IS THERE ROOM FOR ANOTHER HEDONIC MODEL? –THE ADVANTAGES OF THE GAMLSS APPROACH IN REAL ESTATE RESEARCH

DR. MARCELO CAJIAS

ABSTRACT: Hedonic modelling is essential for institutional investors, researchers and urban policy-makers in order to identify the factors affecting the value and future development of rents over time and space. While statistical models in this field have advanced substantially over the last decades, new statistical approaches have emerged expanding the conventional understanding of real estate markets. This paper explores the in-sample explanatory and out-of-sample forecasting accuracy of the Generalized Additive Model for Location, Scale and Shape (GAMLSS) model in contrast to traditional methods in Munich's residential market. The results show that the complexity of asking rents in Munich is more accurately captured by the GAMLSS approach, leading to a significant increase in the out-of-sample forecasting accuracy.

KEYWORDS: Hedonic modelling, Residential real estate, GAMLSS, GAM, out-of-sample forecast.

ACKNOWLEDGEMENTS: I would like to express my appreciation to PATRIZIA Immobilien AG (https://www.patrizia.ag/en) for providing me with the essential data and especially with the large computational requirements needed for the estimation of the large models. All statements of opinion reflect the current estimations of the author and do not necessarily reflect the opinion of PATRIZIA Immobilien AG or its associated companies.

PATRIZIA Immobilien AG | Fuggerstraße 26 | 86150 Augsburg | GERMANY

IS THERE ROOM FOR ANOTHER HEDONIC MODEL? –THE ADVANTAGES OF THE GAMLSS APPROACH IN REAL ESTATE RESEARCH

ABSTRACT: Hedonic modelling is essential for institutional investors, researchers and urban policy-makers in order to identify the factors affecting the value and future development of rents over time and space. While statistical models in this field have advanced substantially over the last decades, new statistical approaches have emerged expanding the conventional understanding of real estate markets. This paper explores the in-sample explanatory and out-of-sample forecasting accuracy of the Generalized Additive Model for Location, Scale and Shape (GAMLSS) model in contrast to traditional methods in Munich's residential market. The results show that the complexity of asking rents in Munich is more accurately captured by the GAMLSS approach, leading to a significant increase in the out-of-sample forecasting accuracy.

KEYWORDS: Hedonic modelling, Residential real estate, GAMLSS, GAM, out-of-sample forecast.

IS THERE ROOM FOR ANOTHER HEDONIC MODEL? –THE ADVANTAGES OF THE GAMLSS APPROACH IN REAL ESTATE RESEARCH

1. INTRODUCTION

The improvement in the field of the statistical modelling of hedonic price equations was enormous during the last three decades. Primarily driven by advances in the field of regression analysis, statistical inference and especially computational speed, the accuracy of hedonic equations has increased considerably, leading to a better understanding of the fundamental factors affecting property rents and prices. These advantages have led econometricians at the same time to carefully decide which model to employ when reproducing real estate markets accurately, as the range of models and their complexity has widened significantly. While contemporary models focus nowadays on the incorporation of non-linear and spatial effect in the hedonic equation, statistical research over the last decade has developed additional methods and instruments to incorporate advanced effects in order to enhance the explanatory power of regression analysis. This paper builds upon this new statistical research and aims at exploring a new approach in hedonic modelling.

In theory, any hedonic model that is designed to decompose the price of a dwelling in a certain market might be able to capture all the underlying factors affecting property prices asymptotically as well as efficiently. But does any hedonic equation considering just non-linear and spatial effects explain real estate prices sufficiently and efficiently? In other words, are there more effects rather than non-linear and spatial effects to consider when estimating hedonic equations? In this context, statistical methods of the "new era" expand the traditional regression by considering further effects such as the distribution of the response as an

optimization criterion. This new approach refers to the Generalized Additive Model for Location, Scale and Shape (GAMLSS) introduced by [20] in 2005, which accounts, beside the traditional spatial and non-linear effects, additionally for "non-normal" effects between the response and the underlying covariates. Although models including spatial and non-linear effects in real estate studies have shown an enhanced out-of-sample performance, see for example [4, 16, 24, 5], this paper aims at exploring the GAMLSS method and its explanatory and forecasting features in hedonic regression equations, which at the same has been rarely employed in a real estate context.

Do traditional models fail at explaining real estate prices accurately? The complexity of real estate prices – in contrast to similar consumer goods – lies in their nature. On the one hand, real estate assets are exposed within a certain market to spatial, temporal and intangible interdependences and on the other hand determined by their own building characteristics. When considering all these three effects simultaneously, any hedonic model should be able to control for intertemporal, spatial and property-specific (auto-) correlations dynamically in such a way that the assumptions behind the chosen estimator remain asymptotically valid and the explanatory level reaches a suitable level. As the modelling of these three effects is quite difficult, econometricians chose in general the isolation of one of these factors (so called fixed effects) in order to reproduce the remaining effects accurately. However, research over the last decade has provided evidence that additional – partly intangible – factors such as submarket heterogeneity, local amenities or access to public transport affect real estate prices significantly, see for example [2, 3, 17, 1]. New approaches, such as the GAMLSS, aim thus at applying advanced statistical instruments in order to capture these effects in the hedonic equation and thereby improve the explanatory power.

Beyond the three aforementioned effects – space, time and property-specific – two main issues have arisen when modelling real estate prices: the heterogeneous distribution of the

variance of rents and their skewness across space. While the general method of moments – from a strictly econometric point of view – does not require the response to be normally distributed, the explanatory accuracy of any regression model has to deal permanently with extreme non-normally-distributed responses. The non-homogeneity of variance across space indicates that the range in the willingness to pay for rents within certain submarkets varies in the extremes of the distribution. And although this sounds reasonable, its statistical modelling is difficult to capture by traditional dummies variables in the spirit of Ordinary Least Squares (OLS). The latter effect, the skewness of rents, instead provides information that rents within a market are empirically – under any probability density function – not normally distributed and that at the same time after controlling for the available covariates a share of rents tend to be under- or overpriced leading to excess residual heteroscedasticity. Although this phenomenon might be isolated by censoring the response through its quantile distribution or by robust variance-covariance-estimators, it still affecting the accuracy of the hedonic equation.

In order to account for these anomalies, the GAMLSS approach proposes the expansion of traditional Generalized Linear Models (GLM) by modelling the parameters of the response as semiparametric functions of the covariates, overcoming thus the restrictions of traditional methods. In simple words, the GAMLSS approach fits a relationship without involving strong assumptions between the response and the covariates. The consideration of these new effects into a hedonic equation are yet expected to lead to a more accurate estimation of the underlying data generating process of real estate prices and to an enhancement in the accuracy of out-of-sample forecasts.

Having said this, this paper aims at modelling the hedonic equation considering the non-homogeneity of variance over space and skewness in the distribution of a sample of ca. 25k asking rents in Munich, Germany. In doing so, I employ the GAMLSS approach which has

been rarely employed in the field of real estate. In contrast, its statistical accuracy in capturing locational effects under different distributions has shown extraordinary results in science areas such as biology, biosciences, energy economics, fisheries, food consumption, growth curves estimation, marine research, medicine, meteorology, rainfall, among others[i]. The main aim of the paper is therefore to explore whether the incorporation of spatial varying variances and consideration of skewed distributions in Munich's hedonic equation via GAMLSS reduces out-of-sample error variances and leads vis-a-vis to more precise forecasts than traditional hedonic regression models.

2. THE GENERALIZED ADDITIVE MODEL FOR LOCATION, SCALE AND SHAPE

The GAMLSS model is a semiparametric regression method, in which all the parameters of observed distribution for the response can be modelled as additive or non-linear functions of the explanatory variables. The four moments of the response – the mean, variance, skewness and the kurtosis – are generated by the observed variable and explicitly accounted by the GAMLSS statistical approach.

--- Please insert Figure 1 here ----

While the traditional OLS estimator $\hat{\beta} = (X'X)^{-1}X'Y$ is restricted in the incorporation of spatial and distributional effects[ii], the generalized linear model (GLM) is able to capture the distribution of the response (Y) only in the mean equation (μ), omitting however the interdependence with the underlying explanatory variables[iii]. A further assumption for unbiasedness of the traditional OLS estimator is that the distribution of the sample is centered about the estimator $\hat{\beta}$, see Figure 1. In other words, the expected conditional variance of the errors is expected to be homoscedastic distributed across the entire sample. The GAMLSS framework instead allows the flexible modelling of both non-linear effects across the several parameters of the response and the distribution parameters of the endogenous and exogenous variables simultaneously.

The GAMLSS hedonic framework depends on the imposed probability distribution (D), the link function applied to each k distribution parameters g and most importantly on the parameterization of equations for the mean (μ), variance (σ), skewness (ν) and the kurtosis (τ) of the response. The single modelling of the mean (μ) equation without considering any distribution corresponds to the OLS estimator. A GAMLSS equation can be expressed thus as:

$$D(Y) = D(Y|\mu, \sigma, \nu, \tau) = \begin{Bmatrix} g_1(\mu) = \boldsymbol{X}^{\mu}\beta^{\mu} + f_{1k}(\dot{X}_k^{\mu})\gamma_k^{\mu} \\ g_2(\sigma) = \boldsymbol{X}^{\sigma}\beta^{\sigma} + f_{2k}(\dot{X}_k^{\sigma})\gamma_k^{\sigma} \\ g_3(\nu) = \boldsymbol{X}^{\nu}\beta^{\nu} + f_{3k}(\dot{X}_k^{\nu})\gamma_k^{\nu} \\ g_4(\tau) = \boldsymbol{X}^{\tau}\beta^{\tau} + f_{4k}(\dot{X}_k^{\tau})\gamma_k^{\tau} \end{Bmatrix} \quad \text{(I)}$$

where the linear effects $\boldsymbol{X}$ and non-parametric or non-linear effects $f(\dot{\boldsymbol{X}})$ of the k endogenous variables need to be parameterized for each of the moments μ, σ, ν, τ of the response. The GAMLSS optimization model is fitted by maximum penalized likelihood estimation l under the assumption that the response is independent for each of the moments. The penalty term for the optimization in a GAMLSS model including non-parametric effects $f(\dot{\boldsymbol{X}})$ is given thus by

$$l = \sum_{i=1}^{n} \log(f(y_i|\mu_i, \sigma_i, \upsilon_i, \tau_i)) - \frac{1}{2}\sum_{k=1}^{4}\sum_{j=1}^{J_k} \lambda_{kj}\gamma_{kj}\boldsymbol{G}_{kj}\gamma_{kj} \quad \text{(II)}$$

where the first term of the equation represents the likelihood of the linear effects $\boldsymbol{X}$ and the second term represents the likelihood of the penalties with respect to the non-parametric or non-linear effects $f(\dot{\boldsymbol{X}})$. The GAMLSS methodology optimizes the likelihood l with respect to β and γ for a fixed hyper parameter λ based on the space spanned by the matrix of penalties $\boldsymbol{G}$. [20] propose two optimization algorithms, the CG and the RS algorithm, which both lead asymptotically to the maximum penalised log likelihood estimated for $\hat{\beta}$ and $\hat{\gamma}$. When estimating the hedonic models, I use for simplicity the standard procedure RS implemented in the gamlss package in R [18]. A more detailed explanation of the underlying algorithm steps and optimization can be found on the manual published by [23], which at the same time is a very suitable introduction on GAMLSS models in R.

In order to provide some stylized facts of the GAMLSS modelling technique, in this case on asking rents in the Munich market, I compare graphically a OLS and a GAMLSS model of asking rents in €/m² per month (p.m.) as response in contrast to flat's area in m². While both models include the same explanatory variable without any logarithmic transformation in the mean equation, the GAMLSS considers a cubic spline of size in the last three moments, allowing the variance, skewness and kurtosis to vary at different values of the covariate.

--- Please insert Figure 2 here ----

Both scatterplots in Figure 2 show the regression "line" of floor space on asking rents as a single regressor. While the left side of the figure shows that the linear OLS model is clearly not able to capture a significant share of heterogeneity of asking prices and that a linear approach might be poor on forecasting the response, the GAMLSS model shows a more accurate and flexible understating of asking rents in Munich's residential market. Very important in this single analysis is the nature of GAMLSS models to account for the different distributions of rents across the increasing values of the covariate. This is reflected by the varying width (variance, skewness and partly kurtosis) of the distribution lines of rents across the different size classes. Yet, the incorporation of different moments into the hedonic equation enables not only a higher understanding rents, but most importantly provides useful information on the different marginal willingness to pay for flats in dependence on flat's size in our simple case. The results based on the AIC rather than the R^2 criterion confirm finally that rents are more accurately modelled by a GAMLSS model rather than by a traditional estimator.

Three different papers – to my knowledge – focus on the usage of the GAMLSS approach in real estate hedonic research. [15] as part of the GAMLSS developing team, present a new algorithm to speed up the optimization of variable selection in GAMLSS environments. Their results confirm that – on the basis of the new developed algorithm – the explanatory power of

the GAMLSS approach in Munich's rental market is superior in contrast to the GAM approach as measured by the MSE. The seminal study of [8] employs the GAMLSS model for estimating lot values in Aracaju in Brazil and compares its explanatory performance in contrast to OLS, GLM and a series of GAMLSS specifications. Based on R^2 and gAIC as evaluation criterion, the results point to an increase in the R^2 of ca. 15 percentage points and a decline of the AIC of ca. 3.4 % of the GAMLSS relative to the GLM approach. Finally, [19] employ the GAMLSS approach for deriving collateral values of house prices in Austria. While the study focusses on advanced statistical modelling techniques of quantile, Bayesian and Markow-chain (spatial) effects, the authors provide a deep inside on the contribution of the GAMLSS approach in modelling real estate prices. On the basis of these results, I expect a substantial reduction in forecasting bias when employing the GAMLSS approach in Munich's rental market.

3. MODEL PARAMETERIZATION AND OUT-OF-SAMPLE APPROACH

In a first step, I start with a traditional OLS hedonic equation for Munich's residential market of the form:

$$R_{i,j,t} = \boldsymbol{X}\beta + u_{i,j,t} = \boldsymbol{X_{i,j,t}}\beta + \boldsymbol{W}_j\phi + \boldsymbol{Z}_i\theta + \boldsymbol{\psi}_t\alpha_t + \boldsymbol{\psi}_j\alpha_j + u_{i,j,t} \quad \text{(III)}$$

where the response R corresponds to dwelling's i asking rent in quarter t and ZIP area j, $\boldsymbol{X_{i,j,t}}$ in bold correspond to the matrix of dwelling-specific characteristics, $\boldsymbol{W}_j$ accounts for j ZIP-area-specific covariates and the $\boldsymbol{Z}_i$ is a matrix of distances of each dwelling to general amenities such as schools, supermarkets, etc. $\boldsymbol{\psi}_t$ and $\boldsymbol{\psi}_j$ account for time-trend and ZIP-spatial fixed effects, the error term u is set to be iid and i=1,…,26'775; j=1,…,75 and t=2013Q1,…, 2015Q4. In a second step, I model the metric covariates in $\boldsymbol{X}$ and $\boldsymbol{Z}$ in the OLS equation (III) as penalized B-splines $\boldsymbol{f}()$ and optimize the equation as a Generalized Additive Model (GAM) model via the backfitting algorithm [11, 14] based on the following equation:

$$R_{i,j,t} = \boldsymbol{f}(\boldsymbol{X})_{i,j,t} + \boldsymbol{f}(\boldsymbol{Z})_i + \boldsymbol{W}_j\phi + \boldsymbol{\psi}_t\alpha_t + \boldsymbol{\psi}_j\alpha_j + e_{i,j,t} \quad \text{(IV)}$$

The GAMLSS approach allows, besides the incorporation of non-linear and spatial effects, the dynamic modelling of a series of parameterizations of the four moments of the response with regard to the response variable. However, two main problems arise when modelling GAMLSS equations: There is no sufficient evidence in the field of real estate on which distribution to use when explaining real estate prices and most importantly, on how to parameterize the g parameters in μ, σ, ν, τ. When trying to put all these parameters together, the combinations increase exponentially requiring several months of estimation[iv]. In order to overcome with these problems, I simplify the estimation as follows: Firstly, I define the set of covariates, both linear and penalized B-splines, for the μ equation and set them as the initial values in the variance, skewness and kurtosis equation under the normal distribution[v]. Secondly, I optimize the parameterization of the model iteratively based on the procedure developed by Rigby and Stasinopoulos denominated "stepGAICCAll.A" in R in order to select the set of optimal covariates for each single equation[vi]. stepGAICCAll.A is a strategy for selecting the covariates using the $gAIC$. In simple words, the procedure starts with a fixed distribution and selects an appropriate model for μ with fixed σ, ν, τ; afterwards it optimizes the σ model holding ν, τ fixed and so on until τ is optimized. The procedure optimizes the model also backwards, i.e. select τ and hold μ, σ, ν fixed until μ is optimized. Finally, stepGAICCAll.A compares the forward and backward models and choses the optimal set of covariates. In a last step, I re-estimate the optimized parameterized equations and provide the results for a model with only the μ equation and for a model with the four optimized parameterized μ, σ, ν, τ. The in-sample evaluation of the explanatory power of each model is completed via the generalized Akaike criterion ($gAIC$) and the unconstrained R-squared (R^2), see [8], whereas the $gAIC$ is defined as the negative likelihood plus a fixed penalty factor k multiplied by the total degrees of freedom df:

$$gAIC = -2\hat{l} + (k.df) \qquad \text{(V)}$$

In order to assess the forecasting accuracy of the GAMLSS approach, I examine the out-of-sample forecasting accuracy with a bootstrap procedure. I estimate the models excluding 1'300 observations (4.86 %) randomly[vii] and obtain the predicted functional form. Afterwards, I predict the remaining 4.86 % of the sample and calculate the error variance (EV), root mean squared error (RMSE), mean absolute error (MAE), mean percentage error (MPE) and mean absolute percentage error (MAPE). Finally, I repeat the procedure 600 times with replacement and save the results, see [7].

An example of the loop implemented in R to run the simulation is as follows. On a first step, it defines the randomly in- and out-of-sample iteratively with replacement, i.e. 95.14 % and 4.86 % of the sample. In a second step, it estimates the GAMLSS model for the "estimation sample" based on the mu, sigma, nu and tau equations using the RS algorithm. In a final step, it predicts the responses of the "forecast sample" based on the estimated GAMLSS-parameters. Finally, it replaces the sample, repeats the procedure 600 times and saves the results[viii].

```
set.seed(1234);for(i in 1:600){
Estimation_Sample     <- Sample[-sample(1:dim(Sample)[1],1300,replace=T),]
Forecast_Sample       <- Sample[ sample(1:dim(Sample)[1],1300,replace=T),]
Model                 <- gamlss(Mu_Equation,data= Estimation_Sample,
                         sigma.formula = ~ Sigma_Equation,
                         nu.formula    = ~ Nu_Equation,
                         tau.formula   = ~ Tau_Equation, family = NO, method = RS)
Y_hat[[i]]            <- predict(Model, newdata= Forecast_Sample, type="response")
print(i)       }
```

4. DATA AND MARKET DESCRIPTION

Since the sample size is a very important factor either in parametric or semi-parametric or nearly any kind of empirical analysis, it might be worth taking a look at the datasets of the studies focussing on hedonic estimation. In the considered literature, [15] employ ca. 3k dwellings for estimating the hedonic equation in Munich in 2007, whereas [8] focus on ca. 2k lots from 2006 until 2007. Correspondingly, [19] employs ca. 3k data points on single-family houses from 1997 until 2009. For this study in contrast, I merged three different databases. Firstly, I gathered 26'775 observations from multiple listing services (MLS) in Munich from

2013-Q1 until 2015-Q4 as collected by the empirica system database[ix], which contain the most important MLS providers such as Immoscout, Immonet and Immowelt as well as seven others. After filtering and deleting duplicates, the empirica system database provides geographically referenced data with 20 hedonic characteristics. In order to avoid a large drop in sample size due to missing binary hedonic attributes such as wooden floor, sauna or laminate floor, I only include 12 relevant hedonic characteristics. Secondly, I merged two socioeconomic variables: purchasing power per household and the number of inhabitants per households both on a ZIP-code level and yearly basis from the GfK-database[x]. Finally, I gathered the geographical location of relevant amenities from open street map and estimate the lowest Euclidean distance between the amenities and the dwellings[xi]. The matrix $\boldsymbol{Z}_i$ includes the distance vectors in Km. The final database consists finally on 26'775 residential flats, each with a vector of 12 hedonic characteristics, 12 distance variables and 2 socioeconomic variables[xii]. The sample includes only flats rather than single and multi family houses in order to avoid sample bias.

Needless to say, the real estate data employed in this paper measures asking rents rather than transaction or contract rents. As opposed to other European countries, the size of the rental market in Germany – and specifically in Munich – is large, which points to an active use of ML services by landlords and tenants as a traditional marketing channel. In contrast to registry or mortgage approval databases, the advantage of MLS databases such as the empirica database relies on the fast access to real data when estimating hedonic models. Although the data fails at capturing contract rents, the deviation is not expected to lead to a error bias, especially after controlling for 12 hedonic characteristics as explained by [22, 13].

Munich is the capital of the state of Bavaria and with approximately 1.5 million inhabitants the largest city in Bavaria and the third largest city of Germany. Munich is one of the most powerful economic centres of Germany and besides Frankfurt a very important financial

centre with important insurance, biotechnological and media companies. Over the last five years, it has developed as one of the most active residential markets in Germany as the demand for living space has been driven by the strong economic growth, strong competition of large companies and mainly by the stable labour demand. The city employs nowadays more people than workers living in the city, i.e. the commuter rate from workers from outside of Munich is very high. Because of its economic strength and excellent infrastructure, Munich recorded a steady population growth, e.g. during 2002 and 2014 the population increased by around 256k inhabitants, whereas at the same time only approximately 60k new apartments were built. With an average household size of 1.8 it is clear that there are some frictions between supply and demand and as it can be assumed that the population in Munich will increase further, the pressure on the housing market remains high leading to rising real estate prices and rents.

During the observation period, the mean asking rent was ca. 15 €/m²/p.m., whereas the lowest and highest rents ranged from ca. 10 €/m²/p.m. to 21 €/m²/p.m. respectably, as table 1 shows. While the average dwelling in the sample is ca. 80 m² and accounts for ca. 3 rooms, the average distances to selected amenities shows that supermarkets or restaurants are accessible within less than 340 m. The density of theatres, fire stations and swimming pools instead is low and accessible within ca. 3 Km. from an average dwelling. An average household in Munich has a purchasing power of ca. 52k €/p.a. The lowest 5 % of Munich's population has a yearly purchasing power of approx. 43'784 €/p.a., which is remarkably 4.8 % higher than the German purchasing power average.

------- Please insert Table 1 here -------

------- Please insert Figure 3 here -------

One of the main improvements of the GAMLSS approach is the incorporation of variances of the response across space and time into the estimation model. In order to validate this

preliminary assumption, Figure 3 shows the standard deviation of rents divided by the mean rents in each ZIP area (relative standard deviation). The higher the relative the standard deviation is the higher is the span of rents within the observed ZIP area. The map shows that the deviation of rents relative to their ZIP's mean decreases for rising distance to city centre, providing evidence of a heterogeneous distribution of rents across space regardless of the rent level. Thus, the span of rents diverges by more than ± 19 % from the mean rent in the city centre, whereas dwellings in the east and west of Munich present a more consolidated rent range clearly below ± 12 % from the mean rent.

------- Please insert Table 2 here -------

A further contribution of the GAMLSS model is the consideration of skewed responses. The development of rents as presented in table 2 shows a steady growth path of rents over time with remarkable variations in its moments. Just like in the spatial case, the cross-sectional variation of rents relative to its mean points to a widening of the distribution over time as it increased by almost 4 percentage points to 23.9 % during the entire observation sample. While the skewness of asking rents is stable and positive indicating a significant concentration of observations on the left with a longer tail on the right, the kurtosis shows strong deviations from the univariate normal distribution. On average, asking rents tend to be leptokurtic with fatter tails, which implies that the extreme values in the tails approximate to zero slower than the normal distribution, i.e. outliers are more likely. Yet, in an OLS context, the data would require the usage of robust estimators or large preliminary adjustments, e.g. censoring. Since the descriptive statistics provide evidence for a heterogeneous distribution of variances in both the location and the shape of rents, the GAMLSS approach is expected to capture these anomalies and lead to higher in- and out-of-sample understanding of rents in comparison to traditional methods.

5. IN-SAMPLE EXPLANATORY AND OUT-OF-SAMPLE ACCURACY RESULTS

The paper aims at showing the explanatory as well as forecasting accuracy of the GAMLSS approach in hedonic modelling relative to traditional models such as the OLS and GAM. For this purpose, I gathered 26'775 observations in Munich, Germany, and test the forecasting accuracy of the models leaving ca. 5 percent of the observations with replacement iteratively 600 times and evaluate the results by the error variance (EV), root mean squared error (RMSE), mean absolute error (MAE), mean percentage error (MPE) and mean absolute percentage error (MAPE). While the OLS and GAM approach do not adapt the distribution of the response variable to the underlying covariates, the GAMLSS model allows a more dynamic incorporation of these effects as it considers the four underlying moments of the response separately. I evaluate the in-sample regression models via generalized $gAIC$ criterion and unrestricted R^2, rather than the estimated coefficients since the comparability of estimated coefficients is restricted due to the non-linear modelling of metric covariates. While the OLS approach allows a direct interpretation of the estimated elasticities, non-linear coefficients are merely evaluated by a significance test.

------- Please insert Table 3 here -------

Table 3 shows the $gAIC$, the R^2 and the corresponding degrees of freedom (DF) of each model based on the in-sample sample for Munich (overall sample) as well as the evaluation indicators of the out-of-sample simulation. Firstly, the results confirm that traditional approaches such as the OLS or GAM do not explain the underlying factors of asking rents in Munich's residential market sufficiently based on the $gAIC$ and the R^2 criterion. While there is a significant increase in the explanatory power of the GAM approach relative to the OLS model, the GAMLSS approach does capture the explanatory power of the covariates remarkably, regardless of the moment equations included, i.e. the GAMLSS model with or without the μ, σ, ν, τ parameters. Although the GAM model outperforms the OLS approach, the GAMLSS models show a higher understanding of the underlying factors based on the

$gAIC$ and the R^2. The highest increase in the R^2 of approx. 25 percentage points was obtained by the GAMLSS model including linear and non-linear covariates in the last tree moments of the response relative to the OLS approach. The explanatory power of the full GAMLSS model reflects in the relative decrease of the $gAIC$ by 9.0 % and by 3.3 % relative to the OLS and GAM models respectively.

When looking at the out-of-sample explanatory power in Table 3 and Figure 4, the results confirm that the GAMLSS method has indeed an enhanced understating of Munich's rental market. This understanding is at the same time translated into the forecasting results as the GAMLSS models outperform the traditional models. Based on the EV, the GAMLSS models outperform the OLS model by almost –2 percentage points and the GAM model by almost –30 BP. Yet, based on 600 bootstrap loops, the mean absolute error of the full GAMLSS model has a mean of ca. 1.6 €/m² which corresponds to an improvement of about –3.5 % (–0.06 €/m²) and of –17.2 % (–0.33 €/m²) in contrast to the OLS and the GAM model, respectively. These results indicate that the GAMLSS forecasting errors tend to lie in a closer corridor towards the true values and that extreme forecast are less likely. When looking further into the MPE, the results show that the forecasted rents are positive skewed over all models, i.e. the models forecast rents higher than the actual values, leading to a overestimation of rents. This bias decreases when the estimator controls for skewness in the distribution of the response, i.e. the GAMLSS approach. Finally, the mean absolute percentage error between actual and forecasted asking rents decreases in the full GAMLSS model by ca. 2.26 percentage points in contrast to the OLS model. Thus, the GAMLSS model presents the highest forecasting accuracy, providing evidence that rents are asymptotically more precisely forecasted by the GAMLSS approach, especially when dealing with extreme values and spatial varying rents.

------- Please insert Figure 4 here -------

6. CONCLUSION AND RECOMMENDATIONS

Hedonic models are very useful instruments for institutional investors, researchers and policy makers in order to determine the underlying drivers of rents and prices within a market and consequently to identify future market developments or possible investment opportunities. In view of the advantages in statistical inference and also ascribed to the progress in computational speed, the empirical estimation of hedonic models has faced large improvements during the last three decades, leading at the same time to an increasing number of hedonic functional specifications. Despite of the remarkable statistical improvements, but especially due to the complexity of real estate, hedonic models are still limited in capturing real estate relevant effects such spatially varying prices or non-normal responses. Estimation methodologies of the "new era" have emerged, conceptualizing and expanding the assumptions behind traditional hedonic models in order to maximize the explanatory power and minimize forecasting errors. A very well-known framework of the "new era" – but rarely used in real estate studies – is the Generalized Additive Model for Location, Scale and Shape (GAMLSS), which allows the distribution of the response to vary according to both their own four moments – mean, variance, skewness and kurtosis – and the covariates. While the GAMLSS approach has led to an enormous increase in the explanatory power of models in natural and medical sciences, its advantages in the field of hedonic modelling remain scarce and partly new. This paper explored therefore the main advantages of the GAMLSS approach in the context of hedonic real estate research based on more than 25k observation in Munich, Germany.

The results can be explored from two perspectives. The computational requirements necessary to estimate GAMLSS models are high and might constitute one of the limiting factors for (institutional) researchers, despite the large detailed knowledge on statistical inference and programming. However, the empirical results confirm that Munich's residential market can be more accurately explained by means of the GAMLSS model than by traditional models such as the OLS or the GAM. Furthermore, the out-of-sample forecasting simulation based 600

loops with replacement confirmed that the complexity of the GAMLSS models does pay off as measured by traditional forecasting evaluation indicators. The mean absolute error fell by almost 17 % and 4 % in the GAMLSS approach in contrast to the OLS and GAM models, respectably.

Overall, the results show that the theoretical and empirical complexity of the GAMLSS approach in estimating Munich's residential market do pay off in view of the increased explanatory power and primarily in view of the substantial increase in the out-of-sample forecasting accuracy. For policy-makers, the advantages of more accurate hedonic models might lead to a more precise market control and to a better understanding of the local factors affecting rents. For (institutional) researchers, instead, the GAMLSS approach offers a new area of investigation since the framework offers a large number of calibrations, depending on the observed market, data deepness and variables' behaviour. Finally, it is – not surprisingly – to expect that further research might find an even higher forecasting accuracy in view of the large potential of the GAMLSS framework.

7. LITERATURE

[1] F. Angjellari-Dajci, and R. J. Cebula, *The impact of historic district designation on the prices of single-family homes in the oldest city in the United States, St. Augustine, Florida*, Journal of Property Research (2016), DOI: 10.1080/09599916.2016.1151918.

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

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

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

[5] J. Cohen, C. Coughlin and J. Clapp, *Local Polynomial Regressions versus OLS for Generating Location Value Estimates: Which is More Efficient in Out-of-Sample Forecasts?*, Federal Reserve Bank of St. Louis – Research Division - Working Paper Series (2015), No. 2015-014A

[6] R. Davidson and J.G. MacKinnon, *Econometric Theory and Methods, second edition*, New York, Oxford University Press (2003).

[7] F.X. Diebold, *Elements of forecasting*, Thomson South-Western (2004), pp. 297-299.

[8] L. Florencio, F. Cribari-Neto and R. Ospina, *Real estate appraisal of land lots using GAMLSS models*, Cornell University Library (2011), arXiv preprint arXiv:1102.2015

[9] A.S. Fotheringham, C. Brunsdon and M. Charlton, *Geographically Weighted Regression – The analysis of spatially varying relationships*, John Wiley & Sons (2002).

[10] L. I. U. Hao and C. H. E. N. Lang-man, *Non-parametric and Non-linear analysis of stock liquidity with higher frequency data in China on GAMLSS model*, Journal of Shanxi Finance and Economics University (2011), 4, pp. 5.

[11] T. Hastie and R. Tibshirani, *Generalized Additive Models, Monographs on statistics and applied probability*, 43, 1. ed., Chapman and Hall/CRC (1990)

[12] I.L. Hudson, S.W. Kim and Keatley, *Climate effects and temperature thresholds for Eucalypt flowering – A GAMLSS ZIP approach*, 19th International congress on Modelling and Simulation (2011).

[13] R.C. Lyons, *Price signals in illiquid markets – The case of residential properties in Ireland, 2006-2012*, Trinity Economics Papers (2013), TEP No. 0613.

[14] C, Mason and J.M. Quigley, *Non-parametric Hedonic Housing prices*, Housing Studies, 11(3) (1996), pp. 373–385.

[15] A. Mayr, N. Fenske, B. Hofner, T. Kneib and M. Schmid, *GAMLSS for high-dimensional data – a flexible approach based on boosting*, Technical Report Number 098,2010 (2010), University of Munich.

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

[17] T.-C. Peng and Y.-H Chiang, *The non-linearity of hospitals' proximity on property prices: experiences from Taipei, Taiwan*, Journal of Property Research, 32 (4) (2015), 341-361, DOI: 10.1080/09599916.2015.1089923

[18] R Development Core Team. R., *A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing(2016). ISBN 3-900051-07-0.

[19] A. Razen, W. Brunauer, N. Klein, T. Kneib, S. Lang and N. Umlauf, *Statistical risk analysis for real estate collateral valuation using Bayesian distributional and quantile regression*, Working papers in economics and statistics(2014).

[20] R.A. Rigby and D.M. Stasinopoulos, *Generalized additive models for location, scale and shape*. Applied Statistics (2005), 54:507–554.

[21] F. Serinaldi, *Distributional modelling and short-term forecasting of electricity prices by Generalized Additive Models for Location, Scale and Shape",* Energy Economics, 33(6) (2011), pp. 1216-1226.

[22] C. Schimizu, K.G. Nishimura. and T. Watanabe, *House prices from magazines, realtors and the land registry*, Property markets and financial stability, Bank for international Settlements BIS Papers No. 64(2012), pp. 29-36.

[23] M. Stasinopoulos, B. Rigby, G. Heller, V. Voudouris and F. De Bastiani, *Flexible regression and smoothing – Using GAMLSS in R,* (2017), Chapman & Hall, CRC Press.

[24] M. Widłak, J. Waszczuk and K. Olszewski, *Spatial and hedonic analysis of house price dynamics in Warsaw*, National Bank of Poland (NBP) (2015), Working Papers No. 197.

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

Endnotes

[i] See: Hudson, I.L., Kim, S.W. and Keatley 2011; Serinaldi, F. 2011; Hao, L. I. U. and Langman, C. H. E. N. 2011; among others. See also http://www.gamlss.org/?page_id=1050.

[ii] Restricted in the incorporation of spatial and temporal terms in the sense that they only are included as dummy terms or a weighting matrix.

[iii] The GLM model allows a more flexible modelling of the hedonic equation, however the underlying estimator does not consider any interdependence of the endogenous variables.

[iv] Based on e.g. 15 covariates the dynamic combinations between 4 models, different distribution and non-linear effects exploit exponentially.

[v] The GAMLSS approach allows the usage of a series of distributions. However, the optimization of the models under distributions such as student skewed, gama, skewed exponential or sin-arcsinh was very instable, leading to rising likelihood values, variances and partly failures in the likelihood estimations. For simplicity, I focus therefore merely on the normal distribution, which at the same time led to the best results.

[vi] This procedure was very time consuming, it took almost tree weeks to estimate the models based on a RAM of 16 GB.

[vii] The random generator was specifies as set.seed(1234).

[viii] The number of out-of-sample observations was chosen based on the number of quarters.

[ix] www.empirica-systeme.de

[x] www.gfk.com

[xi] www.openstreetmap.com/. I only focus on amenities that might be related to rents rather than including amenities such as speed cameras, pub with darts, flower shop or hair salon.

[xii] The computational requirements for the estimation of the models were very large. The programmed loop in R used continuously a RAM of ca. 14 GB over several days.

Table 1: Descriptive statistics of variables

Variable	Mean	SD	Q5%	Q30%	Q50%	Q70%	Q95%
Metric and binary covariates							
Asking rent €/m²/p.a.	14.823	3.171	10.87	13.06	14.26	15.76	20.833
Living area	79.456	38.38	31.5	58.49	73	90	150
Age relative to 2016	41.323	35.378	1	18	39	51	115
Number of rooms	2.609	1.059	1	2	2.5	3	4
Central heating system (0=else)	0.774	0.418	0	0	1	1	1
Individual heating system (0=else)	0.007	0.083	0	0	0	0	1
Floor heating system (0=else)	0.063	0.242	0	0	0	0	1
Built-in kitchen (0=else)	0.668	0.471	0	0	1	1	1
Balcony (0=else)	0.796	0.403	0	1	1	1	1
Refurbished (0=else)	0.254	0.441	0	0	0	0	1
As-good as new (0=else)	0.108	0.311	0	0	0	0	1
Longitude	11.561	0.059	11.458	11.528	11.563	11.589	11.662
Latitutde	48.137	0.030	48.088	48.118	48.138	48.155	48.185
Inhabitants per household	1.795	0.089	1.655	1.735	1.78	1.822	1.948
Purchasing power per household €/p.a.	53‘806	6‘019	45‘890	49‘562	52‘463	56‘211	64‘586
Distance covariates in Km.							
Theatre	1.783	1.189	0.28	1.018	1.588	2.261	4.009
Swimming pool	3.972	2.411	0.686	2.236	3.617	4.951	8.347
Supermarket	0.333	0.259	0.043	0.184	0.273	0.386	0.774
Subway entrance	0.970	1.182	0.099	0.317	0.518	0.913	3.47
School	0.848	0.627	0.146	0.405	0.646	1.027	2.073
Restaurant	0.247	0.196	0.035	0.111	0.189	0.304	0.667
Pub	0.570	0.449	0.066	0.251	0.441	0.71	1.438
Museum	2.045	1.236	0.336	1.242	1.855	2.551	4.392
Memorial	0.755	0.518	0.168	0.441	0.613	0.901	1.873
Kindergarten	0.421	0.253	0.078	0.264	0.378	0.516	0.891
Fire station	3.488	1.723	0.63	2.406	3.511	4.582	6.151
Biergarten	0.906	0.483	0.215	0.587	0.815	1.146	1.739

Table 2: Descriptive statistics of asking rents over time

Rents in €/m²/p.m.		2013				2014				2015				Overall
		Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	Q1	Q2	Q3	Q4	
Mean μ_i	Level	13.954	13.743	14.167	14.437	14.505	14.573	14.785	15.080	15.240	15.449	15.687	15.547	14.823
	yoy%	/	/	/	/	4.0%	6.0%	4.4%	4.5%	5.1%	6.0%	6.1%	3.1%	11.4%
Standard deviation σ_i	Level	2.782	2.554	2.691	2.924	3.147	2.996	3.013	3.023	3.186	3.170	3.530	3.712	3.171
	% of mean	19.9%	18.6%	19.0%	20.3%	21.7%	20.6%	20.4%	20.0%	20.9%	20.5%	22.5%	23.9%	21.4%
	yoy%	/	/	/	/	13,1%	17,3%	12,0%	3,4%	1,2%	5,8%	17,2%	22,8%	33.4%
Skewness υ_i		1,383	0.980	1.639	1.432	1.492	1.315	1.220	1.114	1.314	1.082	1.412	1.042	1.323
Kurtosis τ_i		4,315	1.783	8.540	6.109	7.730	4.845	3.666	2.661	3.477	2.262	4.038	2.846	4.367
N	n	2‘259	1‘591	1‘326	1‘621	3‘162	2‘622	2‘393	2‘553	2‘110	1‘954	2‘983	2‘201	26‘775
	% of total	8.4%	5.9%	5.0%	6.1%	11.8%	9.8%	8.9%	9.5%	7.9%	7.3%	11.1%	8.2%	100%

Table 3: In- and out-of-sample model accuracy

In-sample estimation							Out-of-sample forecasting evaluation after 600 loops	
Estimation method	σ_i, ν_i, τ_i parameters	Generalized AIC			R^2	Df		
		Absolute	Relative to					
			OLS	GAM				
OLS	/	121’517.8	/	/	27.9%	39	EV	7.2117
							RMSE	2.6855
							MAE	1.9318
							MPE	-3.1496
							MAPE	13.4010
GAM	/	114’436.5	-5.8%	/	45.6%	98	EV	5.5645
							RMSE	2.3589
							MAE	1.6580
							MPE	-2.5102
							MAPE	11.5589
GAMLSS	–	113’339.3	-6.7%	-1.0%	48.0%	114	EV	5.2257
							RMSE	2.2860
							MAE	1.6036
							MPE	-2.3451
							MAPE	11.1916
GAMLSS	+	110’641.3	-9.0%	-3.3%	53.3%	163	EV	5.2799
							RMSE	2.2979
							MAE	1.6001
							MPE	-2.2868
							MAPE	11.1371

Notes: Table provides the out-of-sample forecasting results after 600 loops excluding ca. 5% of the observations with replacement. Error variance (EV), root mean squared error (RMSE), mean absolute error (MAE), mean percentage error (MPE) and mean absolute percentage error (MAPE).

Figure 1: Regression line under homoscedasticity

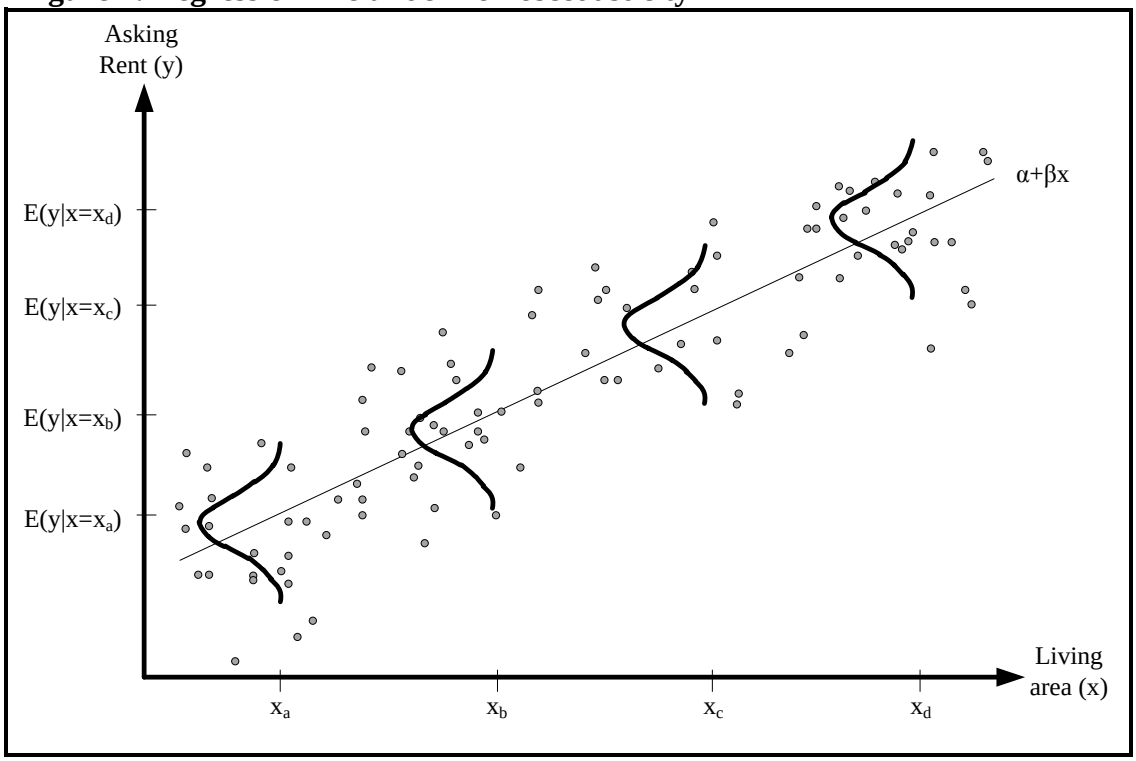
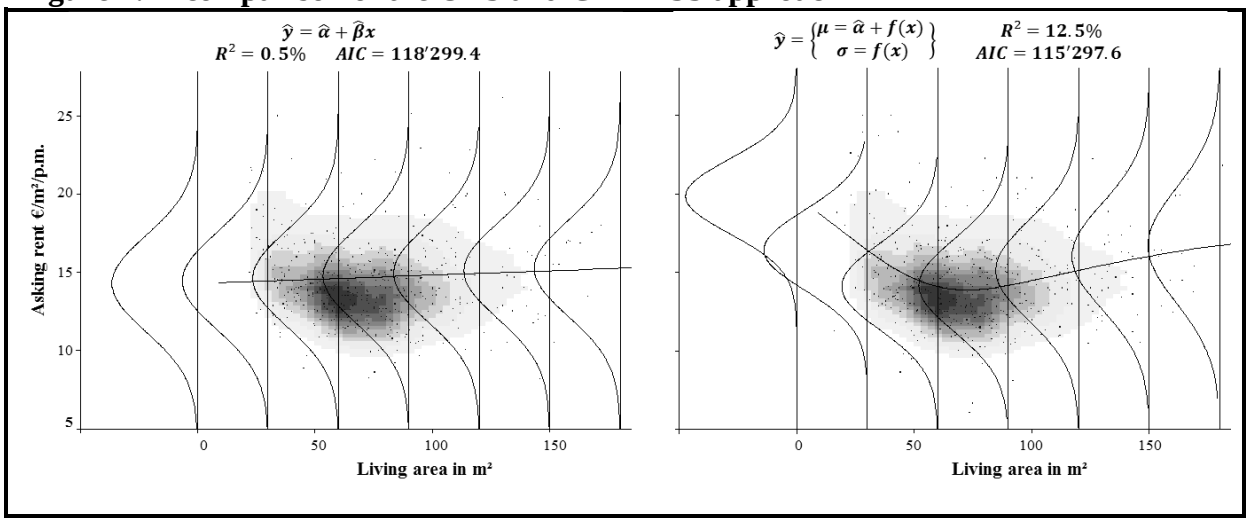
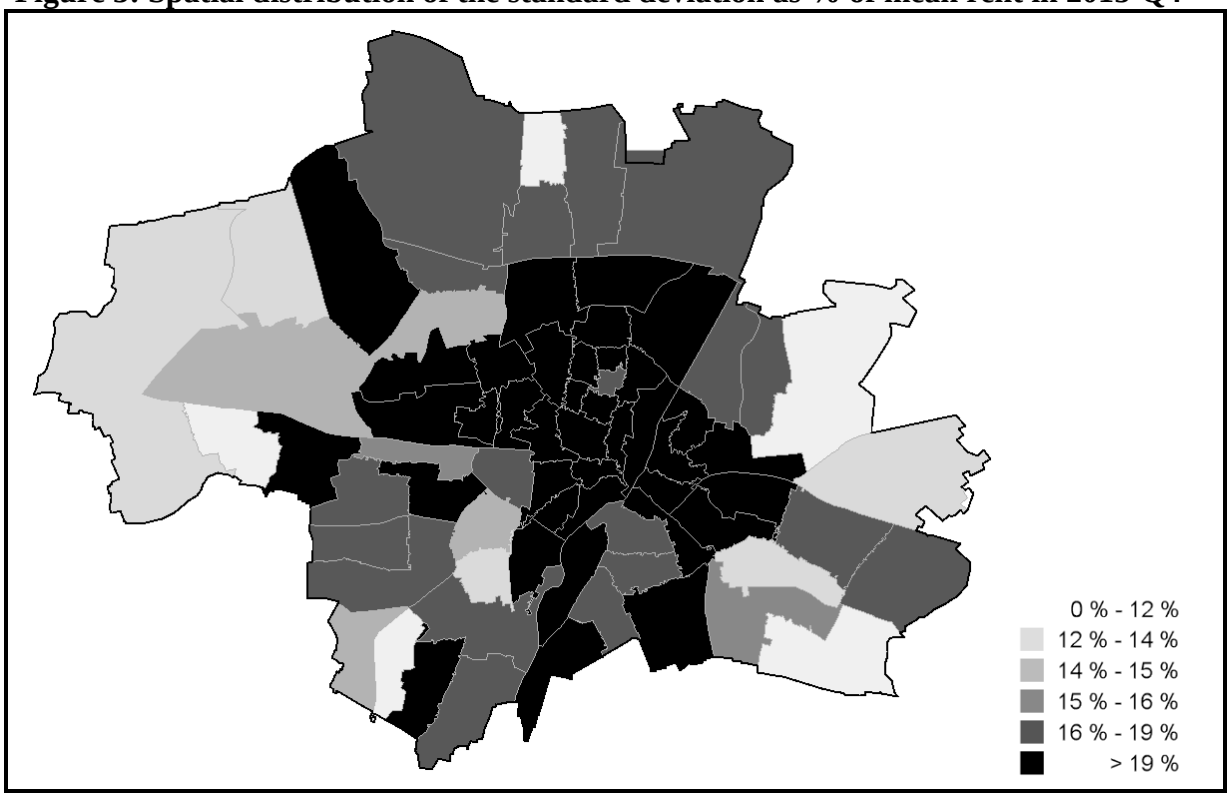


Figure 2: A comparison of the OLS and GAMLSS approach



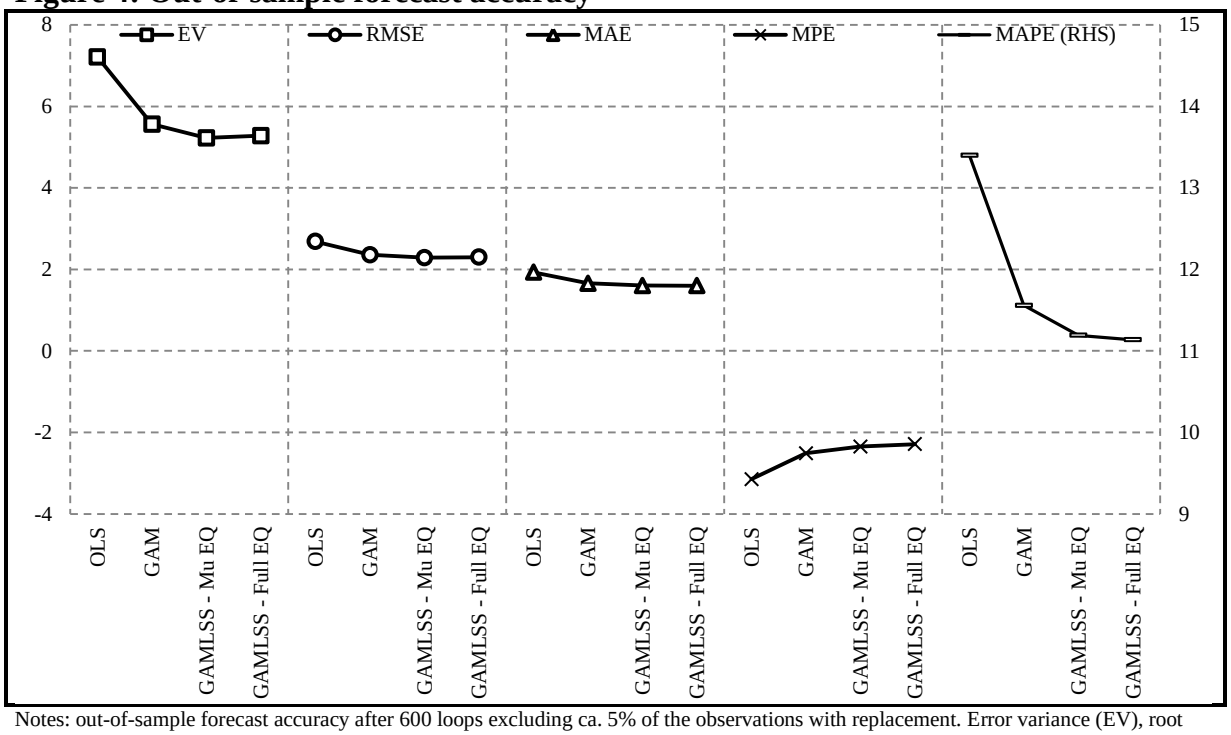
Notes: Models estimated without any transformation in both sides of the equation.

Figure 3: Spatial distribution of the standard deviation as % of mean rent in 2015-Q4



Notes: The map shows the relative standard deviation (standard deviation divided by mean) of asking rents acrossthe ZIP areas of Munich.

Figure 4: Out-of-sample forecast accuracy



Notes: out-of-sample forecast accuracy after 600 loops excluding ca. 5% of the observations with replacement. Error variance (EV), root mean squared error (RMSE), mean absolute error (MAE), mean percentage error (MPE) and mean absolute percentage error (MAPE).