University of Innsbruck

Statistical risk analysis for real estate collateral valuation using Bayesian distributional and quantile regression

Alexander Razen, Wolgang Brunauer, Nadja Klein, Thomas Kneib, Stefan Lang, Nikolaus Umlauf

Working Papers in Economics and Statistics

2014-12

Research Platform
eeecon
Empirical and Experimental Economics
University of Innsbruck
http://eeecon.uibk.ac.at/

University of Innsbruck
Working Papers in Economics and Statistics

The series is jointly edited and published by

- Department of Economics
- Department of Public Finance
- Department of Statistics

Contact Address:
University of Innsbruck
Department of Public Finance
Universitaetsstrasse 15
A-6020 Innsbruck
Austria
Tel: + 43 512 507 7171
Fax: + 43 512 507 2970
E-mail: eeecon@uibk.ac.at

The most recent version of all working papers can be downloaded at
http://eeecon.uibk.ac.at/wopec/

For a list of recent papers see the backpages of this paper.

Statistical Risk Analysis for Real Estate Collateral Valuation using Bayesian Distributional and Quantile Regression

Alexander Razen[1], Wolfgang Brunauer[2], Nadja Klein[3], Thomas Kneib[3], Stefan Lang[1], Nikolaus Umlauf[1]

[1] *University of Innsbruck, Universitätsstr. 15, 6020 Innsbruck, Austria.*

[2] *UniCredit Bank Austria AG, Julius Tandler-Platz 3, 1090 Wien, Austria.*

[3] *Georg August University Göttingen, Platz der Göttinger Sieben 5, 37073 Göttingen, Germany.*

Abstract

The Basel II framework strictly defines the conditions under which financial institutions are authorized to accept real estate as collateral in order to decrease their credit risk. A widely used concept for its valuation is the hedonic approach. It assumes, that a property can be characterized by a bundle of covariates that involves both individual attributes of the building itself and locational attributes of the region where the building is located in. Each of these attributes can be assigned an implicit price, summing up to the value of the entire property.

With respect to value-at-risk concepts financial institutions are often not only interested in the expected value but also in different quantiles of the distribution of real estate prices. To meet these requirements, we develop and compare multilevel structured additive regression models based on GAMLSS type approaches and quantile regression, respectively. Our models involve linear, nonlinear and spatial effects. Nonlinear effects are modeled with P-splines, spatial effects are represented by Gaussian Markov random fields. Due to the high complexity of the models statistical inference is fully Bayesian and based on highly efficient Markov chain Monte Carlo simulation techniques.

Keywords: Bayesian hierarchical models, hedonic pricing models, GAMLSS, distributional regression, quantile regression, multilevel models, MCMC, P-splines, value-at-risk

1 Introduction

The financial crisis of 2008, originated from the U.S. subprime mortgage crisis, impressively revealed the significance to the global economy of housing in general and its reliable valuation in particular. A widely used concept here is the hedonic valuation approach, theoretically developed by Lancaster (1966) and Rosen (1974). It assumes that a property can be characterized by a bundle of covariates that involves both individual attributes of the building itself and locational attributes of the region where the building is located in. Each of these attributes can be assigned an implicit price, summing up to the value of the entire property, see e.g. Sheppard (1999) or Malpezzi (2003).

Hedonic pricing suggests the use of regression models that explain the price in dependence of attributes. The functional form of this dependence should allow for nonlinearities, see Wallace (1996) or Malpezzi (2003). Thus, we draw on generalized structured additive regression (STAR) models, described e.g. in Fahrmeir et al. (2013), where continuous covariates are modeled as P(enalized)-splines, see Eilers and Marx (1996) or Lang and Brezger (2004).

Since the price of a house is considerably influenced by its geographic location, see e.g. Cohen and Coughlin (2008) or Helbich et al. (2014), we include sociodemographic, economic and neighborhood covariates of the regions where the buildings are located in. In doing so, the hierarchical structure of the Austrian political-administrative units, on which these covariates are defined, suggests a multilevel regression model: Single-family homes (level-1) belong to municipalities (level-2), which are nested in district (level-3), which are themselves nested in counties (level-4). Multilevel

regression models are described e.g. in Gelman and Hill (2006) or in Lang et al. (2014) in the context of STAR models.

Brunauer et al. (2013) propose a multilevel STAR model of the form (a more detailed description will be given in Section 3):

$$\begin{aligned}
\text{level-1:} \quad p_{qm} &= f_{1,1}(area) + \ldots + f_{1,q_1}(age) + \mathbf{x}'\boldsymbol{\gamma} + f_{spat_1}(s_1) + \varepsilon_1 \\
\text{level-2:} \quad f_{spat_1}(s_1) &= f_{2,1}(purchase\ power) + \ldots + f_{2,q_2}(education) + f_{spat_2}(s_2) + \varepsilon_2 \\
\text{level-3:} \quad f_{spat_2}(s_2) &= f_{3,1}(price\ index) + f_{spat_3}(s_3) + \varepsilon_3 \\
\text{level-4:} \quad f_{spat_3}(s_3) &= \gamma_0 + \varepsilon_4,
\end{aligned} \tag{1}$$

and analyze the expected value of house prices in Austria. However, the crisis of 2008, if nothing else, has shown the importance of a profound failure analysis, wherefore the expected value obviously is not sufficient. It is rather important to analyze the whole distribution of house prices. Thus, recent studies set the focus to the quantiles of house prices, see e.g. McMillen (2008) or Haupt (2014), with the estimates being based on quantile regression, first introduced by Koenker and Bassett (1978).

In this paper, we extend the work of Brunauer et al. (2013) and estimate conditional quantiles of house prices in Austria with two conceptually different approaches: We apply a number of multilevel STAR models for location scale and shape (GAMLSS type regression), see Rigby and Stasinopoulos (2005) and Klein et al. (2013), and a Bayesian version of quantile regression, see Waldmann et al. (2013).

Statistical inference is fully Bayesian and based on highly efficient Markov chain Monte Carlo (MCMC) simulation techniques. For the estimation we use the R-package **BayesR** (Umlauf et al. (2013)). The final model selection is based on proper scoring rules, see Gneiting and Raftery (2007), and mean weighted errors.

The remainder of the paper is structured as follows: In Section 2, the data set is described in detail. Section 3 presents GAMLSS type regression models and Bayesian quantile regression models in the context of hedonic regression for house prices. Section 4 attends to model selection before the software used for estimation is described in Section 5. Results are presented in Section 6, the final section draws some conclusions.

2 Data description and model specification

Our dataset contains the price as well as different attributes of 3,231 owner-occupied single-family houses in Austria. It has been collected by the UniCredit Bank Austria AG between October 1997 and September 2009 in order to estimate the value of the bank's collateral for mortgages and its associated risk.

The dependent variable in our analysis is the house price per square meter (p_{qm}), which seems, at a first glance, to be approximately lognormally distributed (Figure 1). However, we can see that the mode is not that pronounced than we would expect from the theoretical distribution – an issue we will dwell on in chapter 3.2.

The reason why we examine the prices per square meter instead of the total prices is that the effects of the covariates are typically proportional to the size of the house. Using the prices per square meter implicitly controls for these interactions between the floor area and the remaining house attributes.

The set of explanatory variables can be separated into two groups:

- *Structural covariates* characterize the property itself, e.g. the size, the age or the quality of the house.
- *Spatial covariates* characterize the region where the building is located in. They are defined on different levels (municipal, district or county) and account for sociodemographic, economic or neighborhood attributes.

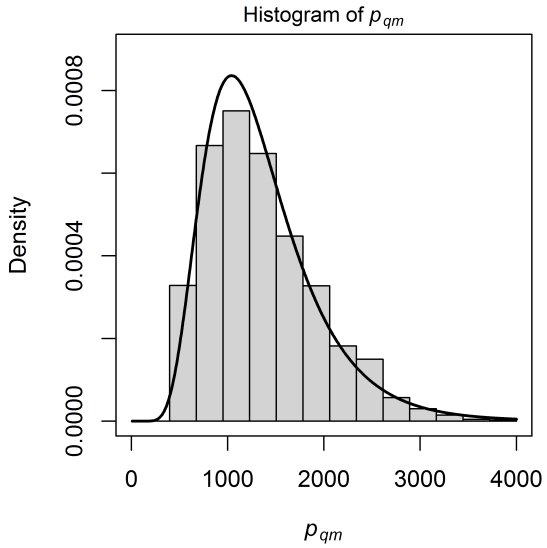

Figure 1: *Histogram of the dependent variable p_{qm} together with the density of a lognormal distribution. The parameters $\mu = 7.1237$ and $\sigma^2 = 0.1763$ determining this density correspond to the empirical values of the data.*

2.1 Structural covariates

The structural covariates involve both continuous and categorical variables. The continuous variables measure the size, the age and the year of purchase, the categorical ones describe the quality and the equipment of the house:

- Continuous covariates: Since we focus on the (logged) prices per square meter the floor area (*area*) should have a decreasing effect due to the law of diminishing marginal utility. For the plot area where the house is built on (*area_plot*) an increasing effect can be assumed. The age of the building at the time of sale (*age*), calculated as the difference between the year of purchase and the year of construction, reflects depreciation over time. Therefore, we expect a decreasing effect. Finally, the year of purchase (*time_index*) incorporates the remaining unexplained temporal heterogeneity, e.g. inflation, indicating an increasing effect.

- Categorical covariates: The quality of the house is measured by its overall condition (*cond_house*), the quality of the heating system (*heat*) and the quality of the bathroom and toilets (*bath*). Obviously, we expect an increasing effect for quality enhancing characteristics. Moreover, the existence of an attic (*attic_dum*), a terrace (*terr_dum*) and a garage (*garage*, further separated into good and bad quality) describe the equipment of the property and should increase its price. However, we have to encode all these categorical covariates according to two slightly different methods that have been employed for data collection (see table 3 in appendix A for details).

2.2 Spatial covariates

Throughout Austria, one can observe considerable spatial variation in house prices, which we want to explain using sociodemographic, economic and neighborhood attributes. These covariates are defined on three different spatial resolutions according to the hierarchical structure of the Austrian political-administrative units: Municipalities are subareas of districts which are themselves nested in counties, as it is illustrated by Figure 2. This hierarchical structure imposes the following multilevel model:

Level-1 is the individual level, on which the prices of 3,231 single-family homes and the corresponding structural covariates are available (see Section 2.1).

Level-2 is the municipal level, where individual observations are available in 946 of the 2379 Austrian municipalities. The following spatial covariates are defined on this level:

- Sociodemographic and economic attributes: The inhabitants' disposable income is reflected by the purchase power index (*pp_ind*) and the share of academics (*educ*) as a proxy for the average level of education. Since the share of academics is strongly positively skewed, it will enter our models logarithmically (denoted by the prefix *"ln"*) in order to avoid volatile estimation results. Both the purchase power index and the share of academics are assumed to have an increasing effect on house prices. In contrast, structural weakness, represented by an excess of the population's age, should affect prices negatively. We measure population's age by an index (*age_ind*) constructed as the population-weighted mean of 20 age cohorts.

- Measures of metropolitan areas and proximity to work: Being a scarce resource land is more valuable in densely populated areas. Therefore, we expect an increasing effect for the population density (*dens*), which will enter the models logarithmically (prefix *"ln"*) for the same reason as before. Furthermore, urban economic theory states that commuting from areas with low economic activity affects prices negatively, while commuting to centers of economic life has an increasing effect. However, close proximity to such centers may also bring along some disamenities for residents, reversing the positive effect. Thus, a high commuter index (*comm*), meaning many employees commute from the respective municipality, should reduce house prices while the effect of a low commuter index is unclear.

Level-3 is the district level, where observations are available in 109 of the 121 Austrian districts. On this level, we incorporate a real estate price index (*wko_ind*) reflecting the local house price level. Additionally, we employ a correlated spatial effect exploiting the neighborhood structure.
Level-4 is the county level, where we only include the global intercept.
Table 3 in appendix A provides a detailed description and summary statistics of all covariates.

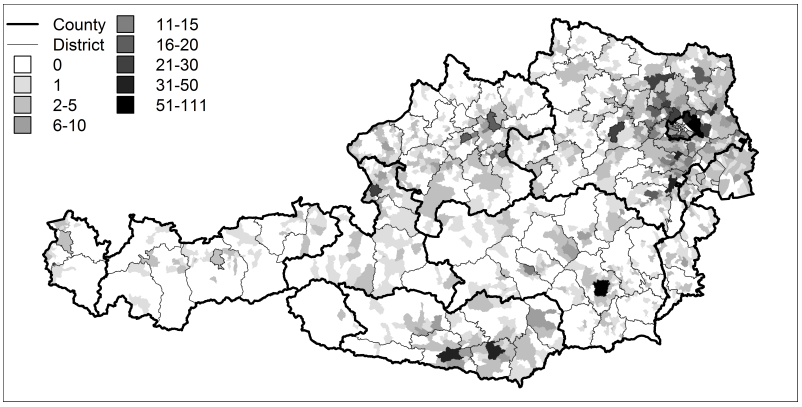

Figure 2: *Number of observations on the municipal level*

3 Methodology

3.1 Multilevel STAR models

Structured additive regression (STAR) models assume that the distribution of the response variable y, given covariates $\mathbf{z}$ and $\mathbf{x}$, belongs to an exponential family, see Fahrmeir et al. (2013) for details. The conditional mean $\mu_i = \mathbb{E}(y_i|\mathbf{z}, \mathbf{x})$ is linked to a structured additive predictor

$$\eta_i = f_1(z_{i1}) + \ldots + f_q(z_{iq}) + \mathbf{x}_i'\boldsymbol{\gamma}, \quad i = 1, \ldots, n,$$

by $\mu_i = h(\eta_i)$. Here, $f_1, \ldots, f_q$ are possibly nonlinear functions of the covariates $\mathbf{z}$, $\mathbf{x}_i'\boldsymbol{\gamma}$ is the usual linear part of the model and h is a known response function. The inverse of the response function $g = h^{-1}$ is called link function.

Normally distributed prices

As the most basic model, we choose the identity for the response function and assume

$$y_i = \eta_i + \epsilon_i = f_1(z_{i1}) + \ldots + f_q(z_{iq}) + \mathbf{x}_i'\boldsymbol{\gamma} + \epsilon_i, \quad i = 1, \ldots, n,$$

with ϵ_i being mutually independent Gaussian with mean 0 and variance σ^2, i.e. $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. The functions f_j comprise usual nonlinear effects of continuous covariates (e.g. the floor area, the age of the building, the year of purchase, etc.) or indicate a spatial index of the region a certain observation belongs to (e.g. municipality, district or county). Using known basis functions B_k, each effect f is approximated by

$$f(z) = \sum_{k=1}^{K} \beta_k B_k(z).$$

with $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)'$ being a vector of unknown regression coefficients to be estimated. Defining the $n \times K$ design matrix $\mathbf{Z}$ with elements $\mathbf{Z}[i,k] = B_k(z_i)$, the vector $\mathbf{f} = (f(z_1), \ldots, f(z_n))'$ of function evaluations can be written in matrix notation as $\mathbf{f} = \mathbf{Z}\boldsymbol{\beta}$. Accordingly, we obtain

$$\mathbf{y} = \boldsymbol{\eta} + \boldsymbol{\varepsilon} = \mathbf{Z}_1\boldsymbol{\beta}_1 + \ldots + \mathbf{Z}_q\boldsymbol{\beta}_q + \mathbf{X}\boldsymbol{\gamma} + \boldsymbol{\varepsilon},$$

where $\mathbf{y} = (y_1, \ldots, y_n)'$, $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)'$ and

$$\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I}).$$

In a multilevel STAR model the regression coefficients $\boldsymbol{\beta}_j$ of a function f_j may themselves obey a regression model with a structured additive predictor, i.e.

$$\boldsymbol{\beta}_j = \boldsymbol{\eta}_j + \boldsymbol{\varepsilon}_j = \mathbf{Z}_{j1}\boldsymbol{\beta}_{j1} + \ldots + \mathbf{Z}_{jq_j}\boldsymbol{\beta}_{jq_j} + \mathbf{X}_j\boldsymbol{\gamma}_j + \boldsymbol{\varepsilon}_j, \tag{2}$$

where the terms $\mathbf{Z}_{j1}\boldsymbol{\beta}_{j1}, \ldots, \mathbf{Z}_{jq_j}\boldsymbol{\beta}_{jq_j}$ represent additional nonlinear functions $f_{j1}, \ldots, f_{jq_j}$, $\mathbf{X}_j\boldsymbol{\gamma}_j$ comprises additional linear effects, and

$$\boldsymbol{\varepsilon}_j \sim \mathcal{N}(\mathbf{0}, \tau_j^2\mathbf{I})$$

is a vector of i.i.d. Gaussian errors, see Lang et al. (2014) for details.

Further levels are possible by assuming that the second level regression parameters $\boldsymbol{\beta}_{jl}$, $l = 1, \ldots, q_j$, again obey a STAR model. In that sense, the model is composed of a hierarchy of complex structured additive regression models, why it is often also called hierarchical STAR model.

In this paper we use the compound prior (2) if a covariate $z_j \in \{1, \ldots, K\}$ is a spatial index and z_{ij} indicates the region observation i pertains to. Then, the design matrix $\mathbf{Z}_j$ is a $n \times K$ incidence matrix with $\mathbf{Z}_j[i,k] = 1$ if the i-th observation belongs to region k and zero else. The $K \times 1$ parameter vector $\boldsymbol{\beta}_j$ is the vector of regression parameters, i.e. the k-th element in $\boldsymbol{\beta}$ corresponds to the regression coefficient of the k-th region. The use of the compound prior (2) allows for further explaining the region specific effect by spatial covariates.

The hierarchical structure of the Austrian political-administrative units suggests a four level regression model: Single-family homes (level-1) belong to municipalities (level-2), which are nested in districts (level-3), which are themselves nested in counties (level-4). Assuming house prices per square meter to be normally distributed ($\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$) leads to the following four level STAR model, which we will call the *Gaussian model*, see Brunauer et al. (2013):

$$
\begin{aligned}
\text{level-1:} \quad \boldsymbol{p_{qm}} &= \mathbf{f}_1(area) + \mathbf{f}_2(area\_plot) + \mathbf{f}_3(age) + \mathbf{f}_4(time\_index) + \\
&\quad \mathbf{f}_5(muni) + \mathbf{X}\boldsymbol{\gamma} + \boldsymbol{\varepsilon} \\
&= \mathbf{Z}_1\boldsymbol{\beta}_1 + \mathbf{Z}_2\boldsymbol{\beta}_2 + \mathbf{Z}_3\boldsymbol{\beta}_3 + \mathbf{Z}_4\boldsymbol{\beta}_4 + \mathbf{Z}_5\boldsymbol{\beta}_5 + \mathbf{X}\boldsymbol{\gamma} + \boldsymbol{\varepsilon} \\
\text{level-2:} \quad \boldsymbol{\beta}_5 &= \mathbf{f}_{5,1}(pp\_ind) + \mathbf{f}_{5,2}(ln\_educ) + \mathbf{f}_{5,3}(age\_ind) + \mathbf{f}_{5,4}(comm) + \\
&\quad \mathbf{f}_{5,5}(ln\_den) + \mathbf{f}_{5,6}(dist) + \boldsymbol{\varepsilon}_5 \\
&= \mathbf{Z}_{5,1}\boldsymbol{\beta}_{5,1} + \mathbf{Z}_{5,2}\boldsymbol{\beta}_{5,2} + \mathbf{Z}_{5,3}\boldsymbol{\beta}_{5,3} + \mathbf{Z}_{5,4}\boldsymbol{\beta}_{5,4} + \\
&\quad \mathbf{Z}_{5,5}\boldsymbol{\beta}_{5,5} + \mathbf{Z}_{5,6}\boldsymbol{\beta}_{5,6} + \boldsymbol{\varepsilon}_5 \\
\text{level-3:} \quad \boldsymbol{\beta}_{5,6} &= \mathbf{f}_{5,6,1}(wko\_ind) + \mathbf{f}^{mrf}_{5,6,2}(dist) + \mathbf{f}_{5,6,3}(county) + \boldsymbol{\varepsilon}_{5,6} \\
&= \mathbf{Z}_{5,6,1}\boldsymbol{\beta}_{5,6,1} + \mathbf{Z}_{5,6,2}\boldsymbol{\beta}_{5,6,2} + \mathbf{Z}_{5,6,3}\boldsymbol{\beta}_{5,6,3} + \boldsymbol{\varepsilon}_{5,6} \\
\text{level-4:} \quad \boldsymbol{\beta}_{5,6,3} &= \mathbb{1}\gamma_0 + \boldsymbol{\varepsilon}_{5,6,3}.
\end{aligned} \tag{3}
$$

The categorical covariates on level-1, describing the quality and equipment of the house, are encoded as dummy variables and subsumed in the design matrix $\mathbf{X}$ with estimated parameters $\boldsymbol{\gamma}$. The possibly nonlinear functions $\mathbf{f}_1, \mathbf{f}_2, \ldots$ are modeled by P-splines (see Section 3.4).

The level-1 equation contains an uncorrelated random municipality effect (*muni*), controlling for unordered spatial heterogeneity. This municipality-specific heterogeneity is modeled through the level-2 equation and is further decomposed into a district and finally into a county level effect (levels 3 and 4). Furthermore, district specific spatial heterogeneity is modeled through the correlated spatial effect *dist* in the level-3 equation by Markov random fields, denoted by the superscript "*mrf*" (see Section 3.4).

From model (3) we immediately get the conditional mean of the house price per square meter p_{qm}. If we are interested in the φ-th conditional quantile of the house price per square meter instead we use the transformation property of the normal distribution

$$Q_\varphi(p_{qm}) = \mathbb{E}(p_{qm}) + \sigma \cdot \Phi^{-1}(\varphi), \tag{4}$$

with $\Phi^{-1}(\varphi)$ being the φ-th quantile of the standard normal distribution. The variance parameter σ^2 (and so the standard deviation σ) can be replaced by a suitable estimator. Equation (4) shows that the quantiles can be received by simply shifting the mean according to the estimated variance.

Due to the additive structure of our model the conditional quantiles of the house price per square meter change additively with changes in values of covariates. Subsequently, the conditional quantiles of the total house price p change proportionally to the floor area of the building:

$$
\begin{aligned}
Q_\varphi(p) &= area \cdot \left(\eta + \sigma \cdot \Phi^{-1}(\varphi)\right) \\
&= area \cdot \left(f_1(area) + \ldots + f_5(muni) + \gamma_1 x_1 + \ldots + \gamma_p x_p + \sigma \cdot \Phi^{-1}(\varphi)\right).
\end{aligned}
$$

Therefore, if for example covariate x_1 changes by one unit, the predictor η – and so the considered quantile of the price per square meter – changes additively by γ_1. The quantile of the total price then changes by $area \cdot \gamma_1$. Thus, the change in the quantiles of the price is proportional to the floor area of the building. Turning to the nonlinear effects, let $f(z)$ be the nonlinear effect of a covariate z, and let $\mathrm{d}f(z) = f(z+1) - f(z)$. Then, analogously, the quantile of the price per square meter changes by $\mathrm{d}f(z)$ and the quantile of the total price changes by $area \cdot \mathrm{d}f(z)$, again being proportional to the size of the house. Furthermore, since $f(.)$ is a nonlinear function, the change differs over the range of z. For the conditional mean we get analog results.

Lognormally distributed prices

In Section 2 we have seen that the empirical house prices per square meter seem to be approximately lognormally distributed instead of being Gaussian. Assuming the response p_{qm} to be lognormally distributed, i.e.

$$p_{qm} \sim \mathcal{LN}(\mu, \sigma^2),$$

leads to normally distributed logged house prices per square meter. Thus, one could guess to improve model (3) by replacing the response p_{qm} by the logged house price per square meter

lnp_{qm}, i.e.

$$\boldsymbol{lnp_{qm}} = \boldsymbol{\eta} + \boldsymbol{\varepsilon}, \tag{5}$$

which we will call the *Loggaussian model*, with the same hierarchical predictor $\boldsymbol{\eta}$ as before and the errors again being mutually independent normally distributed with mean 0 and variance σ^2. We then receive the conditional quantiles of the house price per square meter by

$$\begin{aligned} Q_\varphi(p_{qm}) &= \exp\left(\eta + \sigma \cdot \Phi^{-1}(\varphi)\right) \\ &= \exp\left(f_1(area) + \ldots + f_5(muni) + \gamma_1 x_1 + \ldots + \gamma_p x_p + \sigma \cdot \Phi^{-1}(\varphi)\right) \\ &= \exp\left(f_1(area)\right) \ldots \exp\left(f_5(muni)\right) \exp(\gamma_1 x_1) \ldots \exp(\gamma_p x_p) \exp(\sigma \cdot \Phi^{-1}(\varphi)). \end{aligned}$$

Obviously, the conditional quantiles of the house price per square meter p_{qm} now change multiplicatively with changes in values of covariates. If for example covariate x_1 changes by one unit, the predictor η again changes by γ_1, but the quantiles of the price per square meter now change multiplicatively by the factor $\exp(\gamma_1)$, yielding

$$\begin{aligned} \Delta Q_\varphi(p_{qm}) &= \exp(\eta + \sigma \cdot \Phi^{-1}(\varphi)) \cdot \exp(\gamma_1) - \exp(\eta + \sigma \cdot \Phi^{-1}(\varphi)) \\ &= \exp(\eta + \sigma \cdot \Phi^{-1}(\varphi)) \cdot (\exp(\gamma_1) - 1)\,. \end{aligned}$$

For the conditional quantiles of the total prices we get:

$$\begin{aligned} Q_\varphi(p) &= area \cdot \exp(\eta + \sigma \cdot \Phi^{-1}(\varphi)) \\ &= area \cdot \exp\left(f_1(area)\right) \ldots \exp\left(f_5(muni)\right) \exp(\gamma_1 x_1) \ldots \exp(\gamma_p x_p) \exp(\sigma \cdot \Phi^{-1}(\varphi)). \end{aligned}$$

So, the quantiles of the total price change multiplicatively with changes in values of covariates too. If for example covariate x_1 changes by one unit, the quantiles of the total price change by

$$\Delta Q_\varphi(p) = area \cdot \exp(\eta + \sigma \cdot \Phi^{-1}(\varphi)) \cdot (\exp(\gamma_1) - 1)\,,$$

making the change again proportional to the floor area. Similarly, if covariate z (representing a nonlinear effect) changes by one unit, both the conditional quantiles of prices per square meter and the conditional quantiles of total prices multiplicatively change by the factor $\exp\left(\mathrm{d}f(z)\right)$, since

$$\exp\left(f(z+1)\right) = \exp\left(f(z+1) - f(z) + f(z)\right) = \exp\left(\mathrm{d}f(z)\right) \exp\left(f(z)\right).$$

Therefore, the change in the quantiles of total prices caused by a change in any covariate again is proportional to the floor area of the building.
For the conditional mean of the house price per square meter

$$\mathbb{E}(p_{qm}) = \exp(\eta + \sigma^2/2)$$

we get analog results.

3.2 GAMLSS

STAR models (see Section 3.1) estimate the conditional mean of a response variable whose distribution belongs to an exponential family. A more flexible approach is given by generalized additive models for location, scale and shape (GAMLSS, Rigby and Stasinopoulos 2005). On the one hand, the class of distributions that can be estimated with GAMLSS is not restricted to the exponential family. On the other hand, GAMLSS allow to model not only the conditional mean of the response variable but the whole set of distribution parameters, i.e. all parameters for the location, scale and shape of a distribution can be related to a set of predictor variables, which of course may vary between the different parameters.
For most distribution families a location parameter $\boldsymbol{\mu}$, a scale parameter $\boldsymbol{\sigma}$ (or $\boldsymbol{\sigma}^2$) and a maximum of two shape parameters $\boldsymbol{\nu}$ and $\boldsymbol{\tau}$ are sufficient to fully characterize the respective distribution.

Using response functions $h_1, \ldots, h_4$ each of these parameters can be linked to a structured additive predictor

$$\begin{aligned}
\boldsymbol{\mu} &= h_1(\boldsymbol{\eta}_1) &= h_1(\mathbf{Z}_{11}\boldsymbol{\beta}_{11} + \ldots + \mathbf{Z}_{q1}\boldsymbol{\beta}_{q1} + \mathbf{X}_1\boldsymbol{\gamma}_1), \\
\boldsymbol{\sigma} &= h_2(\boldsymbol{\eta}_2) &= h_2(\mathbf{Z}_{12}\boldsymbol{\beta}_{12} + \ldots + \mathbf{Z}_{q2}\boldsymbol{\beta}_{q2} + \mathbf{X}_2\boldsymbol{\gamma}_2), \\
\boldsymbol{\nu} &= h_3(\boldsymbol{\eta}_3) &= h_3(\mathbf{Z}_{13}\boldsymbol{\beta}_{13} + \ldots + \mathbf{Z}_{q3}\boldsymbol{\beta}_{q3} + \mathbf{X}_3\boldsymbol{\gamma}_3), \\
\boldsymbol{\tau} &= h_4(\boldsymbol{\eta}_4) &= h_4(\mathbf{Z}_{14}\boldsymbol{\beta}_{14} + \ldots + \mathbf{Z}_{q4}\boldsymbol{\beta}_{q4} + \mathbf{X}_4\boldsymbol{\gamma}_4).
\end{aligned} \tag{6}$$

Of course, each of these predictors may have a hierarchical structure. Thus, GAMLSS can be considered as a generalization of the STAR model (3), where we restricted the focus to the location parameter μ. GAMLSS and STAR models therefore can be subsumed to structured additive distributional regression models, see Klein et al. (2013) for details.

Usually, the response functions are chosen to ensure appropriate restrictions on the parameter spaces. We use, for example, the exponential function to ensure positivity of the scale parameter, i.e. $\boldsymbol{\sigma} = \exp(\boldsymbol{\eta}_2)$.

(Log-)Normal distribution

In Section 3.1 we assumed house prices to be (log-)normally distributed and modeled their conditional mean assuming a homoscedastic variance σ^2. However, as already pointed out in Fahrmeir et al. 2004, not only the mean but also the variance of the response may depend on covariates when modeling real estate data. Thus, we consider a GAMLSS with

$$\begin{aligned}
\boldsymbol{\mu} &= h_1(\boldsymbol{\eta}_1) &= \boldsymbol{\eta}_1, \\
\boldsymbol{\sigma}^2 &= h_2(\boldsymbol{\eta}_2) &= \exp(\boldsymbol{\eta}_2),
\end{aligned}$$

and for both predictors set up the same four level hierarchical STAR model as in (3).

Analogous to Section 3.1 choosing as response either the house price per square meter p_{qm} (which we will then call the *HetGaussian model*, indicating a Gaussian model with heteroscedastic variance) or the logged house price per square meter lnp_{qm} (*HetLoggaussian model*) allows for modeling a normal or a lognormal distribution of the house price per square meter. Accordingly, we receive the conditional mean and the conditional quantiles of the house price per square meter by

$$\begin{aligned}
\mathbb{E}(p_{qm_i}) &= \mu_i, \\
Q_\varphi(p_{qm_i}) &= \mu_i + \sigma_i \cdot \Phi^{-1}(\varphi)
\end{aligned}$$

for the normal distribution and

$$\begin{aligned}
\mathbb{E}(p_{qm_i}) &= \exp\left(\mu_i + \sigma_i^2/2\right), \\
Q_\varphi(p_{qm_i}) &= \exp\left(\mu_i + \sigma_i \cdot \Phi^{-1}(\varphi)\right)
\end{aligned}$$

in the case of a lognormal distribution.

Gamma distribution

In Section 2 we have already mentioned that the mode of the observations is not that pronounced than we would expect from a theoretical lognormal distribution. Thus, we additionally consider another two parameter distribution that is more flexible than the lognormal distribution: the gamma distribution with mean parameter $\mu > 0$ and shape parameter $\sigma > 0$. The probability density function is given by

$$f(y_i|\mu_i, \sigma_i) = \left(\frac{\sigma_i}{\mu_i}\right)^{\sigma_i} \cdot \frac{y_i^{\sigma_i - 1}}{\Gamma(\sigma_i)} \cdot \exp\left(-\frac{\sigma_i}{\mu_i} \cdot y_i\right),$$

with $\Gamma(x) = \int_0^\infty u^{x-1} \exp(-u) \mathrm{d}u$ for $x > 0$ being the gamma function. The mean of the gamma distribution corresponds to μ, the variance is given by μ^2/σ. As we can see from Figure 3 the gamma distribution seems to better capture the mode compared to the lognormal distribution.

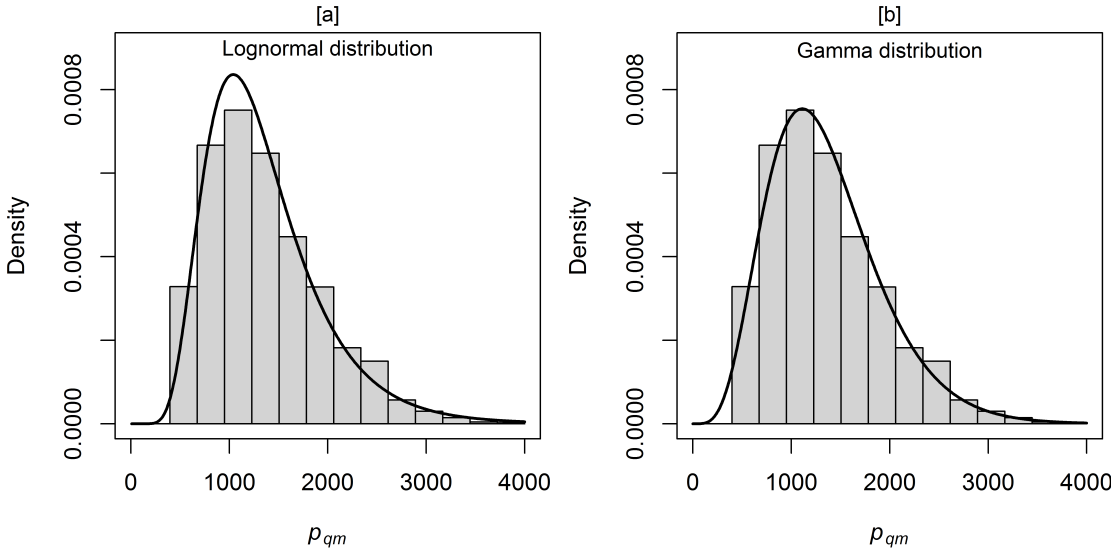

Figure 3: *Histogram of the dependent variable p_{qm} together with the densities of a lognormal* [a] *and a gamma distribution* [b]*, respectively. The mean* 1354.04 *and the variance* $328,858.02$ *determining these densities correspond to the empirical values of the data.*

Setting up a GAMLSS both the mean and the shape parameter are linked to a semiparametric regression predictor – each of them modeled according to (3) – via the exponential function due to the positivity constraints:

$$\begin{aligned} \boldsymbol{\mu} &= h_1(\boldsymbol{\eta}_1) = \exp(\boldsymbol{\eta}_1), \\ \boldsymbol{\sigma} &= h_2(\boldsymbol{\eta}_2) = \exp(\boldsymbol{\eta}_2). \end{aligned}$$

We will call this the *Gamma model*. While we get the conditional mean simply by

$$\mathbb{E}(p_{qm_i}) = \mu_i,$$

there doesn't exist a closed form for the conditional quantiles. However, they can be approximated by a numerical algorithm as it is provided, for example, by the R-function `qgamma` (R Development Core Team (2013)), which is used in this paper.

3.3 Quantile Regression

STAR models as well as GAMLSS assume a specific parametric probability distribution of the response (like the normal, lognormal or gamma distribution) and model some or all of its parameters in dependence of covariates. Quantile regression, in contrast, is a distribution-free approach, trying to directly model the different quantiles of the response as a function of covariates.
In linear quantile regression, as introduced by Koenker and Bassett (1978), we assume

$$q_{\tau,i} = \beta_{\tau,0} + \beta_{\tau,1} x_{i1} + \ldots + \beta_{\tau,p} x_{ip}$$

where q_τ, for $\tau \in (0,1)$, is the τ-quantile of the response distribution. Estimation of the quantile-specific regression coefficients $\boldsymbol{\beta}_\tau$ relies on minimizing the asymmetrically weighted error (AWE) criterion

$$\hat{\boldsymbol{\beta}}_\tau = \arg\min_{\boldsymbol{\beta}_\tau} \left\{ \sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i' \boldsymbol{\beta}_\tau) \right\}, \tag{7}$$

with the loss function ρ_τ defined by

$$\rho_\tau(u) = \begin{cases} u\tau & \text{if } u \geq 0 \\ u(\tau - 1) & \text{if } u < 0, \end{cases}$$

which is also known as the check function. Since there exists no closed form solution for this minimization problem, estimates are typically obtained based on linear programming and modifications of the simplex algorithm, see Koenker and D'Orey (1987) and Koenker (2005) for details. The distribution of the response is implicitly determined by the estimated quantiles q_τ provided that quantiles for a reasonable dense grid of τ-values are estimated.

Generalizations to structured additive predictors are conceptually straightforward. However, estimation is highly challenging and almost impossible for complex hierarchical models, revealing the limits of frequentist quantile regression.

Bayesian quantile regression requires a distributional assumption for the error terms (or equivalently the responses) to be able to set up a likelihood. Following Yu and Moyeed (2001) and Yue and Rue (2011) we will assume independent and identically distributed observations following an asymmetric Laplace distribution with location parameter $\mathbf{x}_i'\boldsymbol{\beta}_\tau$, scale parameter σ^2 and skewness parameter τ,

$$y_i|\boldsymbol{\beta}_\tau, \sigma^2, \tau \overset{\text{iid}}{\sim} \text{ALD}(\mathbf{x}_i'\boldsymbol{\beta}_\tau, \sigma^2, \tau).$$

Then, the density of the responses is given by

$$p(y_i|\boldsymbol{\beta}_\tau, \sigma^2, \tau) = \frac{\tau(1-\tau)}{\sigma^2} \exp\left(-\frac{\rho_\tau(y_i - \mathbf{x}_i'\boldsymbol{\beta}_\tau)}{\sigma^2}\right).$$

Maximizing the corresponding posterior (for fixed σ^2 and τ)

$$\begin{aligned} p(\boldsymbol{\beta}_\tau|\mathbf{y}, \sigma^2, \tau) &\propto \prod_{i=1}^{n} p(y_i|\boldsymbol{\beta}_\tau, \sigma^2) \\ &\propto \exp\left(-\frac{1}{\sigma^2}\sum_{i=1}^{n} \rho_\tau(y_i - \mathbf{x}_i'\boldsymbol{\beta}_\tau)\right) \end{aligned}$$

with respect to $\boldsymbol{\beta}_\tau$ obviously is equivalent to minimizing the AWE criterion (7). However, in contrast to frequentist quantile regression the linear predictor $\eta_{i,\tau} = \mathbf{x}_i'\boldsymbol{\beta}_\tau$ can be replaced by a hierarchical structured additive predictor without any further difficulties.

Since the check function ρ_τ is non-differentiable, inference based on Markov chain Monte Carlo (MCMC) simulations at a first glance seems to be complicated. However, the asymmetric Laplace distribution can be represented as a scaled mixture of normals:

$$Y = \eta + \xi W + \delta Z\sqrt{\sigma^2 W}$$

with $\xi = \frac{1-2\tau}{\tau(1-\tau)}$ and $\delta^2 = \frac{2}{\tau(1-\tau)}$. $W \sim \text{Exp}(\frac{1}{\sigma^2})$ and $Z \sim \mathcal{N}(0,1)$ are independent random variables following an exponential distribution with mean σ^2 and a standard normal distribution, respectively. Thus, using offsets ξW and weights $\delta\sqrt{\sigma^2 W}$ the Bayesian quantile regression problem can be interpreted as a conditionally Gaussian regression model after imputing W as a part of the MCMC sampler, see Yue and Rue (2011) and Waldmann et al. (2013) for details.

3.4 Effect modeling and priors

Effect modeling and priors depend on the covariate or term type. We first describe the general form of basic priors. Then we explain how to model continuous covariate effects and spatial effects using specific design matrices and forms of the basic prior (see Fahrmeir et al. (2013) or Lang et al. (2014) for further covariate types).

General form of basic priors

In a frequentist setting, overfitting of a particular function $\mathbf{f} = \mathbf{Z}\boldsymbol{\beta}$ is avoided by defining a roughness penalty on the regression coefficients, see for instance Fahrmeir et al. (2013) in the context of structured additive regression. The standard are quadratic penalties of the form $\lambda\boldsymbol{\beta}'\mathbf{K}\boldsymbol{\beta}$

where $\mathbf{K}$ is a penalty matrix. The penalty depends on the smoothing parameter λ that governs the amount of smoothness imposed on the function $\mathbf{f}$.

In a Bayesian framework a standard smoothness prior is a (possibly improper) Gaussian prior of the form

$$p(\boldsymbol{\beta}|\tau^2) \propto \left(\frac{1}{\tau^2}\right)^{\text{rk}(\mathbf{K})/2} \exp\left(-\frac{1}{2\tau^2}\boldsymbol{\beta}'\mathbf{K}\boldsymbol{\beta}\right) \cdot I(\mathbf{A}\boldsymbol{\beta} = \mathbf{0}), \tag{8}$$

where $I(\cdot)$ is the indicator function. The key components of the prior are the penalty matrix $\mathbf{K}$, the variance parameter τ_j^2 and the constraint $\mathbf{A}\boldsymbol{\beta} = \mathbf{0}$. Usually the penalty matrix is rank deficient, i.e. $\text{rk}(\mathbf{K}) < K$, resulting in a partially improper prior. The specific structure of $\mathbf{K}$ depends on the covariate type and on prior assumptions about the smoothness of $\mathbf{f}$.

The amount of smoothness is governed by the variance parameter τ^2. A conjugate inverse Gamma prior is employed for τ^2 (as well as for the error variance parameter σ^2 in models with Gaussian responses), i.e. $\tau^2 \sim IG(a, b)$ with small values such as $a = b = 0.001$ for the hyperparameters a and b resulting in an uninformative prior on the log scale. The smoothing parameter λ of the frequentist approach and the variance parameter τ^2 are connected by $\lambda = \frac{\sigma^2}{\tau^2}$.

The term $I(\mathbf{A}\boldsymbol{\beta} = \mathbf{0})$ imposes required identifiability constraints on the parameter vector. A straightforward choice is $\mathbf{A} = (1, \ldots, 1)$, i.e. the regression coefficients are centered around zero.

P-splines

For a continuous covariate z, our basic approach for modeling a smooth function f are P-splines introduced in a frequentist setting by Eilers and Marx (1996) and in a Bayesian version by Lang and Brezger (2004). P-splines assume that the unknown functions can be approximated by a polynomial spline which can be written in terms of a linear combination of B-spline basis functions. Hence, the columns of the design matrix $\mathbf{Z}$ are given by the B-spline basis functions evaluated at the observations z_i. Lang and Brezger (2004) propose to use first or second order random walks as smoothness priors for the regression coefficients, i.e.

$$\beta_k = \beta_{k-1} + u_k, \quad \text{or} \quad \beta_k = 2\beta_{k-1} - \beta_{k-2} + u_k,$$

with Gaussian errors $u_k \sim N(0, \tau^2)$ and diffuse priors $p(\beta_1) \propto const$, or $p(\beta_1)$ and $p(\beta_2) \propto const$, for initial values. This prior is of the form (8) with the penalty given by

$$\sum_{k=d+1}^{K} (\Delta^d \beta_k)^2 = \boldsymbol{\beta}'\mathbf{D}'\mathbf{D}\boldsymbol{\beta} = \boldsymbol{\beta}'\mathbf{K}\boldsymbol{\beta},$$

where Δ^d is the difference operator of order $d = 1$ or $d = 2$ and $\mathbf{D}$ is the corresponding difference matrix.

Markov random fields

The correlated district specific heterogeneity effect $f_{5,6,2}^{mrf}(dist)$ in equation (3) can be modeled by Markov random fields (MRF). Suppose that $z \in \{1, \ldots, K\}$ is the indicator for the district in which a house is located. MRFs define one parameter for every discrete geographical unit (districts in our case), i.e. $f(z) = \beta_z$, and are defined via the conditional distributions of β_z given the parameters β_s of neighboring sites s. We denote the set of neighbors of site z by $N(z)$. Typically sites are assumed to be neighbors if they share a common boundary. MRFs assume that the conditional distribution of β_z given neighboring sites $s \in N(z)$ is Gaussian with

$$\beta_z \,|\, \beta_s, s \neq z \sim N\left(\frac{1}{|N(z)|}\sum_{s \in N(z)} \beta_s, \frac{\tau^2}{|N(z)|}\right),$$

where $|N(z)|$ denotes the number of neighbors of site z.

The joint (prior) distribution of $\boldsymbol{\beta}$ is of the form (8) with penalty matrix $\mathbf{K}$ given by

$$\mathbf{K}[z,s] = \begin{cases} -1 & z \neq s, s \in N(z), \\ 0 & z \neq s, s \notin N(z) \\ |N(z)| & z = s. \end{cases}$$

If a Markov random field is used in the level-1 equation the design matrix $\mathbf{Z}$ is a 0/1 incidence matrix whose entry in the i-th row and k-th column is 1 if the i-th observed house is located in district k and 0 else. In our application the MRF is specified in the level-3 equation to model smooth district specific heterogeneity. In this case the design matrix is the identity matrix, i.e. $\mathbf{Z}_{5,6,2} = \mathbf{I}$.

4 Model selection

In Section 3 we have presented different approaches to model the distribution of house prices. Now, we propose some criteria in order to evaluate the predictive ability of these models and to select a final model. For this purpose, we will determine proper scoring rules (Gneiting and Raftery (2007)) and will consider mean weighted errors.

4.1 Proper scoring rules

Proper scoring rules, as proposed by (Gneiting and Raftery (2007)), are suited to compare the predictive ability of parametric models in terms of probabilistic forecasts based on the predictive distribution and the actual realizations. Thus, we can apply such scoring rules for the five distributional regression models that assume the response to be either normally or lognormally distributed (each with fixed or variable variance) or gamma distributed.

We evaluate the scores for a specific model by means of cross validation, e.g. we randomly divide the data set into five subsets $\Omega_1, \ldots, \Omega_5$ of virtually equal size and estimate the model based on four of those subsets. For the remaining subset, without loss of generality $\Omega_1 = \{y_1, \ldots, y_R\}$, we derive the predictive distributions with densities $f_1, \ldots, f_R$ based on the predictive parameters $\mu_1, \ldots, \mu_R$ and $\sigma_1, \ldots, \sigma_R$ (or $\sigma_1^2, \ldots, \sigma_R^2$). A proper scoring rule S then leads to a score S_{Ω_1} for this subset by summing up the individual contributions

$$S_{\Omega_1} = \frac{1}{R} \sum_{r=1}^{R} S(f_r, y_r).$$

We receive the conclusive score $\mathcal{S}$ as the average score of the five subsets

$$\mathcal{S} = \frac{1}{5} \sum_{i=1}^{5} S_{\Omega_i}.$$

Following Gneiting and Raftery (2007), we consider the logarithmic score (LogS)

$$S(f_r, y_r) = \log\left(f_r(y_r)\right),$$

the quadratic score (QuadS)

$$S(f_r, y_r) = 2f_r(y_r) - ||f_r||_2^2,$$

with $||f_r||_2^2 = \int f_r(\omega)^2 \mathrm{d}\omega$, dominated by the Lebesgue measure, the spherical score (SpherS)

$$S(f_r, y_r) = \frac{f_r(y_r)}{||f_r||_2},$$

as well as the continuous ranked probability score (CRPS)

$$S(F_r, y_r) = -\int_{-\infty}^{\infty} \left(F_r(x) - \mathbb{1}_{\{x \geq y_r\}}\right)^2 \mathrm{d}x,$$

with predictive cumulative distribution function $F_r(x) = \int_{-\infty}^{x} f_r(u)\mathrm{d}u$. Since all of these scores are proper, higher scores yield better probabilistic forecasts when comparing different models.

4.2 Mean weighted error

The τ-quantile of a random variable Y corresponds to the value $\hat{y}$ that minimizes the expected loss

$$\hat{y} = \arg\min_y \{\mathbb{E}(\rho_\tau(Y - y)\},$$

see Koenker (2005) for details. We therefore consider an estimator for the expected loss and select the model with the lowest loss as the final model. For this purpose we revert to the cross validation with subsets $\Omega_1, \ldots, \Omega_5$. For a given quantile τ we estimate each model based on four of these subsets (e.g. $\Omega_2, \ldots, \Omega_5$) and calculate predictions $\hat{y}_{\tau,i}$ for the remaining subset Ω_1. Then, the average loss

$$L_{\Omega_1} = \frac{1}{R} \sum_{i=1}^{R} \rho_\tau (y_i - \hat{y}_{\tau,i}),$$

representing a mean weighted error, yields an estimation for the expected loss. The conclusive loss $\mathcal{L}$ is given as the average loss of the five subsets

$$\mathcal{L} = \frac{1}{5} \sum_{i=1}^{5} L_{\Omega_i}.$$

5 Software

The multilevel structured distributional regression models described above can be estimated with the open source package **BayesX** (Brezger et al. (2005)). An R (R Development Core Team (2013)) implementation for such models is provided in the package **BayesR** (Umlauf et al. (2013)) including a fully interactive interface to the **BayesX** engine. In addition, the package contains infrastructure to conveniently specify complicated multilevel formulas for multiple parameters adopting the usual R "*look & feel*" of model fitting functions. Furthermore, a variety of visualization tools are implemented to explore the estimated functions and perform model specification analysis. A number of extractor functions applied by the common R user, such as `summary()`, `fitted()`, `predict()`, etc., are implemented with multiple parameter support. Model selection can be based on quantile `residuals()`, the `DIC()` and proper scoring rules.
In the following, we exemplify the usage of the software estimating multilevel STAR models with predictors specified in (3). We first load the required packages and data sets.

```
R> library("BayesR")
R> library("spdep")
R> load("HousePrice.rda")
R> load("DistrictsBnd.rda")
```

The file `HousePrice.rda` contains a `data.frame` with the covariates specified in Section 2. The file `DistrictsBnd.rda` contains a boundary map object `DistrictsBnd` that is used to compute the necessary neighborhood structure for estimating the level-3 correlated spatial effect of the districts in Austria. After transforming the class "`bnd`" object to an object of class "`SpatialPolygons`" with

```
R> DistrictsSp <- bnd2sp(DistrictsBnd)
```

the final neighborhood object `DistrictsNb`, which is used for fitting the model, can be generated by

```
R> DistrictsNb <- poly2nb(DistrictsSp)
```

Here, districts are identified as neighbors if they share a common border, but different neighborhood structures can be employed, see e.g. function `dnearneigh()` or `tri2nb()` in package **spdep** (Bivand (2014)). The base model formula that can be used for all parameters is specified as a `list()` by

```
R> f <- list(
+    ## Level 1
+    lnp_qm ~ -1 + heat_o2 + heat_o3 + heat_neu1 + heat_neu2 +
+      bath_o1 + bath_o3 + bath_neu1 + bath_neu2 + garage_1 +
+      garage_2 + marker + attic_dum + cellar_dum + terr_dum +
+      sx(c_area) + sx(c_area_plot) + sx(c_age) + sx(c_time_ind) +
+      sx(municipal, bs = "re"),
+
+    ## Level 2
+    municipal ~ -1 + sx(c_pp_ind) + sx(c_pp_ind) + sx(c_ln_educ) + sx(c_age_ind) +
+      sx(c_comm) + sx(c_ln_dens) + sx(district, bs = "re"),
+
+    ## Level 3
+    district ~ -1 + sx(c_wko_ind) + sx(district, bs = "mrf", map = DistrictsNb) +
+      sx(county, bs = "re"),
+
+    ## Level 4
+    county ~ 1
+  )
```

where the (possibly) nonlinear smooth terms are per default set up using P-splines within the smooth term constructor function `sx()`. The spatially correlated effect is specified by changing the basis type argument of `sx()` to `bs = "mrf"` and providing the neighborhood object to argument `map`. The random effects of the municipals, districts and counties are specified with `bs = "re"`. To estimate a GAMLSS model using the normal distribution with the same formulas for the mean and the variance parameter a named `list()` needs to be created by

```
R> f1 <- list("mu" = f, "sigma2" = f)
```

Hence, each list entry represents one formula object for one parameter of the distribution that is used for modeling. The model is then fitted using the **BayesX** engine by typing

```
R> b1 <- bayesr(f1, family = gaussian2, data = HousePrice, engine = "BayesX")
```

A Gamma model using the untransformed prices is specified in a similar way, i.e., only the base model formula needs to be slightly adapted by exchanging the response

```
R> f2 <- f
R> f2 <- update(f2[[1]], p_qm ~ .)
```

as well as the final formula

```
R> f2 <- list("mu" = f2, "sigma" = f2)
```

since the variance parameter is named `sigma` instead of `sigma2` using the Gamma family object. The model is estimated by

```
R> b2 <- bayesr(f1, family = gamma, data = HousePrice, engine = "BayesX")
```

Using the **BayesX** estimation engine, we can also specify a quantile regression model with the base model formula

```
R> b3 <- bayesr(f, family = quant, data = HousePrice, engine = "BayesX")
```

which estimates a multilevel model for the 50% quantile per default. Model summaries can then be printed by typing e.g.

```
R> summary(b1)
```

which returns the estimation results for all levels and parameters. The estimated smooth and random effects, for e.g. the `mu` parameter of the normal distribution model can be plotted with

```
R> plot(b, model="mu")
```

per default all parameters are plotted. Inspecting the model fit using quantile residuals is possible using the plot method

```
R> plot(b, which = "qq-resid")
```

The resulting scores can be extracted with

```
R> score(b1)
R> score(b2)
```

At the time of writing the **BayesR** software project is still work and progress. A detailed description will be provided, the sources are available at

`https://R-Forge.R-project.org/projects/bayesr/`

the package can be installed within R by typing

```
R> install.packages("BayesR", repos="http://R-Forge.R-project.org")
```

6 Results

We now present the estimation results for the models described in Section 3. The results are based on a final MCMC run with 120,000 iterations and a burn in period of 20,000 iterations. We stored every 100th iteration in order to obtain a sample of 1,000 practically independent draws from the posterior. Computing times for the MCMC sampler ranged between 2 1/2 minutes for the Gaussian model and 55 minutes for the Gamma model on a modern desktop computer (Intel Core i7-3740 Quad-Core, 2.7GHz). Note that no more than 32,000 iterations are typically enough in preliminary MCMC runs to obtain sufficiently exact estimation results. However, we used the comparably large number of iterations in the final run to be absolutely sure about the precision of estimates.

In Section 6.1 we focus on the mean of the house price per square meter. We compare the results from the five parametric models and identify the best of these models with respect to probabilistic forecasts. Afterwards, we will focus on different quantiles of the house price per square meter and compare the results of the selected parametric model with those of the quantile regression (Section 6.2) in order to find a final model.

6.1 Expected value

Structural covariates

Figure 4 shows the effects of the structural continuous covariates. In order to get an impression of the magnitude of effects and make the results comparable, we hold the other continuous structural covariates constant at mean level of attributes and the categorical variables at their mode level and we evaluate all neighborhood covariates and spatial effects at the mode of the municipals (which we will call the *average effect*). If necessary, we additionally transform the functions to natural units (prices in Euro per square meter). Since the effects are quite different in magnitude, we do not show them on the same scale.

In panel [a], the effect of the floor area (variable *area*) is shown. For all models, we find a monotonically decreasing and very pronounced effect of additional floor area on prices per square meter, which is in line with the law of diminishing marginal utility. However, the decreasing effect weakens as the floor area becomes larger. While the results of the skewed distributions (Loggaussian, Lognormal and Gamma model) are virtually the same and cover a range of up to 1,730 Euro, in the Gaussian and the HetGaussian model the effect only accounts for a variation of 1,230 Euro and 1,040 Euro, respectively.

Additional plot area (*area_plot*, panel [b]) yields higher prices per square meter of floor area with the effect becoming weaker as plot area increases. For very large plots, the effect even seems to reverse. However, the data only include very few observations with plot areas larger than 1,300 square meters, leading to wide credible intervals in this area (not shown in this figure). Again, the

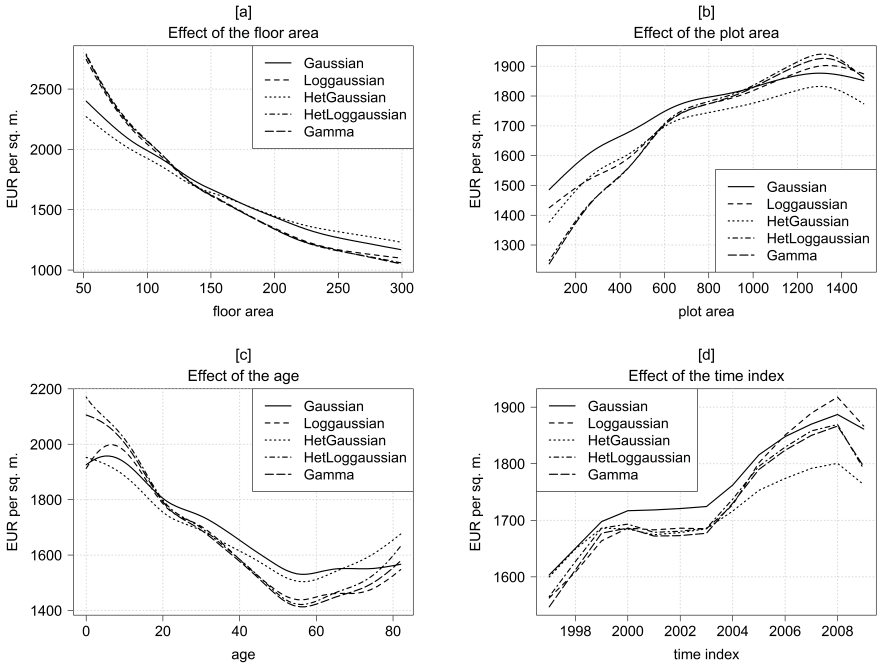

Figure 4: *Effects of the continuous structural covariates.* [a] *Effect of the floor area;* [b] *Effect of the plot area;* [c] *Effect of the age of the building;* [d] *Effect of the time index*

Gaussian and the HetGaussian model yield an effect that is not that pronounced compared to the other models. The results of the Loggaussian model deviate as well from those of the Lognormal and Gamma model, which are again very similar. In total, house prices per square meter change by about 390 Euro (Gaussian model) to 690 Euro (HetLoggaussian model) over the domain of the plot area.

The effect of the *age* of the building, shown in panel [c], can be considered as the rate of depreciation of single family homes. Thus, the initial increase up to an age of 7 years in the results of the Gaussian and the Loggaussian model seems quite unlikely, whereas the more or less linear depreciation (until an age of about 55 years) in the other models is in line with our expectations. However, the decline is much more pronounced in the Lognormal and the Gamma model than in the HetGaussian model. In all models, the effect slightly reverse for old buildings (again with wide credible intervals due to a small number of observations in this area). The age of the house covers a range of about 425 Euro (Gaussian model) to 750 Euro (HetLoggaussian model) per square meter.

The effect of the *time index* (panel [d]) shows the quality controlled development of house prices over time. After a moderate increase from 1997 to 2000, prices almost stay constant until 2003 with similar results for almost all models. Only the Gaussian model predicts slightly higher prices in this period. After 2003 the prices rise until 2008 in a considerably different extent within the five models: While the increase is less marked for the Gaussian and the HetGaussian model, it is more pronounced especially for the Loggaussian model. In the last year of the observation period prices consistently decrease, indicating the effect of the economic crisis of 2008/2009. In total, the time index accounts for variation in a range of 200 Euro (Normal model) to 350 Euro (Loggaussian model).

Spatial covariates

The effects of the spatial covariates are shown in Figure 5, again on the natural scale of prices per square meter. The effect of the purchase power index (*pp_ind*), shown in panel [a], is highly positive between 80 and 130 index points with similar results for the skewed models and slightly weaker effects for the Gaussian and the HetGaussian model. The negative effects for low and high values of the purchase power index is unexpected, but may result from very few municipals with such extreme index points (wide credible intervals). In total, house prices per square meter change

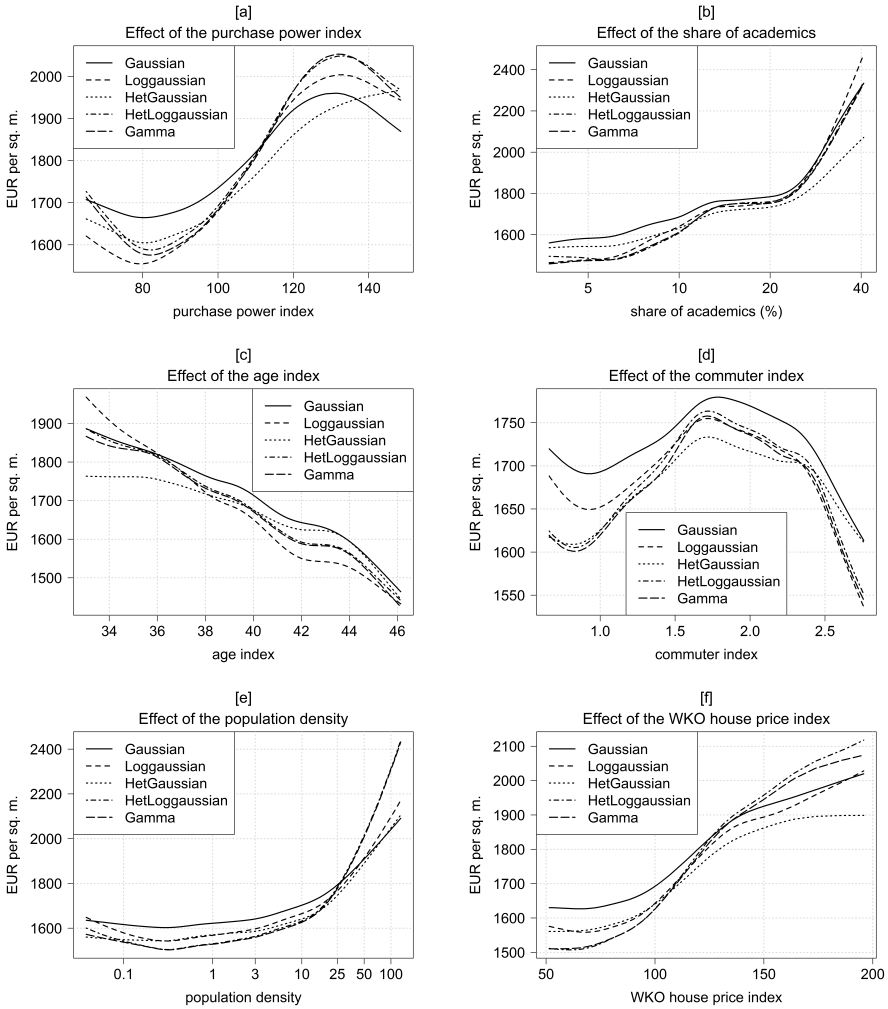

Figure 5: *Effects of the spatial covariates.* [a] *Effect of the purchase power index;* [b] *Effect of the share of academics;* [c] *Effect of the age index;* [d] *Effect of the commuter index;* [e] *Effect of the population density;* [f] *Effect of the WKO house price index*

by about 300 Euro (Gaussian model) to 475 Euro (Gamma model).

Although the share of academics (*ln_educ*, panel [b]) enters the equation logarithmically (see Section 2.2), it is displayed in natural values. The effect is clearly positive, with a pronounced increase starting at a share of approximately 25%. The difference between the five models is rather small, only for municipalities with a very high amount of academics the HetGaussian model seems to underestimate the effect compared to the other models. The share of academics accounts for a variation of up to 1,000 Euro (Loggaussian model).

The effect of the age index (*age_ind*, panel [c]) is more or less linear for all models with the highest slope in the Loggaussian model (bandwidth of about 535 Euro) and the smallest one in the HetGaussian model (bandwidth 320 Euro). The negative direction is in line with our expectations and can be interpreted as a decreasing attractiveness of municipalities that exhibit an excess of age.

The commuter index (*comm*, displayed in panel [d]) has the weakest effect of all continuous covariates with a variation of not more than 220 Euro. The highest values are realized at about 1.7 index points, where the number of people commuting from the municipality almost equals the number of those who enter it. Regions with an unbalanced ratio between commuters from and into the municipality both have slightly lower house prices. Two models seem to overestimate prices compared to the other models, the Gaussian model in the whole range of the index, the Loggaussian model only for low index points.

The results of the population density *ln_dens* (panel [e]) show almost no effect for sparsely inhabited areas and a highly positive effect for densely populated regions. For the latter ones, we can find a considerable difference of up to 350 Euro between the results of the Lognormal and the Gamma model on the one hand and the other models on the other hand. In total, this effect accounts for a variation between 490 Euro (Gaussian model) and 935 Euro (HetLoggaussian model).

Finally, the results for the house price index (*wko_ind*, the only covariate on the district level) are shown in panel [f]. The effect is clearly positive, which is in line with our expectations. However, for index values lower than 90 and higher than 140 the effect is consistently weaker. The bandwidth ranges between 340 Euro (Loggaussian model) and 610 Euro (HetLoggaussian model).

The spatial heterogeneity over Austria, caused by the previous effects, strikingly can be visualized by colored maps. However, for the sake of simplicity, we do not show such maps for all parametric models here, but will do it in the next section when comparing the results of the best parametric model with those of the quantile regression.

Predictive ability

Although the main shapes of the effects of the continuous covariates are very similar for all five parametric models, we have seen several differences when analyzing the effects in detail. For most covariates the results of the skewed models (Loggaussian, Lognormal and Gamma model) considerably differ from those of the models that are based on a normal distribution (Gaussian and HetGaussian model). The empirical distribution of the data (see Section 2) here suggests the skewed models to be superior. Within the skewed models we identified an unexpected behavior of the Loggaussian model for the age of the building as well as further deviations for selected covariates from the other models. Finally, the results of the Lognormal and the Gamma model are very similar for all covariates.

We now compare the predictive ability of these models by means of proper scoring rules (see Section 4.1). Table 1 shows the average logarithmic, quadratic and spherical scores as well as the continuous ranked probability score for the five models. As expected from the considerations above the Lognormal and the Gamma model seem to make the best probabilistic forecasts with the Gamma model to be slightly superior in all scores.

Model	Logarithmic score	Quadratic score	Spherical score	CRPS
Gaussian	-0.5524	0.7234	0.8509	-0.2309
Normal	-0.4635	0.8090	0.8840	-0.2241
Loggaussian	-0.4670	0.8077	0.8860	-0.2250
Lognormal	-0.4358	0.8334	0.8951	-0.2225
Gamma	**-0.4199**	**0.8441**	**0.9020**	**-0.2208**

Table 1: *Comparison of average score contributions of the parametric models obtained from a five-fold cross validation*

6.2 Quantiles

In Section 6.1 we have identified the Gamma model to be the best parametric model with respect to probabilistic forecasts. We now compare the results of this model with those from quantile regression. We consider seven different quantiles (5%-, 10%-, 30%-, 50%-, 70%-, 90%-, 95%-quantile) in order to get a good overview of the whole distribution of house prices. We restrict the discussion to covariates with notable results or major differences between the models. The plots of the remaining covariate effects can be found in appendix 6.

Structural covariates

Figure 6 shows the effects of the floor area and the age of the building with the other covariates again being hold constant at the average effect. In order to facilitate comparability we show the effects of a particular covariate on the same scale for both the Gamma model and the quantile regression.

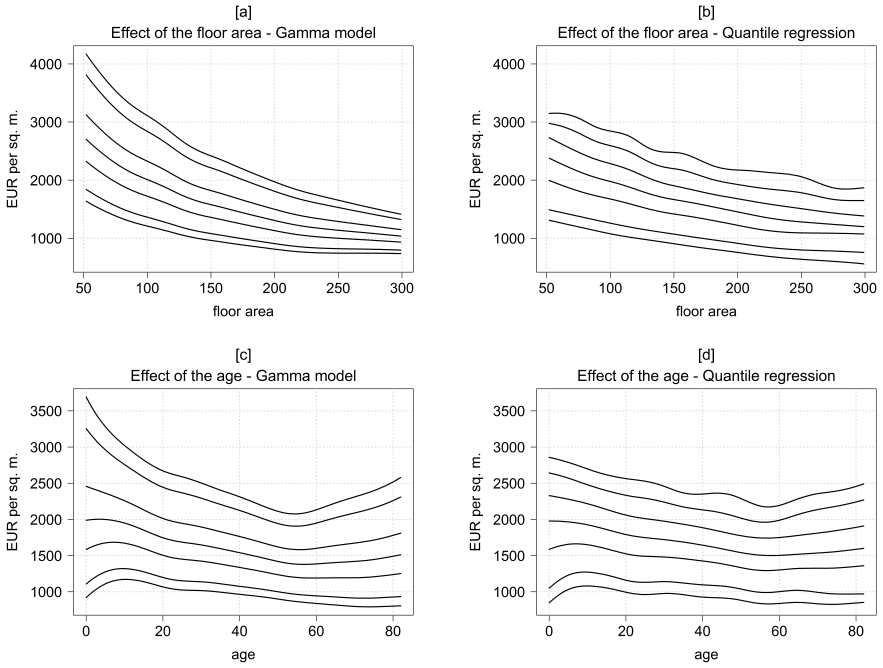

Figure 6: *Effects of selected continuous structural covariates of the Gamma model (left column) and the quantile regression (right column).* [a], [b] *Effect of the floor area;* [c], [d] *Effect of the age of the building*

While the quantiles of the floor area have a convex shape in the Gamma model (panel [a]), the effects are more or less linear in the quantile regression, displayed in panel [b]. Furthermore, the variance considerably differs in the Gamma model with a large variance for small houses and vice versa. The bandwidth between the 95%-quantile and the 5%-quantile decreases from 2,530 Euro for houses with floor areas of 50 square meters to about 675 Euro for buildings with 300 square meters. In contrast, the variance is substantially lower in the quantile regression with the bandwidth only varying between 1,840 Euro (small houses) and 1,310 Euro (large houses).

The results of the age are very similar between the Gamma model (panel [c]) and the quantile regression (panel [d]) for the lower quantiles up to the median. For the 5%- and the 10%-quantile the effects seem to be almost linear and slightly decreasing over the whole range of the age (at least for an age exceeding seven years). The 30%- and the 50%-quantiles are linearly decreasing up to an age of 55 years and almost constant or slightly increasing afterwards. The upper quantiles, in contrast, considerably differ between the two models: In the quantile regression, the effects still decline linearly up to 55 years and reverse hereafter. In the Gamma model the effects more and more tend to a quadratic shape. Thus, the range of house prices is much higher in the Gamma model (up to 2,770 Euro) than in the quantile regression (only 2,000 Euro) especially for new buildings. Finally, the different functional forms of the individual quantiles illustrate a great advantage of using distributional or quantile regression instead of the ordinary mean regression: In the latter one, the quantiles would have the same marginal effect, only being shifted according to the overall variance – a restriction that couldn't be justified economically.

Spatial covariates

Figure 7 shows the estimated quantiles of selected spatial covariates again at the average effect. The 5%- and the 10%-quantiles of the purchase power index (panels [a] and [b]) are almost linearly increasing both for the Gamma model and the quantile regression. For the remaining quantiles we find a tendency towards undulating effects that are more pronounced in the Gamma model. However, as already discussed the decreasing effect for low and high index points shouldn't be overstated due to only a very small number of municipalities with such purchase power indices.

The quantiles of the population density show that over the whole distribution of prices there is

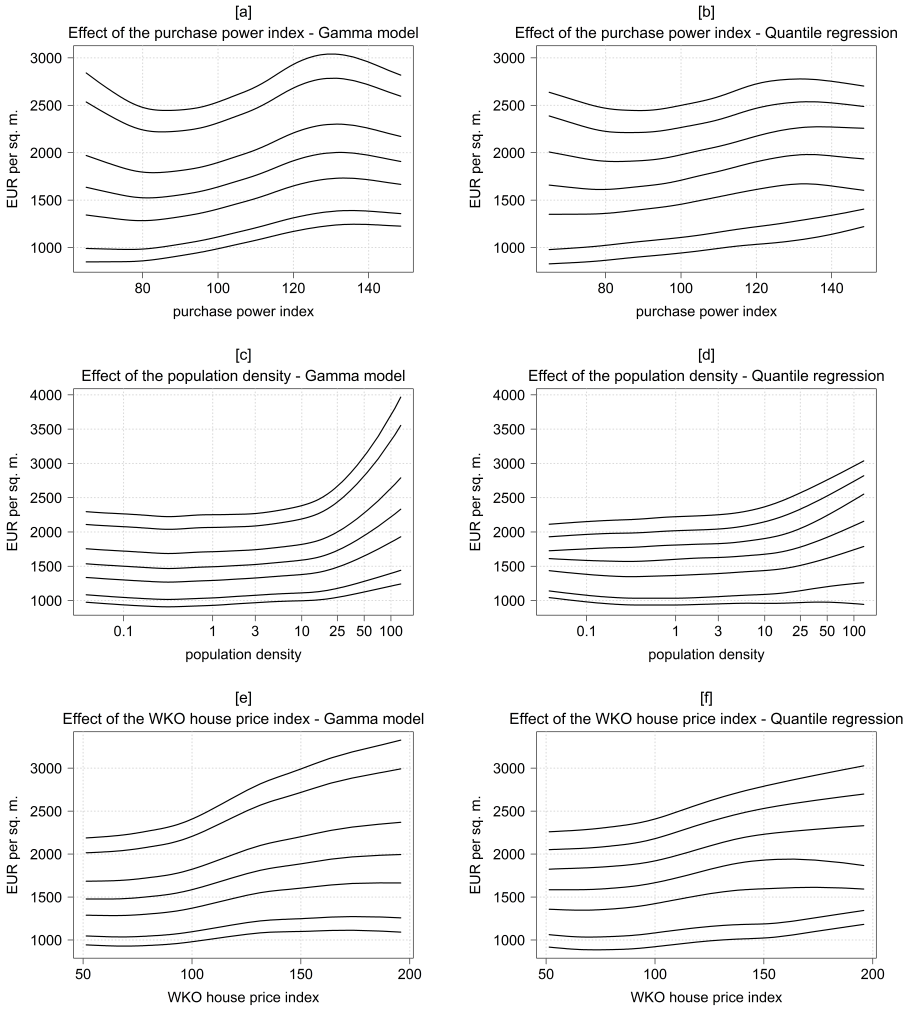

Figure 7: *Effects of selected spatial covariates of the Gamma model (left column) and the quantile regression (right column).* [a], [b] *Effect of the purchase power index;* [c], [d] *Effect of the population density;* [e], [f] *Effect of the WKO house price index*

almost no effect for sparsely inhabited areas, neither in the Gamma model nor in the quantile regression. In contrast, for densely populated areas there is a slightly positive effect for the lower quantiles and an ever-growing positive effect for the upper quantiles. Especially the effects of the 90%- and the 95%-quantile of highly populated areas are much more pronounced in the Gamma model than in the quantile regression leading to a variation of prices between the 5%- and the 95%-quantile of up to 2,725 Euro in the Gamma model and only 2,090 Euro in the quantile regression.

The results of the WKO house price index are quite similar in both models with effects that are more or less linearly increasing for all different quantiles. The slope is rather small for the lower quantiles and somewhat higher for the upper quantiles. For the latter ones the effect is slightly more pronounced in the Gamma model than in the quantile regression.

Distribution of spatial heterogeneity over Austria

The spatial covariates, defined on the municipality- and the district-level, explain spatial heterogeneity to a certain extent, so we call this the *explained* spatial heterogeneity. The remaining i.i.d. spatial random effects $\epsilon_5, \epsilon_{5,6}$ and $\epsilon_{5,6,3}$ as well as the correlated district specific effect $\mathbf{f}^{mrf}_{5,6,2}(dist)$ in (3) account for *unexplained* spatial heterogeneity.

Figure 8 visualizes the distribution of the total spatial heterogeneity over Austria that is composed by the sum of the explained and the unexplained heterogeneity, evaluated again at the average

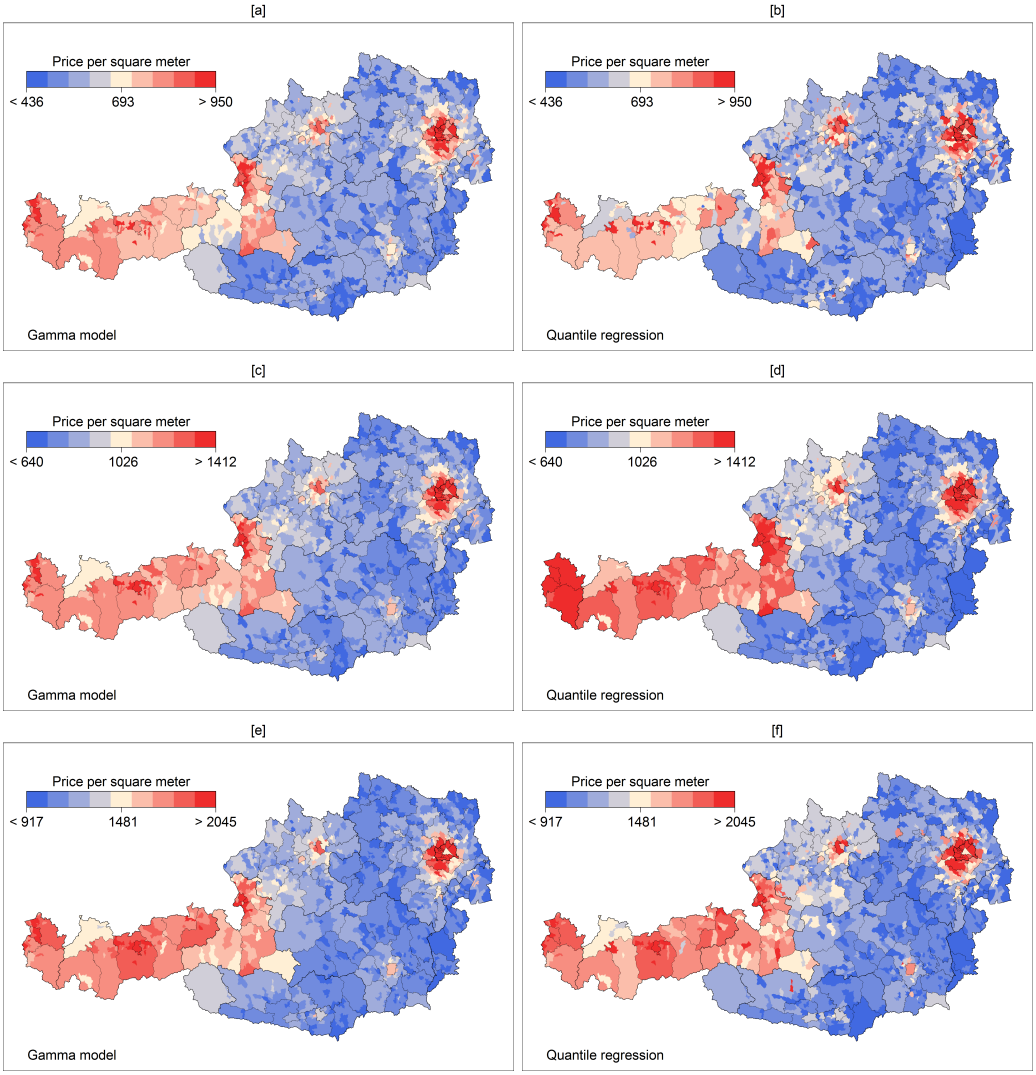

Figure 8: *Distribution of total spatial heterogeneity over Austria.* [a], [c], [e] *Quantiles of the Gamma model;* [b], [d], [f] *Quantile of the quantile regression.*

effect. For the sake of illustration we restrict the analysis to the 10%-, 50%- and 90%-quantiles. We find considerably higher house prices in the western counties of Austria as well as in the city of Linz and the metropolitan area of Vienna. This effect can consistently be observed for all quantiles. Especially for the median, the spatial heterogeneity seems to be more pronounced in the quantile regression compared to the Gamma model.

If we now compare the total spatial heterogeneity with the unexplained spatial heterogeneity, displayed in Figure 9, we can see that the continuous spatial covariates indeed are able to explain a large part of the spatial heterogeneity over Austria, since the variation of the remaining unexplained heterogeneity is considerably lower than the variation of the total spatial heterogeneity. Moreover, the unexplained spatial heterogeneity seems to be more randomly distributed, especially for the 10%-quantile the 90%-quantile in the quantile regression. However, for the median in the quantile regression as well as for all quantiles in the Gamma model, we still find systematically higher prices in the western counties as well as in Vienna and considerably lower prices in Burgenland (far east of Austria) and Carinthia (south of Austria), indicating a spatial effect that is not captured by the covariates involved in the models.

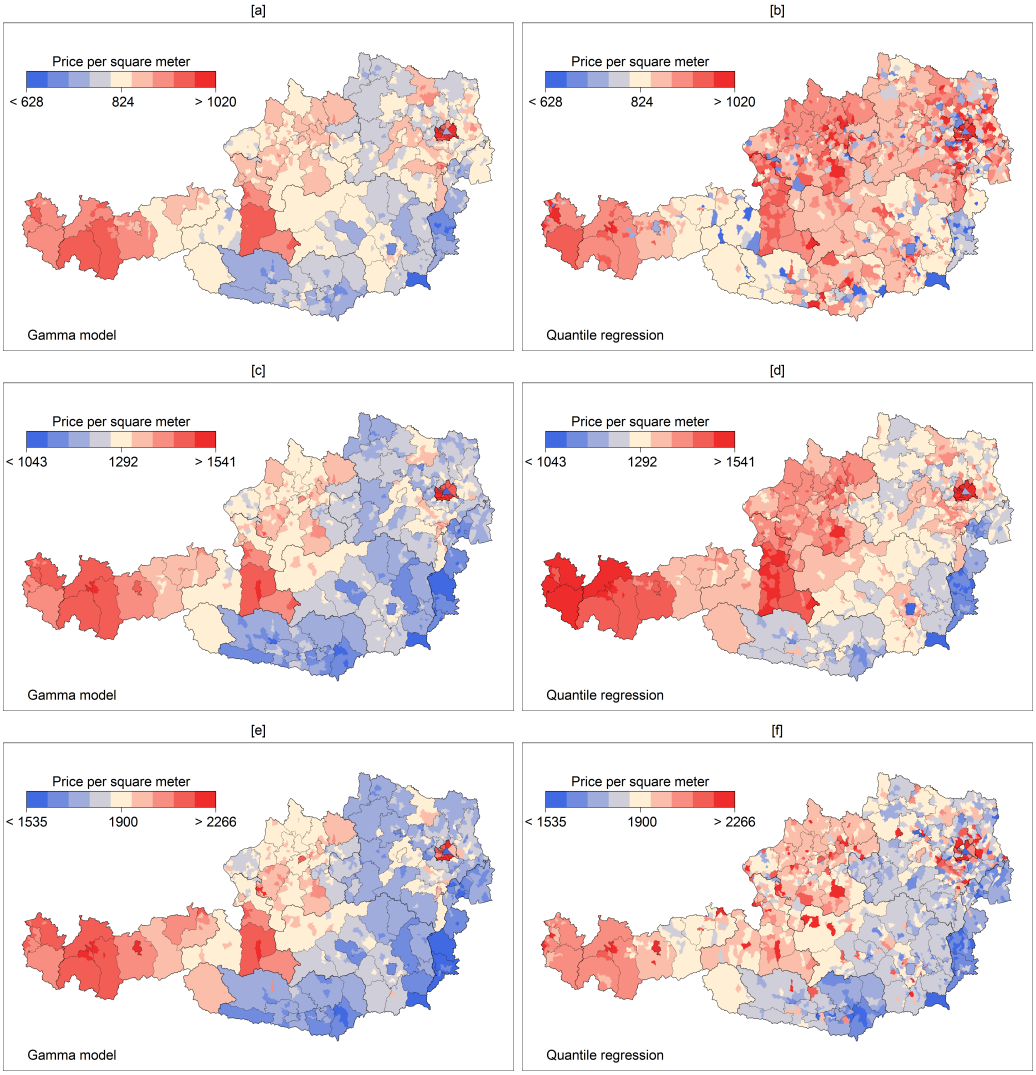

Figure 9: *Distribution of unexplained spatial heterogeneity over Austria.* [a] *Gamma model;* [b] *Quantile regression.*

Predictive ability

The previous results sometimes revealed considerable differences between the Gamma model and the quantile regression. In order to select a final model, we now calculate the mean weighted errors (see Section 4.2) for the different quantiles from a five-fold cross validation.

According to table 2 the Gamma model yields better predictions for all quantiles compared to the quantile regression. At a first glance, this result seems surprising since the mean weighted error basically corresponds to the AWE criterion (7) that the quantile regression tries to minimize. However, there are three possible explanations: First of all, in Bayesian quantile regression the estimation result is given by the posteriori mean which doesn't exactly correspond to the minimization of the AWE criterion (which would be the posteriori mode). Furthermore, the use of smoothness priors penalizes the check-function, leading to estimation results slightly deviating from the actual minimization. Finally, quantile regression tries to minimize the AWE criterion for the observed realizations. The mean weighted error, in contrast, is based on cross validation, meaning that it is calculated for new observations. Thus, the fact that estimation errors can be quite large in quantile regression especially for extreme quantiles, may lead to better predictions for parametric models compared to quantile regression.

Quantile	Model	G1	G2	G3	G4	G5	∅
5%	Gamma Model	36.02	40.33	42.65	37.34	41.82	**39.63**
	Quantile Regression	42.46	47.28	51.83	44.13	49.48	47.04
10%	Gamma Model	63.92	69.31	70.84	64.98	71.20	**68.05**
	Quantile Regression	70.17	74.49	78.89	71.96	76.89	74.48
30%	Gamma Model	135.78	138.99	139.25	133.01	141.77	**137.76**
	Quantile Regression	141.04	142.62	145.98	138.08	143.53	142.25
50%	Gamma Model	160.73	162.75	160.57	154.68	164.99	**160.75**
	Quantile Regression	165.21	165.04	168.82	159.28	166.38	164.95
70%	Gamma Model	144.07	148.11	142.31	136.15	146.51	**143.43**
	Quantile Regression	145.17	146.36	148.32	139.84	145.86	145.11
90%	Gamma Model	80.35	83.46	78.51	75.01	79.99	**79.46**
	Quantile Regression	83.78	84.91	86.30	82.21	84.17	84.27
95%	Gamma Model	52.05	52.99	49.51	46.75	50.53	**50.37**
	Quantile Regression	58.06	58.23	60.17	58.72	58.47	58.73

Table 2: *Mean weighted errors for the different quantiles from a five-fold cross validation with groups G1,. . . ,G5*

7 Conclusion

This paper analyzes the distribution of house prices using Bayesian multilevel structured distributional and quantile regression models. Extending the work of Brunauer et al. (2013) we do not restrict our analyses to the (conditional) mean but additionally concentrate on different (conditional) quantiles of the response in order to summarize its whole distribution. The paper is based on two conceptually different approaches: Distributional regression models, on the one hand, assume a specific parametric probability distribution of the response and model some or all of its parameters in dependence of covariates. Quantile regression, in contrast, directly models the different quantiles of the response as a function of covariates without a specific distribution assumption. Model choice is based on proper scoring rules and mean weighted errors. We identify a model based on the gamma distribution to be most suitable for our data giving new insights into the distribution of house prices. Our findings are of great practical interest, especially with respect to the evaluation of the credit risk of financial institutions that accept real estate as collateral.

The magnitude of the effects considerably differs between the individual covariates. Thus a detailed analysis of confidence bands as well as an automated variable selection for both the particular parameters of the distributional regression models and the individual quantiles of the quantile regression could be a conceivable starting point for further research.

Acknowledgement: This work was supported by funds of the Oesterreichische Nationalbank (Oesterreichische Nationalbank, Anniversary Fund, project number: 15309).

Appendix

A Description of Covariates

Continuous structural covariates

Name	Description [unit]	mean / min. / max.	Exp. Eff.
area	floor area (exc. cellar) [sq. meter]	135 / 44 / 495	+
area_plot	plot space [sq. meter]	742 / 80 / 2500	+
age	age of building [years]	23 / 0 / 82	-
time_index	year of purchase [date]	2005 / 1997 / 2009	o

Categorical structural covariates

Name	Description; categories
cond_house	condition of the house (6 categories); method 1: 1 = (very) good (21.79%), 2 = medium (4.46%), 3 = bad (59.49%); method 2: 4 = (very) good (7.92%), 5 = medium (4.55%), 6 = bad (1.80%)
heat	quality of the heating system (8 categories); method 1: 1 = (very) good (62.46%), 2 = medium (18.85%), 3 = bad (4.43%); method 2: 4 = excellent (4.70%), 5 = very good (4.61%), 6 = good (1.95%), 7 = medium (1.83%), 8 = bad (1.18%)
bath	quality of the bathroom (7 categories); method 1: 1 = (very) good (13.22%), 2 = medium (66.73%), 3 = bad (5.79%); method 2: 4 = very good (7.95%), 5 = good (3.59%), 6 = medium (1.98%), 7 = bad (0.74%)
garage	quality/existence of a garage (3 categories); 1 = high (10.99%), 2 = medium/low (41.23%), 3 = no garage (47.79%)
marker	discrimination between methods (2 categories); 0 = method 1 (85.73%), 1 = method 2 (14.27%)
cellar_dum	existence of a cellar (2 categories); 0 = no cellar (73.23%), 1 = cellar (26.77%)
attic_dum	existence of an attic (2 categories); 0 = no attic (55.87%), 1 = attic (44.13%)
terr_dum	existence of a terrace (2 categories); 0 = no terrace (58.40%), 1 = terrace (41.60%)

Table 3: *Structural attributes of single family homes. The upper part describes continuous covariates and assumptions about the directions of the effects ("+": increasing, "-": decreasing and "o": no strong assumptions), the lower part describes the categorical variables. Covariates cond_house, heat and bath have been collected by two different methods, which makes it necessary to distinguish the respective effects for the two subsamples. Specifically, categories 1,2 and 3 of each of these covariates come from method 1, while the rest of the categories (heat: 4 to 8, bath: 4 to 7 and cond_house: 4 to 6) stems from method 2. Furthermore, a marker discriminating between the two methods of data collection is introduced.*

B Further results

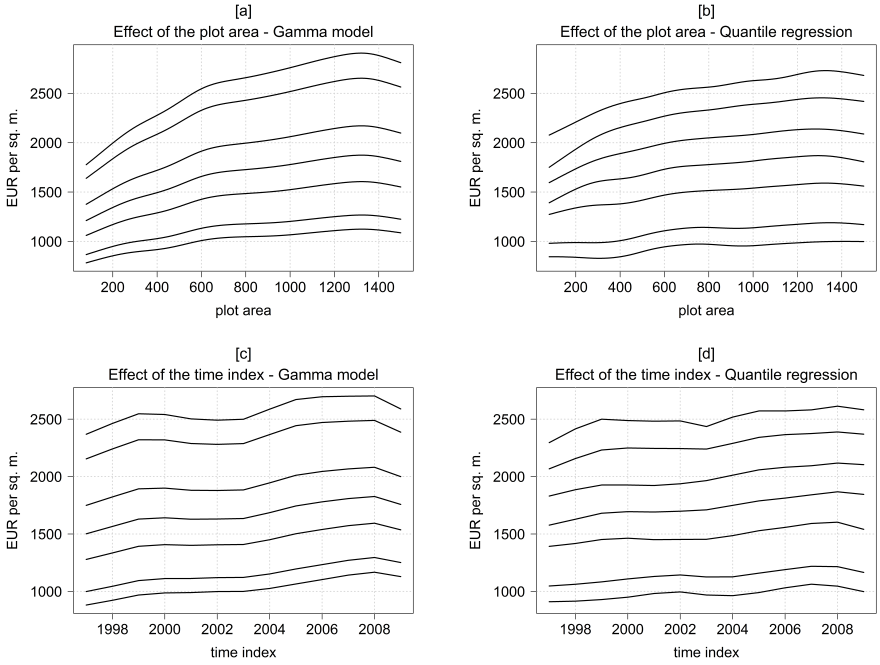

Figure 10: *Effects of the remaining structural covariates of the Gamma model (left column) and the quantile regression (right column).* [a], [b] *Effect of the plot area;* [c], [d] *Effect of the time index*

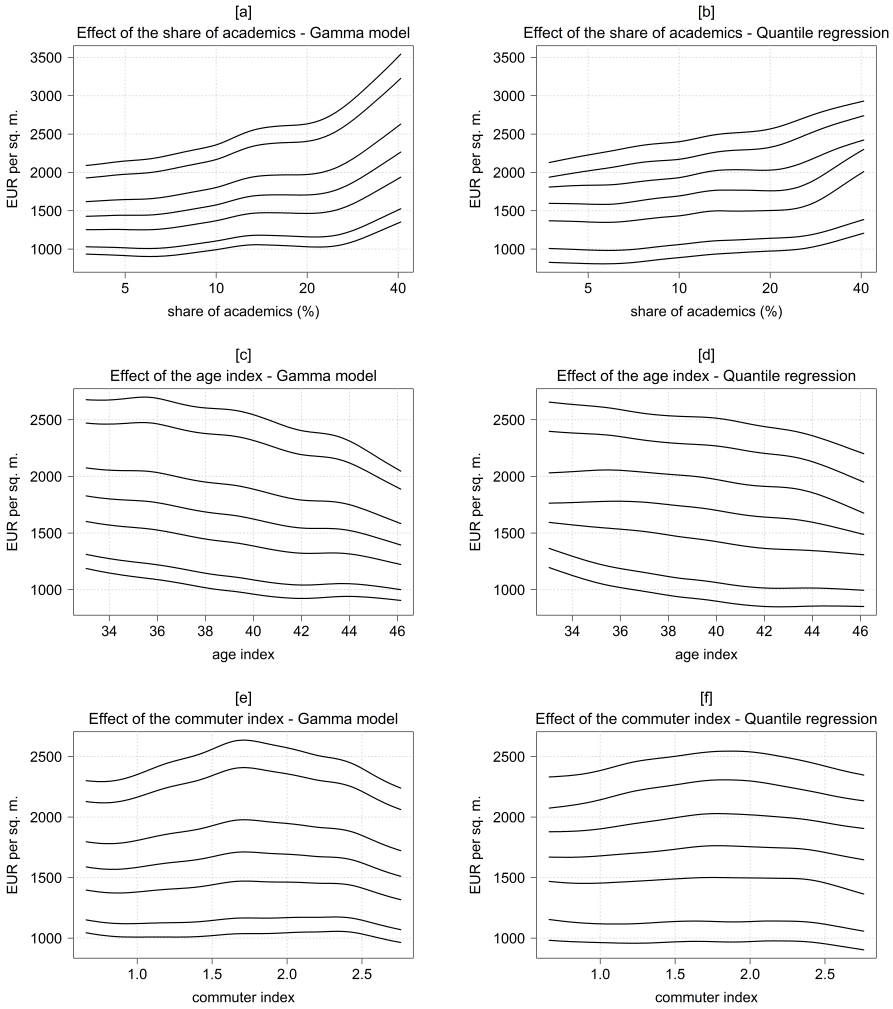

Figure 11: *Effects of the remaining spatial covariates of the Gamma model (left column) and the quantile regression (right column).* [a], [b] *Effect of the share of academics;* [c], [d] *Effect of the age index;* [e], [f] *Effect of the commuter index*

References

Bivand, R. (2014). **spdep***: Spatial dependence: weighting schemes, statistics and models*. R package version 0.5-71.

Brezger, A., T. Kneib, and S. Lang (2005). **BayesX**: Analyzing Bayesian structured additive regression models. *Journal of Statistical Software 14* (11), 1–22.

Brunauer, W. A., S. Lang, and N. Umlauf (2013). Modeling House Prices using Multilevel Structured Additive Regression. *Statistical Modelling 13*, 95–123.

Cohen, J. P. and C. C. Coughlin (2008). Spatial Hedonic Models of Airport Noise, Proximity, and Housing Prices. *Journal of Regional Science 48*, 859–878.

Eilers, P. H. C. and B. D. Marx (1996). Flexible smoothing using b-splines and penalized likelihood. *Statistical Science 11*, 89–121.

Fahrmeir, L., T. Kneib, and S. Lang (2004). Penalized structured additive regression for space-time data: a Bayesian perspective. *Statistica Sinica 14*, 731–761.

Fahrmeir, L., T. Kneib, S. Lang, and B. Marx (2013). *Regression: Models, Methods and Applications*. Springer.

Gelman, A. and J. Hill (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Gneiting, T. and A. E. Raftery (2007). Strictly Proper Scoring Rules, Prediction, and Estimation. *Journal of the American Statistical Association 102*, 359–378.

Haupt, H. (2014). Quantile smoothing spline estimation of urban house price surfaces under conditional price and spatial heterogeneity. To appear in *Advances in Statistical Analysis*.

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

Klein, N., T. Kneib, and S. Lang (2013). Bayesian structured additive distributional regression. Technical report, Department of Statistics, University of Innsbruck. Available at http://eeecon.uibk.ac.at/wopec2/repec/inn/wpaper/2013-23.pdf.

Koenker, R. (2005). *Quantile Regression*. Cambridge University Press, New York.

Koenker, R. and G. Bassett (1978). Regression quantiles. *Econometrica 46*, 33–50.

Koenker, R. and V. D'Orey (1987). Computing regression quantiles. *Journal of the Royal Statistical Society: Series C 36*, 383–393.

Lancaster, K. (1966). A New Approach to Consumer Theory. *Journal of Political Economy 74*, 132–157.

Lang, S. and A. Brezger (2004). Bayesian P-splines. *Journal of Computational and Graphical Statistics 13*, 183–212.

Lang, S., N. Umlauf, P. Wechselberger, K. Harttgen, and T. Kneib (2014). Multilevel structured additive regression. *Statistics and Computing 24*, 223–238.

Malpezzi, S. (2003). Hedonic Pricing Models: A Selective and Applied Review. In T. O'Sullivan and K. Gibb (Eds.), *Housing Economics and Public Policy*, pp. 67–89. Blackwell Science Ltd.

McMillen, D. P. (2008). Changes in the Distribution of House Prices over Time: Structural Characteristics, Neighborhood, or Coefficients? *Journal of Urban Economics 64*, 573–589.

R Development Core Team (2013). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0.

Rigby, R. A. and D. M. Stasinopoulos (2005). Generalized Additive Models for Location, Scale and Shape. *Applied Statistics 54*, 507–554.

Rosen, S. (1974). Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. *Journal of Political Economy 82*, 34–55.

Sheppard, S. (1999). Hedonic analysis of housing markets. In P. C. Cheshire and E. S. Mills (Eds.), *Handbook of Regional and Urban Economics*, Volume 3, pp. 1595–1635. Elsevier Science.

Umlauf, N., N. Klein, A. Zeileis, and S. Lang (2013). *BayesR: Bayesian Regression in R*. R package version 0.1-1.

Waldmann, E., T. Kneib, S. Lang, and Y. Yue (2013). Bayesian Semiparametric Additive Quantile Regression. *Statistical Modelling 13*, 223–252.

Wallace, N. (1996). Hedonic-Based Price Indexes for Housing: Theory, Estimation, and Index Construction. *Federal Reserve Bank of San Francisco Economic Review 3*, 34–48.

Yu, K. and R. A. Moyeed (2001). Bayesian quantile regression. *Statistics & Probability Letters 54*, 437–447.

Yue, Y. and H. Rue (2011). Bayesian inference for additive mixed quantile regression models. *Computational Statistics and Data Analysis 55*, 84–96.

University of Innsbruck - Working Papers in Economics and Statistics
Recent Papers can be accessed on the following webpage:

http://eeecon.uibk.ac.at/wopec/

2014-12 **Alexander Razen, Wolgang Brunauer, Nadja Klein, Thomas Kneib, Stefan Lang, Nikolaus Umlauf:** Statistical risk analysis for real estate collateral valuation using Bayesian distributional and quantile regression

2014-11 **Dennis Dlugosch, Kristian Horn, Mei Wang:** Behavioral determinants of home bias - theory and experiment

2014-10 **Torsten Hothorn, Achim Zeileis:** partykit: A modular toolkit for recursive partytioning in R

2014-09 **Rudi Stracke, Wolfgang Höchtl, Rudolf Kerschbamer, Uwe Sunde:** Incentives and selection in promotion contests: Is it possible to kill two birds with one stone?

2014-08 **Rudi Stracke, Wolfgang Höchtl, Rudolf Kerschbamer, Uwe Sunde:** Optimal prizes in dynamic elimination contests: Theory and experimental evidence

2014-07 **Nikolaos Antonakakis, Max Breitenlechner, Johann Scharler:** How strongly are business cycles and financial cycles linked in the G7 countries?

2014-06 **Burkhard Raunig, Johann Scharler, Friedrich Sindermann:** Do banks lend less in uncertain times?

2014-05 **Julia Auckenthaler, Alexander Kupfer, Rupert Sendlhofer:** The impact of liquidity on inflation-linked bonds: A hypothetical indexed bonds approach

2014-04 **Alice Sanwald, Engelbert Theurl:** What drives out-of pocket health expenditures of private households? - Empirical evidence from the Austrian household budget survey

2014-03 **Tanja Hörtnagl, Rudolf Kerschbamer:** How the value of information shapes the value of commitment or: Why the value of commitment does not vanish

2014-02 **Adrian Beck, Rudolf Kerschbamer, Jianying Qiu, Matthias Sutter:** Car mechanics in the lab - Investigating the behavior of real experts on experimental markets for credence goods

2014-01 **Loukas Balafoutas, Adrian Beck, Rudolf Kerschbamer, Matthias Sutter:** The hidden costs of tax evasion - Collaborative tax evasion in markets for expert services

2013-37 **Reto Stauffer, Georg J. Mayr, Markus Dabernig, Achim Zeileis:** Somewhere over the rainbow: How to make effective use of colors in meteorological visualizations

2013-36 **Hannah Frick, Carolin Strobl, Achim Zeileis:** Rasch mixture models for DIF detection: A comparison of old and new score specifications

2013-35 **Nadja Klein, Thomas Kneib, Stephan Klasen, Stefan Lang:** Bayesian structured additive distributional regression for multivariate responses

2013-34 **Sylvia Kaufmann, Johann Scharler:** Bank-lending standards, loan growth and the business cycle in the Euro area

2013-33 **Ting Wang, Edgar C. Merkle, Achim Zeileis:** Score-based tests of measurement invariance: Use in practice

2013-32 **Jakob W. Messner, Georg J. Mayr, Daniel S. Wilks, Achim Zeileis:** Extending extended logistic regression for ensemble post-processing: Extended vs. separate vs. ordered vs. censored *published in Monthly Weather Review*

2013-31 **Anita Gantner, Kristian Horn, Rudolf Kerschbamer:** Fair division in unanimity bargaining with subjective claims

2013-30 **Anita Gantner, Rudolf Kerschbamer:** Fairness and efficiency in a subjective claims problem

2013-29 **Tanja Hörtnagl, Rudolf Kerschbamer, Rudi Stracke, Uwe Sunde:** Heterogeneity in rent-seeking contests with multiple stages: Theory and experimental evidence

2013-28 **Dominik Erharter:** Promoting coordination in summary-statistic games

2013-27 **Dominik Erharter:** Screening experts' distributional preferences

2013-26 **Loukas Balafoutas, Rudolf Kerschbamer, Matthias Sutter:** Second-degree moral hazard in a real-world credence goods market

2013-25 **Rudolf Kerschbamer:** The geometry of distributional preferences and a non-parametric identification approach

2013-24 **Nadja Klein, Michel Denuit, Stefan Lang, Thomas Kneib:** Nonlife ratemaking and risk management with bayesian additive models for location, scale and shape

2013-23 **Nadja Klein, Thomas Kneib, Stefan Lang:** Bayesian structured additive distributional regression

2013-22 **David Plavcan, Georg J. Mayr, Achim Zeileis:** Automatic and probabilistic foehn diagnosis with a statistical mixture model *published in Journal of Applied Meteorology and Climatology*

2013-21 **Jakob W. Messner, Georg J. Mayr, Achim Zeileis, Daniel S. Wilks:** Extending extended logistic regression to effectively utilize the ensemble spread

2013-20 **Michael Greinecker, Konrad Podczeck:** Liapounoff's vector measure theorem in Banach spaces *forthcoming in Economic Theory Bulletin*

2013-19 **Florian Lindner:** Decision time and steps of reasoning in a competitive market entry game *forthcoming in Economics Letters*

2013-18 **Michael Greinecker, Konrad Podczeck:** Purification and independence

2013-17 **Loukas Balafoutas, Rudolf Kerschbamer, Martin Kocher, Matthias Sutter:** Revealed distributional preferences: Individuals vs. teams *forthcoming in Journal of Economic Behavior and Organization*

2013-16 **Simone Gobien, Björn Vollan:** Playing with the social network: Social cohesion in resettled and non-resettled communities in Cambodia

2013-15 **Björn Vollan, Sebastian Prediger, Markus Frölich:** Co-managing common pool resources: Do formal rules have to be adapted to traditional ecological norms? *published in Ecological Economics*

2013-14 **Björn Vollan, Yexin Zhou, Andreas Landmann, Biliang Hu, Carsten Herrmann-Pillath:** Cooperation under democracy and authoritarian norms

2013-13 **Florian Lindner, Matthias Sutter:** Level-k reasoning and time pressure in the 11-20 money request game *published in Economics Letters*

2013-12 **Nadja Klein, Thomas Kneib, Stefan Lang:** Bayesian generalized additive models for location, scale and shape for zero-inflated and overdispersed count data

2013-11 **Thomas Stöckl:** Price efficiency and trading behavior in limit order markets with competing insiders *forthcoming in Experimental Economics*

2013-10 **Sebastian Prediger, Björn Vollan, Benedikt Herrmann:** Resource scarcity, spite and cooperation

2013-09 **Andreas Exenberger, Simon Hartmann:** How does institutional change coincide with changes in the quality of life? An exemplary case study

2013-08 **E. Glenn Dutcher, Loukas Balafoutas, Florian Lindner, Dmitry Ryvkin, Matthias Sutter:** Strive to be first or avoid being last: An experiment on relative performance incentives.

2013-07 **Daniela Glätzle-Rützler, Matthias Sutter, Achim Zeileis:** No myopic loss aversion in adolescents? An experimental note

2013-06 **Conrad Kobel, Engelbert Theurl:** Hospital specialisation within a DRG-Framework: The Austrian case

2013-05 **Martin Halla, Mario Lackner, Johann Scharler:** Does the welfare state destroy the family? Evidence from OECD member countries

2013-04 **Thomas Stöckl, Jürgen Huber, Michael Kirchler, Florian Lindner:** Hot hand belief and gambler's fallacy in teams: Evidence from investment experiments

2013-03 **Wolfgang Luhan, Johann Scharler:** Monetary policy, inflation illusion and the Taylor principle: An experimental study

2013-02 **Esther Blanco, Maria Claudia Lopez, James M. Walker:** Tensions between the resource damage and the private benefits of appropriation in the commons

2013-01 **Jakob W. Messner, Achim Zeileis, Jochen Broecker, Georg J. Mayr:** Improved probabilistic wind power forecasts with an inverse power curve transformation and censored regression *forthcoming in Wind Energy*

University of Innsbruck

Working Papers in Economics and Statistics

2014-12

Alexander Razen, Wolgang Brunauer, Nadja Klein, Thomas Kneib, Stefan Lang, Nikolaus Umlauf

Statistical risk analysis for real estate collateral valuation using Bayesian distributional and quantile regression

Abstract
The Basel II framework strictly defines the conditions under which financial institutions are authorized to accept real estate as collateral in order to decrease their credit risk. A widely used concept for its valuation is the hedonic approach. It assumes, that a property can be characterized by a bundle of covariates that involves both individual attributes of the building itself and locational attributes of the region where the building is located in. Each of these attributes can be assigned an implicit price, summing up to the value of the entire property. With respect to value-at-risk concepts financial institutions are often not only interested in the expected value but also in different quantiles of the distribution of real estate prices. To meet these requirements, we develop and compare multilevel structured additive regression models based on GAMLSS type approaches and quantile regression, respectively. Our models involve linear, nonlinear and spatial effects. Nonlinear effects are modeled with P-splines, spatial effects are represented by Gaussian Markov random fields. Due to the high complexity of the models statistical inference is fully Bayesian and based on highly efficient Markov chain Monte Carlo simulation techniques.

ISSN 1993-4378 (Print)
ISSN 1993-6885 (Online)