Band 90

Schriften zu Immobilienökonomie und Immobilienrecht

Sebastian Ertl

Three Essays on Estimating and Forecasting Residential Markets

IRE|BS
International Real Estate Business School
Universität Regensburg

Sebastian Ertl

Three Essays on Estimating and Forecasting Residential Markets

Die Deutsche Bibliothek – CIP Einheitsaufnahme
Sebastian Ertl
Three Essays on Estimating and Forecasting Residential Markets
Regensburg: Universitätsbibliothek Regensburg 2018
(Schriften zu Immobilienökonomie und Immobilienrecht; Bd. 90)
Zugl.: Regensburg, Univ. Regensburg, Diss., 2018
ISBN 978-3-88246-402-3

ISBN 978-3-88246-390-3

Verlag: Universitätsbibliothek Regensburg, Regensburg 2018
Zugleich: Dissertation zur Erlangung des Grades eines Doktors der Wirtschaftswissenschaften, eingereicht an der Fakultät für Wirtschaftswissenschaften der Universität Regensburg
Tag der mündlichen Prüfung: 06. November 2018
Berichterstatter: Prof. Gabriel Lee, Ph.D.
Prof. Dr. Kristof Dascher

Contents

List of Figures

List of Tables

Preface

One of the most fundamental human needs is housing. Therefore, residential real estate is one of the most important markets in almost every developed economy. Past has taught us that those markets are neither immune to unforeseen events or other external factors nor are they as predictable as some would make you believe. Therefore, knowing their fundamental functionality as well as the inherent subtleties is crucial. With this in mind, this dissertation focuses on three different aspects of estimating and forecasting residential markets.

One decisive cause of demand for housing is demographic and economic development. But other factors, like employment, income or the cost of living play an important role as well. As a consequence of the financial crisis, a debate emerged on the responsibility of central banks in maintaining price stability and on whether the existing monetary policy framework ensures rational price formation in real estate markets. Interest rates, amongst other things, were blamed to be the cause for the exceptional high fluctuations in house prices across many countries.

Therefore, the first essay *"The sensitivity of house prices under varying monetary regimes – The Nordic scenario"* (chapter 1) – written in collaboration with Marcelo Cajias – aims to examine whether there are differences between the long- and short-term relationship of house prices and interest rates and how the explanatory power of the different determinants can be decomposed.

The elasticity of house prices to monetary policy changes, e.g. via interest rates, is negative from a theoretical perspective and in the long-run. However, house prices adapt in the short-run dynamically to economic, financial, institutional and demographic factors. In this chapter, we focus on the role of monetary policy in contributing to the adjustment of house prices in the long- and short-term across the Nordic housing markets. Thus, we focus explicitly on the relationship between house prices and monetary policy – proxied by short-term interest rates –

in order to examine if house prices present a time-varying (dis-)continuous response to both expansionary and recessionary regimes.

We focus on the Nordic countries Denmark, Finland, Sweden and Norway as they present common similarities like education, health care and social services, but at the same time rather different financial and monetary conditions. While the monetary policy is partly linked to ECB's policy framework, regulatory decisions in the UK, the US and to a certain extent in Russia are of enormous relevance. Furthermore, the Nordic economic model is exposed to different exchange rate regimes. While Finland adopted the Euro and Denmark pegged the Danish Krone to the Euro, Norway and Sweden introduced flexible exchange rates. All in all, the Nordics offer a unique investigation set in order to explore the nature of house prices under varying monetary regimes.

After controlling for economic, institutional and demographic factors, our analysis comes to the (expected) result that housing markets across the Nordics respond negatively in regimes with an expansionary policy, obviously with some differences across the countries. However, our econometric models provide evidence that the impact of monetary shocks on house prices is – different as expected – not constant over time. This holds true especially since the beginning of the financial crisis and the expansionary monetary policy in Europe. When decomposing the explanatory power of the determinants of house prices, the results show that recessionary and expansionary policy regimes play a much more important role in the development of house prices in Finland, Sweden and Norway, than in Denmark. Furthermore, we conclude that the contribution of the individual factors such as short-term interest rates is not constant over time as well. Overall, we confirm that house prices are negatively affected in phases with expansionary regimes in the long-run, but provide evidence of unexpected anti-cyclical effects in the short-run. Consequently, the role of central banks has to be critically examined, since housing markets adjust unevenly to different monetary environments.

In the second essay *"Spatial effects and non-linearity in hedonic modelling – Will large datasets change our assumptions?"* (chapter 2), again co-authored with Marcelo Cajias, we study a different aspect of real estate: Price formation due to location. Location is one of the most important determinants for defining the value of property. But why does spatial heterogeneity matter? The locational immobility of real estate makes its price formation different from traditional commodities. Real estate prices reflect their explicit building attributes, neighbourhood characteristics and finally the share of directly available amenities. Since each region or sub-

market in a city provides a different set of local characteristics like green areas, public schools or shopping facilities, it attracts households according to their own personal preferences. The nearer a property is located to them, the higher (or lower – in case of negative attributes) the benefits for this household and therefore its willingness to pay.

Therefore, chapter 2 analyses the prediction accuracy and explanatory power of three different approaches based on a large dataset with more than 570,000 asking rents across 46 residential rental markets in Germany. This is – to the best of our knowledge – one of the largest datasets used for spatial real estate analysis.

The choice of the functional form in hedonic regression models is crucial when explaining rents within a certain real estate market. Empirical research has thoroughly attested that traditional hedonic models fail to explain the variation of rents accurately, when excluding spatial effects or non-linear relationships. Therefore, the estimation of hedonic regression models has indeed grown substantially over the last years integrating new approaches for modelling spatial heterogeneity, which is essential in the explanation of real estate prices across space. With the list of spatial estimation techniques being very extensive, the Geographically Weighted Regression (GWR) has established itself as a widely used method that expands the restrictive traditional Ordinary Least Squares (OLS) regression by considering spatially varying effects. However, semi-parametric methods like the Generalized Additive Model (GAM) capture spatial effects based on smooth functions and expand the traditional hedonic model by identifying latent nonlinear effects. Since the main goal of any hedonic model is the reduction of misspecification in the estimated coefficients, the GAM model allows covariates to take a nonlinear functional form in order to reduce the error variance and thus enhance the model quality. With GAM models being popular in natural sciences, their usage in the empirical real estate research is very limited.

The results of chapter 2 show that the GWR method, which is a great tool to explore regional factors driving rents within a certain market, is outperformed by the GAM and OLS models. Regarding OLS and GAM, it turns out that the differences in out-of-sample prediction accuracy are not substantial. This results align with several findings of the considered literature. Against expectations, the OLS approach seems to be an equal alternative to (semi-) parametric models. Despite the low discrepancy, our findings match with the results of Mason and Quigley (1996, p. 384) which conclude that the differences between OLS and GAM *"are rather small, though statistically significant"*.

Both of the previous chapters show how real estate prices can be estimated and predicted on the basis of fundamentals and location, respectively. Those variables are specific, quantifiable characteristics of certain properties. But what if there are price movements that can not be explained by changes in fundamental factors? A classical example would be stock markets, where prices frequently react to possible, maybe speculative, future events that obviously do not reflect the current circumstances. As this also applies to real estate markets, we can assume that psychological effects can have an impact on housing prices. But how would one measure this "sentiment"? Sentiment indicators try to capture the "noise" in various markets that cannot be represented by fundamentals, like for example fears or hopes. There are lots of traditional survey-based sentiment indicators, but they might possibly be hard to access, for the wrong region or simply not sufficiently up to date.

In 2006, Google launched a new service called *Google Trends* that allows users to see the interest of all other Google users on certain search terms. Google Trends updates its data so fast that it can be queried on a monthly, weekly, daily or hourly basis and even in real time. The geographical location can be restricted to countries, states and even large cities and there are over 1,000 categories to narrow down the results even more. By doing so, Google offers virtually unlimited, instantaneously available, spatially and textually adjustable and, in addition, free data. This type of data conquered its position in nearly all economic fields, serving as a highly adjustable sentiment indicator that can be used, inter alia, for nowcasting and short-term forecasting. Although Google Trends data can be accessed already since 2008, many interpretation and usage misunderstandings can be found amongst the literature.

Therefore, the third essay *"Pitfalls of using Google Trends data in empirical research – What do microwave baked potatoes tell us about U.S. housing markets?"* (chapter 3) will first give an overview of what Google data actually is and where the potential pitfalls are. Real estate markets appear to be particularly well-suited for search volume related studies, as the "products" of this market involve a large financial commitment, which demands an extensive information gathering process. To the best of my knowledge there is no other paper specifically dealing with the potential pitfalls and disadvantages of Google Trends data in real estate analysis. Secondly, I conduct an empirical analysis to find out, whether the results are still in line with the literature after accounting for those difficulties. For this task, the usual approach in the existing literature would be to simply compare Google models to a baseline model. However, instead of demonstrating only how a very simplistic baseline model can be outperformed, I am more

interested in seeing how the resulting models can compete against comparable "standard" models.

The results show, as expected, that adding search volume data to the estimations leads to an improvement regarding model fit and helps reducing the forecasting errors when compared to a baseline model. However, they also show that there are equally specified "standard" models that fulfill the same requirements and can be used in the same way as the Google models, even with slightly better results.

Especially, when dealing with a "new" type of data, one should know where pitfalls lie and where attention has to be paid. This is not to say that one should not use Google data. In fact, if urgently needed data is not yet available, search volume data can become very useful in terms of delivering meaningful proxies. When monitoring market movements, the delayed publication of various important variables makes nowcasting a necessary task for many researchers. However, search volume data should not be used for the sake of itself. Instead of contrasting it against a simplistic baseline model, it would be more interesting to see how Google models perform compared to or in combination with proven methods that are actually used for this type of task. The results of this study indicate that adding search volume data is not a silver bullet, but at least a useful complement if other data is absent.

Chapter 1 and 2 of this dissertation are published as articles in peer-reviewed academic journals. A slightly adapted version of chapter 3 is accepted for publication.

Chapter 1

The sensitivity of house prices under varying monetary regimes

The Nordic scenario

This chapter is joint work with Marcelo Cajias† and published as:

Marcelo Cajias, Sebastian Ertl, (2017) "The sensitivity of house prices under varying monetary regimes: the Nordic scenario", *International Journal of Housing Markets and Analysis*, Vol. 10 Issue: 1, pp. 4-21, DOI 10.1108/IJHMA-12-2015-0074

Abstract

This paper aims to examine whether there are differences between the long and short-term relationship of house prices and interest rates. The elasticity of house prices to monetary policy changes, e.g. via interest rates, is from a theoretical perspective and in the long-run negative. However, house prices adapt in the short-run dynamically to economic, financial, institutional and demographic factors. In this paper, the authors confirm the aforementioned elasticity for the Nordic housing markets, but provide evidence of drastic deviations from the negative relationship. This is done by employing rolling regressions in search for time-varying betas. The empirical results show that recessionary and expansionary policy regimes play a much more important role in the development of house prices in Finland, Sweden and Norway, than in Denmark. Further it is shown that the relationship between house prices and monetary policy is discontinuous over time, with large deviations from the long-term beta during the last decade. This holds true especially since the beginning of the financial crisis and the expansionary monetary policy in Europe.

† PATRIZIA Immobilien AG, Fuggerstraße 26, 86150 Augsburg, Germany
The authors especially thank PATRIZIA Immobilien AG for contributing the dataset and large computational infrastructure necessary to conduct this study. All statements of opinion reflect the current estimations of the authors and do not necessarily reflect the opinion of PATRIZIA Immobilien AG or its associated companies.

1.1 Introduction

The recent fluctuations in house prices across many European countries have led to an active discussion on the role of central banks in maintaining price stability and on whether the existing monetary policy framework ensures rational price formation. Since house prices are characterized by relatively long adjustment phases, the nature and timing of macroeconomic mechanisms and policies are important when counteracting cyclical price movements. Central banks and governments monitor therefore the development of house prices intensively as booms and busts in housing markets have shown over the last decades to have a large impact on households' debt position, banks' equity ratios and finally on countries' aggregated demand. In this context, an expansionary monetary regime – as the one existing at the moment – causes an immediate fall in interest rates and government bond yields filling financial markets with liquidity. Under such financial conditions, theory would predict a rapid increase in the attractiveness of real estate assets and consequently lead to rising real estate prices. In order to analyze the sensitivity of house prices to changing interest rates regimes, we provide evidence of a time-varying discontinuous response of house prices to expansionary and recessionary monetary regimes in the Nordic housing markets. We focus on the Nordic countries Denmark, Finland, Sweden and Norway as they present a unique financial and monetary environment based on a solid economic output and a stable domestic demand. Furthermore, when decomposing the explanatory power of the determinants of house prices over time, we conclude that the contribution of the individual factors such as short-term interest rates is not constant over time.

The remainder of this paper is organized as follows: Section 1.2 summarizes the literature concerning the relationship between house prices and interest rates as well as the development of this relationship over time. Section 1.3 describes the data and the econometric approach and how relative contributions of the determinants can be calculated in such models. The estimation results are presented in section 1.4. Finally, section 1.5 concludes.

1.2 Literature Review

According to Ma and Liu (2010) there are four general approaches to analyze house price dynamics: The hedonic model, the repeated-sales method, the ripple-effect model and the

fundamental model. The first one assumes the house price to consist of its characteristics and neighborhood information. The second one relies on actual price data over time. The ripple-effect model suggests, that the price formation is caused by shocks in the same or other regions of the housing markets. *"The fundamental model is based on the idea that the housing market variations are driven by the economic factors, such as incomes, gross domestic products, rents, mortgage rates, inflation rates, supplies and demands and so on"* (Ma and Liu, 2010, p. 6). In view of the research question of analyzing the relationship between interest rates and house prices, one has to go with the fundamental approach.

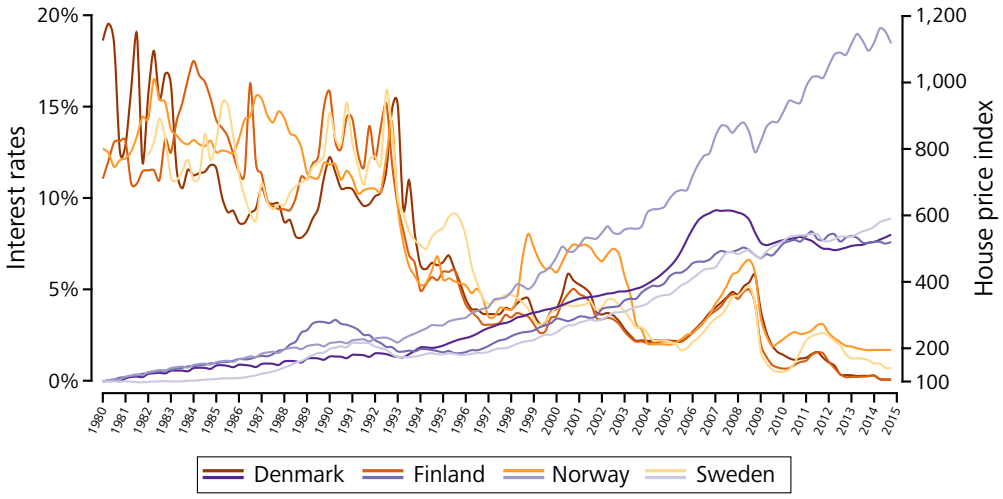
House prices and interest rates are known to be linked in some way. This holds at least for the theory. And there are many theories and models one could argue with why and how the two are related, but this is not the aim of this paper. Instead we follow the line of Demary (2012) as he does not compare different models or test economic theories, but rather gets a deeper understanding of how the transmission channels work and how the interplay of macroeconomic variables evolves over time. The work of Demary (2012) describes, among other things, the effect of a direct interest rate shock on house prices: Imagine a scenario where a (central) bank rises the money market rates. Theory says this would result in higher mortgage rates, since this two markets are strongly connected. Again, this leads to an increase in financing costs which lowers the demand for housing and, as a consequence, the house prices (see Demary, 2012, p. 217). The same transmission mechanism is described in Nastansky (2012) the other way round: A decrease in interest rates would make money market products less attractive to private and institutional investors. Therefore, the demand as well as the prices for other alternative investments such as stocks or real estate would go up (see Nastansky, 2012, p. 167). Consequently, it follows that there should be a negative relationship between house prices and interest rates. A look at figure 1.1 confirms this assumption as the correlation between the house price index and short-term interest rates in the Nordics is below -0.75 in each of the observed countries.[1]

Cardarelli et al. (2008) present a similar argumentation, but go one step further. According to the authors, the connection between monetary policy and the real estate market has changed substantially over time. The wide variety of financing possibilities drives the competition between lenders, which will lead to faster interest rate adjustments. Again, because of the greater range of credit products and the access to these products, together with a relaxation of credit constraints, households or firms are able to finance higher proportions of their

1 A definition of the variables can be found in section 1.3.

investments through credits. Therefore, it is possible that changes in interest rates could have a greater impact on house prices. Their results are mainly consistent with the above mentioned assumption, since they find that monetary policy shocks tend to have bigger effects in countries where housing finance markets are more developed and competitive (see Cardarelli et al., 2008, pp. 118-119, pp. 126-127).

Figure 1.1: House price indices and short-term interest rates development



Note: House prices (index: 1980|Q1=100) (purple lines) on the right axis, short-term interest rates [%] (orange lines) on the left axis.

But what else drives house prices? Nastansky (2012) classifies three groups of determinants: economic, demographic and institutional factors. An example for economic factors are GDP, interest rates or disposable income. Variables like population growth, urbanization or household size would refer to the second group. Finally the financial system, tax legislation or state subsidies fall under the third category (see Nastansky, 2012, p. 169). In this context the Nordic countries – Finland, Denmark, Sweden and Norway – offer a valuable scenario to investigate the effects of the three different determinates on house prices, since they present common similarities but at the same time enormous dissimilarities. In fact, Lujanen (2004, p. 5) states that *"they have similar policies in areas like education, health care and social services, and policy in these areas is based on shared fundamental values also relevant to housing policy. However, the Nordic countries actually display distinctive national differences in many important areas of housing policy"*. The Nordics constitute a group of countries characterized by a unique territorial structure and a heavily polarized population in spatial terms. In contrast to many European countries, they present a remarkable monocentricity with regard to the distribution of their residents, which emphasizes the importance of housing markets in an urban

context. During the last four decades the Nordics faced both a fall in fertility rates and a remarkable increase in life expectancy, due to internal and external migration, leading overall to a negative biased demographic structure. Nevertheless they are in a unique position in a European context as their population will increase on average by 7 percent until 2030 or 0.5 percent p.a. following Oxford Economics; a development not seen anywhere else in Europe.[2] In contrast, the Nordic economic model clearly presents differences across the countries, but one of the main common features is the comprehensive concept of a "welfare state". High taxes ensure an efficient transfer of public services to households in order to ensure and maintain welfare. Beside this, public and private expenditure on human development, education and R&D are extremely high. Finally, the regulation in the labor market via labor and employment associations ensure high employment and short unemployment levels.

The Nordics' financial and monetary conditions, however, are fairly different. While the monetary policy is partly linked to ECB's policy framework, regulatory decisions in the UK, the US and to a certain extent in Russia are of enormous relevance. Furthermore, the Nordic economic model is exposed to different exchange rate regimes. While Finland adopted the Euro and Denmark pegged the Danish Krone to the Euro via ESM-II-Mechanism, Norway and Sweden introduced flexible exchange rates in order to benefit from an independent monetary framework that allows free capital mobility (see Schewe, 2015). See Lujanen (2004) for a very detailed discussion about the similarities and dissimilarities of the Nordic housing markets. All in all, the Nordics offer thus a unique investigation set in order to explore the nature of house prices under varying monetary regimes.

With this in mind, the question arises how to estimate an econometric model that explains the influence of interest rates and further determinants on the development of house prices. A popular approach for this kind of problem is a vector autoregressive (VAR) model with impulse responses. Demary (2012) for example, estimates a VAR model for 10 countries with quarterly data from 1970 to 2005. The author finds that an interest rate shock has a negative effect and explains about 11 percent of the house price variation. He argues that rising interest rates lead to a deteriorated financing situation which decreases the demand for housing and at the same time that this interest rate shock has a negative effect on overall output which intensifies the impact on house prices. Tsatsaronis and Zhu (2004) also estimate a VAR model with quarterly

2 Following the last censuses across many European countries, demographic forecasts changed significantly. In the case of Germany as an example, the last census revealed a population count error of ca. -2 percent or ca. two million inhabitants. In Spain for instance, the demographic forecasts worsen drastically due to rising emigration in response to economic contraction.

data from 1970 to 2003. Their findings are very similar to Cardarelli et al. (2008), Adams and Füss (2010) or Calza et al. (2013) in so far as the impact of short-term interest rates on house prices is shown to be much stronger in countries that use mostly variable interest rates rather than fixed terms. However, Tsatsaronis and Zhu (2004) identify inflation as the largest driver of house prices. They ascribe this due to the fact that real estate is a consumption good but at the same time it acts an investment vehicle with certain liquidity restrictions and high transaction costs. But according to their model monetary variables, like short-term rate, credit-growth or yield-spreads are still able to explain about 30 percent of the house price variation.

All of the papers mentioned above focus on different specifications of VAR models with impulse responses to get the relative contribution of several determinants in explaining house prices. There are, however, certain difficulties or requirements that must be fulfilled when estimating these models. Miles (2014) performs a simple OLS regression on quarterly data from 1973 to 2011 in the US market regressing both the Federal Funds Rate (FFR) as a measure of monetary policy and the 30-year mortgage rate as a long-term interest rate on the house prices. The main difference is the simple presumption that the relationship between house prices and monetary variables could have changed over time. Goodhart and Hofmann (2008), again estimating a panel VAR model, indeed find a significant effect of an interest rate shock on house prices in the housing market, but also assume that the effect of interest rate shocks is larger in times of booming house prices than otherwise. Even though their results are not statistically significant, this is interesting because the authors also suspect a change in the fundamental co-movements between monetary policy and housing markets. Miles (2014) tries to solve this problem by splitting his data in different subsamples to explore whether there are changes regarding the coefficients or significance of the monetary variables. The main difficulty is an objective choice for the break point, which the author attempts to overcome with break point tests. The estimations reveal *"that long-term interest rates have a larger impact on house prices than the FFR, and that the impact of the FFR has fallen into irrelevance in recent years"* (Miles, 2014, p. 56). This stands in contrast to the results of McDonald and Stokes (2013) as they find that the FFR has negative effects on house prices and that the impact has risen over time (see also Miles, 2014, p. 42). Miles (2014), however, calls their interpretations and results into question due to the fact they only use one single regressor in their estimations. One should like to mention that Miles (2014) himself also uses only two regressors, so the results could be questioned as well. Zietz (2012) states that empirical research has to go further than theory. When thinking of a theory one can sure use the 'ceteris paribus' assumption to black

out all other effects that are not important for this topic. This must not be done in an empirical analysis, because the variable being explained is not only depending on this one variable of interest, but on (many) other variables.

Nevertheless, the main question is whether there are fundamental changes in the relationship over time. Just like Zivot and Wang (2006, p. 313) state: *"the economic environment often changes considerably, and it may not be reasonable to assume that a model's parameters are constant."* The classical models yet assume constant parameters over the whole estimation period, so they cannot account for this phenomenon. There are, however, certain econometric techniques that allow parameters to change over time. Guirguis et al. (2005, p. 33) state that particularly in housing markets it is necessary to allow the parameters to vary over time, as there were *"major structural changes and economic fluctuations"* over the last decades and therefore one must account for (sub-)sample instability.[3] They estimate several econometric models, where the coefficients are allowed to change over time and generate forecasts to see which one fits the data best.[4] One thing to mention is that not all of the models are "truly" time varying, but rather rolling versions of constant parameter models. The results show that a specification of the Kalman Filter and rolling GARCH Models outperform all other considered models.[5] A study with similar result is the one of Brown et al. (1997). They compare a constant parameter model (CPM), a recursive OLS, a VAR and a Time Varying Coefficients (TVC) model with Kalman Filter. They conclude that the TVC model outperforms the other constant parameter models.

Leblanc and Bokreta (2009) compare a rolling OLS model and a Kalman Filter approach for the same reason. A direct comparison of the coefficients is difficult, because the two methods differ in their statistical assumptions and estimation approach; the former estimates rolling windows imposing a linearity in the functional form, whereas the latter is a recursive algorithm. The authors conclude that the Kalman Filter estimates are more robust than those from rolling OLS. However, when it comes to forecast accuracy there is no clear winner between the two methods. Only regarding the reaction to changes in the observed data, the Kalman Filter is one step ahead, but it should be recalled that the Kalman Filter is – as the authors say – "a high technology" and therefore not straightforward to apply (see Leblanc and Bokreta, 2009, p. 13).

3 Structural changes in the context of housing markets also refer to changes in the regulatory framework of the private rented sector. For a detailed discussion see Monk et al. (2012) and PATRIZIA research (2015).
4 Rolling VECM, rolling AR, rolling GARCH, Kalman Filter with random walk, Kalman Filter with autoregressions.
5 GARCH: Generalized Autoregressive Conditional Heteroskedastic Model.

In this context, we focus in a first step on a methodology that allows the main determinants of house prices to vary over time as in accordance to Miles (2014) or Leblanc and Bokreta (2009). To do so, we estimate rolling regressions with varying windows for each of the Nordic countries and draw our conclusions with regard to the stability of the relationships between house prices and their fundamental drivers. In a second step we concentrate on the relative explanatory power of the fundamental drivers of house prices within the rolling regression context.

1.3 Data Description and Econometric Models

1.3.1 Data

The data used comes from Oxford Economics via Thomson Reuters Datastream on a quarterly basis and reaches from the first quarter in 1980 to the fourth quarter in 2014.[6] The variables gathered are the house price indices as defined by the official statistical bureaus, short-term interest rates based on three-month money markets, real GDP in local currency, unemployment rate, real personal disposable income, construction activity, (working-) population and harmonized consumer price indices. House price indices consists of a generic measurement for dwellings in the main metropolitan areas of the respective country based on the data collection methods defined by the national statistical offices. The indices account the price development of existing dwellings rather new construction. Most of these variables are expressed in a year-on-year (yoy) growth rate. The unemployment rate, the short-term interest rates and the construction activity are expressed in a yoy (year-over-year) difference. All variables in table 1.1 are stationary according to the augmented Dickey–Fuller and Phillips–Perron tests.

The table shows the descriptive statistics of the variables used for the models. We chose those variables in accordance to Nastansky (2012) to focus on main macroeconomics developments including the labor and construction markets as well as the demographic growth, households income and GDP as main determinants for house price developments. As the data capturing the construction market is not homogenously defined in the Nordics (and also across Europe), we capture construction activities in the respective market as defined by the national statistical offices. While Denmark and Finland capture the number of dwellings started in the housing market, Norway and Sweden focus on the number of dwellings with a building permission.

6 Database: Oxford Economics via Reuters Eikon (up to 2015: Thomson Reuters Datastream).

Table 1.1: Descriptive statistics

	Denmark		Finland		Norway		Sweden	
Variables	mean	sd	mean	sd	mean	sd	mean	sd
HP [% yoy]	5.19	6.57	5.13	8.25	7.40	4.75	5.52	6.60
GDP [% yoy]	1.64	2.38	2.08	3.42	2.47	2.14	2.21	2.41
Construction activity [% yoy]	0.16	25.93	0.83	22.45	0.71	17.29	4.38	32.00
Pers. disp. income [% yoy]	2.17	5.35	2.72	3.93	3.02	2.84	1.87	2.54
Short-term interest rate $\Delta 4$	-0.50	2.11	-0.36	2.04	-0.31	1.87	-0.39	1.97
Unemployment rate $\Delta 4$	-0.05	1.11	0.11	1.56	0.05	0.67	0.15	1.31
Unemployment rate	7.88	2.68	8.30	3.54	3.68	1.16	6.68	2.76
Population [% yoy]	0.28	0.19	0.39	0.11	0.66	0.32	0.44	0.28
Working population $\Delta 4$	0.01	0.32	-0.11	0.20	0.08	0.13	-0.02	0.20
Working population [% yoy]	0.29	0.41	0.22	0.31	0.78	0.30	0.42	0.25
CPI [% yoy]	3.05	2.45	3.25	2.87	3.72	3.12	3.48	3.50

Note: Stationarity test for all variables rejected based on ADF/PP-Tests.

Both measurements may not capture the same effect, they express, however, the magnitude in the respective constructions markets accurately. The yoy-growth in house prices is fairly constant at 5 percent across all four countries, except in Norway with 7.4 percent p.a. The standard deviation shows that the house price yoy-growth rates are – after the construction growth rate - the second most volatile of all variables considered. The sample correlations can be found in table 1.A.1.

The fundamental data is pretty stable across the Nordics. The mean GDP growth is about 2.1 percent p.a., the inflation (CPI) about 3.4 percent p.a. and the personal disposable income growth comes to 2.5 percent p.a., construction growth is about 1.5 percent p.a. and the population growth 0.4 percent p.a. As expected, the unemployment rate reveals bigger deviations, since the employment level in the working age population is high within the EU15, except for Finland.[7] The range reaches from 7.9 percent in Denmark to 3.7 percent in Norway, resulting in a mean of 6.6 percent across the countries. With 4.38 percent there was a remarkably high growth in construction activity in Sweden, compared to the other Nordic countries.

1.3.2 Econometric Models

The fundamental model is a simple linear regression in the form of:

$$\mathrm{HP}_t = X_t\beta + \mathrm{IR}_t\beta_{\mathrm{IR}} + \varepsilon_t, \tag{1.1}$$

7 EU15: Austria, Belgium, Denmark, Finland, France, Germany, Greece, Ireland, Italy, Luxembourg, Netherlands, Portugal, Spain, Sweden, United Kingdom.

where HP_t are the yoy growth rates of house prices as response variable in each country, IR_t the differentiated interest rates in basis points and $X_t = (1, x_1, \ldots, x_p)$ with $\beta = (\beta_0, \beta_1, \ldots, \beta_p)^T$ includes the fundamental explanatory variables. Detailed model specifications can be found in section 1.4. A rolling regression, as mentioned in section 1.2, is a sequence of estimations of one fundamental model, each with a different sample period. There are different forms of rolling regressions. One way is to set a starting point for the first regression and let the sample period increase with each estimation. The other way is to set a fixed window, say 10 years, and roll this window over the whole sample. In this paper the latter is employed. But, like Leblanc and Bokreta (2009, p. 7) said: *"the problematic is to find the optimal rolling window"*. This is true and important since there are a lot of factors to keep in mind: the frequency of the available data, the data itself and especially the estimation purpose. Miles (2014) for example, splits his data in two subsamples, from 1960 to 1979 and from 1970 to 1990. This means the two windows are a period of about 20 years each, which means approximately 80 observations due to quarterly data. Likewise Brown et al. (1997) split their data in two subsamples covering 13 and 11 years. Again with quarterly data this gives 52 and 44 observations respectively. But it is questionable whether two subsamples make a rolling regression. Guirguis et al. (2005) therefore use an increasing window. They start with 40 observations and increase the window every quarter by 1 so that the second estimation covers 41 observations and so on.[8] Leblanc and Bokreta (2009) used a 3-year window with 36 observations because of monthly data. The shorter window is due to the fact that the authors intend to forecast regime switches in the markets. This has to be done in the short-term and therefore a larger estimation window would possibly bias the results. Swanson (1998) - again with monthly data – uses different window specifications, namely a 10-year, a 15-year and an increasing window.

Since the frequency of our data is quarterly, windows smaller than 10 years or 40 observations respectively, might not be reasonable. In contrast a window larger than for example 25 years and therefore 100 observations would not be useful as well, as there are just 140 observations in total. The number of rolling regressions that could be estimated might be too small to get any expressive results. Nevertheless, in this paper many different windows, reaching from 10 to 30 years, were tested and the 15 years or 60 quarters window turned out to be the most suitable with regard to the stability of the results and the overall model inference. The purpose of this rolling regressions is twofold. First it is possible to evaluate the variation in the coefficients over time. If the basic assumption of a linear regression would be fulfilled,

8 Also known as a recursive regression, as seen in Brooks and Tsolacos (2010, p. 185).

the coefficients of most of the rolling regressions and those from the model over the whole period should be nearly the same. There is no such thing as time varying coefficients in the context of linear OLS models. However, if there are any significant changes in the coefficients, either in magnitude or sign, this could be an indication of a structural break in the relationship between the variables. The second issue, when estimating rolling regressions, is the comparison of the contributions of the single variables to the overall explanatory power. This is very similar to an often used technique from VAR models called "variance decomposition". This way it can be checked whether the share of the explained variance of a particular variable changes over time. In this context Groemping (2006, p. 1) states: *"Relative importance refers to the quantification of an individual regressor's contribution to a multiple regression model"*. If and only if all regressors in a multivariate model are uncorrelated, the R^2 is the sum of all R^2 from single, univariate estimations with each regressor of the multivariate model. This could then be seen as the contribution of the variable to the whole model. But, thinking realistically, this won't happen very often, so Groemping (2006) presents six different methods to overcome that problem from which two are recommended. Regarding the other (not recommended) four methods, there are three major problems: First, the decomposition should sum up to the total R^2. Second, there should be no negative contributions. The two methods recommended overcome these first two problems. The third difficulty is a constant contribution while changing the order of the model. In an analysis of variance (anova) sequential sums of squares are calculated. The division of this sequential sum of squares and the total sum of squares reveals the sequential R^2 and the contribution to the total R^2 by each variable. But the key is the word "sequential". Here the order of the variables does matter. If the order of determinants changes, the individual contributions might change significantly.[9] The two recommended metrics built up on the sequential sum of squares, but manage this problem by calculating the average contributions of all possible orders. This is be done either by simple unweighted averages (so called *"lmg"*-method) or averages with data-dependent-weights (so called *"pmvd"*-method), as explained in Groemping (2006, p. 8).

Like mentioned in section 1.2, many of the models used in the literature are VAR models, mostly with some kind of impulse responses to check for the effects of different variables. We do not go with a VAR approach because of the following consideration: VAR models are throughout atheoretical, meaning that they are a purely statistical tool and there are much less possibilities to control for the model specification, since *"let the data decide"* is one of the

9 For more detailed explanations see Groemping (2006).

principles for VAR modelling (see Brooks and Tsolacos, 2010, p. 352). Furthermore, the results and the robustness of the results are highly depending on an accurate transformation of the data and the order in which the determinants go into the model, especially when varying the time horizon. Such kind of model has its advantages for several topics for sure, but won't help us answering the question how the influence of a certain variable is evolving over time.

1.4 Econometric Results

1.4.1 Fundamental housing equation over entire sample

We focus on the sensitivity of house prices to different interest rate regimes and test for the presence of time-varying relationships within a rolling regression framework in the Nordics. Furthermore, we decompose the explanatory power of the house prices equation over time in order to analyse the contribution of individual regressors to the yearly change of house prices. Our fundamental equation regresses macroeconomic variables on the yearly growth rate of house prices from 1980 until 2014 after controlling for necessary condition of stationarity. In order to avoid autocorrelation and heteroskedasticity bias we estimate the standard errors via HAC-variance-covariance-matrix using the procedure suggested by Zeileis (2004).[10]

Table 1.2 shows the detailed model specification for each country. The determinants may differ in the transformation or the lag structure, but the underlying variable occurs in each of the models. The parameterization of each models was performed by stepwise minimization of AIC information criteria. The AIC criterion penalizes the number of regressors and the goodness of the model at the same time, in order to represent the original data generating process of the underlying responses. After parameterizing the models, we estimate the variance inflation factor (VIF) for each of the models to control for latent multicollinearity issues and found VIF-values below the critical values.

The individual regressions for Denmark, Finland, Norway and Sweden in table 1.3 show that over the last 34 years a growth in the macroeconomic output was positively related with an increase in house prices. Thus, a contemporaneous macroeconomic shock of one percent leads for example in Denmark to a rise in house prices of about 1 percent ceteris paribus, whereas in Norway, Sweden and Finland the price-output elasticity is below 0.7 percent. When looking at

10 Heteroskedasticity and autocorrelation consistent (HAC) estimators.

the effects of the construction activity, the coefficients show a positive price elasticity across the Nordics, as an expansion in the supply of dwellings leads ceteris paribus to a contemporaneous increase in house prices. The results show that short-term interest rates – as a proxy for monetary policy framework – are negatively related to house prices across the Nordics, whereas the effect over the last 34 years in Denmark and Sweden is against expectations positive, pointing to an unelastic housing demand to varying financing costs.

Table 1.2: Model specification

Variables	Denmark	Finland	Norway	Sweden
GDP [% yoy]	⊠	⊠	⊠	⊠
Construction activity [% yoy]	□	□	⊠	⊠
Construction activity	⊠	⊠	□	□
Pers. disp. income [% yoy]	□	⊠	□	□
Pers. disp. income [% yoy]$_{(t-4)}$	□	□	□	⊠
Pers. disp. income [% yoy]$_{(t-8)}$	⊠	□	⊠	□
Short-term interest rate $\Delta 4_{(t-2)}$	□	⊠	⊠	□
Short-term interest rate $\Delta 4_{(t-4)}$	⊠	□	□	□
Short-term interest rate $\Delta 4_{(t-8)}$	□	□	□	⊠
Unemployment rate $\Delta 4$	□	⊠	□	□
Unemployment rate	⊠	□	⊠	⊠
Population [% yoy]$_{(t-4)}$	□	⊠	□	□
Working population $\Delta 4$	□	□	□	⊠
Working population [% yoy]$_{(t-4)}$	⊠	□	□	□
Working population [% yoy]$_{(t-8)}$	□	□	⊠	□
CPI [% yoy]	□	□	⊠	□
CPI [% yoy]$_{(t-4)}$	⊠	□	□	□
CPI [% yoy]$_{(t-8)}$	□	⊠	□	⊠

Note: Model parameterization for equation 1.1 based on the minimization of AIC information criterion.

While the estimated elasticity varies in dependence of the lag structure it is ca. -0.7 for Finland and of ca. -0.6 for Norway. Thus, a lagged expansionary shock is transmitted to a greater monetary base and consequently to falling interest rates. In this economic environment house prices rise as financing costs and consequently the demand for housing also rises. Over the last 34 years this relationship holds for the Nordics excluding Denmark and Sweden, where a rise in short-term interest rates of one percentage point has been accompanied with increasing house prices. Looking at the effect of the unemployment rate on house prices, the results show the expected coefficients, again excluding Denmark. Thus, a contraction in the labor market supply is associated with a decrease in house prices holding other fundamental factors fixed. There exist, however, strong differences across the countries, as a 1 percentage point increase in unemployment in Sweden leads to a fall in house prices of ca. 1.8 percent. In contrast, a

deterioration in Finish labor markets of the same magnitude is disproportionally transmitted into house prices as the coefficient is close to 2.4 percent.

Table 1.3: Regression results

Variable	Denmark	Finland	Norway	Sweden
Intercept	−7.37	−5.92	16.27	22.55
	(−5.88) ***	(−2.49) *	(5.67) ***	(18.51) ***
CPI	−1.32	0.26	−0.37	−0.90
	(−5.22) ***	(1.94) .	(−2.18) *	(−8.73) ***
GDP	1.02	0.35	0.54	0.64
	(6.46) ***	(2.23) *	(3.57) ***	(4.39) ***
Construction activity	0.37	5.68	0.03	0.02
	(1.72) .	(9.55) ***	(1.82) .	(2.10) *
Pers. disp. income	0.16	−0.31	0.49	−0.90
	(2.26) *	(−3.29) **	(3.75) ***	(−7.49) ***
Population	−14.26	−21.10	−4.50	6.29
	(−8.62) ***	(−5.62) ***	(−3.18) **	(4.09) ***
Short-term interest rate	0.29	−0.71	−0.58	0.36
	(1.89) .	(−3.60) ***	(−2.72) **	(1.95) .
Unemployment rate	2.11	−2.38	−1.97	−1.79
	(8.37) ***	(−7.41) ***	(−4.86) ***	(−14.45) ***
Adjusted R^2	64.44	83.71	41.58	80.11
AIC	714.99	675.38	689.35	607.43

Note: Significance codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1. Sample from 1980|Q1 to 2014|Q4. Standard errors corrected for auto-correlation using a robust variance-covariance matrix estimator. Parameterization of the models as defined in table 1.2.

Finally, the explanatory power across different specifications are robust and stable. In case of Finland and Sweden, the adjusted R^2 is above 80 percent, whereas 64 percent and 42 percent of the variation of house prices are explained in the case of Denmark and Norway respectively. In order to account for structural breaks the models are tested for structural changes and all the models are re-estimated in a rolling regression framework in the next section.

1.4.2 Fundamental housing equation over rolling cycles

As argued by Miles (2014), the relationship between house prices and fundamental factors may change over time due to vast changes in the financial environment, institutional factors and especially structural breaks, e.g. the introduction of the Euro or relevant shifts in investors' risk-aversion due financial collapses or regulatory changes in the private rented sector. Table 1.4 shows the CUSUM-test, which accounts for structural changes in the development of the response over time. To see whether this is just a problem of the estimation over the whole time

period, the sample was divided in two parts (1980/Q1 to 1996/Q4 and 1997/Q1 to 2014/Q4) and again tested for the possibility of structural breaks. As seen, the results indicate that there might be most likely a problem with the parameter stability and therefore the estimation of separated (rolling) equations should be considered.[11]

Table 1.4: Structural break test

	1980-2014	1980-1996	1997-2014
Denmark	0.79	0.71	1.19
	(0.15)	(0.24)	(0.01) **
Finland	0.70	0.76	1.55
	(0.25)	(0.18)	(0.00) ***
Norway	1.10	0.84	0.40
	(0.02) *	(0.11)	(0.84)
Sweden	1.72	0.82	0.78
	(0.00) ***	(0.12)	(0.15)

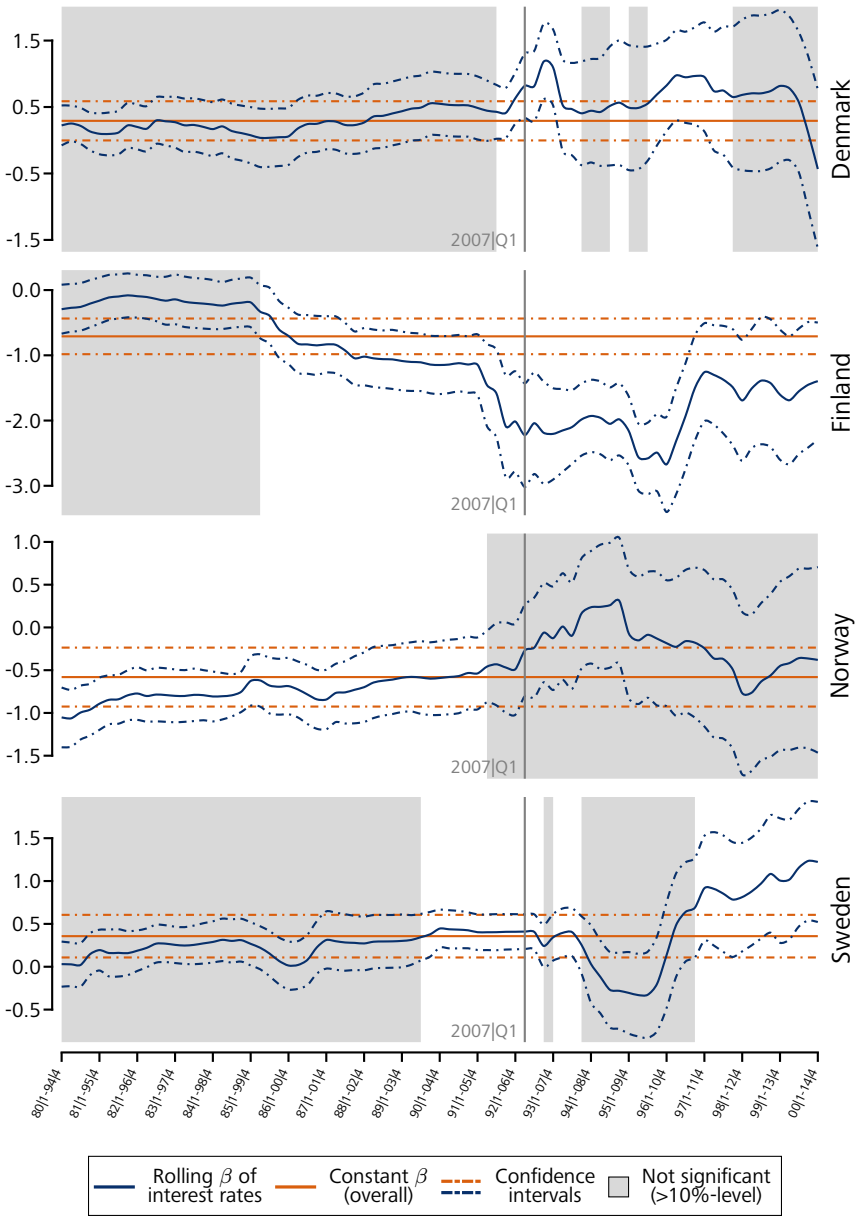
Note: Recursive CUSUM Test. Test statistics with p-values below.

One way to deal with this phenomenon is the introduction of time-dummies in the fundamental equation, which partly fail in capturing cyclical movements and decrease the degrees of freedom dramatically. Another possibility is the estimation of rolling equations. Since we focus on the specific effects of monetary shocks on house prices proxied by short-term interest rates, we re-estimate equation 1.1 respectively the models from table 1.3 in a rolling context including 15 years or 60 quarterly observations, rather than assuming a certain structural breakpoint based on some CUSUM process.

The results of the rolling regressions in figure 1.2 show that the fundamental (theoretical) relationship between short-term interest rates and house prices does not hold over time and display a time-varying development. The constant and the rolling coefficient indeed look pretty similar in the time between 1980 and 2005. This holds especially for Denmark, Norway and Sweden. However, with the beginning of the financial crisis in 2007, the rolling coefficients abruptly change. In Finland however, this effect is visible even earlier. Only two of the four countries, namely Finland and Denmark, are financially close connected to the Eurozone either due to the adoption of the Euro or due to the ESM-mechanism II of the Danish krone. Both countries, faced an abrupt change in the way the ECB and the Danmarks Nationalbank set their monetary policy to control and stabilize the macroeconomic environment.

11 Different specifications of the CUSUM test and other tests for structural breaks were performed, all with very similar results.

Figure 1.2: Rolling coefficients of short-term interest rates



Note: Constant coefficients in the horizontal lines from the estimations in table 1.3. Symmetric confidence intervals after controlling for auto-correlation using a robust variance-covariance matrix estimator.

Despite these results, the effects of the global financial crisis, which led overall to a rapid fall in investment volumes, a shift in risk-return-profiles and enhanced levels of households indebtedness, are clearly visible in all countries. The effect remains stable in the case of Sweden until about 2006, just prior to the global financial crisis, as the demand for real estate assets increased enormously and the Sveriges Riksbank increased its liquidity operations. The Norwegian housing market shows in contrast an uncorrelated relationship to changes in monetary environment between 1985 and 2007. A remarkable result consists in the drastic fall of the

coefficient in Finland after 2006. Following the rolling interest-rate-coefficient, after the central bank filled the markets with liquidity, house prices reacted three times stronger than in the 90s or 80s. In other words, the impact of short-term interest rates, as a proxy for the monetary framework, on the growth of house prices was of ca. -200 basis points after 2006. Thus, since 2006, a rise in interest rates of 1 percentage point was accompanied afterwards by consecutive fall in house prices of ca. 2 percent in Finland. The effect even increased in 2010. This result also shows clearly that the explanatory power of a system with 34 years of information is unable to capture short-term coefficients and the underlying dynamic, or in other words, the rolling regression framework allows the estimation of time-varying relationships.

1.4.3 Relative contribution of fundamental factors in explaining house prices

The explanatory power of the regression can be decomposed by the individual contribution of each of the variables in order to the show time-varying contribution of single regressors on the response as described in section 1.3. This approach allows thus the decomposition of the variation of house prices on single regressors in a rolling estimation context, as shown in figure 1.3.[12]

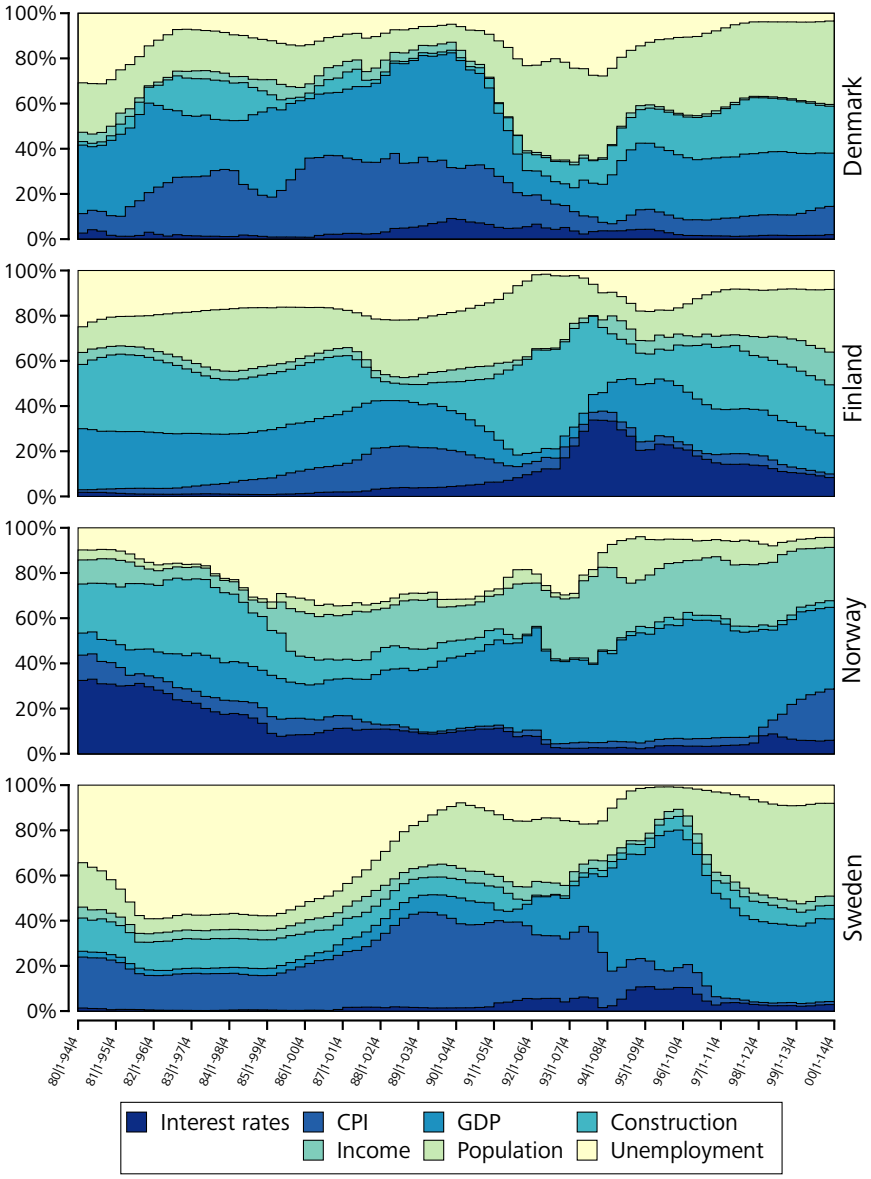
The results of the relative importance of single regressors show useful insights in the interdependence of house prices to their fundamental drivers. In all countries, GDP, population and unemployment rate seem to explain most of the variation in house prices. Table 1.5 shows the mean contributions over all rolling regressions for the different countries. On average – over all countries and rolling regressions – GDP's contribution to the explanatory power is about 24 percent and the unemployment rate as well as the population contribute about 18 percent. Short-term interest rates account for about 7 percent of information, whereas one has to keep in mind that this could be just due to the high contributions of the last third of the regressions in Finland and the first two thirds in Norway. The average just for Sweden and Denmark is about 3 percent.

The contribution of income was comparatively high in Norway since the 90s and construction activity had more influence in Finland, than in the other countries. For Sweden, a remarkable substitution effect between unemployment rate and GDP respectively population is observable when looking at the variation of house prices over time. Prior to the financial crisis almost 60

12 The contributions of the single variables are calculated in percent of the explanatory power (R^2) and therefore sum up to 100 percent.

percent of the variation in house prices was explained by movements in labor markets rather than by aggregated output, shocks or changes in the distribution of households' income.

Figure 1.3: Rolling decomposition of covariates contribution of house price changes



Note: Contributions in percent.

Thus, during the last ten years the relevance of output changes has increased significantly. A contrary effect is observable in Norway, where the contribution of aggregated output growth has decreased over the last years at the cost of an increasing relevance of the consumer price index in explaining house prices.

Looking at Denmark the results suggest a more or less constant relationship between house prices and interest rates and consequently monetary shocks, too, as the relative contribution

was quite constant and on average about 3 percent. In contrast, macroeconomic output, population and changes in the labor supply explain almost 65 percent of house prices' variations.

Table 1.5: Mean contribution of the different determinants

Variable	Denmark	Finland	Norway	Sweden
CPI	17.7 %	6.7 %	5.4 %	19.5 %
GDP	29.2 %	19.0 %	29.8 %	18.5 %
Construction activity	10.8 %	24.9 %	10.9 %	8.2 %
Pers. disp. income	2.4 %	5.0 %	19.0 %	4.4 %
Population	24.5 %	21.3 %	5.1 %	20.3 %
Short-term interest rate	3.0 %	8.8 %	11.2 %	2.8 %
Unemployment rate	12.5 %	14.4 %	18.6 %	26.3 %

Note: Mean contribution over all rolling regression for each country.

Contrary to this development is the increasing importance of monetary shocks in Finland, where the rolling regressions show since the mid-90s a remarkable rise of the importance of ECB's monetary policy in driving the Finnish housing market. For a certain period of time, movements in house prices were up to 30 percent attributable to variations in interest rates pointing to a structural break in the rational formation of prices. Yet, the dependence is going down, but in view of the current expansionary framework of the ECB in terms of the quantitative easing program, a stabilization in the medium-term is rather unlikely.

1.5 Conclusion and Implications

The development of house prices in a country can be described as a dynamic equilibrium between current economic and financial conditions, institutional factors such as taxes or subsidies and finally long-term demographic demand. In this context, strong movements in house prices are interpreted as a direct response to altered conditions in the aforementioned factors and consequently as adjustment periods into a new price-equilibrium. Many of the aforementioned adjustment periods have been seen across many European housing markets during the last decades, mainly as a consequence of fragile economic conditions, increased volatility in financial markets and / or drastic changes in monetary policy. In this paper, we focus on the role of monetary policy in contributing to the adjustment of house prices in the long- and short-term across the Nordic housing markets. Thus, we focus explicitly on the relationship between house prices and monetary policy – proxied by short-term interest rates – in order to deeply examine if house prices present a time-varying (dis-)continuous response to both expansionary and recessionary regimes.

After controlling for economic, institutional and demographic factors, our results come to the (expected) result that housing markets across the Nordics respond negatively in regimes with an expansionary policy, obviously with some differences across the countries. However, our in-depth econometric models provide evidence that the impact of monetary shocks on house prices is – different as expected – not constant over time. On a country level, we found out that the Finish house price sensitiveness to ECB's monetary framework was on its highest level in the last years, whereas house prices in Denmark and Norway did not adjust significantly through the money market. In the case of Finland and Sweden, the results present also evidence that changes in the monetary framework are more and more affecting the drastic changes in house prices, which questions the role of central banks of maintaining price stability. Overall, we confirm that house prices are negatively affected in phases with expansionary regimes in the long-run, but provide evidence of unexpected anti-cyclical effects in the short-run. Consequently, the role of central banks is therefore critically examined, since housing markets adjust unevenly to different monetary environments. Our results are of high concern for policy makers, as they prove evidence that the sensitiveness of housing markets to monetary instruments in the Nordics is playing currently an essential role in the house price formation.

1.A Appendix

Table 1.A.1: Sample correlations

Denmark	I	II	III	IV	V	VI	VII	VIII	IX	X	XI
HP [% yoy]	1.0										
GDP [% yoy]	0.5	1.0									
Construction activity [% yoy]	0.2	0.4	1.0								
Pers. disp. income [% yoy]	0.2	0.0	0.0	1.0							
Short-term interest rate $\Delta 4$	0.0	0.0	−0.2	−0.1	1.0						
Unemployment rate $\Delta 4$	−0.3	−0.5	−0.1	0.2	−0.3	1.0					
Unemployment rate	0.2	0.3	0.1	0.1	−0.2	0.2	1.0				
Population [% yoy]	−0.2	−0.2	−0.1	−0.2	0.1	−0.4	−0.4	1.0			
Working population $\Delta 4$	−0.1	0.1	−0.1	0.2	−0.1	0.3	0.8	−0.5	1.0		
Working population [% yoy]	−0.3	0.0	−0.1	0.1	−0.1	0.2	0.7	−0.2	0.9	1.0	
CPI [% yoy]	0.2	0.1	0.0	0.5	0.0	0.4	0.3	−0.7	0.5	0.3	1.0
Finland	I	II	III	IV	V	VI	VII	VIII	IX	X	XI
HP [% yoy]	1.0										
GDP [% yoy]	0.7	1.0									
Construction activity [% yoy]	0.6	0.3	1.0								
Pers. disp. income [% yoy]	0.3	0.2	0.0	1.0							
Short-term interest rate $\Delta 4$	0.3	0.3	−0.1	0.4	1.0						
Unemployment rate $\Delta 4$	−0.7	−0.7	−0.2	−0.4	−0.4	1.0					
Unemployment rate	−0.4	0.1	0.0	−0.4	−0.4	0.2	1.0				
Population [% yoy]	−0.5	−0.5	−0.3	−0.1	0.1	0.5	−0.2	1.0			
Working population $\Delta 4$	0.2	0.2	0.0	0.2	0.1	−0.1	0.0	−0.2	1.0		
Working population [% yoy]	0.1	0.1	−0.1	0.2	0.1	0.1	−0.1	0.2	0.9	1.0	
CPI [% yoy]	0.3	0.1	−0.2	0.2	0.3	0.0	−0.6	0.4	0.3	0.4	1.0
Norway	I	II	III	IV	V	VI	VII	VIII	IX	X	XI
HP [% yoy]	1.0										
GDP [% yoy]	0.3	1.0									
Construction activity [% yoy]	0.2	0.0	1.0								
Pers. disp. income [% yoy]	−0.2	0.1	0.0	1.0							
Short-term interest rate $\Delta 4$	0.2	−0.1	0.2	−0.1	1.0						
Unemployment rate $\Delta 4$	−0.3	−0.4	−0.2	0.0	−0.3	1.0					
Unemployment rate	−0.2	0.2	−0.1	0.1	−0.4	0.2	1.0				
Population [% yoy]	−0.1	−0.4	0.1	0.2	0.0	−0.1	−0.2	1.0			
Working population $\Delta 4$	0.3	0.1	0.0	−0.1	0.2	−0.1	−0.5	−0.4	1.0		
Working population [% yoy]	0.1	−0.4	0.2	0.1	0.1	−0.2	−0.5	0.8	0.2	1.0	
CPI [% yoy]	0.1	−0.1	−0.1	−0.3	0.2	0.3	−0.5	−0.5	0.5	−0.2	1.0
Sweden	I	II	III	IV	V	VI	VII	VIII	IX	X	XI
HP [% yoy]	1.0										
GDP [% yoy]	0.4	1.0									
Construction activity [% yoy]	0.5	0.3	1.0								
Pers. disp. income [% yoy]	0.2	0.0	0.1	1.0							
Short-term interest rate $\Delta 4$	0.3	0.2	0.2	0.3	1.0						
Unemployment rate $\Delta 4$	−0.6	−0.5	−0.5	−0.3	−0.4	1.0					
Unemployment rate	−0.5	0.1	−0.2	−0.3	−0.3	0.2	1.0				
Population [% yoy]	−0.1	−0.3	0.0	0.2	0.1	0.3	0.0	1.0			
Working population $\Delta 4$	0.4	0.3	0.1	0.0	0.0	−0.3	−0.1	−0.6	1.0		
Working population [% yoy]	0.4	0.1	0.1	0.2	0.2	0.0	−0.1	0.3	0.5	1.0	
CPI [% yoy]	0.1	−0.1	−0.2	−0.1	0.2	0.1	−0.7	0.0	−0.1	−0.1	1.0

Note: Stationarity test for all variables rejected based on ADF/PP-Tests.

Chapter 2

Spatial effects and non-linearity in hedonic modelling

Will large datasets change our assumptions?

This chapter is joint work with Marcelo Cajias† and published as:

Marcelo Cajias, Sebastian Ertl, (2018) "Spatial effects and non-linearity in hedonic modeling: Will large data sets change our assumptions?", *Journal of Property Investment & Finance*, Vol. 36 Issue: 1, pp. 32-49, DOI 10.1108/JPIF-10-2016-0080

Abstract

This paper tests the prediction accuracy and asymptotic properties of two innovative methods proposed along the hedonic debate: The Geographically Weighted Regression (GWR) and the Generalized Additive Model (GAM). We assess the asymptotic properties of linear, spatial and non-linear hedonic models based on a very large dataset in Germany. The results provide evidence for a clear disadvantage of the GWR model in out-of-sample forecasts. There exists a strong out-of-sample discrepancy between the GWR and the GAM models, whereas the simplicity of the OLS approach is not substantially outperformed by the GAM approach. For policy-makers, a more accurate knowledge on market dynamics via hedonic models leads to a more precise market control and to a better understanding of the local factors affecting current and future rents. For institutional researchers, instead, the findings are essential and might be used as a guide when valuing residential portfolios and forecasting cashflows. Sample size is essential when deriving the asymptotic properties of hedonic models. Covering more than 570,000 observations, this study constitutes – to the authors' knowledge – one of the largest datasets used for spatial real estate analysis.

† PATRIZIA Immobilien AG, Fuggerstraße 26, 86150 Augsburg, Germany
The authors especially thank PATRIZIA Immobilien AG for contributing the dataset and large computational infrastructure necessary to conduct this study. All statements of opinion reflect the current estimations of the authors and do not necessarily reflect the opinion of PATRIZIA Immobilien AG or its associated companies.

2.1 Introduction

What are the three most important things when dealing with real estate? *Location, location, location*. This is a pretty common saying about real estate, which makes the statement that the location of a property is one of the most important factors in defining its value. Traditional models for defining the value of properties make use of regression methods in order to decompose the underlying value drivers of properties considering a series of attributes and of course their location within a certain market. The estimation of hedonic regression models has indeed grown substantially over the last years integrating new approaches for modelling spatial heterogeneity, which is essential in the explanation of real estate prices across space. With the list of spatial estimation techniques being very extensive, the Geographically Weighted Regression (GWR) has established itself as a widely used method that expands the restrictive traditional Ordinary Least Squares (OLS) by considering spatially varying effects. Based on the assumption that real estate prices vary over space within a certain market, the GWR method estimates local regressions in order to identify spatially varying parameters and therefore different marginal price functions. The rationale behind the GWR method is plausible since real estate prices are mainly determined by neighbourhood effects, the proximity to common amenities and lastly by households' income distribution. In this context, a major part of the empirical research encourages the assumption that the explanatory power as well as the forecasting accuracy of hedonic models increases when their functional form accounts for spatial effects, thus emphasizing the potentials of the GWR in explaining real estate prices.

Beyond this scope, a series of semiparametric methods which are able to capture spatial effects have been proposed recently and (theoretically) allow a more flexible modelling between the regressor and the predictor without any a priori assumptions regarding the underlying data generating process. In particular methods, like the Generalized Additive Model (GAM), capture spatial effects based on smooth functions and expand the traditional hedonic model by identifying latent nonlinear effects. Since the main goal of any hedonic model is the reduction of misspecification in the estimated coefficients, the GAM model allows covariates to take a nonlinear functional form in order to reduce the error variance and thus enhance the model quality. With GAM models being popular in natural sciences, their usage in the empirical real estate research has been very limited and not been extensively studied.

Given the uncertainty about the statistical advantages of GAM models in hedonic equations, this paper estimates hedonic regressions via OLS, GWR and GAM based on a large dataset

including more than 570,000 observations of rental flats in 46 NUTS3-regions in Germany. The aim of the present study is to test their explanatory power by means of out-of-sample validation approaches. The results show primarily that the explanatory power and predictability of rents in the observed German markets increases significantly when a non-linear and spatially-variant functional form – like the GAM procedure – is chosen.

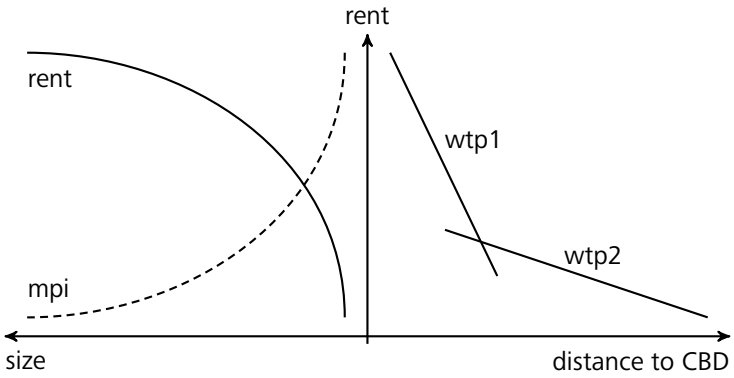
The paper is organized as follows: The upcoming section gives an overview on spatial and non-linear effects in hedonic pricing methods from a theoretical point of view together with empirical evidence. Section 2.3 covers the database, whereas section 2.4 explains the econometric methods used for estimating hedonic prices via OLS, GWR and GAM. The research design and the parameterization of the models is described in section 2.5, as well as the consequential statistical results and implications of the entire analysis. The final section concludes.

2.2 Spatial modelling of real estate prices

Regardless of whether it is building up hedonic real estate indices, forecasting prices or analysing different markets, a significant share of empirical real estate research does not take spatial variables or non-linearity into account. This may be due to different reasons. In the most cases, the lack of the needed data to capture spatial heterogeneity should be the cause. Another possibility may be that spatial models are considered to be complex and difficult to estimate or interpret and that they are not integrated in standard econometric programs.

But why does spatial heterogeneity matter? The locational immobility of real estate makes its price formation different from traditional commodities. Real estate prices theoretically reflect their explicit building attributes, neighbourhood characteristics and finally the share of directly available amenities. Moreover, real estate prices respond to the demand of households for housing, which in turn is based on their disposable income, transport costs and on their own preferences. Spatial variation in rents arises since household's disposable income varies across a city and since some regions or submarkets are able to attract households with higher purchasing power than others. Furthermore, each one of these submarkets provides a different set of local characteristics like green areas, public schools or police departments. The nearer a house is located to them, the higher (or lower – in case of negative attributes) the benefits for this household and therefore its willingness to pay should be.

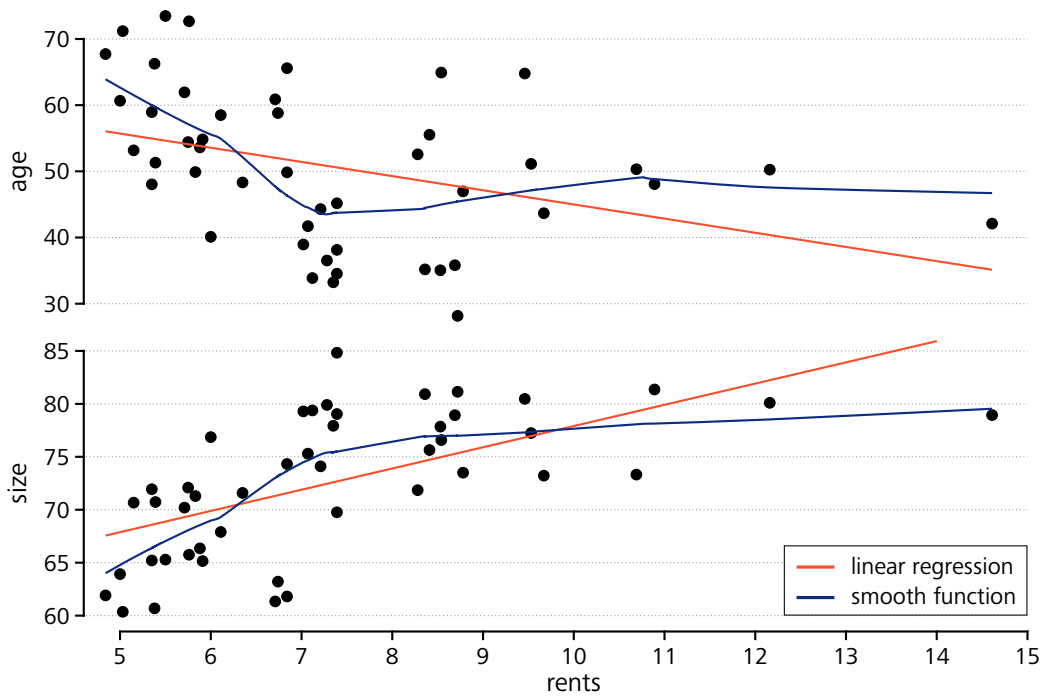
Figure 2.1: The relationship between rents, dwelling size and distance to CBD



Note: Illustration following DiPasquale and Wheaton (1996). Willingness to pay (wtp), marginal price increment (mpi), central business district (CBD).

From a strictly theoretical point of view, household's marginal willingness to pay (directly linked to its income) for one extra unit of housing decreases for additional units of size and their expenditure levels on housing adjust according to the distance to the nearest employment centre, like presented in DiPasquale and Wheaton (1996) and shown in figure 2.1. In this context it is to expect that a nonlinear relationship in the demand functions across the several submarkets within a city takes place. An example for possible non-linear relationships can be found in figure 2.2.

Figure 2.2: Mean rents, dwelling size and age across NUTS3-areas



Note: Aggregated results based on 573,272 asking rents for 46 NUTS3-regions. Mean rents in €/m^2/p.m., mean size in m^2, mean age in years.

Here the mean rents (aggregated over time) are plotted against the mean age and the mean size of the dwellings across all German regions. In both cases it seems that a smooth function might be a better alternative to a simple linear regression line. Of course this is just a simple descriptive figure, but it suggests that the assumption of linear relationships should not be done thoughtless.

In a competitive market without strict regulations, the rent reflects thus an equilibrium between building's characteristics (e.g. quality standards), household's willingness to pay for a housing unit and the availability of amenities nearby. Therefore, hedonic models attempting to decompose rents might consider non-linear relationships and also spatially varying effects. In fact, there are several articles dealing with the issue of spatial dependencies. Many of them are comparing a parametric model – mostly an Ordinary Least Squares (OLS), as a baseline model – to other approaches. The parametric model itself comes either in the form of the common OLS or as an OLS with spatial variables like coordinates, zip-code dummies or others.

McMillen and Redfearn (2010) compare a locally weighted regression, a kernel regression, a conditional regression and a Geographically Weighted Regression (GWR), which is a special case of the Locally Weighted Regression (LWR) method. While analysing the effects of the Chicago's elevated rapid transit line on the surrounding house prices in a case study within their article, it turns out that the spatial models are superior to the parametric model. Further they state that the aim of their paper was not searching for the "best" approach, but rather helping other researchers to get familiar with these kind of models and to help understanding those complex markets even better. To support their statement, they note that there are many different names for the spatial models, but many of them share a common architecture and are – contrary to expectations – easy to implement. Beyond this, they emphasize that the results allow a much deeper insight, because they show more than just an average effect like a linear parametric model would do. Therefore, the criticism on these models of being hard to interpret is unjustified (see McMillen and Redfearn, 2010, p. 731).

Chrostek and Kopczewska (2013) test the prediction quality of different models for the Wroclaw housing market. They use spatial extension models, spatial lag- and error models as well as a GWR, a common OLS and an OLS with geographical coordinates. They conclude that the inclusion of the spatial aspects improves the estimations and that the GWR model fits the data best. Pretty similar results show up in the work of McCord et al. (2014), as they also compare spatial extension approaches, a GWR and an OLS model with different submarkets as spatial

dummies. Just as in Chrostek and Kopczewska (2013), the GWR model performs best, but this time closely followed by the OLS with submarket dummies. They state that the OLS model is as accurate and sometimes more accurate than the geostatistical methods (see McCord et al., 2014, p. 118). The difficulty is to know and/or to determine the submarkets. Widłak et al. (2015) again compare OLS with dummies to a GWR model and get nearly the same results as in McCord et al. (2014), namely that the GWR fits the data slightly better than the OLS.

Empirical research over the last decade has proposed a variety of methods to account for spatial dependencies. A comprehensive and extensive review on spatial hedonic techniques is given by Pace and LeSage (2004), Anselin (2003), Páez et al. (2008), Tse (2002) and Osland (2010). Further methods like Additive Mixed Regression Models and Mixed Geographically Weighted Regressions, exploring both spatially stationary and non-stationary effects on rents have been recently introduced by Brunauer et al. (2010) and Helbich et al. (2014). Several studies – like Sunding and Swoboda (2010), Bitter et al. (2007), Hanink et al. (2012) or Lu et al. (2011) – revealed that rents in large cities respond to a non-stationary functional form that accounts for spatial varying effects. However, following Osland (2010) the GWR framework seems to be very sensitive to multicollinearity in the covariates and at least as good as the traditional OLS. Furthermore, the GWR offers a suitable cartographical examination of the underlying spatial effects on rents. This can be of use for studying market regulation changes or benefits from amenities on rents in cross-sections like Sunding and Swoboda (2010) and Hanink et al. (2012) showed very remarkably.

There is also another type of approach that can be used for spatial analysis, even if this kind of model is not found quite often in the real estate context: the Generalized Additive Model (GAM), introduced by Hastie and Tibshirani (1990). Mason and Quigley (1996) were one of the first to use this kind of model with respect to real estate analysis. In a little example they show that the specification of a hedonic model on the basis of theoretical principles of micro-economic theory can easily be misspecified or even be pointless (see Mason and Quigley, 1996, p. 374). So they state that it is appropriate to take non-parametric procedures into account. They use a GAM approach and a standard hedonic model to construct house price indices for Los Angeles. They conclude that the GAM model has an advantage over the parametric procedure, because of the less rigid assumptions. Although the differences between the models are statistically significant, they admit that they are not very large. Nevertheless, they find that the GAM model has its advantages over a standard parametric model.

Anglin and Gençay (1996) and Gençay and Yang (1996) also compared semi-parametric and parametric models. As one of the first studies in real estate analysis they showed that the spatial semi-parametric models can outperform parametric models in out-of-sample forecast comparisons. Pace (1998) contrasts the forecast accuracy of the GAM approach to parametric and polynomial models. The estimates show that the GAM outperformed all other models used. This matches with the results of Bao and Wan (2004) and Dabrowski and Adamczyk (2010) as they both use the models for forecast comparisons and find that the semi- and non-parametric models outperform parametric models.

There are, however, many other different approaches. Bourassa et al. (2007, 2010) for example demonstrate in two articles many various models to deal with spatial dependencies. In the first paper they use lattice models[1] and two geostatistical methods based on exponential and spherical variograms. The second paper includes a two-stage process with nearest neighbours' residuals and other geostatistical and trend surface models. Again the models of both papers are compared to an OLS with spatial dummies based on their out-of-sample prediction accuracy. The estimations in the first article lead to the conclusion that including submarket variables in an OLS model is of a greater use than applying geostatistical or lattice models. The geostatistical model with disaggregated submarket dummies turns out to give the best results in the second article, whereby the OLS with dummies takes the second place after all, doing better than the geostatistical approach without those spatial dummies.

McGreal and Taltavull de La Paz (2013) employ a Spatio-Temporal Autoregressive (STAR) model, as well as a General Linear Model (GLM), which includes time and space as random factors and calculates interaction effects. Similar to a standard autoregressive model, the STAR-model includes lagged prices (time-component) but also neighbouring prices (space-component). This approach is also used by Clapp (2004) who presents in addition another semi-parametric approach for modelling real estate indices with spatial dependencies: A Local Regression Model (LRM). The model consists of two parts: a standard hedonic model plus a function for the value of space and time which is called Local Polynomial Regression (LPR). This non-parametric part of the model is a data-mining process that seeks to describe the evolution of house prices over space and time (see Clapp, 2004, p. 137). This model again is compared to a baseline OLS model based on out-of-sample forecast errors. They find that the LRM outperforms the OLS as it reduces the forecast error by 11%.[2] Cohen et al. (2015) continue the work of Clapp (2004).

1 Simultaneous Autoregressive (SAR) and Conditional Autoregressive (CAR) models.
2 Out-of-sample mean squared error.

They also use a LPR approach to compare the predictive accuracy against OLS with similar results. They further analyse the density of data needed for more efficient LPR performance. They conclude that the density of data is a key-factor when estimating LPR models.

The considered literature suggests that the most used spatial approach is GWR. The GAM model could not be found that often. Also OLS with spatial variables seems to be a pretty powerful approach when specified correctly. Geniaux and Napoléone (2008) follow the same approach and therefore compare these three models. They state that with a large number of spatial variables, setting up a parametric model might be quite difficult. Especially with a large sample dataset one has to deal with numerous local effects. According to the authors Mixed Geographically Weighted Regression (MGWR) and GAM are capable of managing these difficulties. The MGWR is a special case or rather an extension of the GWR as the GWR is not able handle variables like state indices, environmental zones or the like. The OLS model serves as a baseline model in this article as well. In the end the authors conclude that *"MGWR generally enables a significant gain in model adjustment compared to OLS. However, geoadditive models appear to be even better. GAM fits better than MGWR, is even more flexible in articulating stationary and non-stationary coefficients, works well with a big sample and makes investigating non linearity easy"* (Geniaux and Napoléone, 2008, p. 125).

Based on this, the main objective of the paper is the direct comparison of linear, spatial and semiparametric hedonic methods in predicting rents making use of an extensive dataset with over 570,000 observations for 46 NUTS3-regions in Germany. We expect similar forecasting properties of the three models, but are interested on illustrating their forecasting behaviour under the presence of big data.

2.3 Data description

Since the sample size is a very important factor either in parametric or semi-parametric or nearly any kind of analysis, it might be worth taking a look at the datasets of other studies. In the considered literature there is a pretty wide range. Five of them use a datasets reaching from 440 to 950 observations.[3] One has to admit though that four of the five studies were published in the late 90s, when real estate data was not that easy to get or even available. A pretty good example for the struggles of data search are Chrostek and Kopczewska (2013)

3 Pace (1998), Anglin and Gençay (1996), Chrostek and Kopczewska (2013), Mason and Quigley (1996) and Gençay and Yang (1996).

who initially had 5,600 observations in their database, but had to remove nearly 90 % due to incomplete information. Next there are six studies with a sample size between 2,500 and 5,200 observations, which is the major part of the considered literature.[4] The studies with the largest datasets are Bourassa et al. (2010) with nearly 13,000 observations, Clapp (2004) with 49,500 observations and finally Cohen et al. (2015) with 326,000 records. Even though Cohen et al. (2015) extend the work of Clapp (2004) the employed dataset is a different one.

Germany has one of the largest institutional residential markets in Europe as almost 50 % of the stock is a rental market. In contrast to other European countries, Germany has a polycentric structure with seven main cities (Berlin, Munich, Hamburg, Frankfurt, Dusseldorf, Stuttgart and Cologne) and many secondary as well as tertiary centres surrounding the top 7. The stability of the residential sector has been internationally recognized as the tenure choice model allows labour mobility within the country. Over the last years, the urbanization degree in Germany has increased due to a positive net migration balance from outside the country and especially within the country, leading to rising rents and prices. The rental level is usually negotiated between landlords and tenants but lies within a range dictated by every city depending on location and simple quality groups. Prior to 2015Q2 rent increases were free, but since then some cities regulate subsequent letting agreements to protect tenants and avoid arbitrary rents.

For this study two different databases were merged. On the one hand, 573,272 observations of internet offers of rental flats in Germany were gathered, reaching from 2013-Q1 until 2015-Q2. On the other hand, two socio-economic variables were added: purchasing power per household and the number of inhabitants per households both on a ZIP-code level and yearly basis from the GfK-databank.[5]

The data comes from the empirica system database, which collects and matches internet offers of residential properties from online newspapers and more than ten internet search engines like Immoscout, Immonet, Immowelt and others.[6] After filtering and deleting double enquiries, the empirica system databank provides geographically referenced data on flat offers with more than 30 hedonic characteristics. In order to avoid multicollinearity issues and a large shrinkage of the data due to missing binary hedonic attributes such as wood or laminate floor, only 16 relevant hedonic characteristics from the empirica database were included, which are tabulated

4 Geniaux and Napoléone (2008), McCord et al. (2014), McMillen and Redfearn (2010), Widłak et al. (2015), Bourassa et al. (2007) and Bao and Wan (2004).

5 See `www.gfk.com`

6 See `www.empirica-systeme.de`

together with their descriptive statistics in table 2.1.[7] The final data matrix therefore consists of 573,272 residential flats, each with 16 characteristics across 46 NUTS-3 regions over 10 quarters.

Table 2.1: Variables description and statistics

Variable	Unit	Source	Basis	Mean	SD	Q25%	Q75%
Rent in €/m^2/p.m.	Metric	Empirica	Geographical referenced classification to dwelling	7.98	2.98	5.47	10.00
Area in m^2				72.73	30.38	50.00	90.00
Age in years				53.63	37.08	20.00	87.00
Number of rooms				2.63	0.97	2.00	3.00
Purchasing power per HH		GfK	ZIP-Code	41,360	8,615	33,417	49,123
Inhabitants per HH				1.89	0.17	1.75	2.03
With bathtub	Binary	Empirica	Geographical referenced classification to dwelling	0.55	0.50	0.00	1.00
With built-in-kitchen				0.39	0.49	0.00	1.00
With balcony				0.60	0.49	0.00	1.00
With park slot				0.39	0.49	0.00	1.00
With balcony & terrace				0.70	0.46	0.00	1.00
With terrace				0.15	0.36	0.00	0.00
With elevator				0.25	0.44	0.00	1.00
Heating system				1.61	0.75	1.00	2.00
Brand new dwelling				0.07	0.25	0.00	0.00
Refurbished dwelling				0.18	0.43	0.00	0.00
As-good-as-new dwelling				0.05	0.22	0.00	0.00
Longitude	Geograph. reference	Empirica					
Latitude							
ZIP-code							

Note: The variable heating system corresponds to a trichotome and takes the value of one for floor heating system, two for central heating, tree for room heater and zero otherwise.

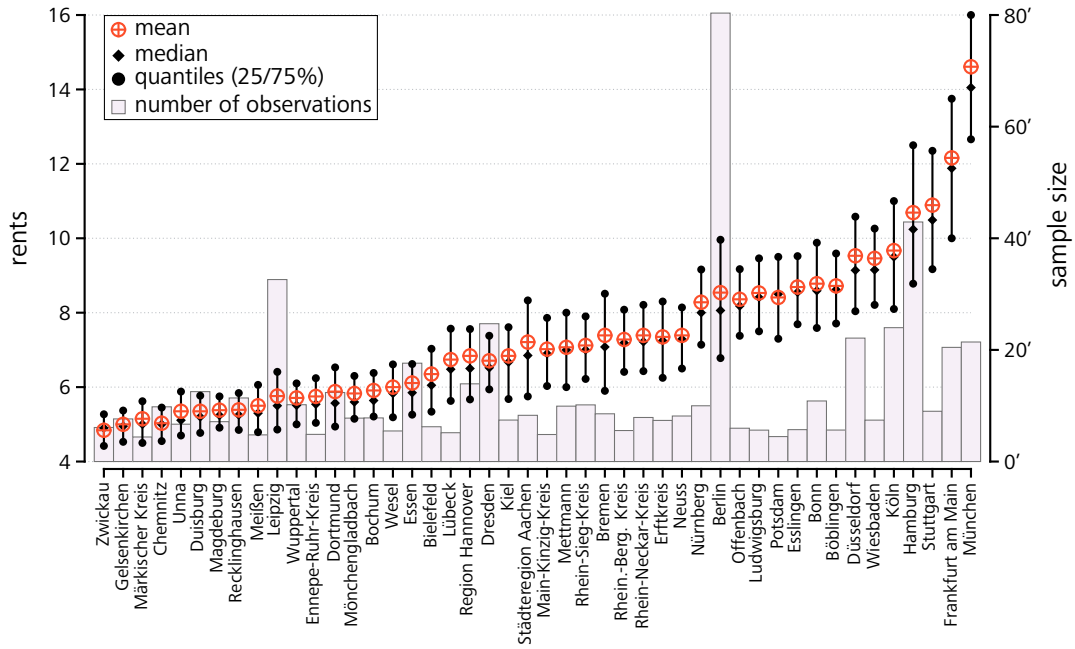
NUTS3 regions correspond to the "Nomenclature of territorial units for statistics", which is a hierarchical system for dividing up the territory in Europe. While the NUTS1 consists on major socio-economic regions, the NUTS3 regions cover small regions like municipalities or counties.[8] We chose NUTS3-areas with more than 300 observations per quarter. Figure 2.3 shows the rent distribution with mean, median and 25/75 percent quantiles and the sample size across those NUTS3-areas. The observed sample size for Berlin is remarkable, but not surprising due to the sheer size of the city. Also of particular note is the wide range of the rents across the cities. Keeping in mind that there are only cities with more than 300 observations per quarter, the mean rents are ranging from 4.8 €/m^2 in Zwickau up to 14.6 €/m^2 in Munich with the 75 percent quantile reaching 16.0 €/m^2. Munich indeed is an expensive city to live in, especially if one takes a look at the second and third most expensive cities, which show

7 Since the Empirica database provides data on asking rents based on a vector of 60 hedonic attributes, we decided to focus on the most important characteristics and omit information such as "with sauna", "laminate flooring" or "with bell" that might be insignificant or lead to multicollinearity.

8 For more information see `www.ec.europa.eu/eurostat/web/nuts/overview`

mean rents of 12.1 €/m^2 in Frankfurt and 10.9 €/m^2 in Stuttgart. While figure 2.3 shows the rents aggregated over time, table 2.2 shows the development of the rents for each quarter aggregated over the different cities. One can see that the observations are almost uniformly distributed as the relative share for each quarter comes to round about 10 percent.

Figure 2.3: Rent distribution and sample size across NUTS3-areas



Note: Mean rents and corresponding quantiles for the different cities can be found on the left axis, the sample size of each city on the right hand side. Rents in €/m^2/month.

Table 2.2: Rent distribution and sample composition

Quarter	Mean rent	SD	Growth	N	relative N	Q25%	Q50%	Q75%
2013/Q1	7.71	2.91	–	56,252	9.81 %	5.52	7.00	9.06
2013/Q2	7.66	2.83	−0.59 %	45,296	7.90 %	5.50	7.00	9.06
2013/Q3	7.64	2.84	−0.29 %	42,591	7.43 %	5.50	7.00	9.00
2013/Q4	7.81	2.95	2.20 %	42,396	7.40 %	5.58	7.10	9.23
2014/Q1	7.98	2.98	2.19 %	72,402	12.63 %	5.74	7.33	9.48
2014/Q2	8.04	2.95	0.67 %	69,235	12.08 %	5.81	7.43	9.50
2014/Q3	8.07	2.97	0.39 %	67,207	11.72 %	5.83	7.45	9.51
2014/Q4	8.20	3.06	1.60 %	62,997	10.99 %	5.89	7.50	9.78
2015/Q1	8.19	3.06	−0.04 %	59,206	10.33 %	5.86	7.51	9.76
2015/Q2	8.27	3.12	0.96 %	55,690	9.71 %	5.88	7.56	9.99
Overall	7.98	2.98	7.26 %	573,272	100.00 %	5.47	7.31	10.00

Note: Mean rent in €/m^2/month. Rental quarter on quarter (qoq) growth in percent. N, number of observations.

2.4 Methods for estimating hedonic price functions

It is not the goal of this section to hand over a complete technical description of the used models. Other authors have done this before.[9] The intention is to give an overview of how those approaches and their methodology work and where the differences are. As mentioned above the traditional hedonic regression, estimated via ordinary least squares, was used for the most part as a baseline model in the considered literature and equation 2.1 shows the approach. Here the rents y depend on various explanatory variables x, like for example property characteristics. Each one of these variables has its coefficient β and can be estimated using equation 2.2. As usual the unobserved variation not captured by the hedonic model remains in ε, which denotes the error term. In this kind of models, there might or might not be an intercept term β_0.

$$y_i = \sum_j X_{ij}\beta_j + \varepsilon_i \tag{2.1}$$

$$\hat{\beta} = (X^T X)^{-1} X^T y \tag{2.2}$$

It is possible to account for spatial variation and nonlinearity in the traditional regression to a certain extent. On the one hand one can run different regressions within the observed market, which might be difficult in view of the sample size needed and might require extensive market knowledge when defining the regions. Another possibility is to include (binary) submarket variables in order to capture geographical effects such as ZIP-codes or city districts. However, as shown by Bourassa et al. (2010) both the definition of boundaries and the number of binary submarket variables are very important since they have a direct impact on the coefficient of determination and prediction accuracy. A further method to expand the traditional linear model is the inclusion of location coordinates and a predefined set of interactions between metric variables and coordinates, the so called spatial expansion method, as seen in Bitter et al. (2007) or Chrostek and Kopczewska (2013).

In this context, the geographical weighted regression proposed by Brunsdon et al. (1996) is based on the fact that the data generating process is non-stationary over space. In this way, it expands the classical linear model by allowing the coefficients to vary over space. As in

9 For a more detailed look at the techniques see Brunsdon et al. (1996, 1998) (GWR), Hastie and Tibshirani (1990) (GAM), Geniaux and Napoléone (2008) (GWR and GAM) or McCord et al. (2014) (GWR and OLS) and of course various textbooks.

equation 2.1 the rents y in equation 2.3 depend on the explanatory variables x, but this time there are no fix, but spatial varying coefficients $\beta(p_i)$, with p_i representing the geographical location in point i. Again there might or might not be an interaction term.

$$y_i = \sum_j X_{ij}\beta_j(p_i) + \varepsilon_i \tag{2.3}$$

To make that estimation work, weighted least squares regressions are necessary. Therefore, different weighting functions like the bi-square in equation 2.4 are available. These kind of functions are called kernel functions and lead to Gaussian distributed weights w. Instead of defining regions a priori, the GWR places a set of windows (or regions) over the space based on an initial bandwidth and finds the optimal bandwidth by minimizing an optimization criterion. The distance between two points is denoted with d and the bandwidth b characterizes the decrease in weight with distance and gives some control over the range of influence of the geographical data. If the distance is greater than the bandwidth the weight is set to zero in this function (see Brunsdon et al., 1998, p. 433; Geniaux and Napoléone, 2008, pp. 115-117).

$$w_{ik} = \begin{cases} \left(1 - \left(\frac{d_{ik}}{h}\right)^2\right)^2 & \text{if } d_{ik} < b, \\ 0 & \text{otherwise.} \end{cases} \tag{2.4}$$

$$W_i = \begin{pmatrix} w_{i1} & 0 & \dots & 0 \\ 0 & w_{i2} & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & w_{iN} \end{pmatrix} \tag{2.5}$$

$$\hat{\beta}_i = (X^T W_i X)^{-1} X^T W_i y \tag{2.6}$$

The calculated and optimized weights from equation 2.4 result in the diagonal weight matrix of equation 2.5. Then the coefficient can be calculated using equation 2.6. But since the coefficients are spatial varying, expression 2.6 is not a single equation, but rather an array of equations, with each $\hat{\beta}_i$ representing the coefficient at a certain location (see Brunsdon et al., 1998, p. 434). This means for every point i all weights for the weighting matrix and the resulting coefficients have to be calculated. Therefore, the computational requirements on a hedonic regression via GWR can be high and very time consuming.

Generalized Additive Models – which were introduced by Hastie and Tibshirani (1990) – can either be semiparametric or non-parametric. In case of a semi-parametric one, the function can be written as seen in equation 2.7. The second part of this equation extends the linear model in the first term by a predefined set of nonlinear functions determined by smoothing functions $f(x)$ and is estimated via a backfitting algorithm. Many different smoothing functions – like for example cubic, cyclic cubic, penalized or thin plate splines – are available (see Geniaux and Napoléone, 2008, pp. 103–107 and Wood, 2006).

$$y_i = \beta X_i + f_1(x_{1i}) + f_2(x_{2i}) + f_3(x_{3i}) + \cdots + \varepsilon_i \tag{2.7}$$

The GAM approach enables thus the combination of location and metric variables capturing both spatial variation and nonlinear effects simultaneously, like for example a nonlinear variation of prices with respect to dwellings size and location. Herein, the choice of the optimal smooth function is very important in order to accurately capture the (expected) nonlinear effect.[10]

A further and more detailed discussion about these models and their nuances would become quickly very technical and therefore not helpful for the aim of this paper. Summarizing the above, it can be said that the traditional linear model is very restrictive in its functional form, while the GWR estimated via weighted OLS is based on the assumption of spatial non-stationary relationships between the predictor and the dependent variables. The semi-parametric approach is estimated via iterative OLS using the backfitting algorithm and enables simultaneous modelling of nonlinear relationships of metric covariates over space.

2.5 Empirical analysis

2.5.1 Model parameterization and forecasting approach

This paper aims at comparing the prediction accuracy and large sample statistical properties of the models mentioned above based on a sample of more than 570,000 asking rents with full hedonic characteristics across 46 NUTS3-regions in German residential markets from 2013-Q1 until 2015-Q2. The response variable of the study is the asking rent in €/m^2/month of a dwelling. In each hedonic model, we include a set of predetermined hedonic characteristics

10 For a pretty detailed look at different smoothing functions and their usage see Wood (2006).

and two socio-economic variables measuring the purchasing power per capita in log and the number of persons per households, both geocoded on a ZIP-code basis. The OLS and the GAM models are estimated with ZIP dummies. The GWR includes all hedonic and socio-economic covariates except for the ZIP-dummies and is estimated using a bi-square spatial bandwidth after the minimization of the cross-validation criterion integrated in the package "mgcv" in the statistical software R, which is also mentioned in Geniaux and Napoléone (2008).[11] The GAM model is parameterized by hedonic as well as socio-economic variables and ZIP-dummies and by a set of smooth terms including the metric covariates flat size, dwelling's age and Gaussian geocoordinates.[12] The estimation procedure and forecasting evaluation were organized as follows:

First: Obtain the predicted hedonic functional form for each regression model, for each NUTS3-area in each quarter. Within this framework there are 3 different model types (OLS, GWR, GAM), 46 different NUTS3-areas (see data description in section 2.3) and 10 quarters (reaching from 2013-Q1 to 2015-Q2). Second: Based on the functional forms, the out-of-sample forecasts of the rents are calculated iteratively. For example, predict the asking rents of $t+1$ based on the functional form obtained in t and compare the results. Third: To measure the performance of the out-of-sample forecasts, forecast evaluation indicators have to be calculated. Two conventional error measurements are the Mean Error (ME) and the Error Variance (EV), which are essentially the mean and the variance of the prediction errors. But since both of them are scale dependent, other indicators are considered in addition. A frequently used error measurement in the literature is the Mean Squared Error (MSE), as seen for example in Bao and Wan (2004), Anglin and Gençay (1996) or Gençay and Yang (1996), although the latter two call it Mean Squared Prediction Error. This is kind of confusing since there is also a common used error measurement called Mean Squared Percentage Error (for example used by Chrostek and Kopczewska, 2013).[13] Therefore, we go with the Mean Error (ME), the Error Variance (EV) as well as the Mean Squared Error (MSE) and the Mean Squared Percentage Error (MSPE). Fourth: In a final step, the forecast evaluation results were aggregated over all quarters and regions for each model type. Finally, their quantile distribution is presented to compare the forecast accuracy of the different approaches.

11 R is a free software environment for statistical computing and graphics. See `www.r-project.org`

12 The penalization term of each smooth is determined by both the automatic procedure implemented in the R-package "mgcv" – whose objective function does not follow an optimization criteria – and a set of manually selected penalization terms above the boundary recommended by Kim and Gu (2004) of $n^{(2/9)}$. While several models were estimated, the penalization term recommended by Kim and Gu (2004) provided the best results.

13 For more details about computation and/or interpretation of this performance indicators see Brooks and Tsolacos (2010, pp. 269–271).

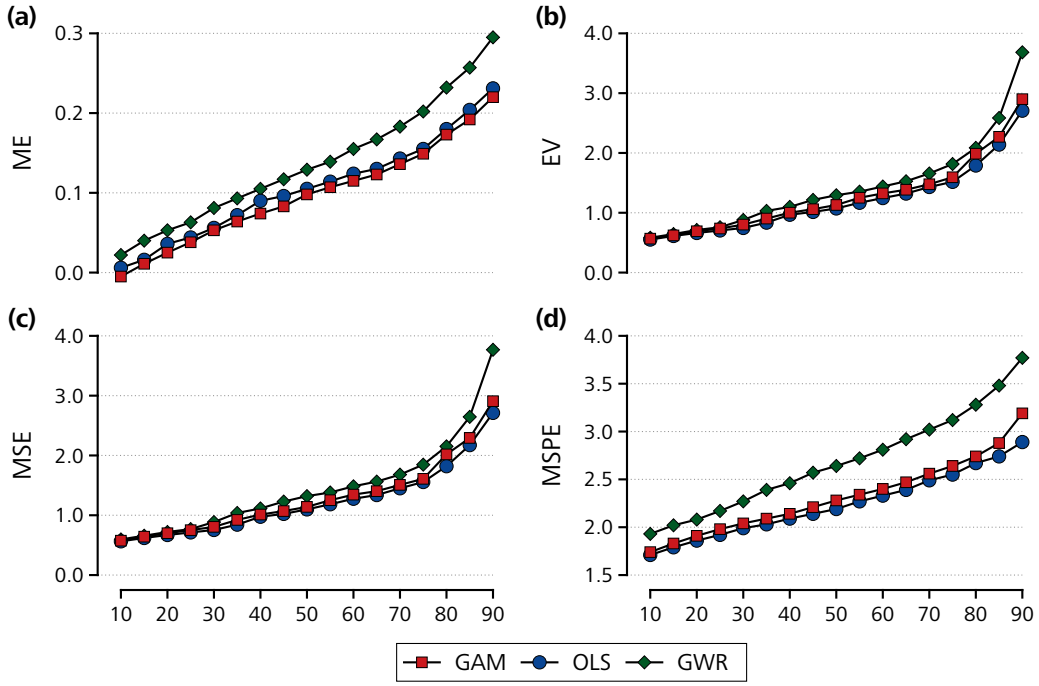
2.5.2 Results and out-of-sample forecasting accuracy

This study aims at evaluating the prediction accuracy of linear, spatial and non-linear models in estimating the hedonic equation for asking rents across several German residential markets. While the employed functional form is based on the OLS, GWR and the GAM, the estimation methodology was chosen to be iterative in forecasting the fitted rents for each quarter based on their 1-quarter-prior functional form. The forecasting accuracy is set to be evaluated by the overall quantile distribution of the mean error, error variance, mean squared error and mean squared percentage error on an aggregated basis regardless of period or the NUTS3-area. Thus, instead of looking at the difference in the estimated coefficients or the patterns of the deviations, we focus merely on the forecasting accuracy of the models based on the four forecasting indicators rather than showing the mean error and error variance of each of the 46 NUTS3-areas and on each quarter.[14]

The top row of figure 2.4 shows the quantile distribution of the mean error (a) and the error variance (b) from the out-of-sample forecasting results for each model OLS, GWR and GAM. While higher forecasting errors correspond to higher error variances for each of the three models – which points to a stable convergence of the estimators – the results show a relative dominance of the GAM model in comparison to the OLS and the GWR. Starting at the lowest quantile on the left side of the exhibit, the GAM model shows a lower mean error for the same error variance level in comparison to the OLS and GWR. Up to the 50 percent quantile in the middle of the exhibit, the GAM model outperforms the OLS and GWR models as the increase in the forecast errors of the latter models is disproportional for each error variance level. Although the models show the same forecast error pattern – where increasing forecast errors are penalized by increasing error variances – the difference between them is strong in the upper quantiles. Thus, while the GAM model has a mean error of ca. 0.1 €/m^2 on an error variance of ca. 1.5 €/m^2, the forecasting accuracy of the GWR model is outnumbered on both the mean error and the error variance. Based merely on the general forecasting criterion, the GAM forecasting is more precise than the OLS or the GWR approach, although the traditional linear model shows remarkable results with an acceptable forecasting inaccuracy.

14 The coefficients proceeding from the OLS, GWR and GAM approach are similar in scale and significance based on several model specifications. The latter is also valid for the explanatory power. However, an examination of the differences in the estimated hedonic functions might not suitable due to the nature of the models and would be outside the scope of the study. Recall that the GWR model estimates (many) local regressions with several coefficients, whereas the OLS and GAM optimize one single equation each with a vector of coefficients.

Figure 2.4: Out-of-sample forecast evaluation



Note: Quantile distribution of (a) Mean Error (ME), (b) Error Variance (EV), (c) Mean Squared Error (MSE), each in €/m^2/month and (d) Mean Squared Percentage Error (MSPE) in percent.

While the mean error and the error variance are general indicators for the forecasting performance of models, the mean squared error and mean squared percentage error are more precise indicators for the evaluation of models, as they take the deviation of the forecast error from the mean error more accurately into account regardless of the sign of deviation, i.e. of positive or negative deviations. The bottom row of figure 2.4 presents the quantile distribution of the mean squared error (c) and the mean squared percentage error (d). While the relationship between MSE and MSPE is positive – e. g. higher squared errors are penalized by higher squared percentage errors – the results show a strong discrimination of the GWR model in contrast to the linear and semi-parametric model. Although the squared error of the models is similar at the lowest quantile on the left side of the exhibit, the GWR has a substantial higher squared percentage error which proportionally increases across the quantiles. In contrast, the GAM and the OLS model remain equivalent up to the 70 percent quantile for both the MSE and MSPE. However, the results show some dominance of the OLS model in being more accurate when forecasting extreme values as the MSE and MSPE are relative lower in comparison to the GAM for the quantiles up to 70 %. In conclusion, the results for the general and the accuracy criterion show that the GAM and OLS models outperform the GWR. The OLS shows stability and in some cases a higher stability as the GWR method.

The results of the single regions are reported in table 2.A.1 in the appendix. This table shows the detailed forecast evaluation of the 46 NUTS3 regions. The reported mean error is calculated over the whole estimation horizon. The last column of the table shows the model with the minimum forecast error for the given NUTS3 area. Summing up, the GWR method could not generate any forecasts better than the other models anywhere. The baseline OLS model performed best in 15 regions (33 %), whereas the GAM method achieved the minimal forecast error in 31 of the 46 areas (67 %).

2.6 Conclusion

The choice of the functional form in hedonic regression models is crucial when explaining rents within a certain real estate market. Empirical research has thoroughly attested that traditional hedonic models fail to explain the variation of rents accurately, mainly due to the exclusion of both spatial effects and non-linear relationships. In the course of the past years the Geographically Weighted Regression (GWR) has established as a suitable hedonic method able to capture spatial effects. The explanatory potentials and statistical advantages of further semiparametric hedonic models – like the Generalized Additive Model (GAM) method – that account for non-linear relationships have not been extensively exploited in empirical real estate research.

This paper analyses the prediction accuracy and explanatory power of three different approaches based on a large dataset of more than 570,000 asking rents across 46 residential rental markets in Germany. Compared to the considered literature, this is to the authors' knowledge one of the largest datasets used for spatial real estate analysis. Rather than attempting to select "the best" model for real estate data with spatial dependencies, the paper focused on comparing the goodness of fit, measured by out-of-sample forecasts. The GWR, which is a great tool to explore regional factors driving rents within a certain market, was outperformed by the GAM and OLS models. In terms of out-of-sample forecasting accuracy it turns out that the differences between OLS and GAM are not substantial.

One could wonder how a GWR model can be outperformed by a OLS model, since the GWR is basically an extended OLS version. In simple words, even if all the additional use of the space-varying parameters from the GWR method would equal to zero, it would give the exact same results as the standard OLS. Even though this holds true for in-sample validation, in out-

of-sample forecasts, this is no longer the case because the GWR weights are static over time, which could therefore lead to less forecasting accuracy.

As this paper performs cross-section estimations and forecasts based on the 1-quarter-prior functional form, there is no space for adding time varying effects in the models. But for further research – possibly in a panel data framework – it could be interesting to see how time varying effects complement the spatial dependencies.

This results align with several findings of the considered literature. Against expectations the OLS approach seems to be an equal alternative to (semi-) parametric models as seen by Osland (2010) or Bourassa et al. (2007, 2010). Despite the low discrepancy between the OLS and GAM, the results of this paper provide evidence for a clear disadvantage of the GWR model in out-of-sample forecasts. Furthermore, the results confirm the findings of Mason and Quigley (1996, p. 384) which conclude that the differences between OLS and GAM *"are rather small, though statistically significant"*.

2.A Appendix

Table 2.A.1: Detailed out-of-sample forecast evaluation

		Mean error			
NUTS3 Name	NUTS3	OLS	GAM	GWR	Min.error
Berlin	DE300	0.197	0.189	0.318	GAM
Bielefeld	DEA41	0.076	0.066	0.107	GAM
Böblingen	DE112	0.068	0.028	0.157	GAM
Bochum	DEA51	0.092	0.072	0.101	GAM
Düsseldorf	DEA11	0.165	0.157	0.210	GAM
Ennepe-Ruhr-Kreis	DEA56	0.057	0.051	0.075	GAM
Erftkreis	DEA27	0.105	0.089	0.130	GAM
Esslingen	DE113	0.124	0.123	0.149	GAM
Frankfurt am Main	DE712	0.215	0.165	0.296	GAM
Hamburg	DE600	0.143	0.138	0.241	GAM
Ludwigsburg	DE115	0.103	0.093	0.153	GAM
Magdeburg	DEE03	0.047	0.043	0.047	GAM
Märkischer Kreis	DEA58	0.068	0.055	0.086	GAM
Meißen	DED2E	0.092	0.079	0.104	GAM
Mettmann	DEA1C	0.082	0.075	0.128	GAM
Mönchengladbach	DEA15	0.055	0.044	0.072	GAM
Neuss	DEA1D	0.104	0.101	0.123	GAM
Offenbach	DE71C	0.115	0.096	0.144	GAM
Potsdam	DE404	0.128	0.104	0.190	GAM
Recklinghausen	DEA36	0.056	0.054	0.074	GAM
Region Hannover	DE929	0.155	0.154	0.194	GAM
Rhein.-Berg. Kreis	DEA2B	0.051	0.038	0.075	GAM
Rhein-Neckar-Kreis	DE128	0.098	0.096	0.164	GAM
Rhein-Sieg-Kreis	DEA2C	0.092	0.083	0.109	GAM
Städteregion Aachen	DEA2D	0.137	0.116	0.178	GAM
Stuttgart	DE111	0.171	0.161	0.221	GAM
Unna	DEA5C	0.066	0.061	0.084	GAM
Wesel	DEA1F	0.068	−0.063	0.096	GAM
Wiesbaden	DE714	0.136	0.124	0.159	GAM
Wuppertal	DEA1A	0.067	0.066	0.087	GAM
Zwickau	DED45	0.050	0.042	0.056	GAM
Bonn	DEA22	0.120	0.122	0.151	OLS
Bremen	DE501	0.085	0.087	0.149	OLS
Chemnitz	DED41	0.026	0.027	0.049	OLS
Dortmund	DEA52	0.103	0.105	0.126	OLS
Dresden	DED21	0.120	−0.299	0.152	OLS
Duisburg	DEA12	0.070	0.070	0.082	OLS
Essen	DEA13	0.087	0.090	0.109	OLS
Gelsenkirchen	DEA32	0.065	0.067	0.069	OLS
Kiel	DEF02	0.095	0.102	0.120	OLS
Köln	DEA23	0.164	0.165	0.194	OLS
Leipzig	DED51	0.122	0.129	0.155	OLS
Lübeck	DEF03	0.095	0.095	0.139	OLS
Main-Kinzig-Kreis	DE719	0.136	−0.341	0.168	OLS
München	DE212	0.315	−0.453	0.410	OLS
Nürnberg	DE254	0.172	0.180	0.205	OLS

Note: This table shows the detailed forecast evaluation of the 46 NUTS3 regions. The reported mean error is calculated over the whole estimation horizon reaching from 2013-Q1 until 2015-Q2. The last column of the table shows the model with the minimum forecast error for the given NUTS3 area.

Chapter 3

Pitfalls of using Google Trends data in empirical research

What do microwave baked potatoes tell us about U.S. housing markets?

A slightly adapted version of this chapter is accepted for publication in *International Journal of Housing Markets and Analysis*, DOI 10.1108/IJHMA-05-2018-0031

Abstract

Google offers virtually unlimited, instantaneously available, spatially and textually adjustable and, in addition, free data. Although Google Trends data can be accessed already since 2008, many interpretation and usage misunderstandings can be found amongst the literature. Therefore, I will focus on two main objectives: Firstly, I will give an overview of what Google data is in the first place and what the potential pitfalls are. Secondly, I will conduct an empirical analysis to find out, whether the results are still in line with the literature after accounting for those difficulties. Additionally, the resulting models are contrasted against other comparable models. The results are in line with the literature. Adding search volume data to the estimations leads to an improvement regarding model fit and helps reducing the forecasting errors compared to a baseline model. However, I will also show that there are equally specified "standard" models that fulfill the same requirements and can be used in the same way as the Google models, even with slightly better results. Real estate markets appear to be particularly well-suited for search volume related studies, as the "products" of this market involve a large financial commitment, which demands an extensive information gathering process. To my knowledge there is no other paper especially dealing with the potential pitfalls and disadvantages of Google Trends data in real estate analysis.

3.1 Introduction

In 2006 Google launched a new service called *Google Trends* that allows users to see the interest of all other Google users on certain search terms. In 2008 Google introduced another service called *Google Insights for Search*, which was heavily inspired by Google Trends, but was actually intended for advertising and market research. They allowed to download the data and added features to compare multiple search terms and filter the data choosing different categories and/or regions.[1] In addition, there is another but quite similar service called *Google Correlate*. This service enables the user to find other queries, similar to a given search term or time series. Google merged Trends and Insights in 2012, while keeping the features of both services.[2] As of today, Google Trends updates its data so fast that it can be queried on a monthly, weekly, daily or hourly basis and even in real time. The geographical location can be restricted to countries, states and even large cities and there are over 1,000 categories to narrow down the results even more. By doing so, Google offers virtually unlimited, instantaneously available, spatially and textually adjustable and, in addition, free data. This type of data conquered its position in nearly all economic fields, serving as a highly adjustable sentiment indicator that can be used, inter alia, for nowcasting and short-term forecasting.

Although Google Trends data can be accessed already since 2008, many interpretation and usage misunderstandings can be found amongst the literature. Therefore I will focus on two main objectives in this paper: Firstly, I will give an overview of what Google data is in the first place and where potential pitfalls and difficulties lie. Real estate markets appear to be particularly well-suited for search volume related studies, as the "products" of this market involve a large financial commitment, which demands an extensive information gathering process. To my knowledge there is, surprisingly, no other paper especially dealing with the potential pitfalls and disadvantages of Google Trends data in real estate analysis. Therefore, I will demonstrate that search terms like `[microwave baked potatoes]` can be valid predictors of US housing prices and can also increase the out-of-sample forecasting accuracy significantly, when ignoring the special aspects of search engine data. Apart from the obvious absurdity of this example, the overall design, presentation and results will still match the results of many other authors.

1 See "Announcing Google Insights for Search", available at: `https://adwords.googleblog.com/2008/08/announcing-google-insights-for-search.html` (accessed 2018-03-12).

2 See "Insights into what the world is searching for – the new Google Trends", available at: `https://search.googleblog.com/2012/09/insights-into-what-world-is-searching.html` (accessed 2018-03-12).

Secondly, I will perform a more serious empirical analysis to find out, whether the results are still in line with the literature after accounting for those characteristics. For this task, the usual approach in the existing literature would be to simply compare Google models with a baseline model. However, instead of demonstrating only how a very simplistic baseline model can be outperformed, I am more interested in seeing how the resulting models can compete against comparable "standard" models.

The results show, as expected, that adding search volume data to the estimations leads to an improvement regarding model fit and helps reducing the forecasting errors when compared to a baseline model. However, they also show that there are equally specified "standard" models that fulfill the same requirements and can be used in the same way as the Google models, even with slightly better results.

The remainder of this paper is structured as follows: In the subsequent section I will give an overview of the existing literature. Section 3.3 shows the advantages as well as the disadvantages and potential pitfalls of search engine data. This findings are then used to conduct an analysis for the US housing markets in section 3.4. The final section concludes.

3.2 Literature Review

The application area of search engine data is truly enormous. Ginsberg et al. (2009) declare that they could estimate and predict influenza epidemics with search query data and they assign their model the name *Google Flu Trends*. They manage it to track the spread of influenza in the US just based on highly correlated search terms. Their method is faster than the reports from the Centers for Disease Control and Prevention (CDC), which collect their data from actual surveillance reports from laboratories. The tracking results from Google Flu Trends have a delay of only one day, whereas it takes one week or more for the CDC (see Harford, 2014). Preis et al. (2010) investigate whether search volume data and financial market fluctuations are linked. They find evidence for correlations between the S&P 500 transaction volumes and the search volume of the corresponding company names. They also find a tendency that search volume and transaction volume show recurring patterns. Therefore, they conclude that search volume reflects the current attractiveness of trading stocks.

It should be rather obvious that this data finds its use not only regarding marketing strategies or market research, but also in nearly any other economic field, whether it be finance, macroe-

conomics, social economics, sales, tourism, automotive industry or real estate markets. Askitas and Zimmermann (2009) or Pavlicek and Kristoufek (2015) for instance analyze the job markets in Germany and in the Visegrad Group countries respectively.[3] Guzman (2011) attempts to set up a measurement for real-time inflation expectations based on search engine data, whereas Vosen and Schmidt (2011) use Google Trends as an indicator for private consumption. But there are also other research areas like Rivera (2016) who estimates and forecasts hotel registrations in Puerto Rico with the help of Google. Goel et al. (2010) on the other hand, utilize the data to predict box-office revenues for movies, the sales of video games or the chart placing of songs. Koop and Onorante (2016) present dynamic model selection methods that improve the nowcasts of nine major monthly US macroeconomic variables with the help of Google data.

Two of the standard references regarding Google Trends research would be Choi and Varian (2012) and Stephens-Davidowitz and Varian (2015). Both papers clearly illustrate the possibilities with and characteristics of Google data for scientific research. Mohebbi et al. (2011) show different applications, specifically for dealing with Google Correlate data. With Hal Varian being the chief economist at Google, those three papers are published from within the Google company. Nevertheless, they give an excellent, yet critical summary of the data itself and show various possible applications.

All of these papers have in common that they use Google data for *nowcasting*. Nowcasting, also known as short-term forecasting, describes the process of estimating the most recent figures of different variables or short: *predicting the present*. Since most macroeconomic data is usually released with a time lag, other currently available data has to be used to estimate those values. Google data is especially suitable for this task, because of its immediate availability. But apart from the availability and customizability, there is another interesting point: the origin of the data. Google Trends' units are not amounts, currencies or prices, but rather interest. As mentioned above, Google Trends measures how often a certain term is searched for, relative to all queries. Therefore, the data could be understood as a kind of sentiment indicator or, like Preis et al. (2010) put it, as the *"collective 'swarm intelligence' of Internet users"*. There are lots of traditional survey-based sentiment indicators, but they are time-consuming, expensive and of course, released with a time lag, as Dietzel (2016) points out. However, sentiment indicators try to capture the "noise" in various markets that cannot be represented by fundamentals, like for example irrational fears, hopes or simply interest. As a matter of fact, there is a strong opinion that this Google sentiment indicator could be more reliable than proven indicators,

3 Visegrad Group: The Czech Republic, Hungary, Poland and Slovakia.

that are constructed from surveys. This is due to the consideration that people, for a variety of reasons, may pretend or not always be honest when answering a survey. Regarding the Google searches, however, a reflection of the "true" sentiment can be expected, because they take place privately on the own phone, tablet or computer without any exogenous pressure (see Wu and Brynjolfsson, 2015; Dietzel, 2016). Heinig et al. (2016) compare different proven sentiment indicators for European commercial real estate markets, whereby there is also one added, calculated from Google Trends data. They find the Google indicator to work better than expected.

Real estate markets seem to be particularly well-suited for search volume related studies, as the "products" of this market involve a large financial commitment. Therefore, people will extensively inform themselves before buying or selling on this market. Google is able to aggregate all of these search queries into a custom-made sentiment indicator. In addition, it should be possible to extract this indicator for any specific or fine-grained research area one is currently working on. Dietzel et al. (2014) find that a combination of Google and macro data helps to improve forecasts significantly. Using VAR models they show that even in models without other data then Google, the base model is outperformed. They state that search volume data can act as an early market indicator. Rochdi and Dietzel (2015) construct different indices from Google Trends data, trying to anticipate REIT market movements. They show that investment strategies based on their index would have outperformed a buy-and-hold strategy by 15 percent. Conducting volatility forecasts of the US REIT market, Braun (2016) uses Generalized Autoregressive Conditional Heteroskedastic (GARCH) models to show that Google models outperform the baseline model, especially in periods of high volatility. While using the search volume data as a proxy for investor sentiment, the author states that Google variables can be used as an early warning system for periods of high volatility.

Hohenstatt et al. (2011) are one of the first to analyze housing markets based on Google search queries. They find Google data alone provides the best goodness-of-fit, but point out that this statement has to be interpreted with caution. During their research, the aftermath of the financial crisis was still prevalent, accompanied by extreme market movements. They state that an analysis under normal economic conditions could lead to the conclusion that a combination of real-world data and Google data performs best. A few years later Hohenstatt and Kaesbauer (2014) analyze the U.K. housing market using a panel VAR framework. Their findings again confirm that Google subcategories, especially *Real Estate Agency*, can serve as an indicator

of transaction volume. Askitas (2016), however, constructs a ratio of "buy and sell-searches", called BUSE index, to get a proxy for the relation of expected home buyers to expected home sellers. He finds this index to have a significant correlation with the US national S&P/Case-Shiller Home Price Index. Since S&P releases its index with a two-month lag but Google data is available almost instantly, the BUSE index can be used for short-term forecasting of housing prices in the US. Further he states that this index can be used to understand the post bubble burst dynamics in the US housing market, as well as it can be utilized as an instrument for monitoring housing market conditions. Dietzel (2016), on the other hand, predicts turning points in the US housing market measured by the Case-Shiller 20-City House Price Index. He states that sentiment plays a significant role in future house price formation, which cannot be explained exclusively by fundamentals. Using a multivariate probit model, the results show that the Google model always predicts the signals for turning points correctly, although the timing of those turning points is not always accurate. Even though, according to the author, this model can be used as an indicator for upcoming changes in house prices because the signals are always early, but never late.

Wu and Brynjolfsson (2015) can be found referenced quite often in the reviewed literature. They demonstrate how search query data can be used to predict a housing price index (HPI) as well as sales volumes. Thereby, they follow the approach of comparing a baseline model to different model specifications including Google Trends categories, using simple linear regressions. Their base model uses only the home sales and the HPI from the past to predict the current home sales and the HPI. Then the Google categories *Real estate agencies* and *Real estate listing* as well as lagged versions of them are added to the base model. Concerning the estimation of quarterly sales, their results show a remarkably good model fit (adjusted R^2) of 0.973 just for the baseline model. By adding the Google predictors, the fit can be improved up to 0.983. They also present a specification, where the sales are only predicted by the Google categories *Real estate agencies* and *Real estate listing* and their respective 1-quarter lag. They find the model fit to be slightly below the base model, but a value of 0.970 can still be considered as "satisfactory". Regarding the HPI, they conduct the same estimations and report a consistent model fit of 0.987 over nearly all specifications, including the specification without any other variables than the Google predictors. Subsequently, they aim to test the forecasting accuracy. Unfortunately, it seems as they used other model specifications for the forecasting than for the estimations. This makes the transparency difficult as they only specify that they used the best-fitted model from the training data set. However, after forecasting home sales as well as

the HPI, the results show that the Google models improve the forecasts for different states in many cases, although the improvement is, as they say, rather modest on average. Further they compare the home sales predictions to the ones from National Association of Realtors (NAR) and find that their predictions for the current home sales indeed are slightly better, yet the difference is not statistically significant. As they run a nowcast as well as a 1-quarter forecast, they find the latter to be considerably better than the forecasts from NAR. They conclude that their methods seem to provide a significant improvement in forecasting, additionally, they outperform not only the base model, but also the predictions from the established experts in the field.

Usually, some researchers have success with a certain method or approach, others do not. With Google Trends data it seems that everyone has success. After reading most of the literature mentioned before, one could think that – putting it exaggerated – Google Trends data is something like a silver bullet. Especially when reading the paper of Wu and Brynjolfsson (2015) the results almost seem to be too good to be true. They show a model fit of over 97 % for a simple base model that is improved even further. Additionally, forecast accuracy measures are reported that do not belong together with the estimations presented in the same chapter. Even if they briefly describe the process of model selection for the forecasts, the underlying models themselves are not shown. Without accusing somebody one could become skeptical at least.

As a result, the question arose whether there are papers that criticize the usage of Google Trends. This brings us back to Google Flu Trends, because the hype about Google Flu Trends did not last very long. Harford (2014) indicates that Google was mainly interested in finding statistical patterns and that they put correlation on the same level as causation. He also states that a theory-free analysis of correlations by itself is inevitably fragile. He alludes to Butler (2013), who found that Google Flu Trends' estimates were almost double the CDC's. Other studies show further deviations from the CDC reports as well. Lazer et al. (2014a) put the overall approach of Google Flu Trends into question and analyze the difficulties that occur with this project. Lazer et al. (2014b) find that Google Flu Trends does not perform significantly better than a simple autoregressive approach using the 2-week-lagged CDC reports. Google Flu Trends was launched in 2008 and updated in 2009, 2013, and 2014. Since 2015 the service has been stopped.[4]

4 See "The Next Chapter for Flu Trends", available at: `https://research.googleblog.com/2015/08/the-next-chapter-for-flu-trends.html` (accessed 2018-03-12) and "Flu Trends model updates for the United States", available at: `https://www.google.org/flutrends/about/` (accessed 2018-03-12).

Apart from Butler (2013) and Lazer et al. (2014a,b), which examine specifically the Google Flu Trends project, there is – to my knowledge – no other paper dealing with the difficulties and potential pitfalls of search engine data in scientific research. Maybe the most informative article is the one from Stephens-Davidowitz and Varian (2015), as their main objective is to present an overview of the data and its usage. Of course, many papers touch on the difficulties, they maybe have come across during their research. Most of them, however, do not put much effort into this task as the results are the center of attention, obviously. Another common characteristic of many papers is that there is only a comparison between a baseline model and models with Google variables. Model specifications without Google predictors are generally not taken into account, although they could deliver similar results. Most likely this method is chosen because many of the possible predictors are not available when needed. Nevertheless, there are variables and model specifications that allow a proper comparison between models with Google variables and equally specified "standard" models.

The advantages that come with Google data are clear and there is no reason not to use it. But the results of many other papers make it tempting to use it without further questioning. Therefore, this paper will focus on two main objectives: The following section gives an overview of what Google data is in the first place and where potential pitfalls lie. Subsequently, I will perform an empirical analysis to find out, whether the results are still in line with the literature after accounting for those difficulties and contrasting them against other comparable models.

3.3 About Google Trends

As seen in the previous section, the interest on using search engine data in research is strong and growing. The first and most common used source is Google Trends. To obtain the data you just go to the website, enter a search term and view or download the time series shown.[5] You can adjust the time horizon, the geographic location and refine your results using different predefined categories. According to the selected time horizon you get either monthly, weekly, daily, hourly or even real-time data. Additionally, you can compare up to five search terms at once. The Google Trends website seems to be designed for a wide range of users, as there are features like "trending stories" or "featured insights" which are top trending search topics graphically illustrated. You can also find the so-called "top charts" which are a kind of summary

5 Google Trends website: `https://trends.google.com`.

on what people were the most interested in at a given time and / or region within a certain category.[6]

However, Google Correlate has a much simpler design.[7] You can enter the search term or time series you want to analyze and get the top 100 highest correlated search terms. The time horizon is set by the entered data. Google states that *"Google Correlate is like Google Trends in reverse"*, but the functionality of how those two services work is different.[8] In Google Trends you obtain a single time series, which represents the interest on the desired search term. In Google Correlate you get one hundred time series based on the correlation with the entered data that each represent the interest in the associated search term.

3.3.1 Correlation is enough

The heading of this section is an extract from the article of Anderson (2008). In the context of big data, he claims: *"'Correlation is enough.' We can stop looking for models. We can analyze the data without hypotheses about what it might show"* and concludes with the very provocative statement: *"Correlation supersedes causation"*. Even if this statement is apparently incorrect, exaggerating and meant to be provoking, if you take a look at the literature, some may think that few authors indeed did worry more about the results rather than the foundation. However, following Anderson's logic, this section shows what can be archived when using Google data blindly.

As in many other papers, the goal of this example is to show whether Google Trends data helps to improve estimations and predictions of house prices, in this case represented by the S&P/Case-Shiller 20-city composite home price index (HPI). The results can be seen in table 3.1. To find predictors we make use of Google Correlate just like Scott and Varian (2015), Varian (2014) or Stephens-Davidowitz and Varian (2015). By doing so, the predictors are going to be single search terms. While those papers show advanced methods for variable selection, we choose the predictors manually, similar to Baker and Fradkin (2011, 2017). After uploading the data to Google Correlate, the top 100 most correlated series are gathered. Those predictors are then smoothed to reduce the impact of short-term fluctuations as Dietzel (2016) suggests. Furthermore, the dataset is split up in an estimation and a prediction set, as we want to analyze

6 Google also reports yearly "top-charts" in the manner of an end-of-the-year review; see "Year in Search", available at: `https://trends.google.com/trends/yis/2017/US` (accessed 2018-03-12).

7 Google Correlate website: `https://www.google.com/trends/correlate`.

8 See "Google Correlate FAQ", available at: `https://www.google.com/trends/correlate/faq` (accessed 2018-03-12).

the out-of-sample prediction performance. The estimation period ranges from March 2004 to December 2016 and the prediction period covers the first three months of 2017.

To keep things simple an autoregressive approach is chosen for the baseline model, just like Choi and Varian (2009, 2012), Mohebbi et al. (2011) or Wu and Brynjolfsson (2015) did, for instance. Following the latter we also include population as a control variable. For the Google predictors, we choose the terms `[magnetic door]`, `[fun videos]` and sure enough `[microwave baked potatoes]`. We do not need to doubt the choice of variables at this moment, as *correlation is enough*. The correlations with the S&P/Case-Shiller index for the estimation period can be seen on the right side in table 3.1.

Table 3.1: "Blind" estimation and prediction results

Dep. var.: HPI	I	II	III	IV	V	cor.
$\text{HPI}_{(t-3)}$	0.975 ***	0.763 ***	0.634 ***	0.526 ***		0.97
	(0.020)	(0.025)	(0.030)	(0.043)		
`[magnetic door]`		6.548 ***	6.313 ***	5.669 ***	6.310 ***	0.87
		(0.623)	(0.551)	(0.568)	(0.796)	
`[fun videos]`			3.848 ***	4.671 ***	9.895 ***	0.87
			(0.581)	(0.615)	(0.618)	
`[microwave baked potatoes]`				2.666 ***	9.967 ***	0.90
				(0.804)	(0.750)	
Controls (I-V)	Intercept, population					
AIC	964.9	871.4	833.6	824.6	928.9	
Adjusted R^2	0.944	0.967	0.975	0.976	0.953	
MAPE	1.020	0.380	0.200	0.140	0.350	
Improv. Adj. R^2		2.5 %	3.3 %	3.4 %	0.9 %	
Improv. MAPE		-62.7 % **	-80.4 % **	-86.3 % **	-65.7 % ***	

Note: Dependent variable: S&P/Case-Shiller 20-city composite home price index (HPI)
Standard errors in parentheses. Correlation with dependent variable (cor.).
Levels of significance: 0 '***' 0.01 '**' 0.05 '*' 0.1 '-' 1
Estimation: 03/2004 – 12/2016 | Forecast: 01/2017 – 03/2017.

As mentioned before, the base model I is an autoregressive approach with a 3-month lag. With this specification we get an adjusted R^2 of 0.944. For model II, the term `[magnetic door]` is added, which improves the fit of the model by 2.5 % to 0.967. In model III `[fun videos]` is added to the specification of model II, which again improves the model fit by 3.3 % compared to the baseline model. Model IV contains all three additional predictors, although it seems that `[microwave baked potatoes]` can't contribute considerably more, regarding model fit. However, the improvement of the adjusted R^2 by 3.4 % is the highest of all specifications. Following many other authors, model IV shows a specification without the autoregressive part. In this model the S&P/Case-Shiller index is estimated solely by Google predictors. Nevertheless,

the results are similar to the baseline model, even with a slight improvement of 0.9 % regarding the model-fit. These findings are very similar to those of Wu and Brynjolfsson (2015).

In a next step the models were used to predict January to March 2017. Like shown in the literature, a major part uses Google variables for short-term forecasting or nowcasting. Therefore, choosing a three-month forecast horizon seems appropriate. Based on this forecast the mean absolute percentage error (MAPE) is calculated to serve as an accuracy measure. As in Pavlicek and Kristoufek (2015), the Diebold-Mariano test is then used to check whether the change in forecasting accuracy is significant (see also Harvey et al., 1997). The improvement of forecasting accuracy is over 60 % for model II and V and over 80 % for model III and IV. All forecasts generated with the help of the Google search terms are significantly better than the forecast of the base model. In summary it can be said that the S&P/Case-Shiller index seems to be affected by the interest on `[magnetic door]`, `[fun videos]` and `[microwave baked potatoes]`. We were able not only to improve the estimations but also to archive significantly better forecasts using Google predictors. The AIC figures support this conclusion.

To make this clear: the results that have been shown here are not meaningful. The correlations and therefore the estimations were obviously spurious. Furthermore, the model selection did neither follow any scientific rules, nor was there an appropriate model diagnostic. But the overall approach and presentation follows many of the aforementioned authors. The only purpose of this section is to demonstrate that there are certain things to keep in mind when dealing with Google data and that not every "significant" outcome is also meaningful.

3.3.2 Interpretation pitfalls

The first important question worth asking is: What does Google Trends measure in the first place? A common misconception is that it reports the absolute number of search queries for a given search term. According to Google, however, each data point is divided by the total searches of the corresponding location and time range to compare relative popularity. Therefore, Google Trends calculates the ratio of the number of searches in relation to the total number of searches conducted at any given time and place. This makes Google Trends a relative index because the base – being the total number of queries for the respective time and place – changes over time. Google Trends itself explains that this relative form is reported because *"otherwise places with the most search volume would always be ranked highest"*. In

a last step the data is scaled on a range from 0 to 100, where the maximum of the series is set to 100.[9]

Further, there are some additional points to keep in mind: Firstly *"Google Trends has an unreported privacy threshold. If total searches are below that threshold, a 0 will be reported"* (Stephens-Davidowitz and Varian, 2015, p. 13). This means that it is more likely to encounter zeros (no data) when analyzing an earlier time period or smaller region. The second issue is that the data comes as a sample from the total Google search database, which can differ slightly from day to day. Researchers who want to get precise data can average the data from different days, although Stephens-Davidowitz and Varian (2015) state that this should not be necessary for the most cases, because the sampling generally gives precise results. Google Trends differs between "real time data" which is a sample of search queries from the last seven days and "non-real time data" which is a sample of the whole Google search database reaching from 2004 up to 36 hours prior to the request.[10]

Another problem one could run into is comparing search terms on the Google Trends website itself. You can compare up to five search terms at once. This can help interpreting the series, because if the value of one data point is twice as high as the value of a second data point – from a different series, but at the same point of time – the number of searches for the first data point was twice as large as for the second data point (see also Stephens-Davidowitz and Varian, 2015). The aforementioned scaling process of the data is done separately for each request, but not separately for each search term. So, for example if one is interested in comparing the relative search activity for a very popular search term and a very unpopular search term it could happen that term 2 shows up as zero, because of this normalization. But if this request is done separately for both search terms you get values other than zero for the cost of not being able to compare them, like you could have done the other way (see Stephens-Davidowitz and Varian, 2015).

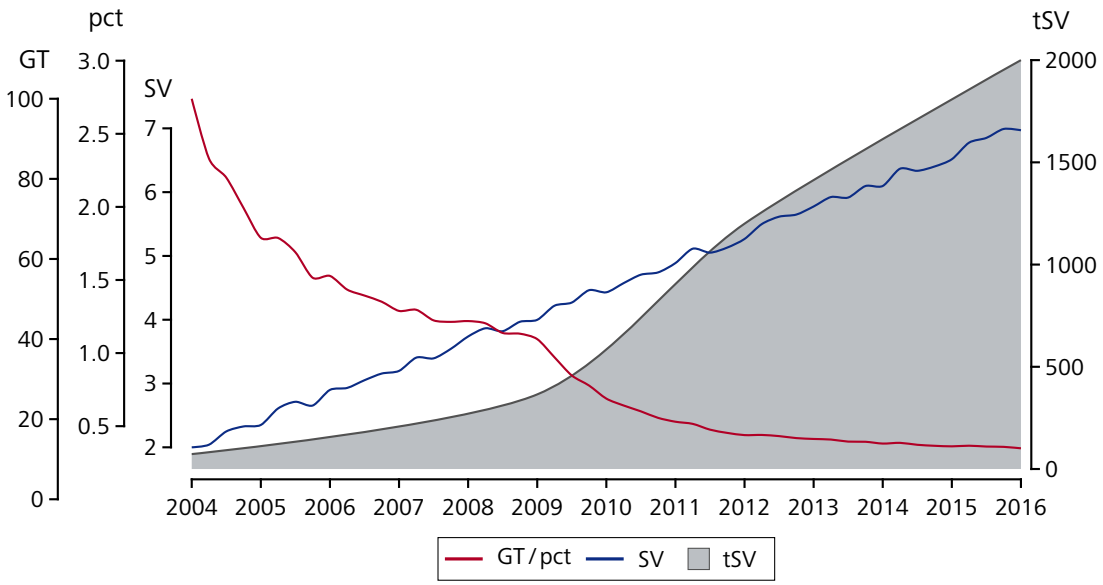
To illustrate the generation of Google Trends data, take a look at figure 3.1. The blue line shows a fictional number of searches for a certain term. These numbers are in billions, so in the beginning of 2004 round about 2 billion searches were made. This search volume (SV) can be read off the inner left axis. The interest on this fictional search term is constantly increasing over time so that in 2016 nearly 7 billion requests were made. The gray curve shows the total

9 See also "How Trends data is adjusted", available at: `https://support.google.com/trends/answer/4365533?hl=en&ref_topic=4365599` (accessed 2018-03-12).

10 See "Where Trends data comes from", available at: `https://support.google.com/trends/answer/4355213?hl=en&ref_topic=6248052` (accessed 2018-03-12).

search volume (tSV) worldwide on the right axis. This curve is derived from numbers released by Google.[11] Actually, it is a quiet difficult task to find reliable numbers concerning Google searches, especially when looking at specific regions or countries. There are indeed reports on the different search engines and their usage for certain countries, but these studies should be treated with care. For example, comScore releases a report containing actual numbers of searches and the market shares of the according search engines for the US-market, but one could quite easily miss the fact that these numbers only include searches from desktop PCs.[12] Mobile phones or tablets are not included, although they should have an enormous impact on the number of search queries. Google started its business already in 1998, but since Google Trends was developed later, figure 3.1 shows the time horizon from 2004 till 2016.

Figure 3.1: Exemplary calculation of Google Trends



Note: Google Trends (GT), percentage (pct), search volume (SV), total search volume (tSV), number of search volume in billions.

In 2004 Google announced that they had 200 million searches per day which brings us to 73 billion searches a year. Between 2004 and 2006 the number of searches was increasing moderately, but since 2007 the search queries rose almost exponentially. In 2009 Google announced that there were more than 1 billion searches each day, which results in 365 billion search queries a year. In the past few years the growth has slowed down a little bit, but all in all there should have been at least around 2 trillion searches in 2016.

11 According to "Google now handles at least 2 trillion searches per year", available at: `https://searchengineland.com/google-now-handles-2-999-trillion-searches-per-year-250247` (accessed 2018-03-14).

12 See "comScore Releases February 2016 U.S. Desktop Search Engine Rankings", available at: `https://www.comscore.com/Insights/Rankings/comScore-Releases-February-2016-US-Desktop-Search-Engine-Rankings` (accessed 2018-03-12).

To realize these numbers is important, as they are the base of the relative index mentioned beforehand. The first step is to calculate the query share or in other words dividing SV by tSV for each data point. This gives the percentage (pct) of the search and is represented by the red line with the middle axis on the left. For example: in 2004 73 billion queries were made in total. These included 2 billion searches concerning the fictional term, which translates into roughly 2.7 %. In other words: 2.7 % of all worldwide searches in 2004 were related to this fictional search term. In a second step the new series is scaled with its maximum being 100. This does not change the curve, simply its scale. This final transformation – the constructed "Google Trends" (GT) data – is shown on the left axis on the left side and again represented by the red line.

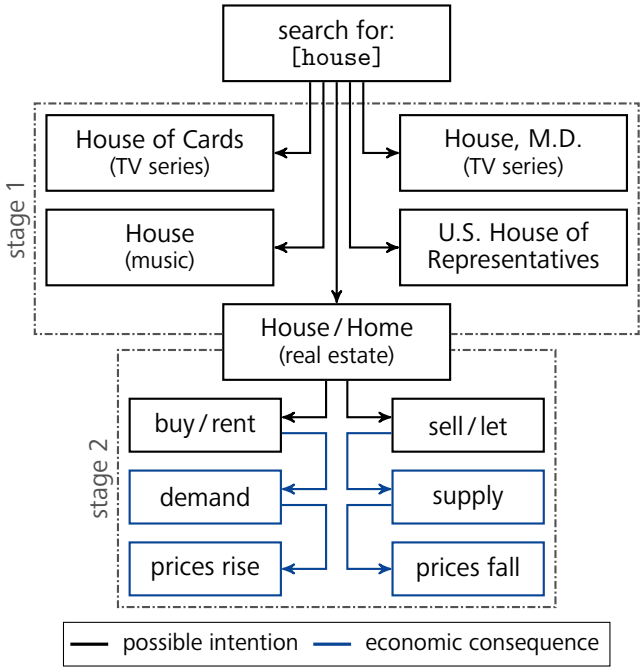
Now a crucial interpretation pitfall is becoming obvious: Although the absolute number of searches was increasing the whole time period, the Google Trends time series is decreasing. That is because of the varying base of this index. It has to be understood that Google Trends data does not report the interest, but the relative interest in a certain search term. This differentiation is important, because although the interest might be increasing, the relative interest can be decreasing. Stephens-Davidowitz and Varian (2015) address this issue giving an example with the search term `[science]`. Since 2004 the relative interest on `[science]` in the US seems to decrease. But this is not due to the fact that less people are interested in this topic. Rather it is because the internet in 2004 was mainly used by colleges, universities or researchers. Today the internet is a stage for a much broader audience with highly diverse interests. Therefore, the relative share of internet users looking for `[science]` is smaller even though the absolute number is increasing. So, interpreting long-term trends based on Google Trends data should be done with caution.

3.3.3 Query selection

When working with Google Trends or Correlate data, query selection will be the number one priority when setting up the dataset, because you might have to deal with ambiguity. It has to be considered very carefully which data you want to choose and why. Of course, in every analysis the composition of the dataset is very important, but with "usual" data ambiguity is rarely a problem. There are just a few things to check, like whether the time horizon fits your study, whether it relates to the correct location and so on. With Google the variable selection or rather query selection is a topic for itself.

Let's suppose you want to analyze the job market because you make the assumption that more / less interest in jobs affects the job market in some way. Therefore, you search for the term `[jobs]` and get a time series on the relative interest in `[jobs]`, which is then used for the study, just like it is done by Baker and Fradkin (2011). But what you don't know is that the resulting time series also contains information on `[steve jobs]`. That is because Google Trends shows all queries containing the word `[jobs]` and – if not excluded – `[steve]` is a part of that. Of course, this time series is biased to an unknown extent. In fact, Baker and Fradkin (2017) account for this issue in a later version of their paper. The example above describes the problems that can occur when dealing with ambiguous search terms pretty well, but to illustrate the problem even further, take a look at figure 3.2.

Figure 3.2: Ambiguity in search terms



Note: Simplified illustration of two stages with ambiguous search terms.

Now let us assume that you want to analyze the impact of search queries on house prices. You suspect that an increasing interest in housing markets could affect the house prices. The underlying idea would be that people that inform themselves frequently about house prices, loans or the situation on the housing market in general are more likely to buy a house. A rising interest in the housing market may indicate that the demand for housing is rising as well and therefore adjustment of house prices should be observable.

Analogous to the aforementioned example you choose the single search term [house] to analyze this presumption. The ambiguity pitfall appears in two stages, as figure 3.2 illustrates. The first stage would be the obvious ambiguity of the search terms. With Google Trends it is only possible to observe what has been searched, but not in which context.[13] By looking at figure 3.2 it cannot be said whether the intention of this query was to search for [house] in the sense of housing market or whether the user was looking for house music or his/her favorite TV show. It is also worth mentioning that Google Trends data is highly affected by seasonal events. During the preparation of this paper a U.S. election took place and the recommended searches suggested *U.S. House of Representatives* when entering [house] in Google. Therefore, attention must be paid to all the different interpretations or meanings of the search terms chosen. Even the point in time can play a significant role. For instance, Google users certainly had different motivations when searching for [Lehman Brothers] before and after the year 2008.

For now, let us assume that we got the desired Google Trends data for [house] in the sense of real estate. Here the second stage comes into play. Unlike stage 1, this stage does not deal with the obvious double meaning of specific words, it rather covers economic consequences that result from the immanent intention of the user, which again is not observable. Looking at figure 3.2, users could search for [house] because they want to buy a house. If the data shows a rising level of relative interest on [house] this would indicate that more people than usual want to buy a house and this could be seen as an increase in demand. In this case the data should (theoretically) support our assumption that the prices rise as well. However, at the same time users could also want to sell when searching for [house], which leads to a contrary interpretation. In this case the supply would rise and therefore the prices are expected to decrease. With such a simple approach like the keyword [house], a solution in the manner of "just use [buy house] or [sell house] as search terms" is obviously tempting (and appropriate), but this straightforward example also holds true for not so obvious phrases.

Taking this one step further we assume that there is an increase in the interest on [house] but this time the rising interest is caused by buyers and sellers at the same time to equal parts. In this case we clearly would observe a rising interest in the Google data but (theoretically) the prices should stay the same, because demand and supply increase to the same extent. So, there is definitively a problem when the demand and supply sides cannot be differentiated. Of

13 Google Trends categories can help to circumvent this kind of problem to a certain degree and they will be discussed later on.

course the search terms have to be chosen in consideration of the research question. If you are analyzing the number of houses sold, it should not matter this much whether the search queries are supply or demand related as both affect the number of houses sold in the same way (see also Wu and Brynjolfsson, 2015).

If fact, Askitas (2016) addressed this problem while constructing a US housing market index based on a buy / sell ratio. In order to archive this, the author uses punctuation filters in Google Trends, which is basically a certain syntax used when entering the search terms. Everyone can do this to avoid ambiguous search results to a certain degree, especially since the "syntax" consists only of four characters, giving you the ability to perform logical operations. A `[ ]` (space) means `AND`, the `[+]` (plus) stands for `OR`, the `[−]` (minus) for `NOT` and terms in quotation marks make sure that only the exact phrase is included in the results. So, if you were to search for `[jobs]` but not caring for `[steve]`, you should type in `[jobs -steve]`.[14]

Of course this is a very simple example, but it illustrates potential pitfalls when dealing with a single search term. Therefore, Google Trends offers different categories, that is, a collection of different aggregations of various search terms on a certain topic. At present there are 25 top level categories and about 1,100 at the second level.[15] These categories avoid the stage 1 ambiguity at least to a certain extent. Each of these categories is a single time series on its own, generated through the aggregation of specific search terms. There are categories for nearly anything, reaching from *Food & Drink* over *Beauty & Fitness* to *Real Estate*. The sub-categories can be used to improve the accuracy for the respective task even further.

The issue with those categories is that you only get their names, but no information on what they are containing. There is no possibility to check the search terms associated with a certain category. Choi and Varian (2012) state that the assignment procedure is made by a natural language classification engine which is probabilistic. Therefore, a query such as `[apple]` could be partially assigned to *Computers & Electronics*, *Food & Drink* and *Entertainment*. This leads back to the stage 2 ambiguity. If you don't know which terms a certain category is containing, you can't be sure which economic consequences you should expect. It is worth pointing out that there is a category called *Real Estate Agencies* which is used in many publications to forecast house prices, like for example Wu and Brynjolfsson (2015) or Bennöhr and Oestmann

14 See "Search tips for Trends", available at: `https://support.google.com/trends/answer/4359582?hl=en&ref_topic=4365530` (accessed 2018-03-12).

15 Number of categories obtained through `R`-package `gtrendsR`. Last checked in March 2018. See also "Google Trends Categories - Category listing", available at: `https://github.com/pat310/google-trends-api/wiki/Google-Trends-Categories` (accessed 2018-03-12).

(2014). Here we have basically the same issue as in figure 3.2: Do users search for agencies because they want to buy or because they want to sell? However, since a category is an aggregation of probably several hundreds of thousands of keywords, they are much less sensitive to ambiguous or wrongly assigned search terms or even seasonal effects of single keywords.[16] Therefore, the categories still have a huge advantage over single search terms.

But not only ambiguity is a serious problem, there is also an issue with spurious correlations as seen in the estimation at the beginning. As these correlations are purely random, there is no causal relationship between the variables and therefore the estimations might indeed give results, but they will not be meaningful. Spurious correlations are an issue especially (but not only) when working with Google Correlate, because the search terms have to be chosen by hand. A good example can be found in Varian (2014), illustrating different ways for variable selection methods he attempts to forecast new home sales with the help of Google Correlate data. It turns out that `[oldie lyrics]` is the second best predictor for new home sales.[17] This is not a big problem, if it is this obvious. It is more difficult if ambiguity hides spurious correlations. As mentioned above, the query selection process is considered particularly important as the research results stand and fall with this choice. Google Trends categories can help to avoid potential pitfalls to a certain degree, but query selection still is a challenging task.

3.3.4 Practical problems

A practical problem one could run into is that Google Trends and Google Correlate data are actually different transformations of the same data generating process. Therefore, the statement that *"Google Correlate is like Google Trends in reverse"* is somehow misleading as the two cannot be compared or converted into each other without further ado. Google Correlate in contrast to Google Trends does not scale the data. Here the data is standardized, so the units are standard deviations above mean.[18] This leads to certain differences regarding interpretation and comparability. Additionally, with Google Correlate there is only the option to specify the country, in contrast to Google Trends, where the desired region can be restricted not only to countries but also to states and big cities.[19] This means also that Google Correlate cannot control the results with categories or any other filters. This is a disadvantage, because

16 This number is a wild guess by me, as there is no information on the composition of the categories available.
17 Needless to say that he excluded this predictor right away.
18 See "Google Correlate FAQ" (footnote 8).
19 The option to choose cities is restricted to the United States.

there is a very high probability of catching spurious correlated search terms, which have to be sorted out by hand.

But the most fatal deficiency is the different time horizon. Google states that Google Correlate contains data from January 2003 to present and is updated weekly.[20] In fact, data returned by Google Correlate has a lag of at least 12 months. When tested in March 2018, the returned correlations reached only until March 2017. In a scenario where current data is needed – which is the key element of the whole search engine data idea – Google Correlate is not an option. To the authors knowledge there is no comment in the FAQs / Blogs indicating why that is the case. Also, it may be possible, that Google Correlate has been shut down completely, without any announcement.

Preis et al. (2010) also make an important consideration when estimating the S&P 500 with the help of Google Trends, as mentioned above. They find clear correlations between the stock index and the search volume of the corresponding company names. On the other hand, they also point out that the most likely reason to search for company names, relating to the S&P 500, would be media coverage. Therefore, Google Trends acts like a proxy variable for news. However, they also find that the current price movements seem to affect the search volumes in the following weeks. Because news frequently comment on current price movements, this could increase the interest in various companies. Therefore, the question of endogeneity has to be considered.

Additionally, there are some econometrical subtleties to keep in mind, especially when dealing with Google data: under- or overfitting and multicollinearity. Underfitting, also known as omitted variable bias, occurs when an important determinant of the dependent variable is omitted from the estimation. The result would be that the coefficients of all other variables would be biased and inconsistent. When dealing with big data in general, usually the problem would be having too many predictors, but it can be the case, for example, if one was to run an estimation with Google Trends variables being the only regressors. Wu and Brynjolfsson (2015), for instance, present a model, where a house price index is estimated only with two Google Trends categories and their respective lags. But here another problem arises. Multicollinearity occurs, when two or more of the predictors are highly correlated. Therefore, using two Google Trends subcategories from the *Real Estate* category alone, could imply a problem

20 See "Google Correlate FAQ" (footnote 8).

with multicollinearity, but when adding the 1-month lag of each series as well, this is almost certainly an issue.

As mentioned, when dealing with big data, usually the problem is to find the appropriate predictors in an appropriate number. Burnham and Anderson (2002, p. 17) state that with many predictors it may be possible to receive *"regression equations with high R^2 values, 'significant' F values, and many 'significant' regression coefficients, […] even if the explanatory variables are independent of y"*. Usually, a very large number of predictors is needed to become a problem. But imagine someone was to estimate a regression with the top 100 correlated search queries as the predictors. This may seem a little far-fetched, but the variable selection process is a serious concern when working with search engine data. Varian (2014, p. 18) states *"that there are billions of queries, so it is hard to determine exactly which queries are the most predictive for a particular purpose"*. Even after classifying these queries in categories, there is still a long list of predictors worth considering so that overfitting and spurious correlation are a serious concern. Scott and Varian (2015) show different approaches for model selection, when dealing with a large number of predictors.

3.3.5 Reliability and replicability

Reliability of data and replicability of results are foundations of empirical research. This should also hold true when working with search engine data. The question is whether Google data is reliable and whether the results are replicable. In the previous sections a sampling process was mentioned that is used by Google Trends and Correlate to generate their data. Strictly speaking this could be a problem of replicability, because there is a new sample every day. But one could get over this, if the sample is precise enough. But there is another side of the sampling error. This time not the one in the data generating process but rather the question whether internet users represent a random sample of the population one is researching on (e.g. home buyers / sellers). Stephens-Davidowitz and Varian (2015, p. 17) state that *"in 2004 the internet was heavily used in colleges and universities […]. By 2014, the internet had a much broader population of users"*. Mohebbi et al. (2011, p. 2) also point out that *"while Internet users do not represent a random sample of the United States population, this population has become increasingly less biased over time"*. Again, it is difficult to get exact numbers on internet usage and depending on the source the numbers may vary. Pew Research Center released a fact sheet on internet usage which shows that 52 % of the American adults used the internet in

2000 whereas the share increased to 89 % in 2018.[21] The share of internet users has indeed increased drastically over the last years, which indicates that the sample truly has become less biased. Wu and Brynjolfsson (2015) state that, according to the National Association of Realtors, 90 percent of home buyers in the US used the Internet to search for a home in 2012. Nevertheless, one has to keep in mind that this question heavily depends on the country, region and topic under investigation. Since Google Trends does not provide information about the underlying volume of search queries, this random sample is not guaranteed when analyzing less popular topics or regions. It would be a different case, if one was to analyze the impact of search queries in Eritrea, where roughly 1 % of the population uses the internet.[22] It is also a different situation whether one is analyzing the housing markets or commercial real estate (like REITs, pension funds and other), where investors might have other sources of information than Google (see Heinig et al., 2016). Furthermore, we have to ask not only whether the internet users represent an unbiased sample, but also whether Google users represent an unbiased sample. Although Google has a dominant position in the search engine markets worldwide, there are a few exceptions where other search engines are more popular, like China or Russia.[23] But the key point is that Google is not necessarily needed, even if you are planning to inform yourself about real estate or any other big investments. Wu and Brynjolfsson (2015, p. 115) evoke that *"some consumers may bypass the search engine all together and go directly to certain websites […] when considering buying and selling a home. Others might have a long-standing relationship with a trusted realtor or do not use the Internet. Using Google Search alone would miss these types of consumers."*

There are, however, much more serious issues one has to think of when dealing with Google data. In the literature section the Google Flu Trends project was mentioned. This is a textbook example for a big and publicly known project. In the meantime, it has been stopped and the following points show possible reasons why it failed. The upcoming pitfalls are of concern for any research efforts, sometimes to a greater, sometimes to a smaller extent. Firstly there is a slight chance for *red team attacks*, like Lazer et al. (2014a) name them. They occur when users intentionally want to manipulate the data through mass-generation of fake search queries. It is not very likely to happen when doing "usual research", but for something as big as Google Flu

21 See "Internet / Broadband Fact Sheet", available at: `http://www.pewinternet.org/fact-sheet/internet-broadband/` (accessed 2018-03-12).

22 See "Country ICT data (until 2016) – Percentage of Individuals using the Internet", available at: `https://www.itu.int/en/ITU-D/Statistics/Pages/facts/default.aspx` (accessed 2018-03-12).

23 See "Share of desktop search traffic originating from Google in selected countries as of October 2017", available at: `https://www.statista.com/statistics/220534/googles-share-of-search-market-in-selected-countries/` (accessed 2018-03-12).

Trends it could become a problem. Additionally, there does not always have to be a malicious intent for distorting the data. Sometimes other circumstances, like media coverage, provoke an "unjustified" increase in search activity. In his article about Google Flu Trends Butler (2013) states that reinforced press reports may have encouraged healthy people to search for flu-related topics. The base of this whole approach is the assumption that search engine data is an aggregation of intention signals, which reflect the unbiased interests of the users. In other words, search behavior is assumed to be endogenous. Google Flu Trends, however, showed that this assumption is at least questionable. People will also search when certain (maybe unknown, unexpected or even undesired) triggers awaken their interest. One can imagine that during and after the financial crisis a large share of the internet users informed themselves solely because they wanted to know what was going on and not because they were somehow affected.

This leads us to the second issue one has to be aware of. It is what Lazer et al. (2014a) call *blue team dynamics*. These dynamics describe adjustments in the Google search algorithm, other changes to the functionality of Google or even changes to the data itself. Google states that there are more than 500 improvements to the search algorithms in a typical year, which gives us approximately 6,500 changes in the time between 2004 and 2016.[24] Taking this one step further, Google introduced *related searches* which are the additional, similar search terms that appear when searching for something.[25] These terms are exogenously generated and therefore the searches induced by these terms are exogenously affected. Consequently, additional search volume is created, that may or may not bias the true "sentiment". From Google's perspective this is of course necessary and desired as they want to improve their services, but in terms of replicability it could complicate things. When working with the Google Trends website and looking for historic data in the US, you will eventually find a note informing that *"an improvement to our geographical assignment was applied from 1/1/2011"*. One could wonder what these improvements would be and currently there is no more information available. There was, however, a short comment in the Google Trends help website – which was documented by Lazer et al. (2014c, p. 15) – that notified the users that *"this update was applied retroactively [to] provide even better geo-location data for search queries [and] may manifest itself in certain queries as a discontinuity in the trend line"*. This, indeed, is a big deal,

24 See "How Search Works: From algorithms to answers", available at: `https://www.google.de/insidesearch/howsearchworks/thestory/` (accessed 2018-03-12).

25 See "Organizing lists of related searches", available at: `https://search.googleblog.com/2011/06/organizing-lists-of-related-searches_16.html` (accessed 2018-03-12).

because Google changed the data retroactively. This means that everybody who used the data before the update will no longer be able to reproduce the results after the update, as the data is not the same anymore. Additionally, this update was not announced in any way, neither was it documented. There was only the short note in the help section of Google Trends and even this note is deleted by now. During the development of this paper another change was applied. Once again the only information concerning this change can be found in a note on the graphs of Google Trends, which states: *"An improvement to our data collection system was applied from 1/1/16"*. The exact date is not known, but it was sometime in the second or third quarter of 2017, which means that once more the data was changed retroactively. Screenshots of both notes can be found in figure 3.A.1 in the appendix. Yet another example can be found in the FAQs of Google Correlate where they inform that they changed the sample size for the US in December 2011. Further they state: *"while this does not have much of an effect on popular queries, it may cause a noticeable increase in variance for queries with lower volumes"*.[26] Again, there had been changes, this time not directly to the data, but to the data generating process. These blue team dynamics are not only of concern for the data itself, but also the classification of the data. Choi and Varian (2012) state that there are 30 categories and approximately 250 subcategories. However, three years earlier Choi and Varian (2009) wrote that there are 27 categories with 241 subcategories. As mentioned above, looking at those categories in 2018 there are 25 main categories with roughly 1,100 subcategories. Therefore, it is safe to say that there have been some drastic changes as well.

Now what do we do with this information? First of all, it seems that those major changes are not the rule but the exception. However, if they occur they can have large consequences. To put it in exaggerated terms: Google Trends data one might currently work with could be gone before the task is completed. Of course, there is a possibility for other data sources to undergo substantial changes as well, but in such case there would be announcements and / or a proper documentation.

3.4 Empirical Analysis

This empirical analysis joins the literature and follows a quasi-nowcasting approach. As mentioned in the beginning, another common characteristic of many papers is that the comparison

26 See "Google Correlate FAQ" (footnote 8).

only happens between the base model and the models with Google variables. Other model specifications (without Google variables) are generally not taken into account, although they could deliver similar results. This method is chosen, presumable due to the fact that most of the possible predictors are not yet available. There are, however, certain variables and model specifications that allow a proper comparison to Google Trends models. The aim of this section is to illustrate whether the results are still in line with the literature after accounting for the possible difficulties mentioned in the last section and compare those results to those of equally specified "standard" models.

As shown above, a large share of the literature uses a very simplistic approach when trying to evaluate the benefits from Google data. For the most part the strategy is to set some kind of base model. A very frequent specification for the base model is an autoregressive process without additional variables, which is then compared to more specified models including the Google variables. Nearly all of these papers evaluate their results, inter alia, on the basis of model fit and some forecast error measurement. The rationale behind this approach is perfectly clear, because it has to be ensured that the data necessary for predictions is available.

$$y_t = \beta_0 + \beta_1 y_{t-3} + \beta_2 x_t + \varepsilon_t \tag{3.1}$$

$$y_{t+1} = \beta_0 + \beta_1 y_{t-2} + \beta_2 x_{t+1} + \varepsilon_t \tag{3.2}$$

$$y_t = \beta_0 + \beta_1 y_{t-3} + \beta_2 x_{t-1} + \varepsilon_t \tag{3.3}$$

$$y_{t+1} = \beta_0 + \beta_1 y_{t-2} + \beta_2 x_t + \varepsilon_t \tag{3.4}$$

Assume the values of x and y to be known at time t. In equation 3.1 variable y_t depends on its own previous values y_{t-3} and some variable x_t. Here all values are known. But if someone would forecast y_{t+1}, like shown in equation 3.2, the value of x_{t+1} would be needed. This forecast could not be calculated, unless x_{t+1} would be predicted beforehand. In equation 3.3, y_t again depends on its own previous value y_{t-3}, but this time on a lagged version of x. This time the forecast of y_{t+1} can indeed be calculated because all values are known, as equation 3.4 shows.

In theory we can assume that the values of x and y are known at time t, but in fact most of the data is published with some delay, which can reach from some days to half a year or even longer. Therefore, the models above would have to be adapted. Because Google data is available almost instantly, it is possible to calculate the forecast of y_{t+1} with models like

equation 3.4 immediately in time t. This would not be possible with other predictors, if they are released with a lag of several months. But not only the release of predictors can be delayed. It is for this reason that the term *nowcasting* goes along with Google Trends data. If anyone wants to estimate the current values of y_t, because they are not yet known, it is possible to do so with models like equation 3.3 or even 3.1.

3.4.1 Dataset

For this empirical analysis two types of data are used. The first one is of course the Google Trends data. The second one is economic data from the FRED database (Federal Reserve Bank of St. Louis).[27] This database was chosen because it offers a broad variety of time series and the data is freely accessible. As the majority of the available data relates to the United States, the geographical focus of this study shall be as well the US residential real estate market. Concerning the Google data this geographical choice can only be beneficial as Google Trends is most developed for the US.

The dependent variable chosen for this study is the seasonally adjusted S&P/Case-Shiller 20-city composite home price index. This index is a common used proxy for the US housing market as it reflects the home prices of 20 metropolitan statistical areas (MSA) all over the US and therefore represents the most important US residential markets.[28] The following dataset contains various time series that are potentially influencing real estate markets in general and this index in particular. A detailed list of the variables and the corresponding identifiers for FRED can be found in table 3.A.1 in the appendix. Some of these variables are somewhat overlapping in the sense that they measure very similar or even the same matters with small deviations (for example: total construction spending vs. total private construction spending). Therefore, they are grouped into different topics. In anticipation of the model selection later on, it is worth noting that variables of the same topic are not allowed to be in the same model specification to prevent potential problems with multicollinearity. Furthermore, not all of these variables are going to be in the final models, although a broad variety of variables is helpful for the model selection process. All in all, there are eight categories and 24 FRED variables for the model specification.

27 See `https://fred.stlouisfed.org`.
28 For more detailed information about calculation or index composition see S&P CoreLogic (2018).

Because many authors in the above-mentioned literature argued that Google Trends data functions as a sentiment indicator, other sentiment indices were included: the *Coincident Economic Activity Index for the United States*, as well as the *Leading Index for the United States*, both released by the Federal Reserve Bank of Philadelphia, the *OECD Indicator for the United States* from the Organization for Economic Co-operation and Development (OECD) and the *Consumer Sentiment index* from the University of Michigan. Assuming that the S&P/Case-Shiller 20-city home price index depends on a kind of economic indicator, the integration of those indices should help to find out whether Google Trends data is an alternative to "established" sentiment indicators (similar to the work of Heinig et al., 2016).

Of course, variables like GDP, CPI and Population should be included in the dataset as well to account for economic and demographic changes. Personal disposable income, consumption expenditures and saving rate might have an influence on the house prices as well. Furthermore, (un)employment indicators – like the employment-population ratio and the unemployment rate itself – as well as interest rate proxies – like the 15/30-year mortgage rate or the 3/6-month treasury bill – were considered. To avoid potential difficulties with a structural break, there is also a dummy variable for the financial crisis that occurred between January 2008 and June 2009. Some housing market indicators were included as well: the median sales price for new houses in the US, the number of building permits for new houses, the number of new houses built and the construction activity. Since Google Trends data is available from 2004 onwards the range of all variables from FRED was chosen accordingly. Most of the time series were on a monthly basis, but some had a higher frequency and therefore were aggregated to monthly values.

This leads us to the Google Trends data. As mentioned above, there are several possibilities to gather this data. The most common used possibility in the present literature is the usage of the different real estate categories offered by Google Trends. As mentioned above, there are 25 top level categories with about 1,100 subcategories to choose from. Luckily, there are only 9 categories relating to real estate. The top level is called *Real Estate* with the subcategories being *Apartments & Residential Rentals*, *Property Development*, *Property Inspections & Appraisals*, *Real Estate Agencies*, *Real Estate Listings*, *Property Management*, *Commercial & Investment Real Estate* and *Timeshares & Vacation Properties*. The main issue with these categories is – as mentioned multiple times – that there is no possibility to get an idea of the included search

terms or the differences between the categories. Nevertheless, with exception of the last three categories, all of the above are going to be considered for the estimations.

The most simple approach to find Google predictors – as suitable for the project as possible – would be to search for single keywords. But the pitfalls of this method should be clear by now. A single keyword is highly vulnerable to seasonality, ambiguity and other problems. But calculating an own index, specifically made for the desired task could be a potential solution. Effectively, this is the same concept as the original Google Trends categories and was done beforehand, for example by Askitas (2016), Heinig et al. (2016) or Askitas and Zimmermann (2009). Here, we follow two different approaches: Firstly, there is a Google keyword index, which is intended to be a very general index that consists of 32 keywords associated with the U.S. real estate market. Those keywords were chosen so that a broad coverage is ensured. The second index is meant to reflect specifically the S&P/Case-Shiller 20-city composite home price index, so the chosen keywords were basically the names of the 20 cities combined with the keyword `[house]`.[29] To avoid ambiguity as much as possible, all search terms where retrieved from within the *Real Estate* category. A list of the used search terms can be found in table 3.A.2 in the appendix.

Having the variables ready one substantial step is to check for stationarity and possible seasonality issues. The latter does not appear to be a problem with the FRED data, as it is available in seasonal adjusted versions. The variables were checked using the Augmented Dickey-Fuller (ADF) test. In case of non-stationarity different transformations of the time series were tested, like for example growth rates, differences or the Hodrick-Prescott-filter. This leaves us with a database, containing 24 FRED time series, 6 Google Trends categories and 2 custom-made Google indices. Additionally, lagged versions of the variables were included as well. The data is reaching from the beginning of 2004 to October 2017 on a monthly basis.

3.4.2 Econometric approach and evaluation

This section addresses the econometric strategy and the model selection procedure. Keeping the last sections in mind, the assumption of this study requires that search activity on Google somehow affects house prices. But it has also been shown that this assumption is not as resistant as some would say. What if media would report about increasing house prices? If people are interested, they most probably will "google" it. In this case the house prices would

29 A list of those cities can be found in S&P CoreLogic (2018, p. 11).

affect the search behavior. Admittedly, this probably would be the case only for very drastic price changes, excessive media coverage or certain incidents like a financial crisis. However, for the purpose of this investigation, this endogeneity problem can be easily solved by using lagged versions of the Google variables. Current house prices most likely will not affect the internet search activity from yesterday.

With the data being stationary, we have a database reaching from the February 2004 to October 2017 which results in 165 observations for each variable, which will be used in a basic Ordinary Least Squares (OLS) framework. The intention is to compare the models on the basis of model fit (adjusted R^2), the Akaike-Information-Criterion (AIC) and an out-of-sample prediction error measurement, namely the Mean Absolute Error (MAE). Therefore, the database has to be divided in an estimation and prediction set. As mentioned above, this empirical example joins the literature and follows a quasi-nowcasting approach. Since the most recent figures of the S&P/Case-Shiller 20-city home price index are typically released with a lag of two to three months, the forecasting horizon is set to be three months. Therefore, the database is split into two parts. The first part, reaching from February 2004 to July 2017, acts as the estimation dataset. The second part, reaching from August to October 2017, represents the prediction set, which is used to evaluate the forecasts.

The design of this study is oriented towards the existing literature to get a better comparison, of course considering the properties of the present data. First of all, this means that there is going to be a baseline model, which will be a simple regression with an autoregressive term as the only predictor. The second step is to select suitable models, containing one of the Google variables. The intention is to find similar models with slightly different specifications to get a valid comparison as well as to check for robustness. In contrast to the existing literature, the goal is also to find "standard" models without any Google variables that can be contrasted against the other models as well as against the base model. Additionally, these models have to be suited for nowcasting, like shown at the beginning of this section. Thus, a restriction of this study is that model specifications which would require data that is usually not available at time t, are not allowed. This simulates a practical application of a nowcasting process and ensures that the "standard" models stay comparable to the Google models. To avoid multicollinearity issues, variables of the same topic are not allowed to be in the same model, as mentioned above. An example would be the 15-year vs. 30-year fixed mortgage rate or the different Google Trends categories. Table 3.2 shows the results.

Table 3.2: Estimation results: Base vs. Google vs. "standard"

	base	
Dep. var.: HPI	Estimate	Std. Error
HPI $_{(t-3)}$	0.8277	0.0509 ***
Controls	Intercept	
Adjusted R^2	0.7009	
MAE	0.0041	
AIC	-1,267.6	

	A-I		B-I	
Dep. var.: HPI	Estimate	Std. Error	Estimate	Std. Error
HPI $_{(t-3)}$	0.7437	0.0653 ***	0.7907	0.0606 ***
GT Property Inspections Appraisals $_{(t-1)}$	-0.0003	0.0001 *		
Coincident Economic Activity Index $_{(t-3)}$			-0.8279	0.3876 **
Controls (A & B)	Intercept, crisis dummy			
Adjusted R^2 / Improvement	0.7056	0.7 %	0.7202	2.8 %
MAE / Improvement	0.0037	-9.8 %	0.0039	-4.9 %
AIC	-1,260.8		-1,276.5	

	A-II		B-II	
Dep. var.: HPI	Estimate	Std. Error	Estimate	Std. Error
HPI $_{(t-3)}$	0.7190	0.0675 ***	0.7699	0.0611 ***
GT Property Inspections Appraisals $_{(t-1)}$	-0.0003	0.0001 **		
Coincident Economic Activity Index $_{(t-3)}$			-0.8188	0.3770 **
US 30-Year Fixed Mortgage Rate	0.0297	0.0148 **	0.0267	0.0133 **
Controls (A & B)	Intercept, crisis dummy			
Adjusted R^2 / Improvement	0.7175	2.4 %	0.7297	4.1 %
MAE / Improvement	0.0038	-7.3 %	0.0040	-2.4 %
AIC	-1,266.6		-1,281.2	

	A-III		B-III	
Dep. var.: HPI	Estimate	Std. Error	Estimate	Std. Error
HPI $_{(t-3)}$	0.7317	0.0657 ***	0.7565	0.0616 ***
GT Property Inspections Appraisals $_{(t-1)}$	-0.0003	0.0001 **		
Coincident Economic Activity Index $_{(t-3)}$			-1.2172	0.4424 ***
US 30-Year Fixed Mortgage Rate	0.0282	0.0144 *	0.0264	0.0131 **
Civilian Unemployment Rate $_{(t-3)}$	0.0049	0.0029 *		
Real Gross Domestic Product $_{(t-9)}$			0.0509	0.0256 **
Controls (A & B)	Intercept, crisis dummy			
Adjusted R^2 / Improvement	0.7227	3.1 %	0.7335	4.7 %
MAE / Improvement	0.0038	-7.3 %	0.0040	-2.4 %
AIC	-1,268.6		-1,282.5	

Note: Dependent variable: S&P/Case-Shiller 20-city composite home price index (HPI)
Base model (base), Google models (A-I–III), "standard" models (B-I–III)
Standard errors calculated using heteroskedasticity and autocorrelation consistent (HAC) estimators; improvement compared to base model.
Levels of significance: 0 '***' 0.01 '**' 0.05 '*' 0.1 '-' 1
Estimation: 02/2004 – 07/2017 | Forecast: 08/2017 – 10/2017.

For the model parameterization, a stepwise regression approach was used over many different specifications, trying to find a compromise between maximizing the goodness of fit and minimizing the forecast error. Additionally, the AIC was used to further support the model selection. It is worth pointing out that there were many different specifications which could be worth considering. The following results are not a definitive list. Nevertheless, there were also many models which did not pass the criteria of this study. These include the custom-made Google Trends indices. However, these variables were mainly added to show the possibilities with and capabilities of Google Trends rather than fully expecting them to work properly.

The base model has a three-month lagged autoregressive term as its only independent variable. The 3-month lag comes with the restriction of the study design. Otherwise, one would not be able to calculate the appropriate nowcasts. The adjusted R^2 with a value of 0.7 is an usual range for an autoregressive model of this sort. The comparison models are structured in two columns: column A presents the Google models and column B the "standard" models. All models have the dummy variable for the financial crisis included. In order to avoid autocorrelation and heteroskedasticity bias, the standard errors are estimated via a HAC-variance-covariance-matrix (see Zeileis, 2004).

For the first set (A-I and B-I), the Google variable *Property Inspections & Appraisals* as well as the *Coincident Economic Activity Index for the United States* are added to the base model, respectively. Both variables act as a kind of economy or market sentiment indicator. The difference between A-I and B-I can be found in the lag of the economic activity index. This three-months lag, again, is needed to maintain the restriction of the study design, as this variable is also published with a lag of two to three months. Comparing both A-I and B-I to the base model, the results show a slightly higher adjusted R^2 as well as a lower forecasting error. The improvement in percentage values can be found next to the respective figure.

Regarding models A-II and B-II, the 30-year fixed mortgage rates were added, as an interest rate measurement seems to be reasonable in a real estate context. Again, the results show an increase in model fit and decreasing forecasting errors, compared to the base model as well as to the first set. It is to note that the mortgage rates are not lagged. In this case, however, this is no problem, because mortgage rate data is released weekly, so the availability is ensured. Nevertheless, if the nowcasting process should be actually performed on a daily basis, the models would have to be changed slightly.[30]

30 For instance, with a lag of 1 week for the mortgage rates.

For the third stage (A-III and B-III) two macroeconomic variables – the unemployment rate and the GDP– were added, respectively. Once again, the model fit and forecasting errors have improved, compared to the prior specifications. It is noticeable, however, that for the Google model A-III the mortgage rate as well as the unemployment rates are now only significant at a 10 % level.

In general, the coefficients and standard errors are quite consistent over all specifications. All of the models show an improvement in both, goodness of fit and forecast error, compared to the base model. The percentage increase in adjusted R^2 is higher for all of the "standard" specifications compared to the Google models. The improvement regarding the forecasting error, however, is higher for the Google models. Again, the Diebold-Mariano test was used to check whether the difference in forecasting accuracy between column A and B is statistically significant, but the tests showed no evidence.

Table 3.3 shows partial F-tests conducted to check whether the inclusion of the different additional variables brings any improvement regarding model error and predictive power. All inclusions seem to improve the model, although the step from the base model to A-I as well as the step from B-II to B-III is significant only at a 10 %-level.

Table 3.3: Partial F-Tests

Restr.	Full	DF (r|f)	RSS-r	RSS-f	F-Stat	P-value
Base	A-I	160|158	0.0039	0.0037	2.892	0.058 *
Base	A-II	160|157	0.0039	0.0035	4.563	0.004 ***
Base	A-III	160|156	0.0039	0.0035	4.465	0.002 ***
A-I	A-II	158|157	0.0037	0.0035	7.662	0.006 ***
A-I	A-III	158|156	0.0037	0.0035	5.859	0.004 ***
A-II	A-III	157|156	0.0035	0.0035	3.914	0.050 **
Base	B-I	160|158	0.0039	0.0036	6.601	0.002 ***
Base	B-II	160|157	0.0039	0.0034	6.838	0.000 ***
Base	B-III	160|156	0.0039	0.0033	6.032	0.000 ***
B-I	B-II	158|157	0.0036	0.0034	6.824	0.010 ***
B-I	B-III	158|156	0.0036	0.0033	5.118	0.007 ***
B-II	B-III	157|156	0.0034	0.0033	3.312	0.071 *

Note: Partial F-Tests for nested models from table 3.2. Restricted (Restr. / r) vs. full (f) models. Degrees of freedom (DF), Residual Sum of Squares (RSS). Levels of significance: 0 '***' 0.01 '**' 0.05 '*' 0.1 '-' 1.

It is striking that not only all "standard" models have smaller AIC values, but also the AIC figures for model A-I and A-II are slightly higher than the base AIC. This means that if someone was to choose the models solely on the AIC measurements, the Google models would be at a disadvantage.

Ignoring the "standard" models in column B and the AIC numbers for one moment, one could state that the Google models outperform the baseline model. Adding search volume data to the estimations leads to an improvement of about 3.1 % regarding model fit and helps reducing the forecasting errors by about 9.8 %. This result is in line with the major part of the literature. Bringing back column B, however, the results seem to tell a different story. Of course, the numbers still hold true, but it has to be admitted that not only Google data is capable of improving the base model. The results indicate that adding search volume data is not a silver bullet. The "standard" models in column B fulfill the same requirements and can be used in the same way as the Google models, even with slightly better results.

3.5 Conclusion

Google offers virtually unlimited, instantaneously available, spatially and textually adjustable and, in addition, free data. The advantages that come with search engine data for empirical research are obvious. It conquered its position in nearly all economic fields, serving as a highly adjustable sentiment indicator that can be used, inter alia, for nowcasting and short-term forecasting. Additionally, nearly all of the authors in the considered literature present outstanding results. Real estate markets appear to be particularly well-suited for search volume related studies, as the "products" of this market involve a large financial commitment, which demands an extensive information gathering process. It seems as the silver bullet, called Google Trends, is indeed able to measure the *"collective 'swarm intelligence' of Internet users"* as Preis et al. (2010) put it.

All of this makes it tempting to use search engine data without further questioning. By doing so, this paper shows that the search term `[microwave baked potatoes]` is not only a valid predictor of US housing prices, but also increases the forecasting accuracy significantly. Apart from the obvious absurdity of this example, the overall design, presentation and results are in line with the major part of the literature. It seems that *correlation is **not** enough*. Therefore, this paper shows the potential pitfalls and difficulties when working with Google data. It starts with the interpretation of the data itself and carries forward to the query selection process. Dealing with ambiguity is a serious concern, because there are many possible, but unknown intentions of the users when searching for a certain term. Google Trends categories can help but do not eliminate this problem. There are also various practical problems, especially when

working in conjunction with Google Correlate. However, I suspect this service to be shut down at this point of time. One can overcome many of these pitfalls, by knowing them, paying attention to them and carefully constructing a consistent study design.

But there are also disadvantages of Googles' search engine data one cannot overcome. Aside from certain restrictions it is possible to construct Google indicators for any location and topic. At the present time, one can assume that internet users in general and Google users in particular, represent a sufficiently random sample. Nevertheless, since Google Trends does not provide information about the underlying volume of search queries, this random sample is not guaranteed when analyzing less popular topics or regions. Apart from the unreported privacy threshold, there is no possibility to check the overall popularity of a single search term or category. But there are other issues that concern reliability and replicability. The main problem is what Lazer et al. (2014a) call *blue team dynamics*. These dynamics describe adjustments in the Google search algorithm, other changes to the functionality of Google or even retroactive changes to the data itself. To the authors knowledge, there have been two changes, the first one happening in July 2011 changing the data back to January 2011 and the second one being approximately at the same time in 2017, changing one and a half year of data back to January 2016. Additionally, there were several changes to the categories of Google Trends between 2009 and 2018. This means that everybody that used the data before the update will no longer be able to reproduce the results exactly, as the data is not the same anymore. These updates were not announced in any way, neither were they documented.

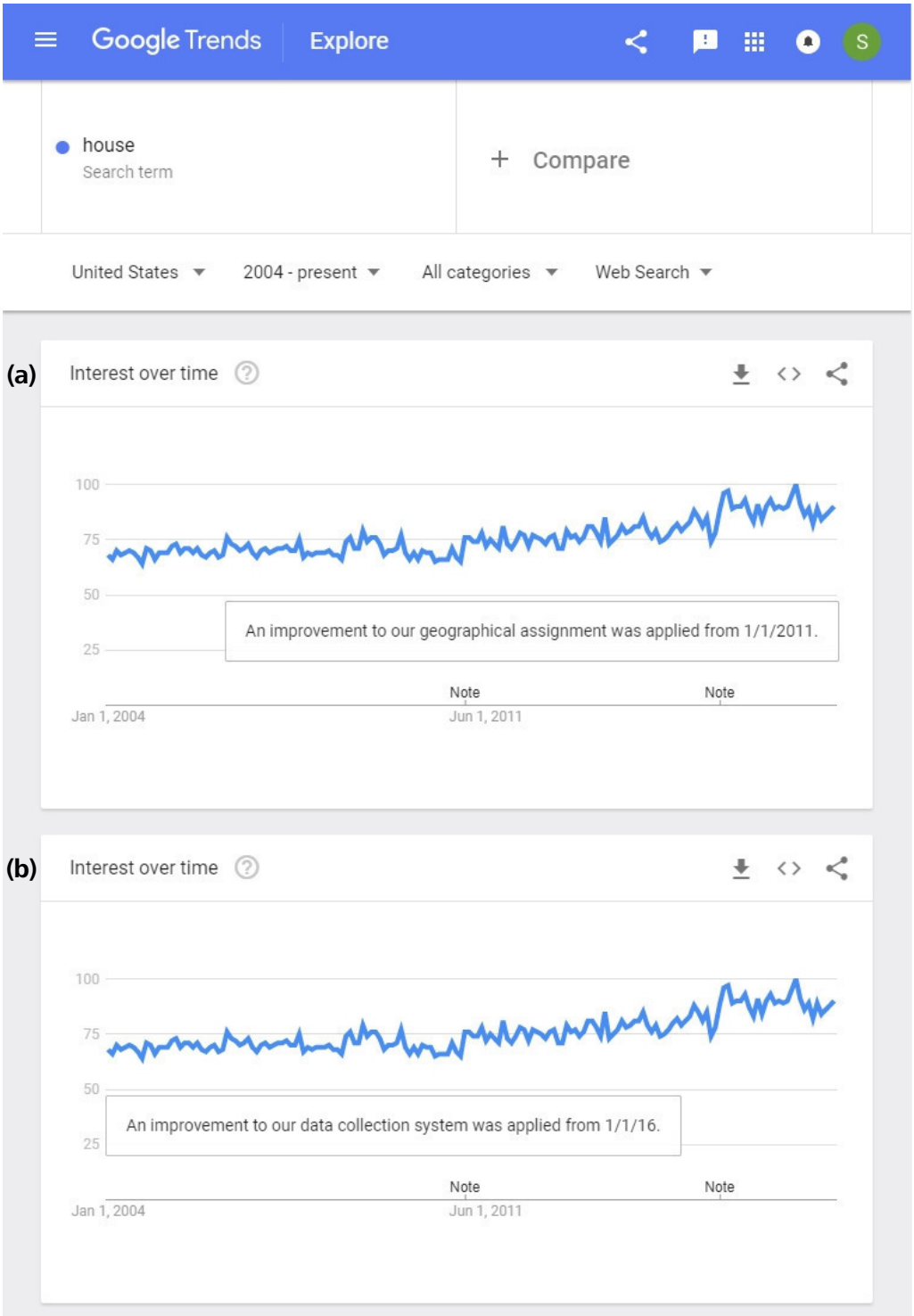
After accounting for the difficulties that can be accounted for, an empirical example for the US housing market is evaluated. The analysis shows different model specifications including Google Trends variables. The results are in line with the literature. Adding search volume data to the estimations leads to an improvement regarding model fit and helps reducing the forecasting errors. However, the major part of the literature only draws a comparison between a base model and models with Google variables. Model specifications without Google predictors are generally not taken into account. Nevertheless, there are certain variables and model specifications that can deliver similar results and allow a proper comparison. This paper shows equally specified "standard" models that fulfill the same requirements and can be used in the same way as the Google models, even with slightly better results.

Especially, when dealing with a "new" type of data, one should know where pitfalls lie and where attention has to be paid. Given that Google Trends data can be accessed already since

2008, many partially severe interpretation and usage misunderstandings can be found amongst the literature. This does not mean that nobody should use Google data. In fact, if urgently needed data is not yet available search volume data can become very useful. When monitoring market movements, the delayed publication of various important variables makes nowcasting a necessary task for many researchers. However, search volume data should not be used for the sake of itself. Instead of using it to show how good it performs against a very simplistic baseline model, it would be more interesting to see how Google models perform compared to or in combination with proven methods that are actually used for this type of task. The results of this study indicate that adding search volume data is not a silver bullet, but at least a useful complement if other data is absent, or as Einav and Levin (2014) put it: *"we don't think that big data will substitute for common sense, economic theory, or the need for careful research designs. Rather, it will complement them. How exactly remains to be seen"*.

3.A Appendix

Figure 3.A.1: Retroactive changes of Google data



Note: Retroactive changes from 1/1/2011 (a) and from 1/1/2016 (b). Screenshots taken and merged on April 30, 2018.

Table 3.A.1: Economic data

Category	Title	Series-ID	Freq	Units
Dependent Variable	S&P/Case-Shiller 20-City Composite Home Price Index	SPCS20RSA	Monthly	Index Jan 2000 = 100, SA
Business Cycle	OECD Indicator for the United States	CSCICP03USM665S	Monthly	Normalised, SA
Business Cycle	University of Michigan: Consumer Sentiment	UMCSENT	Monthly	Index 1966:Q1 = 100, Not SA
Business Cycle	Coincident Economic Activity Index for the United States	USPHCI	Monthly	Index Jul 1992 = 100, SA
Business Cycle	Leading Index for the United States	USSLIND	Monthly	Percent, SA
Employment	Civilian Employment-Population Ratio	EMRATIO	Monthly	Percent, SA
Employment	Unemployment Rate: 20 years and over	LNS14000024	Monthly	Percent, SA
Employment	Civilian Unemployment Rate	UNRATE	Monthly	Percent, SA
Housing	New Privately-Owned Housing Units Completed: Total	COMPUTSA	Monthly	Thousands of Units, SA Annual Rate
Housing	Median Sales Price for New Houses Sold in the United States	MSPNHSUS	Monthly	Dollars, Not SA
Housing	New Private Housing Units Authorized by Building Permits	PERMIT	Monthly	Thousands of Units, SA Annual Rate
Housing	Total Private Construction Spending: Residential	PRRESCONS	Monthly	Millions of Dollars, SA Annual Rate
Housing	Total Construction Spending: Residential	TLRESCONS	Monthly	Millions of Dollars, SA Annual Rate
Interest Rates	6-Month Treasury Bill: Secondary Market Rate	DTB6	Daily	Percent, Not SA
Interest Rates	15-Year Fixed Rate Mortgage Average in the United States	MORTGAGE15US	Weekly	Percent, Not SA
Interest Rates	30-Year Fixed Rate Mortgage Average in the United States	MORTGAGE30US	Weekly	Percent, Not SA
Interest Rates	3-Month Treasury Bill: Secondary Market Rate	TB3MS	Monthly	Percent, Not SA
National Accounts	Real Gross Domestic Product	GDPC1	Quarterly	Billions of Dollars, SA Annual Rate
National Accounts	Real Disposable Personal Income	DSPIC96	Monthly	Billions of Dollars, SA Annual Rate
National Accounts	Real Personal Consumption Expenditures	PCEC96	Monthly	Billions of Dollars, SA Annual Rate
National Accounts	Personal Saving Rate	PSAVERT	Monthly	Percent, SA Annual Rate
National Accounts	Consumer Price Index for All Urban Consumers: All Items	CPIAUCSL	Monthly	Index 1982-1984 = 100, SA
Population	Total Population: All Ages including Armed Forces Overseas	POP	Monthly	Thousands, Not SA
Financial Crisis Dummy	NBER based Recession Indicators for the United States	USREC	Monthly	binary (+1 or 0), Not SA

Note: All data retrieved from "FRED (Federal Reserve Bank of St. Louis) Economic Data", available at: `https://fred.stlouisfed.org` (accessed January 29, 2018). Seasonally Adjusted (SA).

Table 3.A.2: Google keywords

Google Keyword Index	20-City Index
[apartments]	[house atlanta]
[apartments for sale]	[house boston]
[apartments prices]	[house "charlotte nc"]
[buy house]	[house chicago]
[del webb]	[house cleveland]
[dr horton]	[house dallas]
[home listings]	[house denver]
[homefinder]	[house detroit]
[homes]	[house "las vegas"]
[homes prices]	[house "los angeles"]
[homes sale]	[house miami]
[house prices]	[house minneapolis]
[house sale]	[house "new york"]
[houses]	[house phoenix]
[lennar]	[house portland]
[mortgages]	[house "san diego"]
[NAR]	[house "san francisco"]
[new construction]	[house seattle]
[new home]	[house tampa]
[property]	[house washington]
[property for sale]	
[property prices]	
[purchase home]	
[real estate]	
[real estate agency]	
[real estate agent]	
[real estate broker]	
[real estate listings]	
[realtor]	
[residential]	
[residential real estate]	
[zillow]	

Note: Search terms, used for the calculation of specific Google keyword indices. All search terms were retrieved from within the *Real Estate* category.

Conclusion

This dissertation focuses on three different aspects of estimating and forecasting residential markets. In chapter 1 we analyze on the role of monetary policy in contributing to the long- and short-term adjustment of house prices across the Nordic housing markets. We focus explicitly on the relationship between house prices and monetary policy – proxied by short-term interest rates – in order to examine if house prices present a time-varying (dis-)continuous response to both expansionary and recessionary regimes. Furthermore, we analyze how the explanatory power of the different determinants can be decomposed. In this context, the Nordic countries – Denmark, Finland, Sweden and Norway – offer a valuable scenario, as they present common similarities like education, health care and social services, but at the same time rather different financial and monetary conditions. Overall, we confirm that house prices are negatively affected in phases with expansionary regimes in the long-run, but we also provide evidence of unexpected anti-cyclical effects in the short-run. Consequently, the role of central banks has to be critically examined, since housing markets adjust unevenly to different monetary environments.

In chapter 2 we study the effects of spatial heterogeneity on rent prices in Germany. The immobility of real estate makes its price formation different from traditional commodities. As a result, location is one of the most important determinants for defining its value. Therefore, the choice of the functional form for hedonic regression models is crucial. We analyse the prediction accuracy and explanatory power of three different econometrical approaches based on a large dataset with more than 570,000 asking rents across 46 residential rental markets in Germany. This is – to the best of our knowledge – one of the largest datasets used for spatial real estate analysis. With the list of spatial estimation techniques being very extensive, the Geographically Weighted Regression (GWR) has established itself as a widely used method that expands the restrictive traditional Ordinary Least Squares (OLS) by considering spatially varying effects. Semi-parametric methods like the Generalized Additive Model (GAM), however,

capture spatial effects based on smooth functions and expand the traditional hedonic model by identifying latent nonlinear effects. Surprisingly, the results show a clear disadvantage for the GWR model. Regarding OLS and GAM, it turns out that the differences in out-of-sample prediction accuracy are not substantial. Against expectations the OLS approach seems to be an equal alternative to (semi-) parametric models. However, our findings also imply that although the discrepancy between OLS and GAM is small, the GAM model still provides the most accurate predictions for most cases.

In chapter 1 & 2 I show how real estate prices can be estimated and predicted on the basis of fundamentals and location, respectively. Finally, in chapter 3 I consider the question how to estimate house prices if there are price movements that can not be explained by a change in fundamental factors. Sentiment indicators try to capture the "noise" in various markets, like for example fears or hopes. There are lots of traditional survey-based sentiment indicators, but they might possibly be hard to access or simply not sufficiently up to date. Google Trends, however, offers virtually unlimited, instantaneously available, spatially and textually adjustable and, in addition, free search query data. This type of data conquered its position in nearly all economic fields, serving as a highly adjustable sentiment indicator. Although Google Trends data can be accessed already since 2008, many interpretation and usage misunderstandings can be found amongst the literature. Therefore, I give an overview of what Google data actually is and where the potential pitfalls lie. To the best of my knowledge, there is no other paper specifically dealing with the potential pitfalls and disadvantages of Google Trends data in real estate analysis. Secondly, I conduct an empirical analysis to find out, whether the results are still in line with the literature after accounting for those difficulties. For this task, the usual approach in the existing literature would be to compare Google models to a simple baseline model. Admittedly, I find it more interesting to see how Google models perform compared to proven methods that are actually used for this type of task. Therefore, I check how the resulting models can compete against comparable "standard" models. The results show, as expected, that adding search volume data to the estimations leads to an improvement regarding model fit and helps reducing the forecasting errors when compared to a simple baseline model. However, they also show that there are equally specified "standard" models that fulfill the same requirements and can be used in the same way as the Google models, even with slightly better results. The findings indicate that adding search volume data is not a silver bullet, but at least a useful complement if other data is absent.

We live in exciting times for research. An article from 2013 states that 90 % of all the data in the world has been generated over the last two years, which is more data than what was created in the entire history of the human race.[1] Another article assumes that in 2017 we created even more data in one year alone.[2] Whether these numbers are entirely correct or not, the presence, influence and growth of "big data" can not be denied. Despite the (justified) concerns that come with this development, it does and will reshape our lifestyle, work environment and economic research. Of course this includes the real estate sector, as well. In this dissertation, I show various kinds of data. "Traditional" fundamental data is used in chapter 1. Chapter 2 presents a database in which internet offers of residential properties from online newspapers and more than ten internet search engines like Immoscout, Immonet, Immowelt and others are collected and matched. Finally, in chapter 3 I focus on the usage of Google search engine data. New and more detailed data can lead to more accurate estimations and predictions. However, the data alone will not solve any problems if the methods and assumptions are not properly thought through. I personally share the view of Lazer et al. (2014a) who suggest that: *"Instead of focusing on a 'big data revolution,' perhaps it is time we were focused on an 'all data revolution,' […] using data from all traditional and new sources, and providing a deeper, clearer understanding of our world."*

1 See "Big Data – for better or worse", available at: `https://www.sintef.no/en/latest-news/big-data-for-better-or-worse/` or "Data, data everywhere...", available at: `https://www.ibm.com/watson/infographic/discovery/big-data-challenge-opportunity/` (accessed 2018-05-30).

2 See "More data will be created in 2017 than the previous 5,000 years of humanity", available at: `https://appdevelopermagazine.com/4773/2016/12/23/more-data-will-be-created-in-2017-than-the-previous-5,000-years-of-humanity-/` (accessed 2018-05-30).

References

Adams, Z. and R. Füss (2010). Macroeconomic determinants of international housing markets. *Journal of Housing Economics 19*(1), 38–50.

Anderson, C. (June 23, 2008). The End of Theory: The Data Deluge Makes the Scientific Method Obsolete. magazine 16.07, wired. Available at: `http://www.wired.com/2008/06/pb-theory/`.

Anglin, P. and R. Gençay (1996). Semiparametric estimation of a hedonic price function. *Journal of Applied Econometrics 11*(6), 633–648.

Anselin, L. (2003). Spatial Externalities, Spatial Multipliers, and Spatial Econometrics. *International Regional Science Review 26*(2), 153–166.

Askitas, N. (2016). Trend-Spotting in the Housing Market. *Cityscape: A Journal of Policy Development and Research 18*(2), 185–198.

Askitas, N. and K. Zimmermann (2009). Google Econometrics and Unemployment Forecasting. *Applied Economics Quarterly 55*(2), 107–120.

Baker, S. and A. Fradkin (2011). What Drives Job Search? Evidence from Google Search Data. Discussion Paper 10-020, Stanford Institute for Economic Policy Research.

Baker, S. and A. Fradkin (2017). The Impact of Unemployment Insurance on Job Search: Evidence from Google Search Data. *Review of Economics and Statistics 99*(5), 756–768.

Bao, H. and A. Wan (2004). On the Use of Spline Smoothing in Estimating Hedonic Housing Price Models: Empirical Evidence Using Hong Kong Data. *Real Estate Economics 32*(3), 487–507.

Bennöhr, L. and M. Oestmann (2014). Determinants of house price dynamics. What can we learn from search engine data? Discussion Paper 153, Helmut-Schmidt-Universität, Department of Economics.

Bitter, C., G. Mulligan, and S. Dall'erba (2007). Incorporating spatial variation in housing attribute prices: A comparison of geographically weighted regression and the spatial expansion method. *Journal of Geographical Systems 9*(1), 7–27.

Bourassa, S., E. Cantoni, and M. Hoesli (2007). Spatial Dependence, Housing Submarkets, and House Price Prediction. *The Journal of Real Estate Finance and Economics 35*(2), 143–160.

Bourassa, S., E. Cantoni, and M. Hoesli (2010). Predicting House Prices with Spatial Dependence: Impact of Alternatives Submarkets Definitions. *Journal of Real Estate Research 32*(2), 139–159.

Braun, N. (2016). Google search volume sentiment and its impact on REIT market movements. *Journal of Property Investment & Finance 34*(3), 249–262.

Brooks, C. and S. Tsolacos (2010). *Real Estate Modelling and Forecasting*. Cambridge and New York: Cambridge University Press.

Brown, J. P., H. Song, and A. McGillivray (1997). Forecasting UK house prices: A time varying coefficient approach. *Economic Modelling 14*(4), 529–548.

Brunauer, W., S. Lang, P. Wechselberger, and S. Bienert (2010). Additive Hedonic Regression Models with Spatial Scaling Factors: An Application for Rents in Vienna. *The Journal of Real Estate Finance and Economics 41*(4), 390–411.

Brunsdon, C., S. Fotheringham, and M. Charlton (1996). Geographically Weighted Regression: A Method for Exploring Spatial Nonstationarity. *Geographical Analysis 28*(4), 281–298.

Brunsdon, C., S. Fotheringham, and M. Charlton (1998). Geographically Weighted Regression - modelling spatial non-stationarity. *Journal of the Royal Statistical Society: Series D (The Statistician) 47*(3), 431–443.

Burnham, K. and D. Anderson (2002). *Model selection and multimodel inference: A practical information-theoretic approach* (2nd ed.). New York, NY: Springer.

Butler, D. (2013). When Google got flu wrong. *Nature 494*(7436), 155–156.

Cajias, M. and S. Ertl (2017). The sensitivity of house prices under varying monetary regimes: The Nordic scenario. *International Journal of Housing Markets and Analysis 10*(1), 4–21.

Cajias, M. and S. Ertl (2018). Spatial effects and non-linearity in hedonic modeling. *Journal of Property Investment & Finance 36*(1), 32–49.

Calza, A., T. Monacelli, and L. Stracca (2013). Housing Finance and Monetary Policy. *Journal of the European Economic Association 11*(1), 101–122.

Cardarelli, R., D. Igan, and A. Rebucci (2008). The changing housing cycle and the implications for monetary policy: Chap. 3. In *Housing and the business cycle*, World economic outlook, pp. 103–132. Washington, DC: Internat. Monetary Fund.

Choi, H. and H. Varian (2009). Predicting the Present with Google Trends. Working paper, Google Research Blog. Available at: `http://dx.doi.org/10.2139/ssrn.1659302`.

Choi, H. and H. Varian (2012). Predicting the Present with Google Trends. *Economic Record 88*, 2–9.

Chrostek, K. and K. Kopczewska (2013). Spatial Prediction Models for Real Estate Market Analysis. *Ekonomia 35*, 25–43.

Clapp, J. (2004). A Semiparametric Method for Estimating Local House Price Indices. *Real Estate Economics 32*(1), 127–160.

Cohen, J., C. Coughlin, and J. Clapp (2015). Local Polynomial Regressions versus OLS for Generating Location Value Estimates: Which is More Efficient in Out-of-Sample Forecasts? Working paper 2015-014B, Federal Reserve Bank of St. Louis. Available at: `https://research.stlouisfed.org/wp/more/2015-014`.

Dabrowski, J. and T. Adamczyk (2010). Application of GAM Additive Non-Linear Models to Estimate Real Estate Market Value. *Geomatics and Environmental Engineering 4*(2), 55–62.

Demary, M. (2012). Wechselbeziehungen von makroökonomischen Variablen und Immobilienpreisen: Kap. B2. In N. B. Rottke and M. Voigtländer (Eds.), *Immobilienwirtschaftslehre | Band II - Ökonomie*, Immobilienfachwissen, pp. 215–248. Köln: Immobilien Manager Verlag IMV.

Dietzel, M. (2016). Sentiment-based predictions of housing market turning points with Google trends. *International Journal of Housing Markets and Analysis 9*(1), 108–136.

Dietzel, M., N. Braun, and W. Schäfers (2014). Sentiment-based commercial real estate forecasting with Google search volume data. *Journal of Property Investment & Finance 32*(6), 540–569.

DiPasquale, D. and W. Wheaton (1996). *Urban Economics and Real Estate Markets*. Englewood Cliffs, NJ: Prentice-Hall.

Einav, L. and J. Levin (2014). The Data Revolution and Economic Analysis. In J. Lerner and S. Stern (Eds.), *Innovation policy and the economy 14*, NBER Innovation policy and the economy, pp. 1–24. Chicago: The University of Chicago Press.

Gençay, R. and X. Yang (1996). A forecast comparison of residential housing prices by parametric versus semiparametric conditional mean estimators. *Economics Letters 52*(2), 129–135.

Geniaux, G. and C. Napoléone (2008). Semi-Parametric Tools for Spatial Hedonic Models: An Introduction to Mixed Geographically Weighted Regression and Geoadditive Models: Chapter 5. In A. Baranzini, J. Ramirez, C. Schaerer, and P. Thalmann (Eds.), *Hedonic Methods in Housing Markets*, pp. 101–126. New York, NY: Springer New York.

Ginsberg, J., M. Mohebbi, R. Patel, L. Brammer, M. Smolinski, and L. Brilliant (2009). Detecting influenza epidemics using search engine query data. *Nature 457*(7232), 1012–1014.

Goel, S., J. Hofman, S. Lahaie, D. Pennock, and D. Watts (2010). Predicting consumer behavior with Web search. *Proceedings of the National Academy of Sciences of the United States of America 107*(41), 17486–17490.

Goodhart, C. and B. Hofmann (2008). House Prices, Money, Credit and the Macroeconomy. Working paper series 888, European Central Bank. Available at: `http://ssrn.com/abstract=1120162`.

Groemping, U. (2006). Relative Importance for Linear Regression in R: The Package relaimpo. *Journal of Statistical Software 17*(1), 1–27.

Guirguis, H. S., C. I. Giannikos, and R. I. Anderson (2005). The US Housing Market: Asset Pricing Forecasts Using Time Varying Coefficients. *Journal of Real Estate Finance and Economics 30*(1), 33–53.

Guzman, G. (2011). Internet Search Behavior as an Economic Forecasting Tool: The Case of Inflation Expectations. *The Journal of Economic and Social Measurement 36*(3), 119–167.

Hanink, D., R. Cromley, and A. Ebenstein (2012). Spatial Variation in the Determinants of House Prices and Apartment Rents in China. *The Journal of Real Estate Finance and Economics 45*(2), 347–363.

Harford, T. (2014). Big data: are we making a big mistake? *Significance 11*(5), 14–19.

Harvey, D., S. Leybourne, and P. Newbold (1997). Testing the equality of prediction mean squared errors. *International Journal of Forecasting 13*(2), 281–291.

Hastie, T. and R. Tibshirani (1990). *Generalized Additive Models* (1st ed.), Volume 43 of *Monographs on statistics and applied probability*. London: Chapman and Hall/CRC.

Heinig, S., A. Nanda, and S. Tsolacos (2016). Which Sentiment Indicators Matter? An Analysis of the European Commercial Real Estate Market. Discussion Paper ICM-2016-04, Henley Business School, ICMA-Centre.

Helbich, M., W. Brunauer, E. Vaz, and P. Nijkamp (2014). Spatial Heterogeneity in Hedonic House Price Models: The Case of Austria. *Urban Studies 51*(2), 390–411.

Hohenstatt, R. and M. Kaesbauer (2014). GECO's Weather Forecast for the U.K. Housing Market: To What Extent Can We Rely on Google Econometrics? *Journal of Real Estate Research 36*(2), 253–281.

Hohenstatt, R., M. Kaesbauer, and W. Schäfers (2011). "Geco" and its Potential for Real Estate Research: Evidence from the U.S. Housing Market. *Journal of Real Estate Research 33*(4), 471–506.

Kim, Y.-J. and C. Gu (2004). Smoothing spline Gaussian regression: More scalable computation via efficient approximation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 66*(2), 337–356.

Koop, G. and L. Onorante (2016). Macroeconomic Nowcasting Using Google Probabilities. In *Proceedings of the 1st International Conference on Advanced Research Methods and Analytics*, Valencia. Universitat Politècnica València.

Lazer, D., R. Kennedy, G. King, and A. Vespignani (2014a). Big data. The parable of Google Flu: traps in big data analysis. *Science 343*(6176), 1203–1205.

Lazer, D., R. Kennedy, G. King, and A. Vespignani (2014b). Google Flu Trends Still Appears Sick: An Evaluation of the 2013–2014 Flu Season. Working paper, SSRN. Available at: `http://dx.doi.org/10.2139/ssrn.2408560`.

Lazer, D., R. Kennedy, G. King, and A. Vespignani (2014c). Supplementary Materials for: Big data. The parable of Google Flu: traps in big data analysis. *Science 343*(6176), 1203–1205.

Leblanc, M. and R. Bokreta (2009). Analysis of the US Real Estate Market: Time-Varying Estimation and Forecast of the S&P Case-Shiller Composite 20 Cities. Working paper, SSRN. Available at: `http://ssrn.com/abstract=1486682`.

Lu, B., M. Charlton, and S. Fotheringham (2011). Geographically Weighted Regression Using a Non-Euclidean Distance Metric with a Study on London House Price Data. *Procedia Environmental Sciences 7*, 92–97.

Lujanen, M. (2004). *Housing and Housing Policy in the Nordic Countries*. Nordic Council of Ministers.

Ma, L. and C. Liu (2010). The decomposition of housing market variations. *International Journal of Housing Markets and Analysis 3*(1), 6–16.

Mason, C. and J. Quigley (1996). Non–parametric hedonic housing prices. *Housing Studies 11*(3), 373–385.

McCord, M., P. Davis, M. Haran, D. McIlhatton, and J. McCord (2014). Understanding rental prices in the UK: a comparative application of spatial modelling approaches. *International Journal of Housing Markets and Analysis 7*(1), 98–128.

McDonald, J. F. and H. H. Stokes (2013). Monetary Policy and the Housing Bubble. *Journal of Real Estate Finance and Economics 46*(3), 437–451.

McGreal, S. and P. Taltavull de La Paz (2013). Implicit House Prices: Variation over Time and Space in Spain. *Urban Studies 50*(10), 2024–2043.

McMillen, D. and C. Redfearn (2010). Estimation and Hypothesis Testing for Nonparametric Hedonic House Price Functions. *Journal of Regional Science 50*(3), 712–733.

Miles, W. (2014). The housing bubble: How much blame does the fed really deserve? *Journal of Real Estate Research 36*(1), 41–58.

Mohebbi, M., D. Vanderkam, J. Kodysh, R. Schonberger, H. Choi, and S. Kumar (2011). Google Correlate Whitepaper. Working paper, Google. Available at: `https://www.google.com/trends/correlate/whitepaper.pdf`.

Monk, S., C. Tang, C. Whitehead, and S. Markkanen (2012). *The private rented sector in the new century: A comparative approach*. København: Boligøkonomisk Videncenter.

Nastansky, A. (2012). Geldpolitik und Immobilienpreise: Kap. B1. In N. B. Rottke and M. Voigtländer (Eds.), *Immobilienwirtschaftslehre | Band II - Ökonomie*, Immobilienfachwissen, pp. 163–214. Köln: Immobilien Manager Verlag IMV.

Osland, L. (2010). An Application of Spatial Econometrics in Relation to Hedonic House Price Modeling. *Journal of Real Estate Research 32*(3), 289–320.

Pace, K. (1998). Appraisal Using Generalized Additive Models. *Journal of Real Estate Research 15*(1), 77–99.

Pace, K. and J. LeSage (2004). Spatial Statistics and Real Estate. *The Journal of Real Estate Finance and Economics 29*(2), 147–148.

Páez, A., F. Long, and S. Farber (2008). Moving Window Approaches for Hedonic Price Estimation: An Empirical Comparison of Modelling Techniques. *Urban Studies 45*(8), 1565–1581.

PATRIZIA research (2015). European Residential Markets 2015/2016. PATRIZIA INSIGHT, PATRIZIA Immobilien AG. Available at: `http://www.patrizia.ag/en/company/research/research-publications/patrizia-insight/`.

Pavlicek, J. and L. Kristoufek (2015). Nowcasting unemployment rates with Google searches: evidence from the Visegrad Group countries. *PloS one 10*(5), 1–11.

Preis, T., D. Reith, and E. Stanley (2010). Complex dynamics of our economic life on different scales: Insights from search engine query data. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences 368*(1933), 5707–5719.

Rivera, R. (2016). A dynamic linear model to forecast hotel registrations in Puerto Rico using Google Trends data. *Tourism Management 57*, 12–20.

Rochdi, K. and M. Dietzel (2015). Outperforming the benchmark: Online information demand and REIT market performance. *Journal of Property Investment & Finance 33*(2), 169–195.

Schewe, T. (2015). Monetary policy regimes and the Nordic model. Publication Series 14, Buskerud and Vestfold University College.

Scott, S. and H. Varian (2015). Bayesian Variable Selection for Nowcasting Economic Time Series. In A. Goldfarb, S. Greenstein, and C. Tucker (Eds.), *Economic Analysis of the Digital Economy*, National Bureau of Economic Research Conference Report, pp. 119–135. Chicago: University of Chicago Press.

S&P CoreLogic (April 2018). Case-Shiller Home Price Indices. Index Methodology, S&P Dow Jones Indices. Available at: `https://eu.spindices.com/index-family/real-estate/sp-corelogic-case-shiller`.

Stephens-Davidowitz, S. and H. Varian (2015). A Hands-on Guide to Google Data. Working paper, Google. Available at: `http://people.ischool.berkeley.edu/~hal/Papers/2015/primer.pdf`.

Sunding, D. and A. Swoboda (2010). Hedonic analysis with locally weighted regression: An application to the shadow cost of housing regulation in Southern California. *Regional Science and Urban Economics 40*(6), 550–573.

Swanson, N. R. (1998). Money and output viewed through a rolling window. *Journal of Monetary Economics 41*(3), 455–474.

Tsatsaronis, K. and H. Zhu (2004). What Drives Housing Price Dynamics: Cross-Country Evidence. BIS Quarterly Review, Bank for International Settlements. Available at: `http://ssrn.com/abstract=1968425`.

Tse, R. (2002). Estimating Neighbourhood Effects in House Prices: Towards a New Hedonic Model Approach. *Urban Studies 39*(7), 1165–1180.

Varian, H. (2014). Big Data: New Tricks for Econometrics. *Journal of Economic Perspectives 28*(2), 3–28.

Vosen, S. and T. Schmidt (2011). Forecasting private consumption: Survey-based indicators vs. Google trends. *Journal of Forecasting 30*(6), 565–578.

Widłak, M., J. Waszczuk, and K. Olszewski (2015). Spatial and hedonic analysis of house price dynamics in Warsaw. NBP Working Papers 197, Narodowy Bank Polski, Economic Research Department. Available at: `http://EconPapers.repec.org/RePEc:nbp:nbpmis:197`.

Wood, S. (2006). *Generalized additive models: An introduction with R*. Texts in statistical science. Boca Raton, Florida: Chapman & Hall/CRC.

Wu, L. and E. Brynjolfsson (2015). The Future of Prediction: How Google Searches Foreshadow Housing Prices and Sales. In A. Goldfarb, S. Greenstein, and C. Tucker (Eds.), *Economic Analysis of the Digital Economy*, National Bureau of Economic Research Conference Report, pp. 89–118. Chicago: University of Chicago Press.

Zeileis, A. (2004). Econometric Computing with HC and HAC Covariance Matrix Estimators. *Journal of Statistical Software 11*(10), 1–17.

Zietz, J. (2012). Methoden: Ökonometrie und ihre Grenzen: Kap. E1. In N. B. Rottke and M. Voigtländer (Eds.), *Immobilienwirtschaftslehre | Band II - Ökonomie*, Immobilienfachwissen, pp. 765–804. Köln: Immobilien Manager Verlag IMV.

Zivot, E. and J. Wang (2006). *Modeling financial time series with S-PLUS* (2nd ed.). New York, NY: Springer.

Universität Regensburg
IRE|BS Institut für Immobilienwirtschaft
Fakultät für Wirtschaftswissenschaften