1, ...T$. In matrix notation:

$$\ddot{\mathbf{y}}_i = \ddot{\mathbf{X}}_i\beta + \ddot{\mathbf{u}}_i = \ddot{\mathbf{X}}_i\beta + \ddot{\mathbf{v}}_i$$

where $\ddot{\mathbf{X}}_i$ is TxK. The fixed effect estimator is consistent and unbiased under these assumptions:

1. Linear Static model
2. No perfect multicollinearity and all regressors are time varying
3. There is strict exogeneity of the regressors across all time periods, i.e. $E[(v_{it}|X_i, a_i)] = 0$ for $t = 1, 2, ..., T$.

As the FD, also the FE estimator is unbiased and consistent when a_i and X_i are correlated. The beta for this specification can be estimated from the equation shown below:

$$\hat{\beta}_{FE} = (\ddot{\mathbf{X}}'\ddot{\mathbf{X}})^{-1}\ddot{\mathbf{X}}'\ddot{\mathbf{y}} = (\sum_{i=1}^{N} \ddot{\mathbf{X}}_i'\ddot{\mathbf{X}}_i)^{-1}(\sum_{i=1}^{N} \ddot{\mathbf{X}}_i'\ddot{\mathbf{y}}_i) = (\sum_{i=1}^{N}\sum_{i=1}^{T} \ddot{\mathbf{x}}_{it}'\ddot{\mathbf{x}}_{it})^{-1}(\sum_{i=1}^{N}\sum_{t=1}^{T} \ddot{\mathbf{x}}_{it}'\ddot{\mathbf{y}}_{it})$$

By adding two additional assumptions, i.e. homoskedatisticity and no serial correlation across time, the FE estimator is efficient. However, serial correlation in the error term is inversely proportional to T, i.e. $Corr(\ddot{v}_{it}, \ddot{v}_{is}) = -\frac{1}{T-1}$, which means that the assumption of serial correlation is not required as T gets large. However, in our analysis T is quite small thus we cannot exploit this feature.

5.1.4 Random Effects Estimator

Consider the original model with the intercept:

$$y_{it} = \beta_0 + x_{it1}\beta_1 + x_{it2}\beta_2 + ... + x_{itk}\beta_k + u_{it} = \beta_0 + x_{it1}\beta_1 + x_{it2}\beta_2 + ... + x_{itk}\beta_k + a_i + v_{it}$$

with $i = 1, 2, .., N$ and $t = 1, 2, ..., T$ and $k = 1, 2, ..., K$.
The random effect estimator is unbiased and consistent under these assumptions:

1. Linear Static model
2. No perfect multicollinearity
3. There is strict exogeneity of regressors across all time periods, i.e. $E[(v_{it}|X_i, a_i)] = 0$ for $t = 1, 2, ..., T$
4. $Cov(x_{itk}, a_i) = 0$, i.e. the unobserved component a_i must have zero conditional mean and must be uncorrelated with every explanatory variable across all i and t

In other words, what distinguishes RE from the other estimators is that a_i cannot be correlated with the regressors for all time periods and for all cities. However, even when there is correlation applying OLS is not an efficient procedure because of positive serial correlation in the error term u_{it}:

$$E(u_{it}^2) = E(a_i^2) + 2E(a_i v_{it}) + E(v_{it}^2) = \sigma_a^2 + \sigma_v^2$$

$$E(u_{it}u_{is}) = E[(a_i + v_{it})(a_i + v_{is})] = E(a^2) + E(a_i v_{it}) + E(a_i v_{is}) + E(v_{it}v_{is}) = \sigma_a^2$$

Efficiency of the estimator requires that the previous four assumptions along with both homoskedasticity, i.e. $E(v_{it}^2|X_i) = \sigma_v^2$ and no serial correlation in the error term $E(u_{it}u_{is}) = 0$ must hold. Serial correlation can be adjusted by using either GLS transformation or serial correlation robust standard errors. This is derived by defining

$$\lambda = 1 - \left[\frac{\sigma_v^2}{\sigma_v^2 + T\sigma_a^2}\right]^{\frac{1}{2}}$$

then the model is transformed such that:

$$y_{it} - \lambda\bar{y}_i = \beta_0(1 - \lambda) + \beta_1(x_{it1} - \lambda\bar{x}_{it1}) + ... + \beta_k(x_{itK} - \lambda\bar{x}_{itK}) + (v_{it} - \lambda\bar{v}_i)$$

As pointed out by *Croissant and Millo(2008)*, the feasible β_{RE} is obtained by estimating $\hat{\lambda}$ and running an OLS regression on the transformed data.

$$\hat{\lambda} = 1 - \left[\frac{\hat{\sigma}_v^2}{\hat{\sigma}_v^2 + T\hat{\sigma}_a^2}\right]^{\frac{1}{2}}$$

where:

$$\hat{\sigma}_a^2 = \frac{1}{[NT(T-1)/2] - (K+1)} \sum_{i=1}^{N} \sum_{t=1}^{T-1} \sum_{s=t+1}^{T} \hat{u}_{it}\hat{u}_{is}$$

$$\hat{\sigma}_u^2 = \frac{1}{NT - (K+1)} \sum_{i=1}^{N} \sum_{t=1}^{T} \hat{u}_{it}^2$$

$$\hat{\sigma}_v^2 = \hat{\sigma}_u^2 - \hat{\sigma}_a^2$$

The plm package on RStudio provides four different methods to estimate λ:

1. "swar" from *Swamy and Arora (1972)*
2. "walhus" from *Wallace and Hussain (1969)*
3. "amemiya" from *Amemiya (1971)*
4. "nerlove" from *Nerlove (1971)*.

According to *Baltagi (2005)*, *Wallace and Hussain(1969)* suggest using OLS residuals $\hat{u}_{OLS}$ as estimates of the true errors u_{it}. Amemiya(1971), instead, suggests using the LSDV[43] residuals rather than OLS residuals. *Swamy and Aurora (1972)* method is based on obtaining estimates of the variance component from the mean square errors by running a within and a between regression. *Nerlove (1971)* estimates $\sigma_a^2 = \sum_{i=1}^{N} \frac{(\hat{a}_i - \hat{a})^2}{N-1}$ where a_i are dummy coefficients calculated by running a LSDV regression and σ_u^2 is derived from the RSS_{within}[44] divided by NT without correction for the degrees of freedom. In our analysis we opt for Wallace and Hussain's method, which estimates the error components by using OLS residuals.

[43] Least Square Dummy Variable
[44] RSS stands for Residual Sum of Squares

When implementing the serial correlation adjustment, note that a $\lambda = 0$ implies that the model coincides with the original one, while a $\lambda = 1$ implies that RE = FE. Additionally, an important quality of the estimator is that it is consistent and asymptotically normally distributed as N gets large with fixed T. Considering that our dataset is composed by a large number of N and a small number of T, this feature might reveal to be particularly advantageous for our analysis.

5.2 Machine Learning

In this section the random forest machine learning approach is explained. It is an alternative and complementary method to understand the relationship between dependent and independent variables. The process differentiate itself from regression methods, since it does not provide estimates. The procedure we are interested in will deliver a specific result, i.e. the importance of each predictor on real house price returns.
In order to provide an exhaustive explanation of the methodology it is best to first explain decision trees and bagged procedures, which set the foundations for the random forest.

5.2.1 Decision tree

According to *Kuhn and Johnson (2016)*, decision trees help the user split the data in different samples obtaining a specific value of the variable of interest at the end of each node of the tree. Simple regression trees are not iterative procedures, i.e. a single tree should be trained from the dataset we have at our disposal. Following the tree would lead to n leave nodes with a value of housing returns.
A clear limitation of this approach is that the decision tree is a unique procedure which may not provide the best results in some occurrences, so we would have to repeat the decision tree manually many times. This manual process could be interpreted as a sort of check to understand if the results we get in the first tree are consistent or just a random occurrence.

5.2.2 Bagged Trees approach

In the '90s, as pointed out by *Kuhn and Johnson (2016)*, the concept of bagging trees which poses a fix to the poor predictive power of decision trees began to appear. Bagging is an acronym for bootstrapped aggregation and it is basically one of the first instances of ensemble techniques. This methodology applies the bootstrapping procedure to our dataset and repeats it a bunch of times to form different samples. Then each decision tree is one of an "ensemble" of trees and generates predictions for housing returns. Predicted housing returns generated from each tree are then averaged to get a final prediction about the variable of interest.
However, when using such procedure, a shortcoming arises. Although the bootstrapped aggregation method introduces some randomness into the process, bagged trees are not

completely independent because at each tree split all predictors are re-considered. As a result, it is likely that the structure will be similar among trees trained from different bootstrap samples. This characteristic is known as "tree correlation" and represents the main shortcoming of bagging trees structures. Implementing a random forest model overcomes the bootstrapping correlation, therefore we opted for this particular machine learning technique.

5.2.3 Random Forest

According to *Kuhn and Johnson (2016)*, this machine learning technique is an improvement over the bootstrapped aggregation procedure. Indeed, it incorporates some randomness during the variable selection process to correct for tree correlation which prevents bagging from reducing predicted values variance.
For reasons of conciseness, the whole process will be explained for one iteration, i.e. for one randomly selected sample. The algorithm follows precise steps.

First of all, a bootstrapped sample with replacement is generated containing as many rows as in our dataset. Observations drawn in the bootstrap are pooled together in the training dataset. Those observations which are not included in the "bag" form instead the "Out of bag sample"(OOB). Once the training sample is selected, the next step entails a random selection of k variables from the original set of independent variables. *Kuhn and Johnson (2016)* suggest that the best k is the total number of explanatory variables (p) divided by three. For instance, in our case, we consider eight variables so k should be set equal to three. Using k random variables (where k is defined as "tuning parameter") introduces the second randomness component into the procedure and represents the main difference between the bagged tree and random forest approach. Indeed, randomly drawing k out of all the parameters (eight variables in our case) rather than choosing the best informative predictor among all the eight variables allows to solve any problem of tree correlation by preventing the formation of similar structures among different trees.

At this point the first step of the process begins. The key point now is to establish which variable (among the k already randomly selected) should be used to start the tree. In

order to do so the SSE[45] measure is implemented across the randomly selected variables.

$$SSE = \sum_{i \in S_1} (y_i - \overline{y}_1)^2 + \sum_{i \in S_2} (y_i - \overline{y}_2)^2$$

Starting from one regressor, for instance the mortgage rate (it does not matter which one the learner starts from), each value is taken from the training dataset and observations (house price returns) are thus split accordingly into two subsets.

Once these two subsets are created, say S_1 and S_2, SSEs are estimated for each partition for every value of that specific variable. This procedure generates a copious amount of SSEs, one for each value of the mortgage rate, and then the one with the lowest SSE is considered for that specific variable. Once we have finished with the mortgage rate, the same process goes on for the other two randomly selected explanatory variables. When the loop is completed for all the k variables, the starting node of the tree becomes the variable with the lowest SSE (among the three regressors). At this point, the bagged sample (which is now considered as a whole) is split according to the value of the variable corresponding to the lowest SSE. Next, the two subsets just created separately undergo the same process previously explained. The only difference is that the k random variables used to determine the starting node are now selected from (p-1) explanatory variables. The algorithm looks at the factor among k with the lowest SSE and then splits the data once again (at this point the data is already split two times). The procedure is carried out for both subsets until we find a final node value for the house price returns. When the first tree is completed, i.e. there are no more variables to partition the sample, the process starts over with a new bootstrapped sample.

The aim of this long process is to obtain values for house price returns in each generated tree leaf (end node). However, for our purposes we would need a measure describing the relative importance of each variable.

In random forest settings, the percentage increase in MSE serves, indeed, as a calibration mechanism for the predictors. For each bootstrapped sample there is an OOB sample containing all the observations not included in the bagging procedure. This sample provides a control check for the accuracy of the model. According to *Nordhausen (2014)*,

[45] Sum of squared errors

for each tree generated, the first observation is taken from the OOB sample and it is run down each tree considering the set of predictors values in the first row. This leads to a different value for each terminal node of each tree. At this point, the average of these values is computed and compared to house price returns in the first row. This process is looped through all the other observations in order to obtain a response (predicted house price return) for each row n in the OOB sample. Then the overall OOB MSE can be computed as

$$MSE_0 = \frac{1}{n}\sum_{n=1}^{N}(y_{n,OOB} - \hat{y}_{n,OOB})$$

The importance of the variable $i = 1, .., k$ is related to the % increase in MSE obtained when permuting that variable in the OOB sample. This procedure generates a MSE_i such that the percentage increase in Mean Squared Error is:

$$\%incrMSE = \frac{MSE_i - MSE_0}{MSE_0}100$$

The idea is that larger levels of the $\%incrMSE$ are related to greater predictor's importance. This is due to the fact that when permuting a variable, a percentage increase in the errors points out its importance for the response variable.

At the end of the procedure, changing the tuning parameters allows to check about the stability of the model. In our case, as suggested from *Kuhn and Johnson (2016)*, the tuning parameter is set equal to three and changing its value (either to two or four) does not impact our predictions. The insensitiveness of our results to changes in the tuning parameter suggests that the algorithm is stable.

5.3 Bubble persistence, reversion and adjustment coefficients

This section of the methodology entails the last part of our analysis. As previously mentioned, a hint on how to grasp the presence of a housing bubble at city level is provided by *Capozza et al. (2002)*, who in turn elaborate on a model from *Abraham and Hendershott (1994)*. Their analysis relies on a specific equation with three different coefficients capturing housing behaviours. The equation is the following:

$$\Delta P_{i,t} = \alpha \Delta P_{i,t-1} + \beta(P^*_{i,t-1} - P_{i,t-1}) + \gamma \Delta P^*_{i,t}$$

where $\Delta P_{i,t}$ indicates actual returns (due to the log-first difference procedure) and $\Delta P_{i,t-1}$ represents its lag. $P^*_{i,t}$ is the log of predicted house prices and its corresponding lag is $P^*_{i,t-1}$. Ultimately, $\Delta P^*_{i,t}$ represents the fundamental for house price returns. In our case, $P^*_{i,t}$ is equal to the fitted values obtained from running the FE estimator on the following equation:

$$P^*_{i,t} = \alpha_0 + \beta_1 \Delta HS_{i,t} + \beta_2 \Delta POP_{i,t} + \beta_3 \Delta GDP_{i,t} + \beta_4 UN_{i,t} + \beta_5 MR_{i,t} + \beta_6 \Delta EMP_{i,t} + \beta_7 NM_{i,t} + u_{i,t}$$

where the variable urban sprawl is not included as it is constant across time thus it cancels out after the fixed effects transformation.

Needless to say, the most interesting feature of the equation shown above are the RHS coefficients. The α represents the "bubble persistence" coefficient, indicating the level of serial correlation between real returns and their lagged values. A value of α greater than one is defined by *Capozza et al. (2002)* as a sign of explosive behaviour, whereas a value lower than one as a convergence measure.

The next coefficient is β, which could be seen as a bubble "burst coefficient", i.e. capturing the likelihood of the bubble to burst. Intuitively, the larger the β, the larger the mean reversion when an overshooting occurs.

Abraham and Hendershott (1994) provide a few lines on the interpretation and relationship between α and β via the implementation of a sensitivity analysis on those coefficients. When the latter is larger than the former, it implies that potential gains, i.e. the "overpricing premium", are wiped out sooner.

On the other hand, when α is greater than β, the cycle is more persistent and lasts longer. In other words, the β shows that, if there is a considerable discrepancy between the predicted price and the actual one such that the difference is negative, then the coefficient carries a lot of weight, i.e. it is more likely that the bubble will burst. A β coefficient which is negative may indicate that the mean reversion is intense and that probably the bubble has already burst.

While trying to define some standard interpretation rules for the coefficients, it is important to point out that if β is negative together with the difference of its term (fundamental prices are lower than the actual ones), the β will have a positive effect on real house price returns, i.e. a sort of explosive behaviour which *Capozza et al (2002)* associate with a no-cycle period. To our understanding the no-cycle is due to the fact that the β turns into a α leading to a de-facto absence of the mean reversion coefficient. On the contrary, when the difference is positive, i.e. real prices are lower than fundamentals, a negative β would drive prices further down, whereas a positive coefficient would act as a serial correlation term for the LHS. Focusing on α greater than one, we should expect a more than proportional increase of housing returns than the previous year. In this case, a β between zero and one might imply a longer lasting cycle depending on the magnitude of α. When α is between zero and one, effects of a β lower than one are much stronger and the convergence is sharper.
According to what previously said, in order to define evidence of a bubble, we consider $\alpha > 1$ and $\alpha > \beta$ as fundamental conditions.

The third coefficient in the equation derived from *Capozza et al. (2002)* is γ. Specifically, it is defined as the "immediate partial adjustment", i.e. the adjustment rate in real house price returns following a shock to fundamentals. In other words, it can be interpreted as the market response to a change in the underlying asset. Indeed, a positive coefficient would indicate that if fundamentals returns increase, then real returns should be right behind tracking closely that shift. The opposite holds if γ is estimated with a negative coefficient. According to this, in a efficient market we should expect an α equal to 0 and a γ equal to 1.

6 Econometric Analysis

6.1 Preliminary results

Before presenting the results, we believe it is appropriate to spend a few words discussing major variable insights anticipated by *Case and Shiller (2003)*. The first remark concerns the relationship between dependent and independent variables. Given the absence of a clear cause-effect relationship among them, the authors infer that this simultaneity issue might affect the coefficients sign so as to make their interpretation ambiguous.
For instance, the effect of housing starts on home prices is linked to two contrasting explanations. On the one hand, in a inelastic supply environment housing starts may be limited driving up prices. On the other hand, a shock on the demand side could boost house prices incentivizing builders to build more.
Another case for ambiguity presents itself when the variation in employment is considered. The variable, a proxy for demand, may positively impact house prices. However, rising house prices are associated with higher housing costs which could act as a deterrent to attract employees in that specific area.
One last example of this simultaneity bias occurs when considering the mortgage rate. On a general level, low rate levels are seen as a stimulus to the housing market, whereas high levels depress it. However, that is not always the case. Indeed, digging into the behaviour of mortgage rates leads to interesting results. The event of low rates might also be associated in time of economic recession as a response from the Central Bank to a weak demand, likewise high interest rates may be the result of a response to a bullish market.

Now that we have explored some simultaneity issues, we can turn to the preliminary results of our analysis.

In the first four columns of table 1, we present results of the base regression analyzing the relationship between house price returns and the fundamental explanatory variables introduced by *Case and Shiller (2003)*.
We use the plm package to obtain (P)OLS, FD, FE and RE estimates, which are presented in Table 1. At a general level, the variation in employment (Δemployment rate)

has no statistical power in explaining house returns across all regressions, whereas the ΔGDP per capita is consistently statistically significant at the 1% level. Estimates are quite consistent across all four specifications but there are some interesting differences.

First of all, the FD estimator, generally speaking, generates different results when compared to the others in relation to Δhousing starts and unemployment rate.
The former shows the same positive sign across different models but it varies in significance. According to FE, (P)OLS and RE, the null hypothesis of no association between a change in housing starts and actual returns is rejected at least at the 5% level. The latter, instead, shows a coefficient which is statistically significant at 1% in all specifications (except for the FD) with a negative sign. This looks intuitive from an economic perspective: the higher the level of unemployment, the more depressed the housing demand. Moreover, the unemployment rate is often associated with foreclosures which drive down prices as well.
Secondly, although in column (1) the Δpopulation is statistically significant at the 10% level, the variable stands with no statistical power across all other specifications.
As mentioned at the beginning of the section, the presence of simultaneity could be one of the main reasons why specific variables appear to have little or no statistical power.

The analysis proceeds from columns (5) to (8) with the inclusion of net migration. This is the first variable which augments the base model of *Case and Shiller (2003)*. According to our intuition, a positive flow of net migration may be associated with an increase in housing demand. However, first evidence shows that including this variable in the model does not bring significant changes, since the variable is not statistically significant in any of our four specifications. Overall, comparing columns (1) and (5), adding net migration affects the predictive power of the population variable. Indeed, its effect is wiped out when net migration is factored in the model.

In columns (9) and (10), the inclusion of urban sprawl in the base specification brings some results worth discussing. Intuitively, we should expect that the larger the available land for housing construction in a specific city, the more elastic the supply side, the lower the level of housing returns. Indeed, when the variable is included in the model, it turns

out that its coefficient is statistically different from zero at the 5% level in both (P)OLS and RE specifications. The coefficient related to the mortgage rate, instead, increases its statistical significance if compared to the base model.
Of main interest is the effect of urban sprawl on mortgage rate predictive power between columns (1-9), (4-10) and (2). Indeed, considering the nature of the variable, its effect should be entailed in the FE specification. As expected, when the variable is included in (9) and (10), the mortgage rate coefficient becomes statistically significant in a similar way as in (2).

The last two specifications confirm what previously said and the inclusion of both net migration and urban sprawl slightly augments the magnitude of the mortgage rate coefficients.
Generally speaking, almost all specifications capture ca. one third of the house price returns variance. However, the adjusted R-square is low in FE and FD models meaning that we might miss some important variables to help explain the dependent variable.

6.2 Specification Tests

The results in Table 1 are not adjusted for serial correlation and heteroskedasticity. Specifically, the presence of these effects allows estimates to be consistent and unbiased, although standard errors are incorrect and estimators are not B.L.U.E.[46].
In order to understand which estimator best fits our dataset, we run some specification tests. These tests will shed some light on which estimator is optimal according to our dataset.

The first check we are going to perform on the basic specification is the Breusch-Pagan Lagrange Multiplier test (Table 10.1 and 10.2, Appendix) which helps shed some light on (P)OLS and RE/FE models. The null hypothesis is

$$H_0 : \sigma_a = 0$$

implying the absence of serial correlation in the error term, where σ_a is the standard deviation of the unobserved error component. If the null hypothesis holds, the RE model

[46] Best Linear Unbiased Estimator

Table 1: Panel regressions, incorrect standard errors

	Dependent variable: Real house price returns											
	(P)OLS	FE	FD	RE	(P)OLS	FE	FD	RE	(P)OLS	RE	(P)OLS	RE
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
%Δ*Housing starts*	0.0223***	0.0197**	0.0136*	0.0219***	0.0232***	0.0198**	0.0136*	0.0227***	0.0203***	0.0202***	0.0213***	0.0212***
	(0.0076)	(0.0077)	(0.0072)	(0.0075)	(0.0076)	(0.0077)	(0.0072)	(0.0076)	(0.0076)	(0.0076)	(0.0076)	(0.0076)
%Δ*Population*	0.5499*	−0.0984	−0.0247	0.4857	0.5062	−0.1276	−0.0409	0.4363	0.4296	0.4235	0.3553	0.3470
	(0.3167)	(0.3765)	(0.3888)	(0.3206)	(0.3190)	(0.3813)	(0.3892)	(0.3232)	(0.3196)	(0.3200)	(0.3226)	(0.3233)
%Δ*GDP per capita*	0.4699***	0.4283***	0.3488***	0.4671***	0.4626***	0.4259***	0.3433***	0.4596***	0.4739***	0.4736***	0.4643***	0.4638***
	(0.0621)	(0.0627)	(0.0725)	(0.0617)	(0.0624)	(0.0629)	(0.0727)	(0.0620)	(0.0618)	(0.0618)	(0.0621)	(0.0620)
Unemployment rate	−0.0044***	−0.0071***	−0.0027	−0.0045***	−0.0044***	−0.0069***	−0.0018	−0.0045***	−0.0040***	−0.0040***	−0.0039***	−0.0039***
	(0.0007)	(0.0012)	(0.0032)	(0.0007)	(0.0007)	(0.0013)	(0.0034)	(0.0007)	(0.0007)	(0.0007)	(0.0007)	(0.0007)
Mortgage rate	−0.0050	−0.0106*	−0.0170	−0.0051	−0.0055	−0.0111*	−0.0188	−0.0056	−0.0070*	−0.0070*	−0.0080*	−0.0079*
	(0.0041)	(0.0059)	(0.0118)	(0.0042)	(0.0041)	(0.0060)	(0.0120)	(0.0042)	(0.0041)	(0.0041)	(0.0042)	(0.0042)
%Δ*Employment rate*	0.1021	0.1144	−0.0585	0.1036	0.0904	0.1121	−0.0616	0.0926	0.1103	0.1104	0.0951	0.0953
	(0.1219)	(0.1313)	(0.1431)	(0.1222)	(0.1223)	(0.1315)	(0.1431)	(0.1226)	(0.1214)	(0.1214)	(0.1216)	(0.1217)
Net migration					0.0003	0.0002	0.0006	0.0003			0.0004	0.0004
					(0.0002)	(0.0003)	(0.0006)	(0.0002)			(0.0002)	(0.0002)
Urban Sprawl									−0.0004**	−0.0004**	−0.0004**	−0.0004**
									(0.0002)	(0.0002)	(0.0002)	(0.0002)
Constant	0.0622***			0.0639***	0.0633***			0.0651***	0.0731***	0.0731***	0.0758***	0.0759***
	(0.0220)			(0.0227)	(0.0220)			(0.0228)	(0.0224)	(0.0225)	(0.0225)	(0.0226)
Observations	434	434	386	434	434	434	386	434	434	434	434	434
R^2	0.3084	0.2655	0.0905	0.3024	0.3105	0.2660	0.0931	0.3041	0.3168	0.3158	0.3206	0.3193
Adjusted R^2	0.2987	0.1630	0.0785	0.2926	0.2991	0.1614	0.0788	0.2926	0.3055	0.3046	0.3078	0.3065
F Statistic	31.7400***	22.8900***	7.3900***	30.8600***	27.4000***	19.6200***	6.3260***	26.5900***	28.2100***	28.0900***	25.0700***	24.9200***

Note: *p<0.1; **p<0.05; ***p<0.01

reduces to the (P)OLS with no serial correlation in the error term[47]. However, when the null hypothesis is rejected, the test does not give any information about the reliability of other estimators.

In practice, the test gives the possibility to specifically check for individual, time and joint effects. Accordingly, there is no evidence of individual effects, while there are significant time effects (H_0 is not rejected). Testing for both effects together leads to the rejection of the null and the addition of both net migration and urban sprawl to the base model does not change test results.

Considering the "twoways" specification, the null is rejected, therefore the test shows the presence of some panel effects, suggesting that the (P)OLS is an inappropriate estimator for our analysis. For this reason, we move along our analysis by investigating whether RE, FE or FD may be suitable estimators.

According to *Wooldridge (2010)*, performing the Hausman Test (Table 11, Appendix) should help choosing one between RE and FE estimators. The within estimator is consistent if a_i and the explanatory variables are correlated, whereas the RE is not. Under the null, RE and FE estimators are consistent; under the alternative, only the FE estimator is consistent. In this case, the null hypothesis is rejected, so the test suggests that the FE is the appropriate estimator. However, an exception is found when urban sprawl is added to the original model.

Intuitively, urban sprawl is originally included in the a_i term. The correlation between a_i and the regressors decreases after controlling for this additional variable, as it captures the part of the "fixed effect" included in the a_i component, thus the test indicates the RE estimator as the most efficient one. Adding together net migration and urban sprawl, instead, suggests that the best estimator appears still to be the FE one. This last finding has not a clear explanation. The reasons why that is the case might be found in an unknown relationship between net migration and the unobserved error component.

The next step involves a specification test between the FE and (P)OLS estimator. This is done via an F-test where the alternative hypothesis implies the presence of significant effects. In this setting, both individual and time effects can be tested. The individual

[47] Section 5.1.4

effects are weakly significant at 5% and at 10% when augmenting the model with urban sprawl. The intuition behind it is the following: the inclusion of urban sprawl in the analysis, given that the variable itself is a "fixed effect", decreases the significance of individual effects. Additionally, there is strong evidence confirming the presence of time effects. Considering both individual and time effects to our FE and (P)OLS models, the p-value allows us to reject the null hypothesis and conclude that the FE model is the best fit for our data. Detailed results of tests concerning individual, time and both effects are reported in table 12 and 13 (Appendix).

Up to this point, the FE estimator appears to be the more suitable for our data. However, there is a consideration which needs to be made. According to *Croissant and Millo (2008)*, the tests implemented so far are quite general, i.e. for any panel data. Since our dataset presents short time series, it is likely that results might be biased toward rejection. In the literature, there are tests which are specifically optimized for short panels.
One of these is Wooldridge's test for serial correlation in "short" FE models (Table 14, Appendix) defining the null hypothesis on a regression of FE residuals on their one period lag[48]:

$$e_{i,t} = a + \delta e_{i,t-1} + n_{i,t}$$

where the rejection of the null

$$\delta = -\frac{1}{T-1}$$

implies serial correlation. On the contrary, if the null is not rejected, it means that as T increases δ approaches zero eliminating serial correlation. In our case the null holds, therefore there is no serial correlation after FE.
At this point, it seems appropriate a comparison between the FE and FD, given that applying both estimators cancels out fixed effects. Wooldridge's first-difference-based test (Table 15.1 and 15.2, Appendix) proposes a serial correlation test that can also be seen as a specification test comparing FE and FD estimation procedures[49]. On the one hand, if the idiosyncratic errors of the original model u_{it} are uncorrelated, then the errors

[48] *Croissant, Millo (2008)*
[49] *Croissant, Millo (2008)*

of the FD model

$$\Delta u_t = u_{it} - u_{i,t-1}$$

are correlated with correlation

$$cor(\Delta u_{it}, \Delta u_{i,t-1}) = -0.5$$

on the other hand, any time invariant effect will be eliminated due to first differencing. A serial correlation test for models with individual effects of any kind can be based on the following regression:

$$\hat{u}_{i,t} = \delta \hat{u}_{i,t-1} + n_{i,t}$$

and testing the restriction $\delta = -0.5$ corresponding to the null of serial correlation.
In this case, if the null hypothesis cannot be rejected we would conclude that the FE is the more appropriate estimator. On the other hand, if the differenced errors Δu_{it} (not the idiosyncratic ones) are uncorrelated, then u_{it} is a random walk and the most appropriate estimator would be the FD one. Specifying $H_0 =$ "FE" implies no serial correlation in original errors, meaning that due to the FE procedure we removed time-invariant effects (which are by definition correlated), i.e. there is no other sign of serial correlation.
The default hypothesis when conducting this test is "$H_0 = FD$", which tests for serial correlation in FD errors Δu_t. Given that the version "$H_0 = FE$" of the null holds, we would expect the default null hypothesis to be rejected.
In the first specification, we cannot reject the null hypothesis, confirming the results of Wooldridge's test for short FE panels, whereas in the default, as expected, we find evidence of serial correlation in the differenced errors, so we are inclined to prefer the FE estimator.

6.3 Robust Covariance Matrix Estimators

Now we try to replicate again *Case and Shiller (2003)*'s work by resorting to a different approach. It involves applying the function *coeftest* (from the package "sandwich" in R-studio) in order to obtain a specific covariance matrix (*vcov=HC1*). This method gives the user the possibility to cluster standard errors by either group or time. Here, the former option is specified to control for serial correlation and heteroskedasticity, which may arise when dealing with panel data (Table 2).

By comparing the results from Table 1 and 2, it is possible to see that the clustering procedure does not change coefficients magnitude (except for the RE). Indeed, as explained in section 5.1, the conditions of no serial correlation and homoskedasticity are not required to obtain unbiased and consistent estimates when dealing with panel data. However, when these conditions are violated, standard errors are incorrect, estimators are inefficient and inference is wrong. Following the clustering procedure, housing starts standard errors are larger whereas the ones of the population and mortgage rate become smaller. As a consequence, the coefficient of Δhousing starts is less significant than before, whereas the other two variables show a stronger association with the dependent variable.

Observing the first four columns of Table 1 and 2 it is possible to identify some major changes: in (P)OLS and RE housing starts remain statistically significant at 5% rather than 1% as before. According to FD, the variable is not statistically significant anymore. In Table 2, (P)OLS and RE estimates show a population coefficient statistically significant at 5% in (1), (4), (5), (8), (9) and (10) whereas FE and FD draw attention to the mortgage rate, whose null hypothesis of no association with housing returns is rejected as 5%.
When including net migration in columns (5) and (8), the variable mortgage rate is statistically significant also for (P)OLS and RE (at the 10% level). In (6) and (7), instead, the significance level remains at 5% as in the previous specifications (2) and (3).
Considering urban sprawl, in columns (9) and (10) it turns out that the coefficient, significant at 5% in Table 1, loses statistical power (10%). Ultimately, in (11) and (12) the

variable net migration, which is never significantly associated with housing returns in Table 1, shows a coefficient weakly significant at 10%.

At this point, we proceed with an analysis across various model specifications in Table 2. Supporting evidence found in *Case and Shiller (2003)*, ΔGDP per capita appears to be always significant at the 1% level. According to (P)OLS, FE and RE, Δhousing starts and unemployment rate are always statistically significant. Δpopulation, instead, is only significant in (P)OLS and RE estimates, whereas we do not find any evidence of a statistical association between the change in employment rate and house price returns across all the specifications. Of particular interest is the mortgage rate behaviour: the variable is statistically significant in (2) and (3) but not in (1) and (4). When net migration and urban sprawl are added, the regressor becomes significant at 10% confirming that specification (1) and (4) suffer from omitted variable bias.

Adding net migration in specification (5) to (8) does not provide any additional information in the model, since its coefficient is never significant, whereas the null hypothesis that urban sprawl is not associated with real house price returns is rejected at 10%. However, when both variables are included in the model, the former becomes statistically significant at 10% and the latter at 5%.

Looking at coefficients, it is worth mentioning that the RE estimates in columns (10) and (12) are equivalent to those of the (P)OLS in (9) and (11). This fascinating result is due to the λ term being close to zero, which might imply the lack of serial correlation in the error term, specifically $\sigma_a^2 = 0$ (as explained in section 5).
Overall, in almost all specifications, a 1% increase in GDP per capita is associated with an increase in house price returns between 0.3 - 0.4%. This result is intuitive from an economic point of view, since an increase in household income should imply an increase in housing demand. A 10% increase in housing starts is associated with an increase in house price returns of 0.2%. This result suggests that the supply effect takes over the demand effect, i.e. a rise in house returns entails profit opportunities, therefore builders respond by building more. A one unit increase in the unemployment rate is associated with a decrease in returns of approximately 0.4%. When found to be statistically significant, a

mortgage rate decrease (1%) is associated with a 0.5-1.5% increase in returns.

To sum up, we have compared different estimators and found similar estimates for most of the variables included in the model. However, given that differences emerge in relation to some variables, we think that the FE estimator shows the most reliable results.
The reasons why this estimator appears the most appropriate for our analysis are twofolds. First of all, using the FE specification allows to control for those fixed effects not included in the regression. Second, the specification tests implemented in section 6.2 suggest that the FE is the best method to fit the data.

6.4 Comparison with Case and Shiller's results

Given that the FE estimator is the preferred method and that our analysis will depart from the one of *Case and Shiller (2003)*, it seems appropriate to provide few lines of comparison between theirs and our results.
Keeping in mind what has been said at the beginning, i.e. Case and Shiller focus on the most volatile areas in terms of house prices, our analysis will incorporate city effects in a different way. We believe comparing Table 2 and *Case and Shiller (2003)'s* results could give some insights on the main house price predictors differences between the U.S.A. and Europe.

Considering the order of variables following *Case and Shiller (2003)'s* Table 3 (first quadrant), the change in employment appears mainly not statistically significant in six out of the eight most volatile states. When the estimates are significant, the sign is negative. Also in this case, contrary to the general understanding, the effect of employment on house prices is not intuitive, as we might be dealing with the concept of simultaneity. For instance, on the one hand a positive coefficient on Δemployment would imply a boost in demand followed by an increase in house prices. On the other hand, lower prices could attract more demand hence a positive change in employment.

In our Table 2, it is remarkable that according to all panel data estimators the change in employment has no effect on the dependent variable, which supports evidence found by *Case and Shiller (2003)*. The next variable of interest is the change in population which,

Table 2: Panel regressions, clustered standard errors

	Dependent variable: Real house price returns											
	(P)OLS	FE	FD	RE	(P)OLS	FE	FD	RE	(P)OLS	RE	(P)OLS	RE
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
%Δ*Housing starts*	0.0223**	0.0197**	0.0136	0.0223**	0.0232***	0.0198**	0.0136	0.0232***	0.0203**	0.0203**	0.0213**	0.0213**
	(0.0087)	(0.0085)	(0.0085)	(0.0087)	(0.0089)	(0.0085)	(0.0084)	(0.0089)	(0.0084)	(0.0084)	(0.0086)	(0.0086)
%Δ*Population*	0.5499**	−0.0984	−0.0247	0.5499**	0.5062**	−0.1276	−0.0409	0.5062**	0.4296**	0.4296**	0.3553*	0.3553*
	(0.2247)	(0.2584)	(0.2779)	(0.2247)	(0.2273)	(0.2658)	(0.2734)	(0.2273)	(0.2053)	(0.2053)	(0.2090)	(0.2090)
%Δ*GDP per capita*	0.4699***	0.4283***	0.3488***	0.4699***	0.4626***	0.4259***	0.3433***	0.4626***	0.4739***	0.4739***	0.4643***	0.4643***
	(0.0433)	(0.0478)	(0.0446)	(0.0433)	(0.0435)	(0.0480)	(0.0444)	(0.0435)	(0.0437)	(0.0437)	(0.0439)	(0.0439)
Unemployment rate	−0.0044***	−0.0071***	−0.0027	−0.0044***	−0.0044***	−0.0069***	−0.0018	−0.0044***	−0.0040***	−0.0040***	−0.0039***	−0.0039***
	(0.0006)	(0.0009)	(0.0027)	(0.0006)	(0.0006)	(0.0010)	(0.0027)	(0.0006)	(0.0007)	(0.0007)	(0.0007)	(0.0007)
Mortgage rate	−0.0050	−0.0106**	−0.0170**	−0.0050	−0.0055*	−0.0111**	−0.0188**	−0.0055*	−0.0070*	−0.0070*	−0.0080*	−0.0080*
	(0.0032)	(0.0048)	(0.0074)	(0.0032)	(0.0033)	(0.0049)	(0.0080)	(0.0033)	(0.0041)	(0.0041)	(0.0042)	(0.0042)
%Δ*Employment rate*	0.1021	0.1144	−0.0585	0.1021	0.0904	0.1121	−0.0616	0.0904	0.1103	0.1103	0.0951	0.0951
	(0.1510)	(0.1657)	(0.1886)	(0.1510)	(0.1487)	(0.1647)	(0.1865)	(0.1487)	(0.1505)	(0.1505)	(0.1470)	(0.1470)
Net migration					0.0003	0.0002	0.0006	0.0003			0.0004*	0.0004*
					(0.0002)	(0.0002)	(0.0006)	(0.0002)			(0.0002)	(0.0002)
Urban Sprawl									−0.0004*	−0.0004*	−0.0004**	−0.0004**
									(0.0002)	(0.0002)	(0.0002)	(0.0002)
Constant	0.0622***			0.0622***	0.0633***			0.0633***	0.0731***	0.0731***	0.0758***	0.0758***
	(0.0182)			(0.0182)	(0.0181)			(0.0181)	(0.0231)	(0.0231)	(0.0235)	(0.0235)

Note: *p<0.1; **p<0.05; ***p<0.01

when significant, enters Case and Shiller's model with a positive sign in California and Rhode Island and with a negative one in Massachusetts. In our case, the relationship between house price returns and changes in population is positive when significant and consistent with some of the findings of *Case and Shiller (2003)*.

The third regressor is the mortgage rate which surprisingly is never statistically significant in the analysis provided by *Case and Shiller (2003)*. On the contrary, our evidence suggests a negative relationship between the mortgage rate and housing returns. This difference could be attributed to numerous factors. Mainly, the data considered refer to different time periods. In our case, the financial crisis falls in the middle of our time span, which may explain why the ECB response to contrast the recession has had a significant impact on the housing market. At first glance, during the crisis it would make sense to find a positive relationship between the mortgage rates and the housing market[50]. The point is that after the economic downturn ended, the ECB persistency on keeping the mortgage rates close to zero boosted the housing demand from 2013 onwards[51] which could explain the negative relationship found in our data.

Additionally, since some European countries in our dataset are not in the Eurozone system, the presence of different central banks defining interest rates implies higher cross-sectional variation in mortgage rates than the American ones. In the U.S.A. market the 30-year mortgage rate has experienced an uneven declining trend from 1985 to 2002[52]. Although the period considered by Case and Shiller has been subjected to a financial downturn in early 1990s, the recession has not been as strong in magnitude as the one in 2008. However, *Welsh(1993)* shows that restrictive monetary policy of the Federal Reserve represented the main factor causing the economic downturn in early 1990s.

Moving to the unemployment rate, *Case and Shiller (2003)* find ambiguous results. Table 2 in our analysis shows that the variable is negatively associated with housing returns.

Looking at GDP per capita, as anticipated before, *Case and Shiller (2003)* find a positive relationship across most of the eight states. Estimates in table 2 confirm *Case and Shiller (2003)'s* results indicating that the variable has a strong positive association with housing returns.

The last variable of interest is Δhousing starts: our result are consistent with *Case and*

[50]Low rate may be caused by the response of the central bank to weak economic conditions

[51]The only exception to this behaviour is found in the Spanish market

[52]For more information, visit: https://fred.stlouisfed.org/series/MORTGAGE30US

Shiller (2003) suggesting a positive relationship with the respective dependent variable.

6.5 Are there any time, national or local effects?

From previous tests, the FE estimator appears to be the most appropriate method to analyze our dataset. The method, indeed, cancels out all those effects which are constant throughout time, even if they are not specifically included in the model. For this reason, we think that investigating on the nature of these effects could provide significant insights into understanding the movement of housing returns. In order to know more about national and local trends, we consider the (P)OLS methodology which allows to include in the model those effects.

In Table 3, the idea is to capture these relationships by considering time, national or city-level effects. In column (1), the original (P)OLS specification is presented. Columns (2), (3) and (4) consider time, time and country and time and city effects, respectively. Column (5), our last model, considers the interaction term between time and country effects. Table 3 does not present city effects and the interaction term coefficients because of the lack of sufficient space to fit them in a whole page. However, we provide the complete table with each coefficient in Tables 16.1-16.6 (Appendix).

Starting from the variable Δhousing starts, evidence shows that its coefficient does not change across columns (1) to (4) remaining statistically significant at the 5% level. In other words, when controlling for time, national and local effects, the variable maintains the predictive power of the original model, i.e. nothing is captured by these effects.

Talking about the change in population, we can reject the null hypothesis of no association between the independent and the dependent variable only at 10%, as shown in both specification (1) and (2), implying that time effects do not provide any additional information. However, when also country and city effects are included in the model, the variable does not show any level of significance, a sign that most of the variation is absorbed by those effects.

ΔGDP per capita and unemployment rate show a substantial degree of consistency. Despite the addition of the fixed effects, estimates change little in magnitude and remain

statistically significant at the 1% level.

The mortgage rate, instead, is the variable which is affected the most by the addition of these effects. In (1) the coefficient is statistically significant at 10% with a negative sign. Intuitively, lower mortgage rates are a tool to stimulate the housing demand, whereas higher ones are a de-facto deterrent for prospective homebuyers. However, this mechanism may not be instantaneous, meaning that there is some lagged response from central banks which tend to raise rates[53] when the boom may already be at an advanced stage[54].

When considering time effects, the predictive power of the mortgage rate disappears, whereas in (3), when country effects are added in the model, it becomes statistically significant at the 5% level with a positive sign. The inclusion of the country dummy forces the (P)OLS to consider the data not as whole as in (1) but to consider the observations within a country.
It turns out that at the National level the relationship between the mortgage rate and housing returns is positive. With respect to this surprising association, a thorough explanation is provided in section 6.6 together with a graphical illustration (Figure 8). In this occasion, it is possible to appreciate the other side of the mortgage rate, which as a coin entails two facades. In booming times, the authorities should respond with increasing interest rate to control market expansions. During recessions, instead, the lowering of the rate should act as a boost to the market weak economic conditions. These manoeuvres are consistent with a short-term relationship between house prices and mortgage rates at country level.

When city effects are introduced, the (P)OLS work in a similar way but considering the relationship between the mortgage rate and housing returns within a city. The relationship is not statistically significant indicating that the presence of local effects wipes out any association existing at national level between the explanatory and the response variable.

[53] In this case the increase/decrease of rates refers to the interest rate set by a Central Bank. The mortgage rates, however, are linked to interest rates, so a shock to the latter entails some cause-effect relationship to the former

[54] This behaviour of central banks is usually defined as “dovish” as opposed to “hawkish”

Confirming the importance of local effects, net migration remains statistically significant at the 10% level in (1), (2), (3) but loses predictive power in column (4). The urban sprawl, instead, is significant at the 5% level in (1) but its estimated coefficient is less significant in specification (2), underlying the fact that time effects reduce its statistical importance for the model.

By comparing Δhousing starts, ΔGDP per capita and unemployment rate in (2) and (3), it turns out that national effects supersede local ones, i.e. shocking one of these variables has the same impact at national and city level on our dependent variable. The opposite holds in column (4) for mortgage rate and net migration. Given that the mortgage rate is set at national level, it appears reasonable to believe that shocking the variable has certain effects on housing returns when country effects are considered. However, when looking at local level (given that the value of the variable is the same across cities of the same country), there are local factors which may alter that relationship nullifying its predictive power. This is a possible explanation why we find the variable not significant in (4).

The last column includes the interaction term which captures most of the variation in fundamental variables. Indeed, the only ones remaining significant are unemployment and mortgage rate at the 10% and 5%, respectively. Of interest is the latter which, once again, enters with a positive coefficient.

At this point we believe it makes sense to interpret time effects, which are presented in Table 3 along with country effects. Germany is set as the reference country, whereas 2003 is the base year.

The idea is that dummies coefficients will show the change of house price returns in a particular year relatively to the base year, no matter what benchmark year is taken into consideration.
First of all, it is remarkable that the 2005 coefficient is statistically significant at the 1%

level indicating that house price returns across our sample of European countries in that specific year are on average approximately 9.5% larger than in 2003. The same reasoning points out that in 2008 real house price returns are about 4.79% lower than in 2003. The former is a sign of the housing boom preceding the collapse happened two years later. The latter is a direct consequence of the financial recession which in those years crushed global financial markets. These time effects just considered remain consistent also when country and city dummies are added in specification (3) and (4).

Considering specification (3), specifically Belgium, France and Scandinavia dummies show a positive coefficient which indicates that in those areas housing returns are higher on average by 5.8%, 6.4% and 2.8% than the reference country (Germany in this specific case). The time effects are significant in 2005 (10%), 2008 (1%), 2011 (10%) and 2015 (10%). Considering 2005 for instance, it means that overall house price returns are 8.4% higher on average in 2005 than in 2003.

Table 3: Panel regression with time, country and city effects

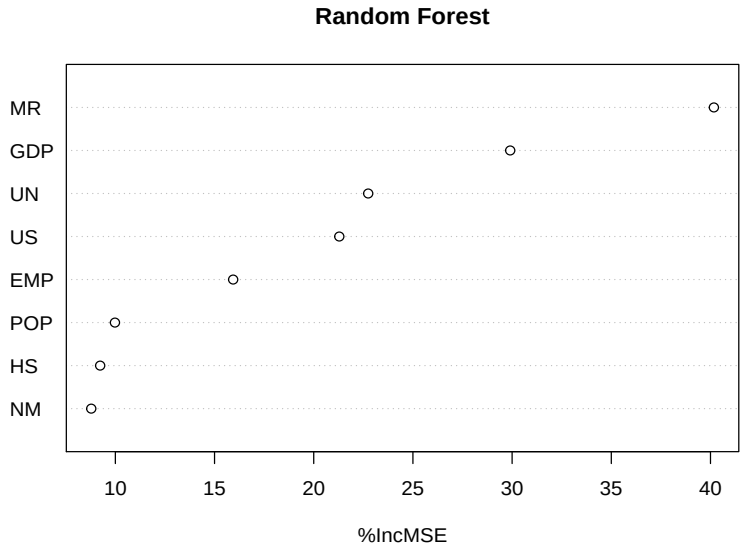
	Dependent variable: Real house price returns				
	Original	Time	Time + Country	Time + City	Time x Country
	(1)	(2)	(3)	(4)	(5)
%Δ*Housing starts*	0.0213**	0.0193**	0.0183**	0.0185**	0.0025
	(0.0086)	(0.0078)	(0.0075)	(0.0079)	(0.0062)
%Δ*Population*	0.3553*	0.4100*	0.0698	−0.1024	0.3147
	(0.2090)	(0.2320)	(0.2519)	(0.2722)	(0.2678)
%Δ*GDP per capita*	0.4643***	0.3879***	0.3080***	0.2982***	−0.0301
	(0.0439)	(0.0609)	(0.0655)	(0.0701)	(0.1130)
Unemployment rate	−0.0039***	−0.0037***	−0.0038***	−0.0076***	−0.0018**
	(0.0007)	(0.0007)	(0.0006)	(0.0012)	(0.0008)
Mortgage rate	−0.0080*	−0.0068	0.0099**	−0.0015	0.0107*
	(0.0042)	(0.0053)	(0.0047)	(0.0079)	(0.0056)
%Δ*Employment rate*	0.0951	0.1883	0.2189	0.2145	0.0138
	(0.1470)	(0.1487)	(0.1734)	(0.1797)	(0.0846)
Net migration	0.0004*	0.0005*	0.0004*	0.0001	0.0001
	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0004)
Urban Sprawl	−0.0004**	−0.0004*	−0.0001	0.0001	−0.0001
	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)
Belgium			0.0586***		0.0239
			(0.0081)		(0.0211)
England			−0.0085		0.0410**
			(0.0083)		(0.0193)
France			0.0649***		−0.0506***
			(0.0072)		(0.0184)
Netherlands			0.0019		−0.0132
			(0.0076)		(0.0155)
Scandinavia			0.0278***		0.0253
			(0.0067)		(0.0232)
Spain			0.0184		0.0310
			(0.0116)		(0.0274)
2004		0.0171	−0.0058	−0.0027	0.0099
		(0.0168)	(0.0173)	(0.0181)	(0.0211)
2005		0.0959**	0.0837**	0.0889**	−0.0281
		(0.0425)	(0.0368)	(0.0391)	(0.0201)
2006		0.0183	0.0165	0.0127	−0.0602***
		(0.0131)	(0.0152)	(0.0174)	(0.0200)
2007		−0.0031	−0.0099	−0.0132	−0.0584***
		(0.0113)	(0.0133)	(0.0157)	(0.0225)
2008		−0.0479***	−0.0609***	−0.0630***	−0.0356
		(0.0127)	(0.0146)	(0.0165)	(0.0220)
2009		0.0078	0.0028	0.0005	−0.0231
		(0.0164)	(0.0185)	(0.0211)	(0.0305)
2010		−0.0029	0.0019	−0.0012	0.0280
		(0.0137)	(0.0168)	(0.0189)	(0.0285)
2011		−0.0274**	−0.0243*	−0.0267	−0.0183
		(0.0117)	(0.0146)	(0.0175)	(0.0162)
2012		−0.0114	−0.0044	−0.0110	0.0544**
		(0.0140)	(0.0174)	(0.0200)	(0.0254)
2013		0.0157	0.0260	0.0187	0.0308
		(0.0130)	(0.0161)	(0.0201)	(0.0226)
2014		−0.0113	0.0066	−0.0036	−0.0082
		(0.0138)	(0.0179)	(0.0234)	(0.0223)
2015		0.0006	0.0343*	0.0185	0.0161
		(0.0145)	(0.0188)	(0.0239)	(0.0270)
Constant	0.0758***	0.0711**	−0.0181	0.0468	−0.0208
	(0.0235)	(0.0312)	(0.0303)	(0.0483)	(0.0321)

Note: *p<0.1; **p<0.05; ***p<0.01

6.6 Random Forest results

Figure 7 shows the results following the implementation of the random forest machine learning approach. The percentage increase in MSE, as anticipated in section 5, is a tool which helps the user understand the importance of each regressor. In relation to this, the random forest model does not give any clue about the association between a predictor and the dependent variable but it gives some insights about which variables are important in the model. The logic is the following: if shuffling values within a variable (this procedure is referred to as "permutation") does not change the informative power explained by the variable itself in the model, then the variable is not important as few changes will occur to MSE.

Figure 7



Alternatively, if the variable has a strong impact on the model, then permuting it will cause the MSE to change significantly. In our case, as it can be seen in Figure 7, the mortgage rate represents the most important predictor of house price returns. Indeed, a change in the variable is associated with an increase in the MSE of roughly 40%. In terms of importance, the change in GDP per capita represents the second predictor according to the percentage increase in MSE. In our case, permuting this variable entails a 30 % increase in MSE, followed by the unemployment rate and urban sprawl, which show an increase in MSE of around 20-25%. The change in employment is also quite

important (15% increase in MSE), whereas the change in population, housing starts and net migration are the least in terms of importance. Indeed, changing their values leads to approximately a 10% increase in MSE.

On the basis of our results, both the change in GDP per capita and the mortgage rate deserve particular consideration. According to *Case and Shiller (2003)*, GDP per capita represents the most important variable in explaining changes in house price returns. Indeed, by looking at their R-squared they conclude that in most of the American states considered GDP explains most of the variance of the dependent variable, whereas other explanatory variables add little information in the model. Our approach goes more in depth into understanding the importance of each variable. Indeed, the implementation of the random forest procedure allows to grasp the relative weight of every variable in the model. Results are consistent with *Case and Shiller (2003)*, confirming that GDP per capita represents one of the main predictors of real house price returns.

With respect to the mortgage rate our results significantly differ from those obtained by *Case and Shiller(2003)*. These results emerging from the machine learning procedure increased our curiosity and brought us to investigate the pattern between the mortgage rate and house price returns. Figure 8 indeed sheds some light on this relationship via four different plots considering time and country trends, effects and the original (P)OLS model. Starting on the up-left, Plot 8.A shows country patterns for housing returns and mortgage rates. The fitted lines across different area points indicates a strong correlation in Spain, Scandinavia, France and Belgium. Concerning the Netherlands, observations indicate a flat relationship, whereas England is the only nation showing a strong negative association.

Shifting the focus on Plot 8.B, the cross-sectional time trend are presented. We suggest to compare plot 8.B with the real price indices displayed in section 4 for the relevant countries.
In 2005, the relationship between mortgage rate and house price returns is displayed as negative. Now, the interpretation for these results may be attributed to Belgium observations, which show that for mortgage rate values around 5% the returns are enormous (as

shown in Figure 6). At the same time, Spain, Germany, the Netherlands and England do not show such huge returns during the same period (see Figures 1- 6). Furthermore, 2005 is considered a boom year, thus it could make sense that the relationship is downward sloping. This hinges on the assumption that in the long-run a prolonged period of low mortgage rates may distort the housing market driving up house prices above their fundamentals, so during the boom it could be plausible to find this sort of relationship. In 2006, the distribution of values shows a less accentuated negative coefficient even though the mortgage rate value is similar to the 2005 level indicating the end of the boom. In 2007, the trend follows closely the one of 2006 but there are some country observations such as the Netherlands, Scandinavia and England dragging down the fitted line, i.e. the 2007 line suffers a bit from downward bias.

2008 as it is well-acknowledged marks a splitting line, i.e. the housing returns are concentrated around zero with large mortgage rates. This may be a sign of the lagged response of the central bank, mentioned before, to deter potential homebuyers but it is hard to say with certainty. Many observations, especially in England, Spain and the Netherlands are very close to the right-edge of the plot with mostly negative returns, a clear signal of the burst of the housing market.

In 2009 the situation reverses, indeed observations are more concentrated just above the zero-return line with some outliers with very low rates and negative returns (Spain, Germany and the Netherlands) and other with slightly positive returns and higher mortgage rates (England). The fitted line, however, is upward sloping indicating a positive relationship between x and y witnessing a phase where economies are still dealing with the consequences of the crisis.

The year 2010 represents another benchmark year, where the relationship is predicted by a line which is flat on the zero level price return. Mortgage rates, however, vary considerably from a low of 2% in Scandinavia to almost 6% in England which might imply different measures taken from central banks to exit the crisis. Other nations' rates are clustered between the 4-5 % and are all indirectly controlled by the ECB. In relation to housing returns, the largest returns are achieved by Scandinavia and Germany in 2010,

close to 20%. The fitted line for 2011 is slightly upward sloping with mortgage rates on average slightly higher than the previous year but there is a lot of cross sectional variation which could bias the line. Observations from Spain show very low rates and negative returns of about 20%. France and Belgium show higher mortgage payments and returns whereas returns in Germany are close to zero with rates around 4.5%. Last but not the least, England shows the highest mortgage rates and negative returns.

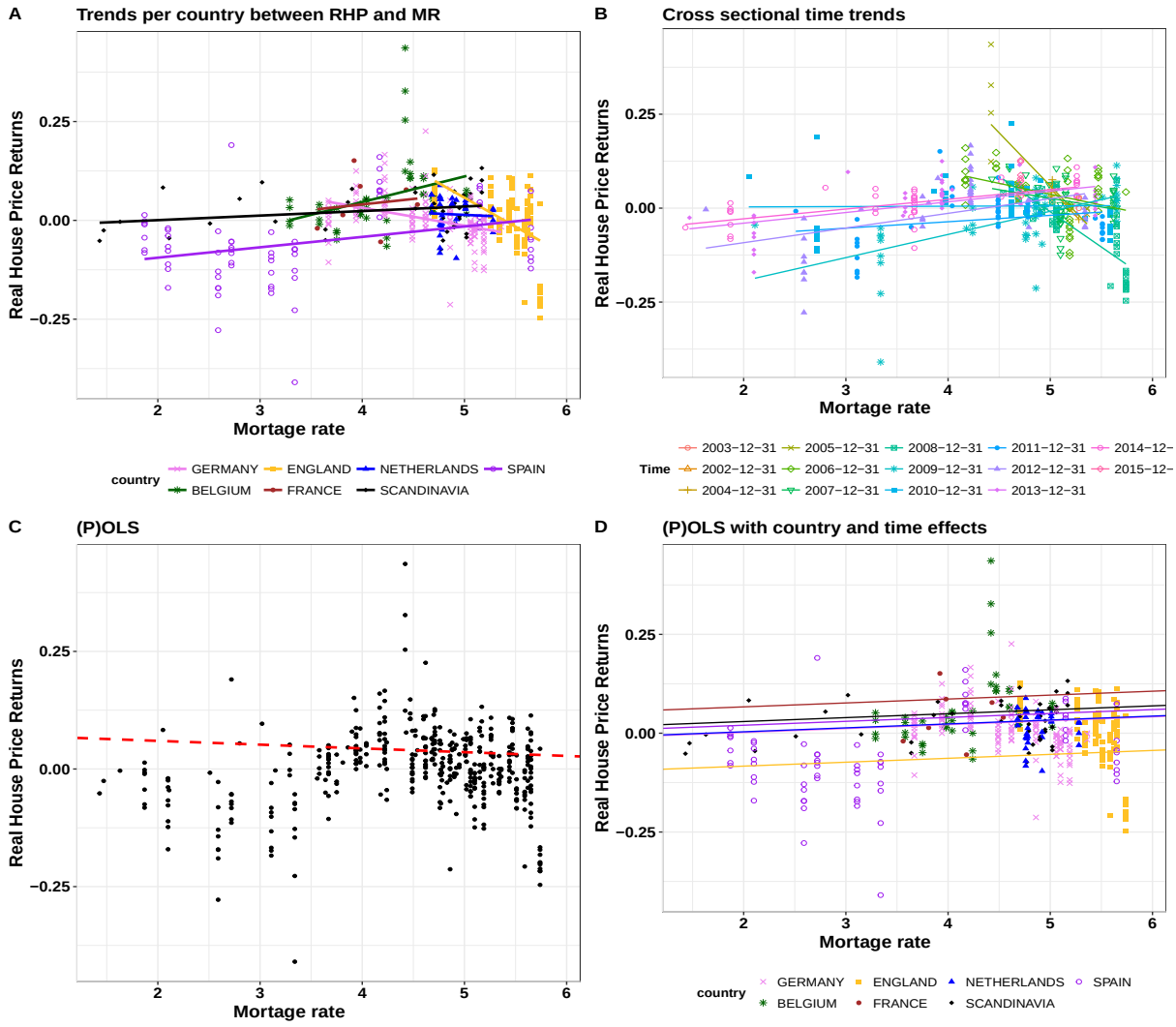
Focusing on 2012, there is an upward relationship with observations on the low-left part of the graph represented by Spain with negative returns and very low rates. Other observations belonging to Germany show returns above 10% and rates close to 4%, whereas England displays high rates but returns which on average tend to zero.

In 2013-2014 the association on the axes is positive. During those years, both the Scandinavian and the Spanish housing market present negative returns with very low mortgage rates. Conversely, Germany shows positive returns and mortgage rates at 4% as well as England which persists with observations showing very high mortgage rates with mainly positive returns.

Concerning the last year available in our analysis, 2015 presents values which are mainly concentrated with positive returns and mortgage rates close to 5% and the fitted line might look like an uptrend but due to the lack of enough observations the line is not shown.

The lower-left graph (Plot 8.C) gives some insights about the relationship between mortgage rate and real house price returns when using the (P)OLS estimator. Data from different cities are pooled together as cross sectional data.
At first glance, the distribution of the data in 8.C may suggest a positive trend between the mortgage rate and real house price returns. The dashed red line, however, representing the OLS estimate of the pooled data, draws a weak downward trend suggesting a negative association with the response variable when controlling for all the other explanatory variables. It seems evident, however, that using (P)OLS does not provide a satisfactory explanation because of the great dispersion in the data.

Figure 8



Plot 8.D shows the association between the mortgage rate and the dependent variable controlling for time and country effects. As described in Table 3, it turns out that the slope becomes positive when controlling for time and country effects, suggesting that the (P)OLS estimator in 8.C is missing relevant information. Indeed, the regression line representing the association between mortgage rate and house price returns show a different slope between graph 8.C and 8.D. However, these results are not surprising, especially when considering the cross-national and temporal variations in the relationship between the dependent and the independent variable outlined in 8.A and 8.B.

7 Serial correlation and potential bubbles

In this section the purpose is to try to understand the emergence of housing bubbles via regression coefficients which should provide some information on the status of the housing market in a specific country and city.
Unfortunately, we are not the first to develop on this kind of research. Indeed, as previously mentioned, strictly on the U.S.A. market, there are numerous studies providing information on how to capture housing market cycles via bubble and mean reversion coefficients. However, our analysis should be quite unique, focusing on the European market, more specifically on European cities.

The procedure on how to measure bubbles is a collection of different methods implemented by previous researchers. Before illustrating in detail our method, we think it is best to provide a short and concise definition of bubble:

> A situation in which news of price increases spurs investor enthusiasm, which spreads by psychological contagion from person to person, in the process amplifying stories that might justify the price increases and bringing in a larger and larger class of investors, who, despite doubts about the real value of an investment, are drawn to it partly through envy of others' successes and partly through a gambler's excitement.
>
> *Shiller, Robert. Irrational Exuberance (p. 240). Princeton University Press*

As previously explained in section 5.3, capturing the behaviour of house prices requires their estimation according to fundamental variables and their comparison with actual values. In our case, the real values are those collected in our spreadsheet, so the only piece of the puzzle we miss to implement the analysis are the predicted values. Generally speaking, we are aware that exploring housing market dynamics is an extremely difficult task, given that there is not a unique way to estimate a fundamental. However, given previous the specification tests, we reached the conclusion that the fitted values of housing returns obtained by using the FE estimator may be a good proxy for house price fundamentals at European, country and city level.

The first step in our study of bubbles involves a generic estimation at the European level to grasp the economic situation in the housing market following the base equation from *Capozza et al (2002)*. Table 4 shows the main results for the European market as a whole.

Table 4: Spotting bubbles, European Countries

	$\Delta P_t = \alpha_0 + \alpha_1 \Delta P_{t-1} + \beta(P^*_{t-1} - P_{t-1}) + \gamma \Delta P^*_t$			
	Intercept	Serial correlation	Bubble Burst	Adjustment Rate
	(1)	(2)	(3)	(4)
European Market	-0.000271	0.192733***	0.505835 ***	0.492002 ***
	(0.002971)	(0.053833)	(0.043761)	(0.074357)
	$\Delta P_{k,t} = \alpha_0 + \alpha_1 \Delta P_{k,t-1} + \beta(P^*_{k,t-1} - P_{t-1}) + \gamma \Delta P^*_{k,t}$			
Belgium	0.05487***	-0.05207	0.53931 ***	0.26985
	(0.00676)	(0.06597)	(0.05330)	(0.18982)
England	-0.01238	0.04978	0.50000 ***	0.52172 ***
	(0.00761)	(0.09734)	(0.08627)	(0.13105)
France	0.0253	0.0325	0.1786	-0.3246
	(0.0264)	(0.2699)	(0.0973)	(1.4621)
Germany	0.01171**	0.08322	0.42496***	-0.05055
	(0.00505)	(0.10373)	(0.06822)	(0.10681)
Netherlands	0.01144*	0.26105**	0.55918 ***	0.93933 ***
	(0.00593)	(0.12246)	(0.12029)	(0.20932)
Scandinavia	0.01802**	0.06387	0.19276	0.50706
	(0.00857)	(0.15268)	(0.13582)	(0.36936)
Spain	-0.0345***	0.1687	0.5133 ***	0.3650 ***
	(0.0110)	(0.1399)	(0.1138)	(0.1330)

Note: *p<0.1; **p<0.05; ***p<0.01

The $\alpha_1 = 0.2$, which is statistically significant, indicates a moderate amount of serial correlation between house price returns. The second coefficient of interest is $\beta = 0.5$ which, according to *Abraham and Hendershott (2002)*, supplies information on the magnitude of mean reversion, which in this case is fairly large, especially larger than α_1 suggesting a quick convergence to the mean. The last term is $\gamma = 0.5$, which shows a rapid adjustment to fundamentals since for any increase in the predicted returns, the actual ones are expected to increase at half pace in the same year.

The next step involves the estimation of the regression at country level. Results are fairly homogeneous in terms of α_1, where only the Netherlands show a coefficient which is statistically significant and about 0.30. Furthermore, they show strong mean reversion and extremely large partial adjustment coefficients.

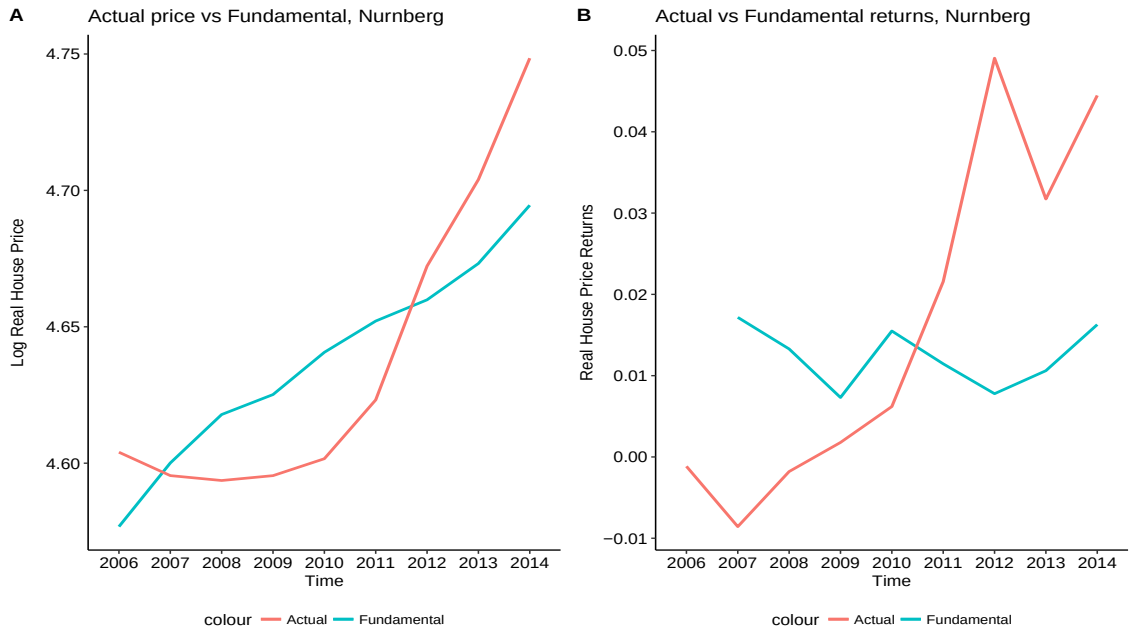
Overall, countries do not appear to experience any serial correlation in housing returns.

This result is quite eye-catching and it will be investigated later on at city level.
Remaining on Table 4, there are some differences across countries concerning the other two coefficients. Belgium, Germany and England show statistically significant mean reversion estimates ranging from 0.4 to 0.5. Moreover, England displays a partial adjustment parameter close to the one found for the entire European market. Ultimately, Spain shows both β and γ are statistically different from zero.
It is important to underline that in those countries where β is statistically significant, its magnitude is close to the one found at European level, indicating that for these estimates the national trend is confirmed.

As anticipated, now we return to the matter of analyzing more in depth why Table 4 results suggest no evidence of serial correlation in most of the countries in our sample. In order to do so, we compute the regression by *Capozza et al. (2002)* on 47 cities. For reasons of conciseness, we describe the results by nation focusing on the main findings.

Overall, German cities present the most striking results. First of all, Nurnberg represents, according to our fundamental conditions (Section 5.3) to detect a bubble, the only city where there is evidence of explosive behaviour. Its α_1 coefficient is greater than one, β is below its national counterpart and γ is equal to zero, implying that a shock to fundamental returns does not affect the behaviour of actual ones. More specifically, the mean reversion coefficient provides additional evidence on the persistence of the bubble as pointed out by *Capozza et al. (2002)*. Indeed, when the serial correlation is greater than the mean reversion coefficient, the run-up in prices takes a long time to converge to the long-run level. In this case, the convergence period, given that the beta's estimate is 0.3, will last many years as an increase in the difference between predicted minus actual log prices of 1% is associated with a 0.3 jump towards the mean. In other words, as displayed in Figure 9, it would take many years to absorb the deviation from fundamentals reached in a single year.
Another German city of interest is Bonn showing a substantial serial correlation coefficient and no evidence of mean reversion, which is what we are looking for since the ultimate purpose is to spot a bubble. Indeed, a beta coefficient which is not statistically different from zero is exactly what we are hoping to find through this analysis.

Figure 9



This would suggest that an increase in prices would persist over time before converging in the long-run to the mean. Up next there is Mannheim showing a similar pattern to Bonn but with a lower alpha. Freiburg presents a slight variation in these patterns. The city exhibits substantial correlation in lagged actual returns and a mean reversion coefficient which is still low in magnitude but statistically different from zero, therefore, an hypothetical increase in prices would converge faster to the fair price.

Hamburg, Dusseldorf and Koln present moderate levels of persistence but larger mean reversion coefficients. These results indicate that actual returns will fluctuate around their theoretical value. In other words, when a run-up in prices occurs, the large beta will deflate it bringing the value in line with its long-run prediction.

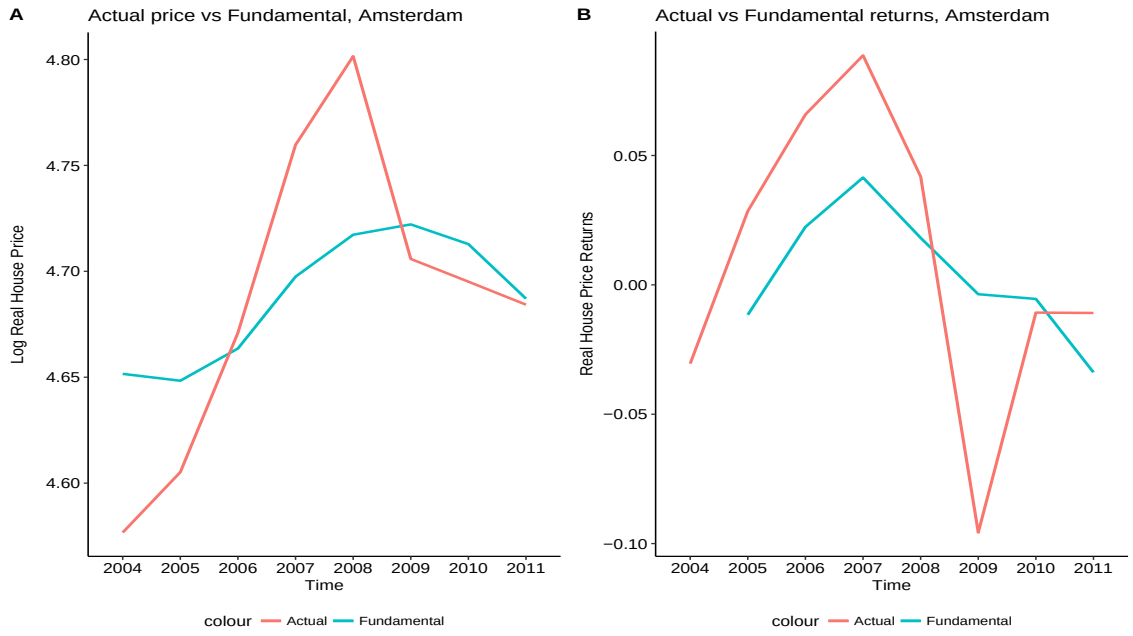
Ultimately for the German market, Dresden shows a quasi-efficient market where the gamma coefficient is close to unity indicating that there is a one-to-one adjustment ratio between actual prices and fundamentals.

According to our opinion, looking at the Spanish market, Bilbao represents the only city with results worth mentioning. The α_1 coefficient is close to unity which may indicate an explosive cycle. However, the β is larger than α_1 in absolute terms signalling a stronger mean reversion effect contrasting the serial correlation.

The next country we focus on is the Netherlands, mainly on the cities of Amsterdam (Figure 10), Rotterdam and The Hague. The capital shows a moderate serial correlation coefficient but a mean reversion estimate which approaches unity. These results should indicate that the latter coefficient takes over the effect of the former. However, given the substantial gamma parameter, which is basically an actual price elasticity of fundamental when a shock to predicted returns occurs, an overshooting is more likely. The Dutch capital exhibits a very elastic response to a change in fundamental returns, which is smoothened by beta. In other words, alpha and gamma would imply overshooting but the high beta dampens the boom effect.

The behaviour in Rotterdam is, on the other hand, much similar to the one found for Dusseldorf, Hamburg and Koln. The Hague is an ordinary example of quasi-efficiency with a partial adjustment value very close to unity.

Figure 10



Considering Belgium, results point out that Liege may be labelled as a quasi-efficient market as there is a strong responsiveness among predicted and actual returns.

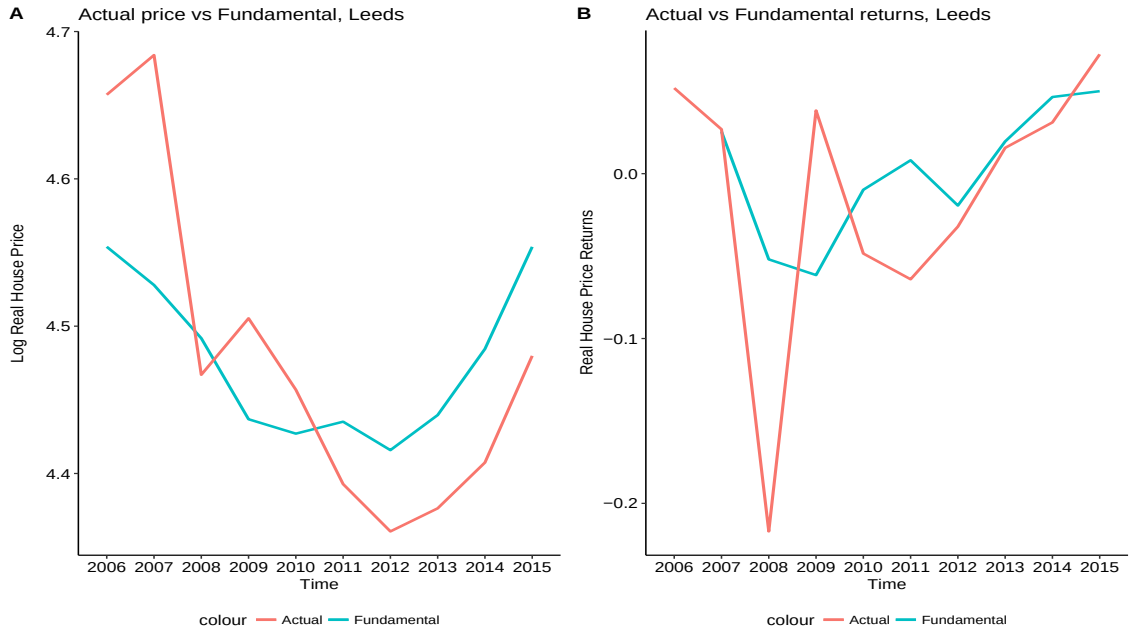
Considering England, the estimates (negative alpha and positive beta) indicate that Brad-

ford faces very short cycles oscillating around the fair price. Bristol and Leeds, shown in Figure 11, present no significant alphas and gammas close to unity also indicating quasi-efficient markets.

Leicester, on the other hand, shows evidence of overreaction to changes in fundamental returns but this reaction is mitigated by a significant negative serial correlation coefficient. The city of Liverpool shows unexpected results (together with Frankfurt), since the gamma is negative and greater than one in absolute terms, i.e. an increase in fundamental returns is followed by a decrease in actual returns. We suggest further research to focus on these markets, since we are not capable of providing a plausible explanation. Ultimately, Nottingham shows little evidence of serial correlation as opposed to the considerable mean reversion coefficient.

Among the Northern European countries, Goteborg deserves particular attention providing shy evidence of overreaction to fundamentals. Copenhagen, instead, given the lack of longer time series, shows meaningless results.

Figure 11



Overall, in this last section, we have investigated bubble patterns and generally we found evidence only in Nurnberg that such irrational exuberance occurs. In other words, looking

at house prices per se (even though we are considering actual prices) is not enough. When examined, cities whose real house prices might suggest a bubblish behaviour[55] show at best evidence of significant convergent serial correlation which by itself is not enough to define what those cities are experiencing as "irrational".

Despite our objective of spotting bubbles reveals to be quite challenging, probably also due to lack of longer time series and the choice of independent variables used to estimate the fundamentals, we found interesting results for the city of Dresden, The Hague, Leeds and Bristol in terms of adjustment to fundamental shocks and for Amsterdam regarding the mean reversion coefficient.

[55]This refer to the section where City and Time trends are presented, pp 31-38.

Table 5.1: Spotting bubbles, European Cities

$$\Delta P_{i,t} = \alpha_0 + \alpha_1 \Delta P_{i,t-1} + \beta(P^*_{i,t-1} - P_{i,t-1}) + \gamma \Delta P^*_{i,t}$$

	Intercept	Serial correlation	Mean Reversion	Adjustment Rate
	(1)	(2)	(3)	(4)
Amsterdam	0.00300	0.45214**	0.95245 **	1.52427 **
	(0.00854)	(0.12579)	(0.21571)	(0.34974)
Anvers	0.05586***	-0.08486	0.60811***	0.22160
	(0.00617)	(0.05420)	(0.03322)	(0.54818)
Barcelona	-0.0281*	0.2854	0.4933*	0.2287*
	(0.0125)	(0.1433)	(0.1999)	(0.0911)
Berlin	0.0209	-0.1650	0.3678**	-0.3524
	(0.0187)	(0.0943)	(0.1093)	(0.2374)
Bilbao	0.00885	0.98094**	1.52685**	0.22600
	(0.01952)	(0.34142)	(0.48946)	(0.37079)
Bonn	0.00516	0.89700***	0.13149	-0.12229
	(0.00592	(0.19948)	(0.14070)	(0.13646)
Bradford	-0.0387***	-0.2880**	0.5456***	0.5613**
	(0.0067)	(0.0892)	(0.1095)	(0.1328)
Bremen	0.01424	0.34759	0.07740*	-0.60313*
	(0.00859)	(0.20020)	(0.03080)	(0.25450)
Bristol	-0.0108	0.3969	1.3387**	0.9719*
	(0.0282)	(0.3119)	(0.4937)	(0.4804)
Bruxelles	0.06137***	-0.00647	0.64513***	0.22860
	(0.01653)	(0.05660)	(0.02474)	(0.55279)
Copenhagen	-0.0137	1.3988	2.6784.	1.8940
	(0.0129)	(0.2495)	(0.4746)	(0.5649)
Dresden	-0.03344**	-0.01169	0.49301***	0.98528**
	(0.00801)	(0.10174)	(0.05273)	(0.23251)
Dusseldorf	0.0126	0.6121***	0.8148***	-0.2824
	(0.0113)	(0.0880)	(0.1295)	(0.8851)
Frankfurt	0.0452*	0.1373	0.9263***	-2.0683**
	(0.0194)	(0.1187)	(0.1295)	(0.5959)
Freiburg	0.00719	0.65538**	0.32897*	-0.15168
	(0.00597)	(0.17669)	(0.13824)	(0.11789)
Gand	0.07288***	-0.20111***	0.42262***	0.03739
	(0.00397)	(0.04470)	(0.02131)	(0.27383)
Goteborg	0.0538***	-0.6152	0.1608	1.1444**
	(0.0106)	(0.3247)	(0.1298)	(0.2524)
Hamburg	0.0114	0.3060*	0.7711**	-0.1046
	(0.0114)	(0.1434)	(0.2587)	(0.4276)
Helsinki	-0.00458	0.50942	1.53417	0.63744
	(0.01449)	(0.40916)	(0.70358)	(1.71210)
Koln	0.02915***	0.31308***	0.42638***	-0.64568
	(0.00255)	(0.06272)	(0.04948)	(0.46928)
Las Palmas	-0.0465**	0.0575	0.5248	0.2306**
	(0.0141)	(0.1035)	(0.2602)	(0.0620)
Leeds	-0.0310	-0.5619	0.3187	1.1906*
	(0.0178)	(0.3379)	(0.3064)	(0.5751)
Leicester	-0.02612**	-0.43889***	0.26988	1.78022***
	(0.00986)	(0.08742)	(0.20841)	(0.32898)

Note: *p<0.1; **p<0.05; ***p<0.01

Table 5.2: Spotting bubbles, European Cities

$$\Delta P_{i,t} = \alpha_0 + \alpha_1 \Delta P_{i,t-1} + \beta(P^*_{i,t-1} - P_{i,t-1}) + \gamma \Delta P^*_{i,t}$$

	Intercept	Serial correlation	Mean Reversion	Adjustment Rate
	(1)	(2)	(3)	(4)
Leipzig	-0.0137	-0.1805	0.3692***	0.1890
	(0.0123)	(0.1457)	(0.0802)	(0.1565)
Liege	0.03115***	0.12111	0.33456***	0.90882**
	(0.00715)	(0.15099)	(0.06078)	(0.26183)
Liverpool	-0.0393	0.6431*	0.3646	-1.6045**
	(0.0170)	(0.2706)	(0.1717)	(0.4450)
London	0.0516	-0.4075	-0.2709	0.4931
	(0.0491)	(0.3257)	(0.7551)	(0.6047)
Madrid	-0.0469**	-0.0984	0.4637**	0.4675**
	(0.0112)	(0.1206)	(0.1350)	(0.1164)
Malaga	-0.0490	0.2111	0.6085**	0.1235
	(0.0249)	(0.2341)	(0.1435)	(0.3351)
Malmo	0.00429	0.04234	0.66833	0.56175
	(0.01570)	(0.18671)	(0.47269)	(0.62882)
Manchester	-0.00736	0.13497	0.19590	0.23264
	(0.02771)	(0.20579)	(0.34078)	(0.36537)
Mannheim	0.00475	0.37321*	0.19096	0.07815
	(0.00432)	(0.15598)	(0.12284)	(0.08393)
Munich	0.0395	0.4915	0.1492	1.1707
	(0.0390)	(0.2624)	(0.1501)	(2.0873)
Newcastle	-0.0288*	-0.0107	0.3654	0.5635
	(0.0142)	(0.3029)	(0.1926)	(0.5251)
Nottingham	-0.0137	0.2670*	0.4956**	0.3472**
	(0.0126)	(0.1131)	(0.1666)	(0.0934)
Nurnberg	0.0040	1.0508***	0.2907**	-0.0738
	(0.0041)	(0.1694)	(0.0998)	(0.1110)
Paris	0.0253	0.0325	0.1786	-0.3246
	(0.0245)	(0.2699)	(0.0973)	(1.4621)
Portsmouth	-0.0337	-0.3128	0.6905	1.6473
	(0.0245)	(0.2099)	(0.3104)	(0.5261)
Rotterdam	0.00873	0.50658**	0.62394**	0.76000
	(0.00666)	(0.14844)	(0.15966)	(0.36652)
Sevilla	-0.0142	0.1337	0.7038*	0.4685
	(0.0215)	(0.1148)	(0.2615)	(0.2820)
Sheffield	-0.0198	-0.0601	0.3145	0.8401*
	(0.0151)	(0.1512)	(0.2307)	(0.2675)
Stockholm	0.0491*	-0.0171	-0.1187	0.8292
	(0.0224)	(0.3731)	(0.2269)	(1.7571)
Stuttgart	0.0406	-0.0345	0.0909	-0.3547
	(0.0259)	(0.4070)	(0.2820)	(0.3829)
The Hague	0.0208	0.1314	0.5445*	1.0708**
	(0.0168)	(0.3015)	(0.2091)	(0.3334)
Utrecht	0.01163	0.12090	0.37400	0.89833
	(0.00694)	(0.19656)	(0.23610)	(0.45816)
Valencia	-0.0220	0.2873	1.5130**	0.0915
	(0.0615)	(0.1940)	(0.3888)	(1.0395)
Zaragoza	-0.0568**	0.3087	0.0869	0.5040
	(0.0189)	(0.1559)	(0.1534)	(0.2565)

Note: *p<0.1; **p<0.05; ***p<0.01

8 Conclusions

This paper provides an original framework for the European housing market, which has been overlooked by previous research.

The econometric analysis with panel data shows that variations in GDP per capita, housing starts and unemployment rate are statistically significant in approximately all specifications. The inclusion of two additional variables, such as net migration and urban sprawl, brings along interesting results. We find little evidence that the former has a positive effect on returns, whereas the latter exhibits a strong negative association with the dependent variable.

Considering that the FE estimator appears to be the best fit for our dataset, we investigate those effects by augmenting a standard (P)OLS with time, country and local dummies. This transformation leads us to find evidence of time and country effects which decrease the statistical significance of certain explanatory variables.

To dig deeper into the matter, a machine learning technique is implemented. The random forest process represents an innovative introduction when studying the real estate market, never implemented before. Indeed, the process provides a reliable measure of importance of each predictor with respect to real returns, which is considerably different from the level of information provided by econometric analysis. Additionally, we are able to reconcile these two complementary methods which add on one another leading to a thorough understanding of house price dynamics. Overall, the mortgage rate emerges as the most important parameter for the model. However, the presence of cross-time and national variation may cause the econometric interpretation of the parameter to be quite misleading. This is the reason why, we implement an exhaustive analysis with respect to the matter.

The last step in our analysis concerns estimation of an equation by *Capozza et al. (2002)* which should give information of potential bubbles and boom-and-burst cycles in the housing market at city level. The results of the city of Nurnberg are worth mentioning as they represent the only case of "Shiller's irrational exuberance" across the 47 cities considered. Unexpected results are found in Frankfurt and Liverpool, i.e. their meaning is not of easy interpretation.

In the German market, Bonn, Mannheim and Freiburg present uptrends in house prices

with elevated serial correlation and low mean reversion coefficients, i.e. interested parties should keep a close eye on them in the near future.

Other cities, namely Dresden, Leeds and Bristol provide evidence of market efficiency, signalling no profitable opportunities due to the lack of predictability in returns. We find evidence of moderate serial correlation and an even larger mean reversion effect in some cities, such as Amsterdam, Nottingham, Hamburg, Koln and Dusseldorf, indicating a life cycle which is to some extent forecastable. Ultimately, we acknowledge that our analysis presents some strong limitations. Mainly the lack of longer time series, which however, is hardly our fault and the presence of heterogeneity in the housing price metrics is not optimal when dealing with panel data. A sequel of our analysis could be carried out with better data to investigate further some patters which we are not currently able to explain. For instance, the city of Liverpool and Frankfurt should be looked closely to understand what is happening in those local markets. The same goes for the city of Copenhagen whose results appear of unclear interpretation. Other trends worth analyzing are those where according to the initial plots in Section 4 a bubble may appear to be building up but our study could not support these surges with concrete evidence. The reason for this may partially be found in the absence of time series for the last 2-3 years for most cities. Spacing away to other methodologies, the implementation of a Vector Autoregression (VAR) procedure at city level would certainly provide additional insights on the association between response and explanatory variables. For instance, it would be interesting to investigate the impact of a shock to the mortgage rate on housing returns. Indeed, we would follow exactly these additional steps if we had longer time series available. The intuition is provided by *McDonald and Stokes (2011, 2013)* who have already used such technique to analyze the impact of the federal funds rate on U.S. housing markets during the financial crisis. Similarly, exploring the impact of monetary policy changes by the European Central Bank on the Eurozone system may shed some light on issues which otherwise would be overlooked.

As a concluding remark we believe that the European market has not been sufficiently explored yet, and we are confident that with our analysis we are able to foster further research with respect to the topic.

Appendix

Table 6 - Sources: UK, Germany and Spain			
Variables	**UK**	**Germany**	**Spain**
House prices	UK Government	Statista, Bulwiengesa AG	Ministero de Fomento
Housing starts	UK Government	Die Regionaldatenbank Deutschland	Ministero de Fomento
Population	Eurostat	Eurostat	Eurostat
GDP per capita	Eurostat	Eurostat	Eurostat
Unemployment	Eurostat,OECD	OECD	OECD
Employment	Eurostat, OECD	OECD	OECD
Mortgage rate	ECB	ECB	ECB
Net Migration	UK, Office for National Statistics	Destatis.de	Eurostat
Urban sprawl	OECD, Metropolitan database	OECD, Metropolitan database	OECD, Metropolitan database
CPI	UK Government	Destatis.de	Instituto Nacional de Estadistica

Table 7 - Sources: Denmark, Sweden and Finland			
Variables	**Denmark**	**Sweden**	**Finland**
House prices	Statistics Denmark	Statistics Sweden	Statistics Finland
Housing starts	Statistics Denmark	Statistics Sweden	Statistics Finland
Population	Eurostat	Statistics Sweden	Statistics Finland
GDP per capita	Eurostat	Eurostat	Eurostat
Unemployment	OECD	OECD	OECD
Employment	OECD	OECD	OECD
Mortgage rate	ECB	ECB	ECB
Net Migration	Statistics Denmark	Statistics Sweden	Statistics Finland
Urban sprawl	OECD, Metropolitan database	OECD, Metropolitan database	OECD, Metropolitan database
CPI	Statistics Denmark	Statistics Sweden	Statistics Finland

Table 8 - Sources: Belgium, France and Norway				
Variables	**Netherlands**	**Belgium**	**France**	**Norway**
House prices	Centraal Bureau voor de Statistiek	StatBel	INSEE	Statistics Norway
Housing starts	Centraal Bureau voor de Statistiek	StatBel	Ministre de la Transition Ecologique et solidaire	Statistics Norway
Population	Centraal Bureau voor de Statistiek	StatBel	Eurostat	Eurostat
GDP per capita	Eurostat	Eurostat	Eurostat	Eurostat
Unemployment	OECD	OECD	OECD	OECD
Employment	OECD	OECD	OECD	OECD
Mortgage rate	ECB	ECB	ECB	Statistics Norway
Net Migration	Centraal Bureau voor de Statistiek	Eurostat	Eurostat	Statistics Norway
Urban sprawl	OECD	OECD	OECD	OECD
CPI	Centraal Bureau voor de Statistiek	National Bank of Belgium	INSEE	Statistics Norway

Table 9 - Type of property, housing starts and mortgage rate			
Countries	**House prices**	**Housing starts**	**Mortgage rate**
UK	Detached, semidet, terraced, flats	Net additional dwellings	Interest rate on loans over 5 years for house purchase
Germany	Detached and semidetached houses	Building permits	Interest rate on loans over 5 years for house purchase
Spain	Vivienda Libre	Vivienda iniciada	Interest rate on loans over 5 years for house purchase
Denmark	Single family homes	Building permits	Interest rate on loans over 5 years for house purchase
Sweden	1-2 dwelling buildings	Building permits	Interest rate on loans over 5 years for house purchase
Finland	Single-family home	Building permits	Interest rate on loans over 5 years for house purchase
Netherlands	Existing own homes	Building permits	Interest rate on loans over 5 years for house purchase
Belgium	All residential dwellings	Building permits	Interest rate is on all loans for house purchase
France	Logements	Building permits	Interest rate on loans over 5 years for house purchase
Norway	Detached house	Building permits	Interest rate on loans

Table 10.1: Breusch-Pagan Lagrange Multiplier Test: (P)OLS vs RE/FE

RHPR $\sim \Delta HS + \Delta POP + \Delta GDP +$ UN + MR + ΔEMP

Time effects
Chisq = 48, df = 1, p-value <0.00000000004
Alternative hypothesis: significant effects
Individual effects
Chisq = 0.72, df = 1, p-value <0.4
Alternative hypothesis: significant effects
Twoways effects
Chisq = 49, df = 2, p-value <0.00000000000002
Alternative hypothesis: significant effects

RHPR $\sim \Delta HS + \Delta POP + \Delta GDP +$ UN + MR + $\Delta EMP +$ NM

Time effects
Chisq = 50, df = 1, p-value <0.000000000001
Alternative hypothesis: significant effects
Individual effects
Chisq = 0.72, df = 1, p-value <0.4
Alternative hypothesis: significant effects
Twoways effects
Chisq = 51, df = 2, p-value <0.000000000008
Alternative hypothesis: significant effects

Table 10.2: Breusch-Pagan Lagrange Multiplier Test: (P)OLS vs RE/FE

RHPR $\sim \Delta HS + \Delta POP + \Delta GDP$ + UN + MR + ΔEMP+US
Time effects
Chisq = 46, df = 1, p-value <0.00000000001
Alternative hypothesis: significant effects
Individual effects
Chisq = 0.0043, df = 1, p-value <0.9
Alternative hypothesis: significant effects
Twoways effects
Chisq = 46, df = 2, p-value <0.0000000001
Alternative hypothesis: significant effects

RHPR $\sim \Delta HS + \Delta POP + \Delta GDP$ + UN + MR + ΔEMP+NM+US
Time effects
Chisq = 48, df = 1, p-value <0.000000000004
Alternative hypothesis: significant effects
Individual effects
Chisq = 0.0024, df = 1, p-value <1
Alternative hypothesis: significant effects
Twoways effects
Chisq = 48, df = 2, p-value <0.00000000004
Alternative hypothesis: significant effects

Table 11: Hausman Test

RHPR $\sim \Delta HS + \Delta POP + \Delta GDP +$ UN + MR + ΔEMP
Chisq = 3400, df = 6, p-value <0.0000000000000002 Alternative hypothesis: one model is inconsistent
RHPR $\sim \Delta HS + \Delta POP + \Delta GDP +$ UN + MR + $\Delta EMP +$ NM
Chisq = 31, df = 7, p-value = 0.00006 Alternative hypothesis: one model is inconsistent
RHPR $\sim \Delta HS + \Delta POP + \Delta GDP +$ UN + MR + $\Delta EMP +$ US
Chisq = 8.7, df = 6, p-value = 0.2 Alternative hypothesis: one model is inconsistent
RHPR $\sim \Delta HS + \Delta POP + \Delta GDP +$ UN + MR + $\Delta EMP +$ NM + US
Chisq = 21, df = 7, p-value = 0.004 Alternative hypothesis: one model is inconsistent

Table 12: F-test: (P)OLS vs FE

RHPR $\sim \Delta HS + \Delta POP + \Delta GDP +$ UN + MR + ΔEMP
Time effects
Chisq = 8.8, df1 = 12, df2 = 420, p-value <0.0000000000000005
Alternative hypothesis: significant effects
Individual effects
Chisq = 1.4, df1 = 47, df2 = 380, p-value <0.05
Alternative hypothesis: significant effects
Twoways effects
Chisq =3.2, df1 = 59, df2 = 370, p-value <0.00000000001
Alternative hypothesis: significant effects

RHPR $\sim \Delta HS + \Delta POP + \Delta GDP +$ UN + MR + $\Delta EMP +$ NM
Time effects
Chisq = 8.9, df1 = 12, df2 = 410, p-value <0.0000000000000004
Alternative hypothesis: significant effects
Individual effects
Chisq = 1.4, df1 = 47, df2 = 380, p-value <0.05
Alternative hypothesis: significant effects
Twoways effects
Chisq = 3.2, df1 = 59, df2 = 370, p-value <0.00000000001
Alternative hypothesis: significant effects

Table 13: F-test: (P)OLS vs FE

RHPR $\sim \Delta HS + \Delta POP + \Delta GDP$ + UN + MR + ΔEMP+US
Time effects
Chisq = 8.4, df1 = 12, df2 = 410, p-value <0.000000000003
Alternative hypothesis: significant effects
Individual effects
Chisq = 1.3, df1 = 46, df2 = 380, p-value <0.1
Alternative hypothesis: significant effects
Twoways effects
Chisq = 3.1, df1 = 58, df2 = 370 , p-value <0.00000000005
Alternative hypothesis: significant effects

RHPR $\sim \Delta HS + \Delta POP + \Delta GDP$ + UN + MR + ΔEMP+NM+US
Time effects
Chisq = 8.5, df1 = 12, df2 = 410, p-value <0.0000000000000002
Alternative hypothesis: significant effects
Individual effects
Chisq = 1.3, df1 = 46, df2 = 380, p-value <0.1
Alternative hypothesis: significant effects
Twoways effects
Chisq = 3.1, df1 = 58, df2 = 370, p-value <0.00000000006
Alternative hypothesis: significant effects

Table 14: Wooldridge's test for serial correlation in short FE models

RHPR $\sim \Delta HS + \Delta POP + \Delta GDP +$ UN + MR + ΔEMP
F = 1.3, df1 = 1, df2 = 380, p-value = 0.3 Alternative hypothesis: serial correlation
RHPR $\sim \Delta HS + \Delta POP + \Delta GDP +$ UN + MR + $\Delta EMP +$ NM
F = 1.2, df1 = 1, df2 = 380, p-value = 0.3 Alternative hypothesis: serial correlation
RHPR $\sim \Delta HS + \Delta POP + \Delta GDP +$ UN + MR + $\Delta EMP +$US
F = 1.3, df1 = 1, df2 = 380, p-value = 0.3 Alternative hypothesis: serial correlation
RHPR $\sim \Delta HS + \Delta POP + \Delta GDP +$ UN + MR + $\Delta EMP +$ NM + US
F = 1.2, df1 = 1, df2 = 380, p-value = 0.3 Alternative hypothesis: serial correlation

Table 15.1 : W's first-difference test for serial correlation in panels

RHPR $\sim \Delta HS + \Delta POP + \Delta GDP +$ UN + MR + ΔEMP
F = 0.0048, df1 = 1, df2 = 340, p-value = 0.9 Alternative hypothesis: serial correlation in original errors
RHPR $\sim \Delta HS + \Delta POP + \Delta GDP +$ UN + MR + $\Delta EMP +$ NM
F = 0.02, df1 = 1, df2 = 340, p-value = 0.9 Alternative hypothesis: serial correlation in original errors
RHPR $\sim \Delta HS + \Delta POP + \Delta GDP +$ UN + MR + $\Delta EMP +$US
F = 0.0048, df1 = 1, df2 = 340, p-value = 0.9 Alternative hypothesis: serial correlation in original errors
RHPR $\sim \Delta HS + \Delta POP + \Delta GDP +$ UN + MR + $\Delta EMP +$ NM + US
F = 0.02, df1 = 1, df2 = 340, p-value = 0.9 Alternative hypothesis: serial correlation in original errors

Table 15.2: W's first-difference test for serial correlation in panels

RHPR $\sim \Delta HS + \Delta POP + \Delta GDP +$ UN + MR + ΔEMP
F = 180, df1 = 1, df2 = 340, p-value <0.0000000000000002 Alternative hypothesis: serial correlation in differenced errors
RHPR $\sim \Delta HS + \Delta POP + \Delta GDP +$ UN + MR + $\Delta EMP +$ NM
F = 170, df1 = 1, df2 = 340, p-value <0.0000000000000002 Alternative hypothesis: serial correlation in differenced errors
RHPR $\sim \Delta HS + \Delta POP + \Delta GDP +$ UN + MR + $\Delta EMP +$ NM + US
F = 180, df1 = 1, df2 = 340, p-value <0.0000000000000002 Alternative hypothesis: serial correlation in differenced errors
RHPR $\sim \Delta HS + \Delta POP + \Delta GDP +$ UN + MR + $\Delta EMP +$ NM + US
CF = 170, df1 = 1, df2 = 340, p-value <0.0000000000000002 Alternative hypothesis: serial correlation in differenced errors

Table 16.1: Panel regression, House price fundamentals with time, national and local effects

	Dependent variable: Real house price returns				
	Original	Time	Time + Country	Time + City	Time x Country
	(1)	(2)	(3)	(4)	(5)
%Δ*Housing starts*	0.0213**	0.0193**	0.0183**	0.0185**	0.0025
	(0.0086)	(0.0078)	(0.0075)	(0.0079)	(0.0062)
%Δ*Population*	0.3553*	0.4100*	0.0698	−0.1024	0.3147
	(0.2090)	(0.2320)	(0.2519)	(0.2722)	(0.2678)
%Δ*GDP per capita*	0.4643***	0.3879***	0.3080***	0.2982***	−0.0301
	(0.0439)	(0.0609)	(0.0655)	(0.0701)	(0.1130)
Unemployment rate	−0.0039***	−0.0037***	−0.0038***	−0.0076***	−0.0018**
	(0.0007)	(0.0007)	(0.0006)	(0.0012)	(0.0008)
Mortgage rate	−0.0080*	−0.0068	0.0099**	−0.0015	0.0107*
	(0.0042)	(0.0053)	(0.0047)	(0.0079)	(0.0056)
%Δ*Employment rate*	0.0951	0.1883	0.2189	0.2145	0.0138
	(0.1470)	(0.1487)	(0.1734)	(0.1797)	(0.0846)
Net migration	0.0004*	0.0005*	0.0004*	0.0001	0.0001
	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0004)
Urban Sprawl	−0.0004**	−0.0004*	−0.0001	0.0001	−0.0001
	(0.0002)	(0.0002)	(0.0002)	(0.0002)	(0.0002)
Belgium			0.0586***		0.0239
			(0.0081)		(0.0211)
England			−0.0085		0.0410**
			(0.0083)		(0.0193)
France			0.0649***		−0.0506***
			(0.0072)		(0.0184)
Netherlands			0.0019		−0.0132
			(0.0076)		(0.0155)
Scandinavia			0.0278***		0.0253
			(0.0067)		(0.0232)
Spain			0.0184		0.0310
			(0.0116)		(0.0274)
Anvers				0.0573***	
				(0.0063)	
Barcelona				0.0501***	
				(0.0176)	
Berlin				0.0505***	
				(0.0156)	
Bilbao				0.0277***	
				(0.0041)	
Bonn				0.0100**	
				(0.0045)	
Bradford				0.0270	
				(0.0194)	

Note: *p<0.1; **p<0.05; ***p<0.01

Table 16.2: Panel regression, House price fundamentals with time, national and local effects

	Dependent variable: Real house price returns				
	Original	Time	Time + Country	Time + City	Time x Country
	(1)	(2)	(3)	(4)	(5)
Bremen				0.0090	
				(0.0078)	
Bruxelles				0.1073***	
				(0.0116)	
City of Bristol				0.0281**	
				(0.0116)	
City of Nottingham				0.0343**	
				(0.0156)	
Copenhagen				0.0217***	
				(0.0080)	
Dresden				0.0328***	
				(0.0107)	
Dusseldorf				0.0206**	
				(0.0081)	
Frankfurt				0.0010	
				(0.0095)	
Freiburg				0.0008	
				(0.0043)	
Gand				0.0597***	
				(0.0062)	
Goteborg				0.0658***	
				(0.0100)	
Hamburg				0.0176**	
				(0.0080)	
Helsinki				0.0162	
				(0.0173)	
Koln				0.0177**	
				(0.0080)	
Las Palmas				0.1266***	
				(0.0188)	
Leeds				0.0160	
				(0.0120)	
Leicester				0.0133	
				(0.0116)	
Leipzig				0.0359***	
				(0.0112)	
Liege				0.0909***	
				(0.0102)	
Liverpool				0.0167	
				(0.0135)	

Note: *p<0.1; **p<0.05; ***p<0.01

Table 16.3: Panel regression, House price fundamentals with time, national and local effects

	Dependent variable: Real house price returns				
	Original	Time	Time + Country	Time + City	Time x Country
	(1)	(2)	(3)	(4)	(5)
London				0.0486***	
				(0.0153)	
Madrid				0.0165	
				(0.0118)	
Malaga				0.1072***	
				(0.0179)	
Malmo				0.0575***	
				(0.0108)	
Manchester				0.0493***	
				(0.0152)	
Mannheim				0.0009	
				(0.0054)	
Munich				0.0325***	
				(0.0075)	
Nurnberg				0.0186**	
				(0.0075)	
Newcastle upon Tyne				0.0118	
				(0.0122)	
Oslo				0.0140	
				(0.0121)	
Paris				0.0792***	
				(0.0093)	
Portsmouth				0.0207*	
				(0.0114)	
Rotterdam				0.0066***	
				(0.0023)	
Sevilla				0.1163***	
				(0.0151)	
Sheffield				0.0305**	
				(0.0140)	
Stockholm				0.0616***	
				(0.0095)	
Stuttgart				0.0178**	
				(0.0081)	
The Hague				0.0161***	
				(0.0024)	
Utrecht				0.0095***	
				(0.0033)	
Valencia				0.0570***	
				(0.0170)	

Note: *p<0.1; **p<0.05; ***p<0.01

Table 16.4: Panel regression, House price fundamentals with time, national and local effects

	Dependent variable: Real house price returns				
	Original	Time	Time + Country	Time + City	Time x Country
	(1)	(2)	(3)	(4)	(5)
2004		0.0171	−0.0058	−0.0027	0.0099
		(0.0168)	(0.0173)	(0.0181)	(0.0211)
2005		0.0959**	0.0837**	0.0889**	−0.0281
		(0.0425)	(0.0368)	(0.0391)	(0.0201)
2006		0.0183	0.0165	0.0127	−0.0602***
		(0.0131)	(0.0152)	(0.0174)	(0.0200)
2007		−0.0031	−0.0099	−0.0132	−0.0584***
		(0.0113)	(0.0133)	(0.0157)	(0.0225)
2008		−0.0479***	−0.0609***	−0.0630***	−0.0356
		(0.0127)	(0.0146)	(0.0165)	(0.0220)
2009		0.0078	0.0028	0.0005	−0.0231
		(0.0164)	(0.0185)	(0.0211)	(0.0305)
2010		−0.0029	0.0019	−0.0012	0.0280
		(0.0137)	(0.0168)	(0.0189)	(0.0285)
2011		−0.0274**	−0.0243*	−0.0267	−0.0183
		(0.0117)	(0.0146)	(0.0175)	(0.0162)
2012		−0.0114	−0.0044	−0.0110	0.0544**
		(0.0140)	(0.0174)	(0.0200)	(0.0254)
2013		0.0157	0.0260	0.0187	0.0308
		(0.0130)	(0.0161)	(0.0201)	(0.0226)
2014		−0.0113	0.0066	−0.0036	−0.0082
		(0.0138)	(0.0179)	(0.0234)	(0.0223)
2015		0.0006	0.0343*	0.0185	0.0161
		(0.0145)	(0.0188)	(0.0239)	(0.0270)
Belgium, 2004					−0.0006
					(0.0341)
Belgium, 2005					0.2744***
					(0.0633)
Netherlands, 2005					0.0566**
					(0.0236)
Belgium, 2006					0.1396***
					(0.0175)
England, 2006					0.0499**
					(0.0242)
Netherlands, 2006					0.1065***
					(0.0191)
Scandinavia, 2006					0.1093***
					(0.0325)
Spain, 2006					0.1096***
					(0.0374)

Note: *p<0.1; **p<0.05; ***p<0.01

Table 16.5: Panel regression, House price fundamentals with time, national and local effects

	Dependent variable: Real house price returns				
	Original	Time	Time + Country	Time + City	Time x Country
	(1)	(2)	(3)	(4)	(5)
Belgium, 2007					0.1032***
					(0.0255)
England, 2007					0.0068
					(0.0217)
France, 2007					0.1717***
					(0.0207)
Netherlands, 2007					0.1030***
					(0.0227)
Scandinavia, 2007					0.0800***
					(0.0225)
Spain, 2007					0.0464
					(0.0337)
Belgium, 2008					0.0278
					(0.0178)
England, 2008					−0.2457***
					(0.0354)
France, 2008					0.1127***
					(0.0228)
Netherlands, 2008					0.0391
					(0.0246)
Scandinavia, 2008					−0.0412
					(0.0264)
Spain, 2008					−0.0652*
					(0.0338)
Belgium, 2009					−0.0297
					(0.0450)
England, 2009					−0.0186
					(0.0319)
France, 2009					0.0092
					(0.0238)
Netherlands, 2009					−0.0647**
					(0.0257)
Scandinavia, 2009					−0.0354
					(0.0365)
Spain, 2009					−0.1329***
					(0.0433)
Belgium, 2010					−0.0437
					(0.0353)
England, 2010					−0.1345***
					(0.0305)

Note: *p<0.1; **p<0.05; ***p<0.01

Table 16.6: Panel regression, House price fundamentals with time, national and local effects

	Dependent variable: Real house price returns				
	Original	Time	Time + Country	Time + City	Time x Country
	(1)	(2)	(3)	(4)	(5)
France, 2010					0.1030***
					(0.0230)
Netherlands, 2010					−0.0345
					(0.0349)
Scandinavia, 2010					−0.0083
					(0.0380)
Spain, 2010					−0.0703
					(0.0605)
Belgium, 2011					0.0152
					(0.0158)
England, 2011					−0.1047***
					(0.0220)
France 2011					0.2125***
					(0.0161)
Scandinavia, 2011					−0.0398
					(0.0297)
Spain, 2011					−0.0997***
					(0.0328)
Belgium, 2012					−0.1062***
					(0.0377)
England, 2012					−0.1373***
					(0.0274)
France, 2012					0.0062
					(0.0170)
Scandinavia, 2012					−0.0822**
					(0.0392)
Spain, 2012					−0.1898***
					(0.0418)
Belgium 2013					−0.0443
					(0.0291)
England, 2013					−0.0743***
					(0.0265)
Scandinavia, 2013					−0.0150
					(0.0216)
Spain, 2013					−0.0872**
					(0.0345)
Constant	0.0758***	0.0711**	−0.0181	0.0468	−0.0208
	(0.0235)	(0.0312)	(0.0303)	(0.0483)	(0.0321)

Note: *p<0.1; **p<0.05; ***p<0.01

References

- Abraham, J.M. and Hendershott, P.H. (1994): *Bubbles in Metropolitan Housing Markets*, NBER Working Papers 4774, National Bureau of Economic Research, Inc.

- Amemiya, T. (1971): *The estimation of the variances in a variance-components model*, International Economic Review, Department of Economics, University of Pennsylvania and Osaka University Institute of Social and Economic Research Association, vol. 12(1),1-13

- Baltagi, B. H. (2005): *Econometric analysis of panel data*, Wiley.

- Blanchard, O.J. and Watson, M.W. (1982): *Bubbles, Rational Expectations and Financial markets*, NBER Working Papers 0945, National Bureau of Economic Research, Inc.

- Card, D. (2007): *How immigration affects U.S. cities*, Centre for Research and Analysis of Migration Department of Economics, University College London Drayton House

- Case, K. E. and Shiller, R. J. (1987): *Prices of single-family homes since 1970: New indices for four cities*. Cowles Foundation for Research in Economics, Yale University

- Case, K. E. and Shiller, R. J. (1989): *The efficiency of the market for single-family homes*, American Economic Review 79(1):125-37

- Case, K. E. and Shiller, R. J. (2003): *Is there a bubble in the Housing Market?*. Cowles Foundation Paper. Yale University

- Capozza, Hendershott, Mack and Mayer (2002): *Determinants of real house price dynamics*, NBER Working Papers 9262, National Bureau of Economic Research, Inc.

- Croissant, Y. and Millo, G. (2008): *Panel data econometrics in R: The plm package*, Journal of Statistical Software, Foundation for Open Access Statistics, vol. 27(i02)

- Dipasquale, D. and Wheaton, W. (1994): *Housing Market Dynamics and the future of Housing Prices*, Journal of Urban Economics, vol. 35, issue 1, 1-27
- Doan, T. A., Litterman, R. B. and Sims, C. A. (1984): *Forecasting and Conditional Projections Using Realistic Prior Distributions.* Econometric Reviews, 3(1), 1-100
- Girouard, N. et al. (2006): *Recent house price developments: the role of fundamentals"*, OECD Economics Department Working Papers, No. 475, OECD Publishing, Paris. Available at: http://dx.doi.org/10.1787/864035447847
- Glaeser, E.L. and Gyourko, J. (2007): *Arbitrage in Housing Markets*, NBER Working Papers 13704, National Bureau of Economic Research, Inc.
- Gupta, R. and Miller, S.M. (2009): *The Time-Series Properties on Housing Prices: A Case Study of the Southern California Market*, Working Papers 0912, University of Nevada, Las Vegas , Department of Economics, revised Dec 2009.
- Gyourko, J. and Voith, R. (1992): *Local Market and National Components in House Price Appreciation*, Journal of Urban Economics 32, 52-69
- Jud, G.D. and Winkler, D.T. *The dynamics of metropolitan housing prices*, Journal of Real Estate Research, vol. 23, no. 1-2, 2002, 29-45.
- Kuhn, M. and Johnson, K (2016): *Applied Predictive Modelling*, Chapter 8, pages 172-203.
- LeSage, J. P. and Pan, Z. (1995): *Using Spatial Contiguity as Bayesian Prior Information in Regional Forecasting Models*, International Regional Science Review, 18(1), 33-53
- Litterman, R.B. (1981): *A Bayesian Procedure for Forecasting with Vector Autoregressions.* Working Paper, Federal Reserve Bank of Minneapolis.
- Litterman, R. B. (1986). *Forecasting with Bayesian Vector Autoregressions ? Five Years of Experience.* Journal of Business and Economic Statistics, 4(1), 25-38.
- Malpezzi, S. (1999): *A simple error correction model of house prices.* The Center for Urban Land Economics Research, School of Business, University of Wisconsin, 975 University Avenue, Madison, Wisconsin 53705;

- Mayer, J.M. and Sommerville, C.T. (2000): *Land use regulation and new construction*, Regional Science and Urban Economics, Elsevier, Vol. 30(6), 639-662,
- Meese, R. and Wallace, N. (1991): *Non parametric estimation of Dynamic Hedonic Price Models and the construction of residential house price indices*, AREUEA Journal, Vol 19, No. 3;
- McCarthy, J. and Peach, R. W. (2004): *Are home prices the next bubble?*,Federal Reserve Bank of New York, 1-17.
- McDonald, J.F. and Stokes, H.H. (2011): *Monetary Policy and the Housing Bubble*, The Journal of Real Estate Finance and Economics, 46, 437-451
- McDonald, J.F. and Stokes, H.H. (2013): *Monetary policy, mortgage rates and the housing bubble*, Economics and Finance Research Vol. 1, 82-91
- Meen, G. (2002): *The Time-Series Behavior of House Prices: A Transatlantic Divide?*, Journal of Housing Economics, Elsevier, vol. 11(1), 1-23
- Mikhed, V. and Zemcik, P. (2009): *Testing for Bubbles in Housing Markets*: A Panel Data Approach, The Journal of Real Estate Finance and Economics, Vol. 38,Issue 4, 366-386
- Muellbauer, J. and Murphy, A. (1997): *Boom and bust in the UK housing market*, Economic Journal, Royal Economic Society, Vol. 107(445), 1701-1727
- Nerlove, M., (1971): *Further evidence on the estimation of dynamic economic relations from a time-series of cross-sections*, Econometrica 39, 359-382.
- Nordhausen, K. (2014): *An Introduction to Statistical Learning - with Applications in R by Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani* International Statistical Review, International Statistical Institute, vol. 82(1), Chapter 8, 303-335.
- Poterba, J. M. (1984): *Tax subsidies to owner-occupied housing: An asset-market approach.* The Quarterly Journal of Economics, 99(4), 729-752
- Poterba, J. M. (1991): *House Price Dynamics: The Role of Tax Policy and Demography*, Massachusetts Institute of Technology (MIT)

- Saiz, A. (2006): *Immigration and housing rents in American cities.* Wharton School, University of Pennsylvania and IZA Bonn
- Spencer, D. E. (1993): *Developing a Bayesian Vector Autoregression Model.* International Journal of Forecasting, 9(3), 407-421.
- Swamy, P.A.V.B. and Arora, S.S. (1972): *The exact finite sample properties of the estimators of coefficients in the error components regression models*, Econometrica 40, 261-275
- Van der Wal, E. and Tamminga, W. (2008): *Why the average dwelling purchase price is not an indicator*, Statistics Netherlands;
- Wallace, T.D. and Hussain, A.(1969): *The use of error components models in combining cross-section and time-series data*, Econometrica 37, 55-72.
- Welsh, C. E. (1993): *What caused the 1990-1991 recession?* Economic Review, Federal Reserve Bank of San Francisco.
- Wooldridge, J. (2010): *Econometric Analysis of Cross Section and Panel Data*, The MIT Press, Chapters 7.2,7.3,7.8.1,10 and 11.2;
- Wooldridge, J. (2015): *Introductory Econometrics*, 6th Edition, South-Western, Chapters 13-14;
- Zeileis, A. (2006): *Object oriented computation of sandwich estimators*, Journal of Statistical Software, Wirtschaftsuniversitaet Wien;