University of Innsbruck

Working Papers
in
Economics and Statistics

Modeling House Prices using Multilevel Structured Additive Regression

Wolfgang Brunauer, Stefan Lang and Nikolaus Umlauf

2010-19

Modeling House Prices using Multilevel Structured Additive Regression

Wolfgang Brunauer[1], Stefan Lang[2], Nikolaus Umlauf[2]

[1] *Immobilien Rating GmbH, Taborstr. 1–3, 1020 Vienna, Austria; e-mail:* wolfgang.brunauer@irg.at
[2] *University of Innsbruck, Universitätsstr. 15, 6020 Innsbruck, Austria.*

Abstract

This paper analyzes house price data belonging to three hierarchical levels of spatial units. House selling prices with associated individual attributes (the elementary level-1) are grouped within municipalities (level-2), which form districts (level-3), which are themselves nested in counties (level-4). Additionally to individual attributes, explanatory covariates with possibly nonlinear effects are available on two of these spatial resolutions. We apply a multilevel version of structured additive regression (STAR) models to regress house prices on individual attributes and locational neighborhood characteristics in a four level hierarchical model. In multilevel STAR models the regression coefficients of a particular nonlinear term may themselves obey a regression model with structured additive predictor. The framework thus allows to incorporate nonlinear covariate effects and time trends, smooth spatial effects and complex interactions at every level of the hierarchy of the multilevel model. Moreover we are able to decompose the spatial heterogeneity effect and investigate its magnitude at different spatial resolutions allowing for improved predictive quality even in the case of unobserved spatial units. Statistical inference is fully Bayesian and based on highly efficient Markov chain Monte Carlo simulation techniques that take advantage of the hierarchical structure in the data.

Keywords: Bayesian hierarchical models, hedonic pricing models, multilevel models, MCMC, P-splines

1 Introduction

In economics, housing is usually treated as a heterogeneous good, defined by a bundle of utility-bearing characteristics, such as structural (physical) characteristics, like floor space area, constructional condition, age etc., and neighborhood (locational) characteristics, like the proximity to places of work, the social composition of the neighborhood etc. A housing transaction can therefore be considered as a tied sale of a set of these characteristics. One way to address this situation are hedonic pricing models, where the price of a housing unit is decomposed into *implicit prices* of the characteristics which are estimated in a regression analysis of price against characteristics. Originally developed for automobiles by Court (1939), the foundations of hedonic price theory have been developed by Lancaster (1966), focusing on the demand side of the market, and Rosen (1974), focusing on the interaction of bid and offer functions. Another often cited reference is Griliches (1971). Reviews of hedonic price theory in a real estate context are provided e.g. in Follain and Jimenez (1985), Sheppard (1999) and Malpezzi (2003).

Typically, residential properties belong to several levels of spatial (administrative) units, which turns the hedonic pricing model into a *multilevel* or *hierarchical* regression problem (see e.g. Goldstein 2003 and Gelman and Hill 2007). In our case, house selling prices with associated individual attributes (the elementary level-1) are grouped in municipalities (level-2), which form districts (level-3), which are themselves nested in counties (level-4). Available neighborhood covariates on either of these spatial resolutions that might be important for predicting house prices should be accounted for, and it is furthermore reasonable to assume that unmeasured neighborhood characteristics such as local policy and infrastructure affect individual house prices.

Another major problem in hedonic price modeling is that economic theory does not provide clear guidance concerning the functional form of the dependence of price on characteristics, which suggests that hedonic pricing models should allow for nonlinearity in the price functions (see e.g. Wallace 1996 or Ekeland et al. 2004). The most commonly used specification to address this problem is the semi-log form (see e.g. Malpezzi 2003 or Sirmans et al. 2005), but this only seems to mitigate the problem of possible nonlinear relationships to some extent. Therefore, Anglin and Gencay (1996) or Martins-Filho and Bin (2005) demand the use of semi- or nonparametric specifications for this situation. Other examples of semi- and nonparametric approaches for real estate can be found e.g. in Mason and Quigley (1996), Pace (1998) or Bontemps et al. (2008).

A particularly broad and rich framework for semiparametric modeling is provided by generalized structured additive regression (STAR) models introduced in Fahrmeir et al. (2004) and Brezger and Lang (2006). In STAR-models, continuous covariates are modeled as P(enalized)-splines as introduced by Eilers and Marx (1996), see also Wood (2006). Furthermore, random effects for spatial indexes, smooth functions of two dimensional surfaces and (spatially) varying coefficient terms may also be estimated using this methodology.

In this paper we apply a *multilevel version of STAR models* recently developed in Lang et al. (2010) for modeling the dependence of house prices on structural and locational house characteristics. In multilevel STAR models the regression coefficients of nonlinear terms may obey another regression model with structured additive predictor. In that sense, the model is composed of a hierarchy of complex structured additive regression models.

We use a dataset of 3231 owner-occupied single family homes in Austria to estimate a multilevel STAR model of the form (a more detailed description will be given in sections 2 and 3):

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

The top level equation is a STAR-model for logged home sales prices per square meter, lnp_{qm}, with possibly nonlinear effects $f_{1,1}, \ldots, f_{1,q_1}$ of continuous structural house characteristics such as the floor space (*area*) or the age of the building (*age*) and the usual linear part $\mathbf{x}'\boldsymbol{\gamma}$ (e.g. dummy variables for the condition and equipment of the house). While most studies examine the effects of these characteristics on the log of total prices, our rationale to examine the effects on logged *prices per sq. m.* is that effects of structural and locational covariates are typically *proportional to the size of the house*, whereas models based on the (log)-house price a priori assume a fixed effect of characteristics independent of the house size. In statistical terminology we implicitly assume an interaction between floor space and the remaining house characteristics. Spatial heterogeneity is modeled through the spatial random effect $f_{spat_1}(s_1)$ of municipalities s_1, which is further decomposed into a district and county level effect (spatial indexes s_2 and s_3). At levels 2 and 3 further possibly nonlinear effects $f_{2,1}, \ldots, f_{2,q_2}$ and $f_{3,1}$ of locational characteristics are included.

Our approach for hedonic house price modeling has the following key features:

- The *hierarchical structure* of the data is exploited for sophisticated modeling of spatial heterogeneity of house prices. In this way we are able to *decompose* the spatial heterogeneity effect and investigate its magnitude at different spatial resolutions.
- At each level of the hierarchy, *nonlinear covariate effects* can be incorporated using P-splines. This provides the advantage of an economic interpretation of spatial heterogeneity on the one hand and considerable improvement of predictions into spatial units without any observations but with known neighborhood characteristics.
- The hierarchical structure of the spatial effect furthermore allows for *improved predictive quality*, particularly in the case of missing spatial units. For instance, prediction for a new house located in a municipality with no observations is greatly enhanced by borrowing strength from the level-2 covariate effects and the level-3 and level-4 spatial effects.

Statistical inference is fully Bayesian and based on highly efficient Markov chain Monte Carlo (MCMC) simulation techniques that take advantage of the hierarchical structure in the data. The methodology allows for very fast computations with several ten-thousand iterations within a few minutes. We use an implementation of multilevel STAR models in the open source software package BayesX for estimation.

The remainder of this article is structured as follows: In the next section 2 the working data set is described. Section 3 presents multilevel STAR models in the context of hedonic regression for house prices. Results are presented in section 4, and the final section draws some conclusions.

2 Data description and model specification

We have a dataset of owner-occupied single-family homes in Austria at our disposal which exhibits a quite typical structure for real estate data:

- The set of explanatory variables consists of covariates characterizing the house, namely the size, age, year of sale, quality and equipment of the building, which we call *structural attributes/covariates.*
- Individual observations are linked to municipality codes, which allows association with covariates accounting for sociodemographic, economic and neighborhood attributes. Following e.g. Can (1998), we will call these *neighborhood attributes/covariates.*

2.1 Structural attributes

The dataset containing dated house prices together with the housing attributes has been collected in order to estimate the value of the collateral for mortgages by the UniCredit Bank Austria AG from October 1997 to September 2009. Two slightly different instructions for data collection have been employed, which is why the structural covariates affected thereof are encoded accordingly (see table 3 in appendix A for a detailed description). We use continuous variables measuring the size and age as well as the time of sale, and categorical variables that describe the quality of the house. We expect the following directions of the effects:

- Continuous covariates/attributes: As we regress the structural covariates on logged prices per sq. m., a decreasing effect of the floor size of the building due to decreasing marginal returns of additional floor size (*area*) and an increasing effect of the size of the plot it is built on (*area_plot*) can be assumed. The age of the building (*age*), which is calculated as the difference between the year of valuation and the year of construction (i.e., the age at the time of sale), reflects depreciation over time and should therefore have a decreasing effect. The time index (*time_index*, the year of purchase of the house) can be considered as the remaining unexplained temporal heterogeneity and is a measure for the quality adjusted development of house prices over time.
- Categorical covariates/attributes: A good condition of the house (*cond_house*), a high quality of the heating system (*heat*) and of the bathroom and toilets (*bath*) should have an increasing effect on house prices. Furthermore, the existence of an attic (*attic_dum*), a terrace (*terr_dum*) and a garage (*garage*, further separated into good and bad quality) should rise house prices.

2.2 Spatial resolutions and neighborhood attributes

The hierarchical structure of the hedonic pricing model is displayed graphically in figure 1. House prices with structural attributes are nested within 3 spatial resolutions and hence associated with the respective neighborhood attributes, which we use on the most detailed level available. We use various socioeconomic and -demographic attributes as well as measures of proximity to work and metropolitan areas, obtained from the sources described in table 4 in appendix A, to explain spatial variation in house prices per sq. m.

Level-1 is the individual level, on which house prices and housing attributes are measured (see subsection 2.1). In total, 3231 observations are available on the individual level after validation.

Level-2 is the municipal level. Observations are available in 946 of the 2379 Austrian municipalities. On level-2, we employ the following covariates:

- Socioeconomic / -demographic characteristics of the neighborhood: On the one hand, we use the purchase power index (*pp_ind*), the average level of education, indicated by the share of academics (*educ*), which both reflect disposable income and should therefore affect prices positively. On the other hand, we use an age index (*age_ind*), constructed as a population-weighted mean of 20 age cohorts, which measures the average age of inhabitants. A high population age index, reflecting excess of age, serves as a proxy for structural weakness and should have a negative effect on house prices.

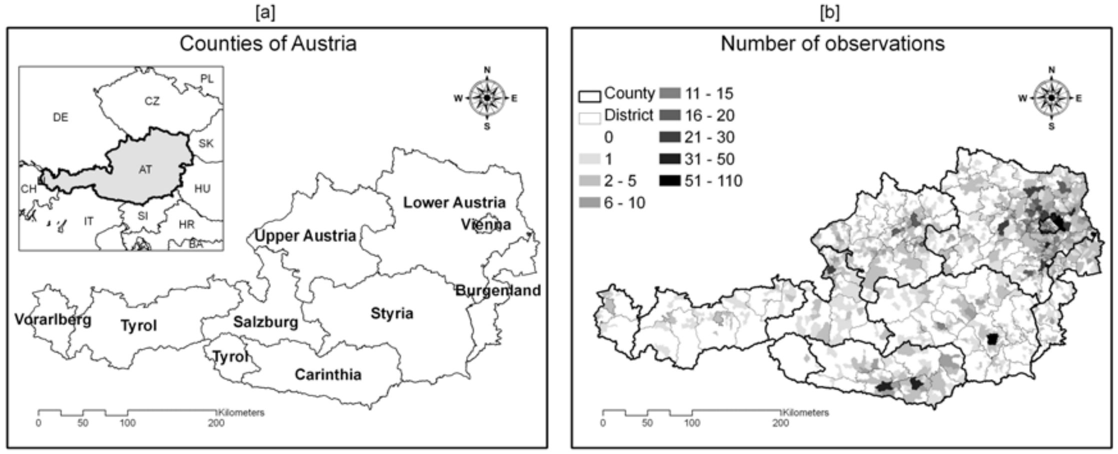

Figure 1: *Levels of hierarchy and distribution of observations.* [a] *The 9 counties of Austria with associated names (level-4) and surrounding countries (small picture).* [b] *Number of observations (level-1) per municipality (level-2); municipalities without observations are left hollow, the shade of municipalities becomes darker as the number of observations increases; additionally shown are the district (level-3) with thin lines and county borders (level-4) with thick lines.*

- Measures of proximity to work and metropolitan areas: Urban economic theory states that commuting to centers of economic activity gives rise to a location rent, which is why a high commuter index (*comm*), i.e. many employees commuting from the municipality, should tend to affect prices negatively. For an overview of urban economics see e.g. DiPasquale and Wheaton (1996). However, close proximity to these centers also provides certain disamenities, as the local infrastructure tends to match the needs of residential use worse. Therefore, the effect of a low commuter index is unclear. Furthermore, as a measure of centrality, we employ population density (*dens*). In densely populated areas, land becomes more valuable, which is why we expect a positive effect of this covariate.

Level-3 is the district level. Individual observations are available on 109 of 121 districts, only the inner districts of Vienna are missing. As each of these districts has neighboring units, spatial effects can be regularized using the neighborhood structure. On this level, an externally provided home price index indicating the neighboring house price level, *wko_ind*, is available.

Level-4 is the county level (9 counties); we do not employ any further explanatory covariates on this level.

A more detailed description of the covariates used at the various levels together with summary statistics is given in table 4 in appendix A.

3 Methodology

3.1 Hierarchical STAR models

Suppose that observations $(y_i, \mathbf{z}_i, \mathbf{x}_i)$, $i = 1, \dots, n$, are given, where y_i is a continuous response variable, and $\mathbf{z}_i = (z_{i1}, \dots, z_{iq})'$ and $\mathbf{x}_i = (x_{i1}, \dots, x_{ip})'$ are vectors of covariates. For the variables in $\mathbf{z}$ possibly nonlinear effects are assumed whereas the variables in $\mathbf{x}$ are modeled in the usual linear way. The components of $\mathbf{z}$ are not necessarily continuous covariates. A component may also indicate a time scale, a cluster- or a spatial index (e.g. municipality, district or county) a certain observation pertains to. Moreover, the components of $\mathbf{z}$ may be two- or even three dimensional in order to model interactions between covariates. We assume an additive decomposition of the effects of z_{ij} (and x_{ij}) and obtain the model

$$y_i = f_1(z_{i1}) + \ldots + f_q(z_{iq}) + \mathbf{x}_i'\boldsymbol{\gamma} + \varepsilon_i. \tag{2}$$

Here, $f_1, \dots, f_q$ are nonlinear functions of the covariates $\mathbf{z}_i$ and $\mathbf{x}_i'\boldsymbol{\gamma}$ is the usual linear part of the model. The errors ε_i are assumed to be mutually independent Gaussian with mean 0 and variance

σ^2, i.e. $\varepsilon_i \sim N(0, \sigma^2)$.

The nonlinear effects in (2) are modeled by a basis functions approach, i.e. a particular function f of covariate z is approximated by a linear combination of basis or indicator functions

$$f(z) = \sum_{k=1}^{K} \beta_k B_k(z). \tag{3}$$

The B_k's are known basis functions and $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)'$ is 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

$$\mathbf{y} = \boldsymbol{\eta} + \boldsymbol{\varepsilon} = \mathbf{Z}_1\boldsymbol{\beta}_1 + \ldots + \mathbf{Z}_q\boldsymbol{\beta}_q + \mathbf{X}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}, \tag{4}$$

where $\mathbf{y} = (y_1, \ldots, y_n)'$, $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_n)'$ and $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$.

In this paper we apply a hierarchical version of STAR models, i.e. the regression coefficients $\boldsymbol{\beta}_j$ of a term f_j may themselves obey a regression model with structured additive predictor. More specifically, we obtain

$$\boldsymbol{\beta}_j = \boldsymbol{\eta}_j + \boldsymbol{\varepsilon}_j = \mathbf{Z}_{j1}\boldsymbol{\beta}_{j1} + \ldots + \mathbf{Z}_{jq_j}\boldsymbol{\beta}_{jq_j} + \mathbf{X}_j\boldsymbol{\gamma}_j + \boldsymbol{\varepsilon}_j, \tag{5}$$

where the terms $\mathbf{Z}_{j1}\boldsymbol{\beta}_{j1}, \ldots, \mathbf{Z}_{jq_j}\boldsymbol{\beta}_{jq_j}$ correspond to additional nonlinear functions $f_{j1}, \ldots, f_{jq_j}$, $\mathbf{X}_j\boldsymbol{\gamma}_j$ comprises additional linear effects, and $\boldsymbol{\varepsilon}_j \sim N(\mathbf{0}, \tau_j^2\mathbf{I})$ is a vector of i.i.d. Gaussian errors.

A third or fourth level in the hierarchy is possible by assuming that the second level regression parameters $\boldsymbol{\beta}_{jl}$, $l = 1, \ldots, q_j$, obey again a STAR model. In that sense, the model is composed of a hierarchy of complex structured additive regression models.

Typically, the compound prior (5) is used if a covariate $z_j \in \{1, \ldots, K\}$ is a unit- or cluster index and z_{ij} indicates the cluster 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 cluster 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 cluster. Using the compound prior (5) we obtain an additive decomposition of the cluster specific effect. The covariates z_{jl}, $l = 1, \ldots, q_j$, in (5) are cluster specific covariates with possible nonlinear cluster effect. By allowing a full STAR predictor (as in the level-1 equation), a rather complex decomposition of the cluster effect $\boldsymbol{\beta}_j$ including interactions is possible. A special case arises if cluster specific covariates are not available. Then the prior for $\boldsymbol{\beta}_j$ collapses to $\boldsymbol{\beta}_j = \boldsymbol{\varepsilon}_j \sim N(\mathbf{0}, \tau_j^2\mathbf{I})$ and we obtain a simple i.i.d. Gaussian cluster specific random effect with variance parameter τ_j^2.

In the model described in section 1 we distinguish four levels: 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). Now the model sketched in (1) can be written as the following four level hierarchical STAR model:

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

On all levels, for continuous covariates possibly nonlinear functions $\mathbf{f}_1, \mathbf{f}_2, \ldots$ modeled by P-splines (see subsection 3.2) are assumed. The categorical covariates on level-1, describing the quality and condition of the house, are encoded as dummy variables and subsumed in the design matrix $\mathbf{X}$ with estimated parameters $\boldsymbol{\gamma}$.

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. Two of the covariates on this level enter the equation logarithmically (denoted by the prefix "*ln_*"), namely the share of academics and the population density. The reason for this is that the distributions of these covariates are strongly positively skewed, which results in volatile estimation results on the natural scale.

The municipality random effect is further decomposed into a district and county level effect (levels 3 and 4). District specific spatial heterogeneity is modeled through the correlated spatial effect *dist* in the level-3 equation by Markov random fields (see the next subsection). We denote this by the superscript "*mrf*". Spatial heterogeneity beyond what can be explained on the district level is modeled through a county specific spatial effect (*county*). For technical reasons the global intercept γ_0 is included on the lowest county level-4.

The fact that we take the logs of house prices *per square meter* results in a special form for the (conditional) mean of house prices. Assuming Gaussian errors for lnp_{qm} results in log-normally distributed prices per square meter p_{qm} and the conditional mean of the total house price p changes multiplicatively with changes in values of covariates, and proportionally to the floor area of the house:

$$\begin{aligned} E(p) &= area \times \exp(\eta + \sigma^2/2) \\ &= area \times \exp(f_1(z_1)) \ldots \exp(f_q(z_q)) \exp(\gamma_0) \exp(\gamma_1 x_1) \ldots \exp(\gamma_p x_p) \exp(\sigma^2/2). \end{aligned}$$

Therefore, if for example covariate x_1 changes by one unit, the predictor η changes by the factor $\exp(\gamma_1)$ and expected total prices change by the factor $area \times \exp(\gamma_1)$. So the change in expected prices is proportional to the floor area of the house. Turning to the nonlinear effects, let $f(z_1)$ be the nonlinear effect of the covariate z_1, and let $\mathrm{d}f(z_1) = f(z_1+1) - f(z_1)$. Then

$$\exp(f(z_1+1)) = \exp(f(z_1+1) - f(z_1) + f(z_1)) = \exp(f(z_1)) \exp(\mathrm{d}f(z_1)),$$

so the expected total house price changes by the factor $area \times \exp(\mathrm{d}f(z_1))$. As $f(.)$ is a nonlinear function, the changes differ over the range of z_1 and are again proportional to the size of the house.

3.2 Priors for the regression coefficient

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 Belitz and Lang (2008). The standard are quadratic penalties of the form $\lambda\boldsymbol{\beta}'\mathbf{K}\boldsymbol{\beta}$ where $\mathbf{K}$ is a $K \times K$ 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

$$p(\boldsymbol{\beta}|\tau^2) \propto \left(\frac{1}{\tau^2}\right)^{rk(\mathbf{K})/2} \exp\left(-\frac{1}{2\tau^2}\boldsymbol{\beta}'\mathbf{K}\boldsymbol{\beta}\right) \cdot I(\mathbf{A}\boldsymbol{\beta} = \mathbf{0}), \tag{7}$$

where $I(\cdot)$ is the indicator function. The key components of the prior are the penalty matrix $\mathbf{K}$, the variance parameter τ^2 and the constraint $\mathbf{A}\boldsymbol{\beta} = \mathbf{0}$.

The structure of the penalty or prior precision matrix $\mathbf{K}$ depends on the covariate type and on our prior assumptions about smoothness of $\mathbf{f}$. Typically the penalty matrix in our examples is rank deficient, i.e. $rk(\mathbf{K}) < K$, resulting in a partially improper prior.

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 overall variance parameter σ^2), 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 $\lambda = \sigma^2/\tau^2$.

The term $I(\mathbf{A}\boldsymbol{\beta} = \mathbf{0})$ imposes required identifiability constraints on the parameter vector. A straightforward choice is $\mathbf{A} = (1, \dots, 1)$, i.e. the regression coefficients are centered around zero. A better choice in terms of interpretability and mixing of the resulting Markov chains is to use a weighted average of regression coefficients, i.e. $\mathbf{A} = (c_1, \dots, c_K)$. As a standard we use $c_k = \sum_{i=1}^{n} B_k(z_i)$ resulting in the more natural constraint $\sum_{i=1}^{n} f(z_i) = 0$.

Particular examples for nonlinear terms in our application are P-splines for modeling nonlinear effects of continuous covariates and Gaussian Markov random fields for modeling smooth spatial effects. Further examples not employed in this paper can be found in Brezger and Lang (2006).

P-splines

For P-splines the design matrix $\mathbf{Z}$ consists of B-spline basis functions evaluated at the observations. The penalty is given by

$$\sum_{k=d+1}^{K} \left(\Delta^d \beta_k\right)^2 = \boldsymbol{\beta}'\mathbf{D}'\mathbf{D}\boldsymbol{\beta} = \boldsymbol{\beta}'\mathbf{K}\boldsymbol{\beta}, \tag{8}$$

were Δ^d is the difference operator of order d and $\mathbf{D}$ is the corresponding difference matrix. The default for d in most implementations is $d = 2$. For more details on Bayesian P-splines see Lang and Brezger (2004).

Markov random fields

The correlated district specific heterogeneity effect $f_{5,6,2}^{mrf}(dist)$ in equation (6) is modeled by Markov random fields (MRF). Suppose that $z \in \{1, \dots, K\}$ 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 $\beta_{z'}$ of neighboring sites z'. Typically sites are assumed to be neighbors if they share a common boundary. We denote the set of neighbors of site z by $N(z)$. MRFs assume that the conditional distribution of β_z given neighboring sites $z' \in N(z)$ is Gaussian with

$$z \,|\, z', z' \neq z \sim N\left(\frac{1}{|N(z)|} \sum_{z' \in N(z)} \beta'_z, \frac{\tau^2}{|N(z)|}\right),$$

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

The joint (prior) distribution of $\boldsymbol{\beta}$ is of the form (7) with penalty matrix $\mathbf{K}$ given by

$$\mathbf{K}[z, z'] = \begin{cases} -1 & z \neq z', z' \in N(z), \\ 0 & z \neq z', z' \notin N(z) \\ |N(z)| & z = z'. \end{cases} \tag{9}$$

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. $\mathbf{Z}_{5,6,2} = \mathbf{I}$.

3.3 Sketch of MCMC Inference

In the following, we will sketch a Gibbs sampler for models with Gaussian errors. For the sake of simplicity we restrict the presentation to a two level hierarchical model with one level-2 equation for the regression coefficients of the first term $\mathbf{Z}_1\boldsymbol{\beta}_1$. That is, the level-1 equation is $\mathbf{y} = \boldsymbol{\eta} + \boldsymbol{\varepsilon}$ as in (4). The level-2 equation is of the form (5) with $j = 1$.

The parameters are updated in blocks where each vector of regression coefficients $\boldsymbol{\beta}_j$ ($\boldsymbol{\beta}_{1l}$ in a second level of the hierarchy) of a particular term is updated in one (possibly large) block followed by updating the regression coefficients $\boldsymbol{\gamma}$, $\boldsymbol{\gamma}_1$ of linear effects and the variance components τ_j^2, τ_{1l}^2, σ^2. The next subsection sketches updates of regression coefficients $\boldsymbol{\beta}_j$, $\boldsymbol{\beta}_{1l}$ of nonlinear terms. Updates of the remaining parameters are straightforward. Full details can be found in Lang et al. (2010).

Full conditionals for regression coefficients of nonlinear terms

The full conditionals for the regression coefficients $\boldsymbol{\beta}_1$ with the compound prior (5) and the coefficients $\boldsymbol{\beta}_j$, $j = 2, \dots, q$, $\boldsymbol{\beta}_{1l}$, $l = 1, \dots, q_1$ with the basic prior (7) are all multivariate Gaussian. The respective posterior precision $\boldsymbol{\Sigma}^{-1}$ and mean $\boldsymbol{\mu}$ is given by

$$
\begin{aligned}
\boldsymbol{\Sigma}^{-1} &= \tfrac{1}{\sigma^2}\left(\mathbf{Z}_1'\mathbf{W}\mathbf{Z}_1 + \tfrac{\sigma^2}{\tau_1^2}\mathbf{I}\right), & \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} &= \tfrac{1}{\sigma^2}\mathbf{Z}_1'\mathbf{W}\mathbf{r} + \tfrac{1}{\tau_1^2}\boldsymbol{\eta}_1, & (\boldsymbol{\beta}_1),\\
\boldsymbol{\Sigma}^{-1} &= \tfrac{1}{\sigma^2}\left(\mathbf{Z}_j'\mathbf{W}\mathbf{Z}_j + \tfrac{\sigma^2}{\tau_j^2}\mathbf{K}_j\right), & \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} &= \tfrac{1}{\sigma^2}\mathbf{Z}_j'\mathbf{W}\,\mathbf{r}, & (\boldsymbol{\beta}_j),\\
\boldsymbol{\Sigma}^{-1} &= \tfrac{1}{\tau_1^2}\left(\mathbf{Z}_{1l}'\mathbf{Z}_{1l} + \tfrac{\tau_1^2}{\tau_{1l}^2}\mathbf{K}_{1l}\right), & \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} &= \tfrac{1}{\tau_1^2}\mathbf{Z}_{1l}'\,\mathbf{r}_1, & (\boldsymbol{\beta}_{1l}),
\end{aligned}
\tag{10}
$$

where $\mathbf{r}$ is the current partial residual and $\mathbf{r}_1$ is the "partial residual" of the level-2 equation. More precisely, $\mathbf{r}_1 = \boldsymbol{\beta}_1 - \tilde{\boldsymbol{\eta}}_1$ and $\tilde{\boldsymbol{\eta}}_1$ is the predictor of the level-2 equation excluding the current effect of z_{1l}.

MCMC updates of the regression coefficients takes advantage of the following key features:

Sparsity: Design matrices $\mathbf{Z}_j, \mathbf{Z}_{1l}$ and penalty matrices $\mathbf{K}_j, \mathbf{K}_{1l}$ and with it cross products $\mathbf{Z}_j'\mathbf{W}\mathbf{Z}_j, \mathbf{Z}_{1l}'\mathbf{Z}_{1l}$ and posterior precision matrices in (10) are often sparse. The sparsity can be exploited for highly efficient computation of cross products, Cholesky decompositions of posterior precision matrices and for fast solving of relevant linear equation systems.

Reduced complexity in the second or third stage of the hierarchy: Updating the regression coefficients $\boldsymbol{\beta}_{1l}$, $l = 1, \dots, q_1$, in the second (or third) level is done conditionally on the parameter vector $\boldsymbol{\beta}_1$. This facilitates updating the parameters for two reasons. First the number of "observations" in the level-2 equation is equal to the length of the vector $\boldsymbol{\beta}_1$ and therefore much less than the actual number of observations n. Second the full conditionals for $\boldsymbol{\beta}_{1l}$ are Gaussian regardless of the response distribution in the first level of the hierarchy.

Number of different observations smaller than sample size: In most cases the number m_j of different observations $z_{(1)}, \dots, z_{(m_j)}$ in $\mathbf{Z}_j$ (or m_{1l} in $\mathbf{Z}_{1l}$ in the level-2 equation) is much smaller than the total number n of observations. For instance, the age of the house in our application has only 80 different values whereas there are more than 3000 individual observations. The fact that $m_j \ll n$ may be utilized to considerably speed up computations of the cross products $\mathbf{Z}_j'\mathbf{W}\mathbf{Z}_j$, $\mathbf{Z}_{1l}'\mathbf{Z}_{1l}$, the vectors $\mathbf{Z}_j'\mathbf{W}\,\mathbf{r}$, $\mathbf{Z}_{1l}'\,\mathbf{r}_1$ and finally the updated vectors of function evaluations $\mathbf{f}_j = \mathbf{Z}_j\boldsymbol{\beta}_j$, $\mathbf{f}_{1l} = \mathbf{Z}_{1l}\boldsymbol{\beta}_{1l}$.

Full details of the MCMC techniques can be found in Lang et al. (2010).

3.4 Software

We use an implementation in the open source software package BayesX for the estimation of hierarchical STAR models. BayesX is publicly available at

http://www.stat.uni-muenchen.de/ bayesx/bayesx.html,

see Brezger et al. (2005) and Brezger et al. (2009). The homepage of BayesX contains also a number of tutorials. The following code fragment exemplifies the usage of BayesX in the context of hierarchical STAR models:

```
dataset data_county;
data_county.infile using c:\data\counties.raw;

dataset data_dist;
data_dist.infile using c:\data\districts.raw;
data_dist.generate dist_mrf = dist;

dataset data_muni;
data_muni.infile using c:\data\municipalities.raw;

dataset data_homes;
data_homes.infile using c:\data\single_family_homes.raw;
```

```
map map_dist.infile using c:\maps\districts.bnd;

mcmcreg hier_STAR;

hier_STAR.hregress
county = const, family=gaussian_re hlevel=2 iterations=32000 step=30 burnin=2000
using data_county;

hier_STAR.hregress
dist = wko_ind(pspline) + dist_mrf(spatial,map=map_dist) + county(hrandom),
family=gaussian_re hlevel=2 using data_dist;

hier_STAR.hregress
muni = pp_ind(pspline) + ... + ln_educ(pspline) + dist(hrandom),
family=gaussian_re hlevel=2 using data_muni;

hier_STAR.hregress
lnp_qm = area(pspline) + ... + age(pspline) + cellar_dum + ... + muni(hrandom),
hlevel=1 family=loggaussian using data_homes;
```

In a first step, dataset objects for each level have to be defined using the `dataset` command, and the data is read from ASCII files. Note that duplicates with respect to the spatial indices have to be dropped on levels 2, 3 and 4, and that the (continuous) covariates should be centered in advance in order to further improve mixing.

Next, we create a `map`-object and read the geographical information of a boundary file for the districts, `districts.bnd`. Based on the boundary information, the `map`-object automatically computes the neighborhood structure of the districts. Note that in order to match the district information with the `map` object on level-3, we generate a copy of the district code in the dataset object `data_dist` which corresponds to the spatial index of the map.

We then define a `mcmcreg`-object and apply the method `hregress` to fit our hierarchical STAR model. We have to set up the models for each level in reverse order, starting with the lowest level.

In three of the four hierarchical levels, we define P-splines with second order difference penalties (`pspline`). On the individual level we also specify linear effects. For technical reasons, the global intercept (`const`) is included on level-4. We have uncorrelated municipality, district and county effects (`hrandom`) on levels 1 to 3. The spatial effect of the district furthermore consists of a correlated part included in the district level equation (`dist_mrf(spatial,map=map_dist)`).

The option `hlevel=1` distinguishes the level-1 equation from the lower level equations (`hlevel=2`). As our response is logarithmically transformed, we define the `family=loggaussian` on level-1. On the lower levels, the (pseudo) responses are Gaussian random effects (`family=gaussian_re`).

Finally, we define in the first equation the number of MCMC-`iterations` and the number of `burnin`-iterations as well as the thinning parameter for the MCMC simulation, `step`.

4 Results

We now present the estimation results for the base model (6). The results are based on a final MCMC run with 502000 iterations and a burn in period of only 2000 iterations. We stored every 500th iteration resulting in a sample of 1000 practically independent draws from the posterior. Computing time for the MCMC sampler was approximately 30 minutes on a moderately modern desktop computer (Intel core duo processor 2.8GHz). Note that no more than 32000 iterations are typically enough in preliminary MCMC runs to obtain sufficiently exact estimation results. The run time for these preliminary runs is only 110 seconds. The comparably large number of iterations in the final run was used to be absolutely sure about the precision of estimates.

We first show in subsection 4.1 the effects of the continuous covariates on all levels (the linear effects estimation results are presented in table 5 of appendix B). Next, we will describe the spatial effects in more detail (section 4.2). The last subsection 4.3 is devoted to model diagnostics and possible improvements of the base model.

4.1 Continuous covariate effects

Structural covariates

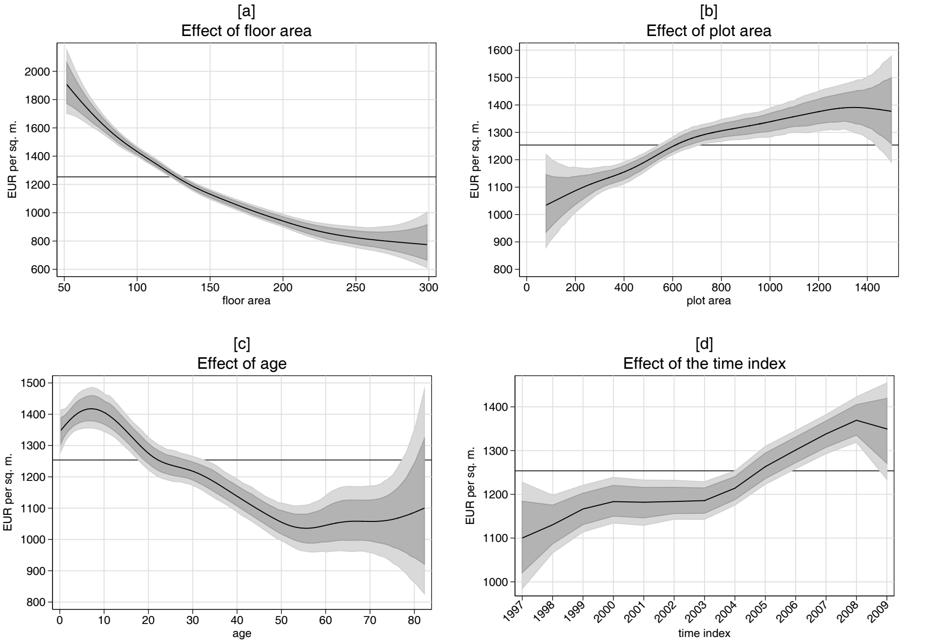

Figure 2: *Effects of the continuous structural covariates of level-1.* [a] *Effect of the floor size area (variable area);* [b] *Effect of the plot area (areaplot).* [c] *Effect of the age of the building (age);* [d] *Effect of the time index (time_index). Shown are the posterior mean estimates with pointwise (dark grey) and simultaneous (light grey) 95% credible intervals.*

Figure 2 shows the effects of the structural continuous covariates together with pointwise and simultaneous 95% credible intervals, see Krivobokova et al. (2010) for their construction. In order to get an impression of the magnitude of effects, we transform the functions to natural units (prices in Euro per sq. m.), where all other covariates are held constant at mean level of attributes (we call this the *average effect*). Since the effects are quite different in magnitude, we do not show them on the same scale.

The effect of the floor area (variable *area*, panel [a]) is very pronounced, it covers a range of more than 1100 Euro. However, the decreasing effect of additional floor area on prices per square meter weakens as the floor area becomes larger.

In panel [b], the effect of the plot area (*areaplot*) is shown. We find that additional plot area yields higher prices per square meter of floor area, although this effect becomes weaker as plot area increases and levels off at around 1200 sq. m. House prices per sq. m. change by more than 350 Euro over the domain of plot area.

As the effect of the *age* of the building (panel [c] of figure 2) can be considered as the rate of depreciation of single family homes, the initial increase up to an age of 3 years is not in line with our expectations; we will come back to this issue in section 4.3. House prices then depreciate nearly linearly until an age of about 50 years and stay constant afterwards. The age of the building accounts for a variation of more than 380 Euro per sq. m.

The effect of the time index, shown in panel [d] of figure 2, shows the quality controlled development of house prices over time. After a slight increase from 1997 to 2000, prices stay constant until 2003 and rise afterwards until 2008. In the last year of the observation period, prices slightly decrease. Although this decrease is within wide confidence bands (due to the small number of observations in 2009), this could be the result of the economic crisis of 2008/2009. In total, the time index accounts for variation in a range of around 260 Euro.

Neighborhood covariates

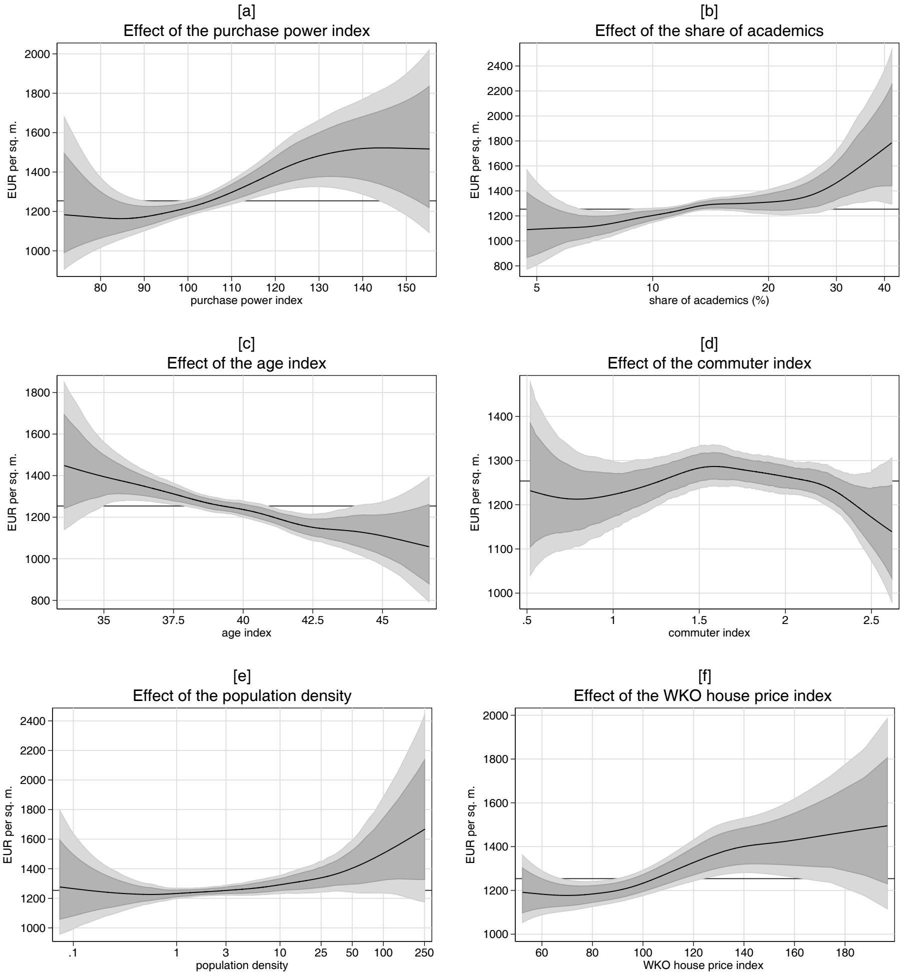

Figure 3: *Effects of the neighborhood covariates. First row: Effect of the purchase power index (pp_ind)* [a] *and the share of academics (educ)* [b]*. Second row: Effect of the age index (age_ind)* [c] *and the commuter index (comm)* [d]*. Third row: Effect of the log of population density (ln_dens)* [e] *and the house price index (wko_ind)* [f]*. Shown are the posterior mean estimates with pointwise (dark grey) and simultaneous (light grey) 95% credible intervals.*

In figure 3, the neighborhood effects are displayed, again on a natural scale of prices per sq. m., together with pointwise and simultaneous 95% credible intervals. In the top row, the effect of the purchase power index (*pp_ind*) is displayed in panel [a]. While for low and high values of this index the effect is negligible, there is a pronounced increase in house prices between 90 and 130 index points. The total effect has a bandwidth of 355 Euro.

As noted in section 3.1, the share of academics (*ln_educ*) enters the equation logarithmically, but is also displayed in natural values in panel [b]. This effect is stronger on the peripheral range of this covariate than in the interior, with a pronounced increase starting at a share of approximately 27% (although within very wide credible intervals). The share of academics accounts for a variation of nearly 690 Euro.

The effect of the age index (*age_ind*, displayed in panel [c]) is nearly linear. The negative direction of this effect could be interpreted as a decreasing attractiveness of municipalities that exhibit an excess of age, which could be expected from our considerations in section 2.2. Prices per sq. m. decrease by nearly 380 Euro or from the lowest to the highest age index.

The effect of the commuter index *comm*, which indicates proximity to work, is displayed in panel [d]. With a range of 146 Euro, the commuter index has the weakest effect of the continuous covariates discussed here and is insignificant in the sense that simultaneous credible intervals cover the average effect over the whole range of the covariate. Although no positive effect of closeness to centers of economic activity is traceable, there are some interesting tendencies: The maximum lies at 1.5, where there is a roughly equal number of commuters from and into the municipality, and decreases afterwards. Considering that this index was constructed as a population-weighted mean of 4 categories (where 0 denotes commuters into the municipality, 1 stands for non-commuters, 2 for commuters within the municipality and 3 into other municipalities), this means that, *ceteris paribus*, the highest value can be obtained by municipalities neither dominated by commuters into nor commuters out of the municipality.

The effect of population density *ln_dens*, displayed in panel [e], shows a tendency toward higher house prices in more densely populated areas, although it is not significant in a strict sense either. This effect accounts for a variation of nearly 440 Euro per sq. m.

Finally, the externally provided house price index *wko_ind* (the only covariate on level-3) is shown in panel [f]. As expected, this effect is increasing, although for index values of more than 140 it becomes weaker and more volatile. This index accounts for a variation of more than 310 Euro per sq. m.

The deviance information criterion (DIC, Spiegelhalter et al. 2002) for the base model is 1564, see also model no. 1 in table 1. Excluding the two "insignificant" variables *comm* and *ln_dens* from the base model yields a DIC of 1562 (model no. 2). A difference of only two units implies that both models are not substantially different in terms of goodness of fit. We therefore keep both variables in the model.

		with outliers		**without outliers**	
No.	**Model**	**Deviance**	**DIC**	**Deviance**	**DIC**
1	Base model	1229	1564	878	1215
2	Base model, insignificant covariates removed	1185	1562	828	1219
3	Reference model	998	1661	614	1366
4	Additional dummy for Vienna	1228	1563	874	1214
5	Model with interaction terms	1158	1511	822	1180

Table 1: *Unstandardized deviance and Deviance information criterion (DIC) for model specifications presented in this paper in the order of appearance. The last two columns provide results for the re-estimated models without outliers as described in subsection 4.3*

4.2 Spatial effects

The total amount of spatial heterogeneity is composed of spatial effects on municipal (level-2), district (level-3) and county level (level-4). Continuous neighborhood effects explain spatial heterogeneity explicitly to a certain extent on two of these levels, we call this *explained* spatial heterogeneity. The remaining i.i.d. spatial random effects ε_5, $\varepsilon_{5,6}$ and $\varepsilon_{5,6,3}$ as well as the correlated district specific effect $f_{5,6,2}(dist)$ in (6) account for *unexplained* spatial heterogeneity.

Our focus in the presentation of the spatial effects is twofold. First, we analyze the *distribution of spatial heterogeneity* over Austria. Second, we discuss the *hierarchical decomposition of unexplained spatial heterogeneity*. For the sake of illustration, we compare the results of the *base model* (6) to those of a model without any explanatory neighborhood covariates which we call *reference model*.

Distribution of spatial heterogeneity over Austria

Figure 4 visualizes the posterior mean of the spatial effect over Austria. Panels [a] and [b] compare the *base model* to the *reference model* with respect to total spatial heterogeneity (explained plus unexplained heterogeneity). In panel [c] we show the amount of unexplained spatial heterogeneity in the *base model*, and we are going to compare it to total spatial heterogeneity in panel [a]. To get

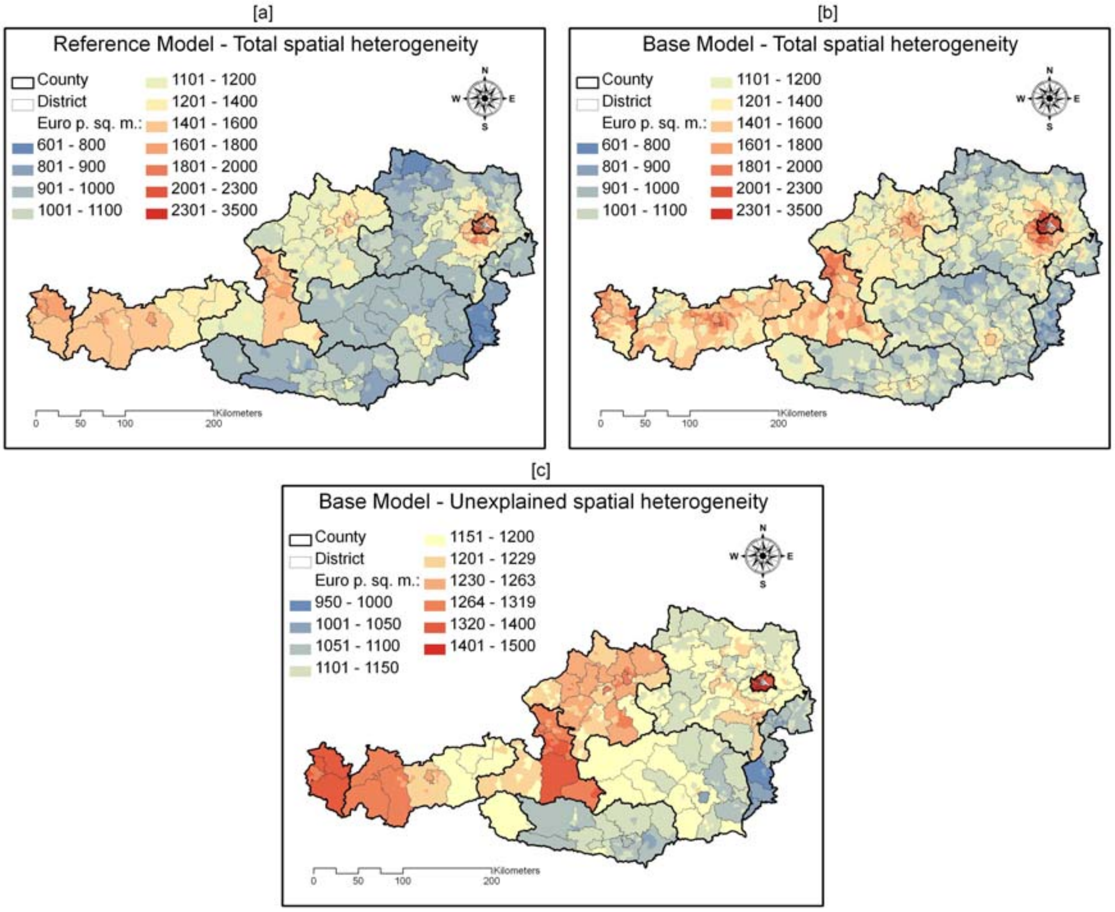

Figure 4: *Distribution of spatial heterogeneity (evaluated at the average effect).* [a] *Total spatial heterogeneity in the reference model.* [b] *Total spatial heterogeneity in the base model (including neighborhood covariate effects).* [c] *Remaining unexplained spatial heterogeneity in the base model (scale differs from above).*

a better intuition for the size of the spatial effects we present the same comparisons in the form of kernel densities of the posterior means in figure 5. Panel [a] of this figure corresponds to panels [a] and [b] in figure 4 and panel [b] to panels [a] and [c] in figure 4.

Interestingly, the distribution of total spatial heterogeneity is similar in the *base model* and the *reference model* (panels [a] and [b] of figure 4). Also the size of the heterogeneity effect is comparable (panel [a] of figure 5). This implies that the *reference model* is able to capture the missing neighborhood covariate effects through the various correlated and uncorrelated random effects. However, it is still worthwhile to include neighborhood covariates for several reasons:

- First of all the spatial pattern in the *base model* provides a more differentiated or “scattered” picture of Austria than the reference model. Overall, the bandwidth of total spatial heterogeneity for the *base model* is wider than for the *reference model* (2301 vs. 1595 Euro per sq. m.). The reason for this is that spatial effects are modeled explicitly, strong heterogeneity may occur where neighborhood covariates have pronounced effects, while in a model where there is only unexplained spatial heterogeneity the shrinkage property of random effect estimators prevents this for small sample sizes in the respective region.

- Neighborhood covariates give spatial effects an economic interpretation and tend to produce stronger heterogeneity if this is theoretically justified.

- The prediction for municipalities without observations borrows strength from both the nonlinear neighborhood effects and the level-3 and level-4 spatial effects. For observed municipalities, unexplained spatial effects adjust neighborhood covariate effects.

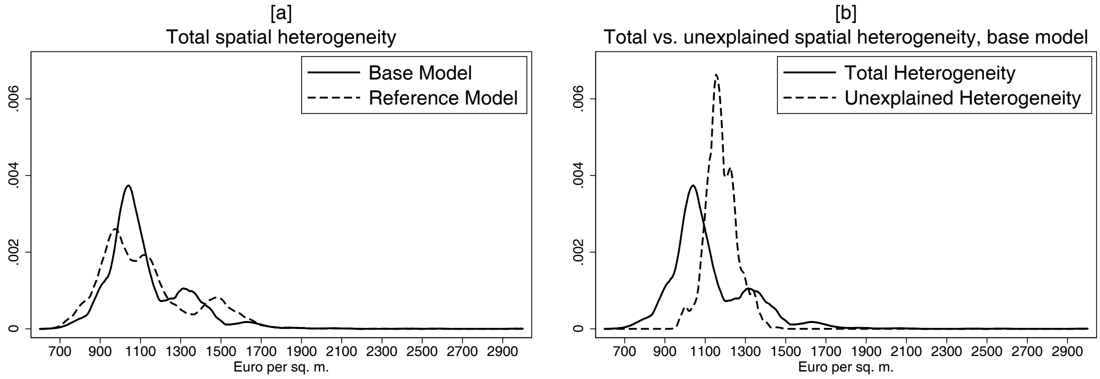

Figure 5: [a] *Kernel densities for total spatial heterogeneity of the base model (solid line) and the reference model (dashed line).* [b] *Kernel densities for total spatial heterogeneity (solid line) and unexplained spatial heterogeneity (dashed line) in the base model.*

- The *base model* specification reduces the deviance information criterion DIC from 1660 in the *reference model* to 1564, or by approximately 100 points, see models no. 1 and 3 in table 1.

Inspecting panels [a] and [c] of figure 4 and panel [b] of figure 5 shows that unexplained heterogeneity is reduced dramatically and only accounts for a variation of 505 Euro per sq. m., which is 22% of total heterogeneity in the *base model.*

Careful analysis of the distribution of unexplained heterogeneity sometimes exhibits interesting patterns and provides suggestions for improving the models. Panel [c] of figure 4 shows that prices are considerably below what can be explained with neighborhood covariates in Burgenland (far east of Austria) and the adjacent parts of Styria as well as Carinthia (south of Austria). In the west of Austria, i.e. parts of Salzburg, Tyrol and Vorarlberg, house prices are above what can be explained with the covariates we have available, probably because these are the classical winter tourism regions in Austria. Yet, the strongest positive effects can be found in the county Vienna, a "spatial outlier". This suggests modeling this effect explicitly. Extending the *base model* and integrating a Vienna dummy in the level-4 equation we obtain a further reduction of unexplained heterogeneity to a range of 407 Euro per sq. m. Although the Vienna dummy is highly significant, the DIC of the improved model stays more or less constant (model no. 4 in table 1).

Decomposition of unexplained spatial heterogeneity

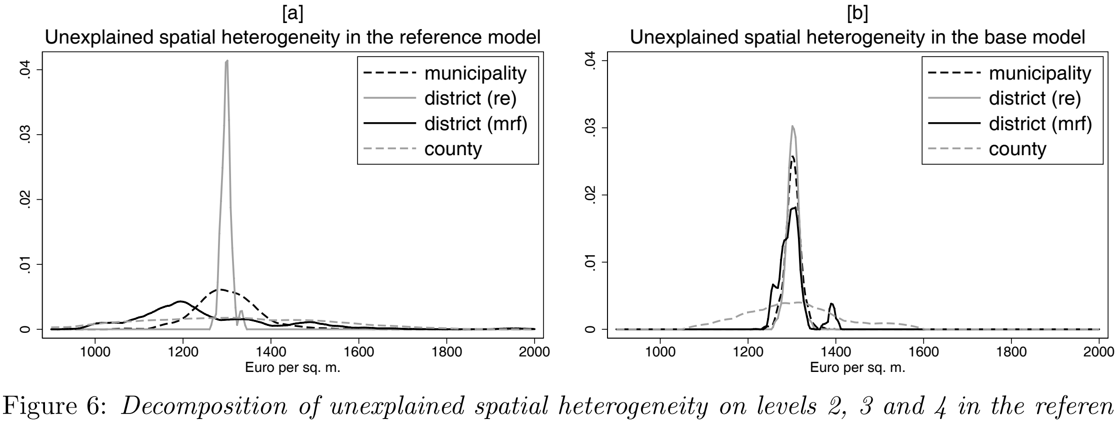

Figure 6: *Decomposition of unexplained spatial heterogeneity on levels 2, 3 and 4 in the reference model* [a] *and in the base model* [b]. *Shown are kernel density estimates of the respective random effects.*

Figure 6 shows the distribution of unexplained heterogeneity attributed to the municipality level-2 (black dashed line), the district level-3 (where the uncorrelated random effect displayed as grey solid line, the Markov random field as black solid line) and the county level-4 (grey dashed line) for the *reference model* in panel [a] and the *base model* in panel [b].

On the municipality level the spatial effect ranges from 976 to 1513 Euro per sq. m. in the reference model, a difference of 537 Euro per sq. m. between municipalities. Unexplained municipal effects can be largely explained by the neighborhood covariates and reduced to a range of approximately 142 Euro per sq. m in the base model.

Unexplained heterogeneity on district level is split into a spatially correlated Markov random field and an uncorrelated random district effect. While the latter is rather weak (in both models it has a range of about 65 Euro per sq. m.), the effect of the Markov random field is very pronounced in the reference model: It ranges from 899 to 1811 Euro per sq. m., accounting for more than 911 Euro in spatial variation. Again, integrating neighborhood covariates leads to a significant reduction of unexplained spatial heterogeneity on this level, which ranges now from 1125 to 1299 Euro per sq. m.

On the county level, unexplained heterogeneity ranges from 928 to 1505 Euro per sq. m., accounting for 578 Euro in variation in the *reference model.* If we model spatial heterogeneity explicitly, the unexplained part is reduced to approximately 341 Euro per sq. m.

In summary, it can be stated that inclusion of the neighborhood covariates explains a great deal of the spatial heterogeneity but a certain proportion remains unexplained, providing starting points for further exploration.

4.3 Model Diagnostics and possible model improvements

In a Bayesian framework, systematic differences between the data and the estimated model can be detected with the aid of posterior predictive checks as advocated in Gelman and Hill (2007). Specifically, we compare the empirical distribution of logged house prices per sq. m. with the simulated posterior distributions of house prices obtained from our base model. In panel [a] of figure 7 we display the empirical distribution of logged prices per sq. m. (black line) together with 1000 replicated samples from the base model (grey lines). The samples are easily obtained as a by product of the MCMC sampler. Panel [b] additionally displays a scatter plot of observed against posterior mean predicted prices per sq. m.

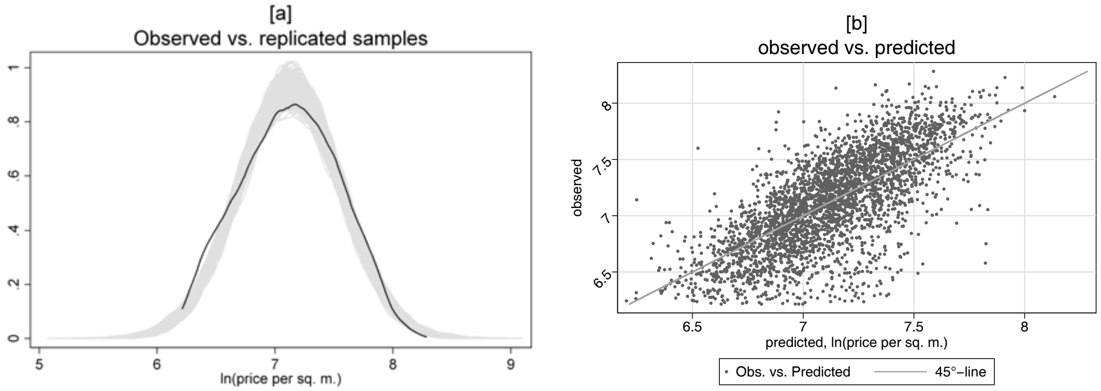

Figure 7: [a] *kernel densities for the distribution of observed house prices (black line) vs. simulated prices according to the base model (grey line),* [b] *scatter plot of observed vs. predicted log house prices per sq. m.*

The predictive checks in panel [a] indicate some misspecification as the simulated responses are sampled in a wider range and are somewhat more concentrated around the mean. While the mean, the standard deviation and most quantiles of the observed logged prices per sq. m. are well within the range of the corresponding sampled model quantities, the extreme quantiles often fall outside the range, see also table 2. This is also supported by panel [b] which shows that predictions tend to be too high for many observations with very low observed logged price per sq. m.

Close inspection of the "problematic observations" shows that the corresponding houses are mostly

	Replicated		**Observed**
	Min	**Max**	
mean	7.10	7.15	7.12
std.dev.	0.40	0.44	0.42
min	5.07	5.94	6.22
1% quantile	6.04	6.23	6.27
5% quantile	6.37	6.48	6.42
25% quantile	6.80	6.87	6.81
50% quantile	7.10	7.16	7.13
75% quantile	7.38	7.45	7.43
95% quantile	7.76	7.87	7.81
99% quantile	8.02	8.20	8.01
max	8.26	9.10	8.28

Table 2: *Mean, standard deviation and quantiles of simulated data from the base model vs. observed data.*

in a group with age less than three years. Recall that the age effect in panel [c] of figure 2 is not in line with our expectations for buildings of an age of less than three years. Improved models are obtained by the following steps:

- *Remove outliers:* The dataset contains a number of "new" houses with implausibly low observed prices per sq. m. below 650 Euro (in total, 43 observations). The reason for these low prices is that for some of the "new" houses the price might have been paid for only partly or even undeveloped land. Removing the outliers results in the expected monotonically decreasing age effect. Moreover, as the last column of table 1 shows, the deviance and with it the DIC decreases dramatically for all model specifications.

- *Include interactions with age:* To find possible interactions with age we generated subsamples for three different house age groups ($age \leq 2$, $2 < age \leq 9$ and $age > 9$) and fitted separate models to them. A careful comparison of the three submodels shows that there are two main effects notably varying over these submodels, namely the effect of the plot area and the effect of the time index. We therefore integrated interaction effects for the plot area with age group 1 (≤ 2 years) and one for the time index with age groups 3 (> 9 years). Again the DIC decreased remarkably by 50 units if outliers are not removed and still by 35 units if outliers are additionally removed (model no. 5 in table 1)

5 Conclusions

This paper analyzes house prices using multilevel structured additive regression models. The proposed modeling framework is particularly useful to model house prices as the models are able to appropriately consider the typical hierarchical structure of the data. In our case, house selling prices with associated individual attributes (level-1) are grouped in municipalities (level-2), which form districts (level-3), which are themselves nested in counties (level-4). At each level of the hierarchy multilevel STAR models allow to incorporate nonlinear effects of continuous covariates, correlated spatial random effects as well as complex interactions. The hierarchical structure of the model can also be utilized for highly efficient MCMC simulation schemes for Bayesian inference allowing for several ten thousand iterations within one or two minutes. Model choice is based on the deviance information criterion, simultaneous credible intervals and posterior predictive checks to detect discrepancies between the data and the model.

Several directions for future research are conceivable: This paper primarily models the (conditional) mean of the responses. In the context of hedonic house price regression joint modeling of the mean and the variance as e.g. provided by Rigby and Stasinopoulos (2005) is of particular practical interest. Models of this kind allow for more precise prediction intervals and with it more reliable risk management. In a similar direction goes quantile regression, see Yue and Rue (2010) in the context of our modeling framework. Finally, more automated model choice and variable selection would be highly interesting. A promising approach for state space models, which are close to our models, has been recently developed by Frühwirth-Schnatter and Wagner (2010).

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.*

The dataset providing the neighborhood covariates described in table 4 comes from three different sources:

1. We use data from the Austrian Federal Bureau of Statistics (*Statistics Austria*) on municipal level (2001), including age cohorts of inhabitants, level of education, and commuting.

2. Data on purchase power (*pp_ind*) and population, both on municipal level and for the year 2009, come from *Michael Bauer Research.*

3. Finally, we use an external home price index as explanatory variables, the home price index published by the Austrian Federal Economic Chamber (*Wirtschaftskammer Oesterreich, WKO*) on a district level (2008). We call this *wko_ind.*

Level 2: Municipal

Name	Description (year) [unit]; source; mean / min. / max.	Exp. Eff.
pp_ind	purchase power index on municiptal level (2009) [n.a.]; source: Michael Bauer Research; 103.13 / 65.0 / 148.45	+
educ	share of academics source: Austrian Federal Bureau of Statistics (2001) [n.a.]; 15.26 / 3.72 / 40.79	+
age_ind	age index, calculated from 5-year age cohorts (2001) [years]; source: Austrian Federal Bureau of Statistics; 39.32 / 33.04 / 46.14	-
comm	commuter index to/from municipality, calculated from categories: 0 = commuting into the municipality, 1 = non-commuters, 2 = commuting within the municipality, 3 = commuting to other municipalities (2001) [n.a.]; source: Austrian Federal Bureau of Statistics; 1.82 / 0.66/ 2.76	o

dens	population density (2001) [inhabitants / hectare];
	source: Austrian Federal Bureau of Statistics;
	5.46 / 0.04 / 128.65
		+

Level 3: District

Name	Description (year) [unit]; source; mean / min. / max.	Exp. Eff.
wko_ind	WKO house price index, calculated as percentage of district-specific price per sq. m. by Austrian average (2008) [n.a.]; source: Autrian Federal Economic Chamber; 100 / 48.01 / 187.6	+

Table 4: *Neighborhood covariates on three levels with assumptions about the directions of the effects ("+": increasing, "-": decreasing and "o": no strong assumptions).*

The variables provided by these different sources have been combined and aggregated for suitable statistical analysis:

- Normalization (on hectares) in the case of population density (*dens*) and as a share of total population with the share of academics (*educ*).
- Constructing indexes as weighted means over different categories, where the age index (*age_ind*) is a weighted mean of 20 age cohorts, and the commuter index (*comm*) is a population weighted mean of 4 different commuting behaviors.

B Linear Effects

Name	cat.	Post. Mean	Std.-Dev.	95% CI	
cond_house					
method 1					
	medium**	-0.033	0.015	-0.003	0.049
	bad	0.009	0.028	-0.045	0.062
method 2					
	very good***	0.253	0.055	0.149	0.366
	good***	0.140	0.050	0.045	0.240
heat					
method 1					
	medium*	-0.026	0.015	-0.054	0.002
	bad***	-0.117	0.030	-0.178	-0.059
method 2					
	very good*	-0.069	0.038	-0.147	0.005
	good	-0.067	0.050	-0.162	0.035
	medium**	-0.114	0.056	-0.224	-0.006
	bad*	-0.125	0.065	-0.247	0.002
bath					
method 1					
	(very) good***	0.064	0.019	0.025	0.101
	bad**	-0.057	0.026	-0.108	-0.006
method 2					
	good	-0.065	0.041	-0.148	0.010
	medium	-0.079	0.054	-0.188	0.029
	bad**	-0.154	0.074	-0.302	-0.004
garage					
	high***	0.104	0.019	0.065	0.139
	medium***	0.052	0.013	0.027	0.076
marker		-0.108	0.069	-0.234	0.0164
attic_dum**		-0.026	0.011	-0.049	-0.003
cellar_dum***		0.098	0.015	0.069	0.125
terr_dum***		0.059	0.013	0.033	0.083
constant***		7.013	0.039	6.929	7.095

Table 5: *Estimation results of linear effects in the base model. Shown are the posterior means with standard deviations and 95% credible intervals. Significance is indicated as follows: * (significant at a 10% level), ** (significant at a 5% level), *** (significant at a 1% level). The signs and sizes of the effects are in line with expectations. Due to the logarithmic transformation of the response, the estimated effect can be approximately interpreted as semi-elasticity, i.e. the percentage change of price per sq. m. by the absolute change of the covariate, see Greene (2003).*

References

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

Belitz, C. and S. Lang (2008). Simultaneous selection of variables and smoothing parameters in structured additive regression models. *Computational Statistics and Data Analysis 53*, 61–81.

Bontemps, C., M. Simioni, and Y. Surry (2008). Semiparametric hedonic price models: assessing the effects of agricultural nonpoint source pollution. *Journal of Applied Econometrics 23*(6), 825–842.

Brezger, A., T. Kneib, and S. Lang (2005). BayesX: Analyzing Bayesian Structural Additive Regression Models. *Journal of Statistical Software 14*(11), 1–22.

Brezger, A., T. Kneib, and S. Lang (2009). BayesX manuals.

Brezger, A. and S. Lang (2006). Generalized structured additive regression based on Bayesian P-splines. *Computational Statistics and Data Analysis 50*, 947–991.

Can, A. (1998). GIS and Spatial Analysis of Housing and Mortgage Markets. *Journal of Housing Research 9*(1), 61–86.

Court, A. (1939). Hedonic price indexes with automotive examples. In A. S. Association (Ed.), *The Dynamics of Automobile Demand*, pp. 99–117. New York: General Motors.

DiPasquale, D. and W. C. Wheaton (1996). *Urban economics and real estate markets.* Upper Saddle River, New Jersey: Prentice Hall.

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

Ekeland, I., J. Heckman, and L. Nesheim (2004). Identication and Estimation of Hedonic Models. *Journal of Political Economy 112*(1), 60–109.

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

Follain, J. and E. Jimenez (1985). Estimating the Demand for Housing Characteristics: A Survey and Critique. *Regional Science and Urban Economics 15*(1), 77–107.

Frühwirth-Schnatter, S. and H. Wagner (2010). Stochastic model specification search for gaussian and partial non-gaussian state space models. *Journal of Econometrics 154*(1), 85 – 100.

Gelman, A. and J. Hill (2007). *Data Analysis Using Regression and Multilevel/Hierarchical Models.* New York: Cambridge University Press.

Goldstein, H. (2003). *Multilevel Statistical Models* (3 ed.). London: Hodder Arnold.

Greene, W. H. (2003). *Econometric analysis.* Prentice Hall Internat.

Griliches, Z. (1971). *Price Indexes and Quality Change Studies in New Methods of Measurement.* Cambridge, MA: Harvard University Press.

Krivobokova, T., T. Kneib, and G. Claeskens (2010). Simultaneous Confidence Bands for Penalized Spline Estimators. *Journal of the American Statistical Association In press.*

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. Hartgen, and T. Kneib (2010). Multilevel Structured Additive Regression. *Working paper*.

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

Martins-Filho, C. and O. Bin (2005). Estimation of hedonic price functions via additive nonparametric regression. *Empirical Economics* (30), 93–114.

Mason, C. and J. M. Quigley (1996). Non–parametric Hedonic Housing Prices. *Housing Studies 11*(3), 373–385.

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

Rigby, R. A. and D. M. Stasinopoulos (2005). Generalized additive models for location, scale and shape. *Applied Statistics 54*(3), 507–554.

Rosen, S. (1974). Hedonic Prices and Implicit Markets Product Differentiation in Pure Competition. *Journal of Political Economy 82*(1), 34–55.

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

Sirmans, G., D. Macpherson, and E. Zietz (2005). The Composition of Hedonic Pricing Models. *Journal of Real Estate Literature 13*(1), 3–43.

Spiegelhalter, D. J., N. G. Best, B. P. Carlin, and A. van der Linde (2002). Bayesian measures of model complexity and fit. *Journal of the Royal Statistical Society B 65*, 583–639.

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.

Wood, S. (2006). *An Introduction to Generalized Additive Models with R.* Boca Raton: Chapman and Hall.

Yue, Y. R. and H. Rue (2010). Bayesian inference for additive mixed quantile regression models. *Computational Statistics & Data Analysis In Press, Corrected Proof*, –.

University of Innsbruck – Working Papers in Economics and Statistics
Recent papers

2010-19 **Wolfgang Brunauer, Stefan Lang and Nikolaus Umlauf:** Modeling House Prices using Multilevel Structured Additive Regression

2010-18 **Martin Gächter and Engelbert Theurl:** Socioeconomic Environment and Mortality: A two-level Decomposition by Sex and Cause of Death

2010-17 **Boris Maciejovsky, Matthias Sutter, David V. Budescu and Patrick Bernau:** Teams Make You Smarter: Learning and Knowledge Transfer in Auctions and Markets by Teams and Individuals

2010-16 **Martin Gächter, Peter Schwazer and Engelbert Theurl:** Stronger sex but earlier death: A multi-level socioeconomic analysis of gender differences in mortality in Austria

2010-15 **Simon Czermak, Francesco Feri, Daniela Rützler and Matthias Sutter:** Strategic sophistication of adolescents – Evidence from experimental normal-form games

2010-14 **Matthias Sutter and Daniela Rützler:** Gender differences in competition emerge early in life

2010-13 **Matthias Sutter, Francesco Feri, Martin G. Kocher, Peter Martinsson, Katarina Nordblom and Daniela Rützler:** Social preferences in childhood and adolescence – A large-scale experiment

2010-12 **Loukas Balafoutas and Matthias Sutter:** Gender, competition and the efficiency of policy interventions

2010-11 **Alexander Strasak, Nikolaus Umlauf, Ruth Pfeiffer and Stefan Lang:** Comparing Penalized Splines and Fractional Polynomials for Flexible Modelling of the Effects of Continuous Predictor Variables

2010-10 **Wolfgang A. Brunauer, Sebastian Keiler and Stefan Lang:** Trading strategies and trading profits in experimental asset markets with cumulative information

2010-09 **Thomas Stöckl and Michael Kirchler:** Trading strategies and trading profits in experimental asset markets with cumulative information

2010-08 **Martin G. Kocher, Marc V. Lenz and Matthias Sutter:** Psychological pressure in competitive environments: Evidence from a randomized natural experiment: Comment

2010-07 **Michael Hanke and Michael Kirchler:** Football Championships and Jersey Sponsors' Stock Prices: An Empirical Investigation

2010-06 **Adrian Beck, Rudolf Kerschbamer, Jianying Qiu and Matthias Sutter:** Guilt from Promise-Breaking and Trust in Markets for Expert Services - Theory and Experiment

2010-05 **Martin Gächter, David A. Savage and Benno Torgler:** Retaining the Thin Blue Line: What Shapes Workers' Intentions not to Quit the Current Work Environment

2010-04 **Martin Gächter, David A. Savage and Benno Torgler:** The relationship between Stress, Strain and Social Capital

2010-03 **Paul A. Raschky, Reimund Schwarze, Manijeh Schwindt and Ferdinand Zahn:** Uncertainty of Governmental Relief and the Crowding out of Insurance

2010-02 **Matthias Sutter, Simon Czermak and Francesco Feri:** Strategic sophistication of individuals and teams in experimental normal-form games

2010-01 **Stefan Lang and Nikolaus Umlauf:** Applications of Multilevel Structured Additive Regression Models to Insurance Data

---

2009-29 **Loukas Balafoutas:** How much income redistribution? An explanation based on vote-buying and corruption. *Revised version forthcoming in Public Choice.*

2009-28 **Rudolf Kerschbamer, Matthias Sutter and Uwe Dulleck:** The Impact of Distributional Preferences on (Experimental) Markets for Expert Services

2009-27 **Adrian Beck, Rudolf Kerschbamer, Jianying Qiu and Matthias Sutter:** Car Mechanics in the Lab - Investigating the Behavior of Real Experts on Experimental Markets for Credence Goods

2009-26 **Michael Kirchler, Jürgen Huber and Thomas Stöckl:** Bubble or no Bubble - The Impact of Market Model on the Formation of Price Bubbles in Experimental Asset Markets

2009-25 **Rupert Sausgruber and Jean-Robert Tyran:** Tax Salience, Voting, and Deliberation

2009-24 **Gerald J. Pruckner and Rupert Sausgruber:** Honesty on the Streets - A Natural Field Experiment on Newspaper Purchasing

2009-23 **Gerlinde Fellner, Rupert Sausgruber and Christian Traxler:** Testing Enforcement Strategies in the Field: Legal Threat, Moral Appeal and Social Information

2009-22 **Ralph-C. Bayer, Elke Renner and Rupert Sausgruber:** Confusion and Reinforcement Learning in Experimental Public Goods Games

2009-21 **Sven P. Jost:** Transfer Pricing Risk Awareness of Multinational Corporations - Evidence from a Global Survey

2009-20 **Andrea M. Leiter and Engelbert Theurl:** The Convergence of Health Care Financing Structures: Empirical Evidence from OECD-Countries. *Revised version forthcoming in The European Journal of Health Economics.*

2009-19 **Francesco Feri and Miguel A. Meléndez-Jiménez:** Coordination in Evolving Networks with Endogenous Decay

2009-18 **Harald Oberhofer:** Firm growth, European industry dynamics and domestic business cycles

2009-17 **Jesus Crespo Cuaresma and Martin Feldkircher:** Spatial Filtering, Model Uncertainty and the Speed of Income Convergence in Europe

2009-16 **Paul A. Raschky and Manijeh Schwindt:** On the Channel and Type of International Disaster Aid

2009-15 **Jianying Qiu:** Loss aversion and mental accounting: The favorite-longshot bias in parimutuel betting

2009-14 **Siegfried Berninghaus, Werner Güth, M. Vittoria Levati and Jianying Qiu:** Satisficing in sales competition: experimental evidence

2009-13 **Tobias Bruenner, Rene Levinský and Jianying Qiu:** Skewness preferences and asset selection: An experimental study

2009-12 **Jianying Qiu and Prashanth Mahagaonkar:** Testing the Modigliani-Miller theorem directly in the lab: a general equilibrium approach

2009-11 **Jianying Qiu and Eva-Maria Steiger:** Understanding Risk Attitudes in two Dimensions: An Experimental Analysis

2009-10 **Erwann Michel-Kerjan, Paul A. Raschky and Howard C. Kunreuther:** Corporate Demand for Insurance: An Empirical Analysis of the U.S. Market for Catastrophe and Non-Catastrophe Risks

2009-09 **Fredrik Carlsson, Peter Martinsson, Ping Qin and Matthias Sutter:** Household decision making and the influence of spouses' income, education, and communist party membership: A field experiment in rural China

2009-08 **Matthias Sutter, Peter Lindner and Daniela Platsch:** Social norms, third-party observation and third-party reward

2009-07 **Michael Pfaffermayr:** Spatial Convergence of Regions Revisited: A Spatial Maximum Likelihood Systems Approach

2009-06 **Reimund Schwarze and Gert G. Wagner:** Natural Hazards Insurance in Europe – Tailored Responses to Climate Change Needed

2009-05 **Robert Jiro Netzer and Matthias Sutter:** Intercultural trust. An experiment in Austria and Japan

2009-04 **Andrea M. Leiter, Arno Parolini and Hannes Winner:** Environmental Regulation and Investment: Evidence from European Industries

2009-03 **Uwe Dulleck, Rudolf Kerschbamer and Matthias Sutter:** The Economics of Credence Goods: On the Role of Liability, Verifiability, Reputation and Competition. *Revised version forthcoming in American Economic Review.*

2009-02 **Harald Oberhofer and Michael Pfaffermayr:** Fractional Response Models - A Replication Exercise of Papke and Wooldridge (1996)

2009-01 **Loukas Balafoutas:** How do third parties matter? Theory and evidence in a dynamic psychological game.

2008-27 **Matthias Sutter, Ronald Bosman, Martin Kocher and Frans van Winden:** Gender pairing and bargaining – Beware the same sex! *Revised version published in Experimental Economics, Vol. 12 (2009): 318-331.*

2008-26 **Jesus Crespo Cuaresma, Gernot Doppelhofer and Martin Feldkircher:** The Determinants of Economic Growth in European Regions.

2008-25 **Maria Fernanda Rivas and Matthias Sutter:** The dos and don'ts of leadership in sequential public goods experiments.

2008-24 **Jesus Crespo Cuaresma, Harald Oberhofer and Paul Raschky:** Oil and the duration of dictatorships.

2008-23 **Matthias Sutter:** Individual behavior and group membership: Comment. *Revised Version published in American Economic Review, Vol.99 (2009): 2247-2257.*

2008-22 **Francesco Feri, Bernd Irlenbusch and Matthias Sutter:** Efficiency Gains from Team-Based Coordination – Large-Scale Experimental Evidence. *Revised and extended version forthcoming in American Economic Review.*

2008-21 **Francesco Feri, Miguel A. Meléndez-Jiménez, Giovanni Ponti and Fernando Vega Redondo:** Error Cascades in Observational Learning: An Experiment on the Chinos Game.

2008-20 **Matthias Sutter, Jürgen Huber and Michael Kirchler:** Bubbles and information: An experiment.

2008-19 **Michael Kirchler:** Curse of Mediocrity - On the Value of Asymmetric Fundamental Information in Asset Markets.

2008-18 **Jürgen Huber and Michael Kirchler:** Corporate Campaign Contributions as a Predictor for Abnormal Stock Returns after Presidential Elections.

2008-17 **Wolfgang Brunauer, Stefan Lang, Peter Wechselberger and Sven Bienert:** Additive Hedonic Regression Models with Spatial Scaling Factors: An Application for Rents in Vienna.

2008-16 **Harald Oberhofer, Tassilo Philippovich:** Distance Matters! Evidence from Professional Team Sports. *Extended and revised version forthcoming in Journal of Economic Psychology.*

2008-15 **Maria Fernanda Rivas and Matthias Sutter:** Wage dispersion and workers' effort.

2008-14 **Stefan Borsky and Paul A. Raschky:** Estimating the Option Value of Exercising Risk-taking Behavior with the Hedonic Market Approach. *Revised version forthcoming in Kyklos.*

2008-13 **Sergio Currarini and Francesco Feri:** Information Sharing Networks in Oligopoly.

2008-12 **Andrea M. Leiter:** Age effects in monetary valuation of mortality risks - The relevance of individual risk exposure. *Revised version forthcoming in The European Journal of Health Economics.*

2008-11 **Andrea M. Leiter and Gerald J. Pruckner:** Dying in an Avalanche: Current Risks and their Valuation.

2008-10 **Harald Oberhofer and Michael Pfaffermayr:** Firm Growth in Multinational Corporate Groups.

2008-09 **Michael Pfaffermayr, Matthias Stöckl and Hannes Winner:** Capital Structure, Corporate Taxation and Firm Age.

2008-08 **Jesus Crespo Cuaresma and Andreas Breitenfellner:** Crude Oil Prices and the Euro-Dollar Exchange Rate: A Forecasting Exercise.

2008-07 **Matthias Sutter, Stefan Haigner and Martin Kocher:** Choosing the carrot or the stick? – Endogenous institutional choice in social dilemma situations. Revised version forthcoming in *Review of Economic Studies*.

2008-06 **Paul A. Raschky and Manijeh Schwindt:** Aid, Catastrophes and the Samaritan's Dilemma.

2008-05 **Marcela Ibanez, Simon Czermak and Matthias Sutter:** Searching for a better deal – On the influence of group decision making, time pressure and gender in a search experiment. *Revised version published in Journal of Economic Psychology, Vol. 30 (2009): 1-10.*

2008-04 **Martin G. Kocher, Ganna Pogrebna and Matthias Sutter:** The Determinants of Managerial Decisions Under Risk.

2008-03 **Jesus Crespo Cuaresma and Tomas Slacik:** On the determinants of currency crises: The role of model uncertainty. *Revised version accepted for publication in Journal of Macroeconomics.*

2008-02 **Francesco Feri:** Information, Social Mobility and the Demand for Redistribution.

2008-01 **Gerlinde Fellner and Matthias Sutter:** Causes, consequences, and cures of myopic loss aversion - An experimental investigation. *Revised version published in The Economic Journal, Vol. 119 (2009), 900-916.*

2007-31 **Andreas Exenberger and Simon Hartmann:** The Dark Side of Globalization. The Vicious Cycle of Exploitation from World Market Integration: Lesson from the Congo.

2007-30 **Andrea M. Leiter and Gerald J. Pruckner:** Proportionality of willingness to pay to small changes in risk - The impact of attitudinal factors in scope tests. *Revised version forthcoming in Environmental and Resource Economics.*

2007-29 **Paul Raschky and Hannelore Weck-Hannemann:** Who is going to save us now? Bureaucrats, Politicians and Risky Tasks.

2007-28 **Harald Oberhofer and Michael Pfaffermayr:** FDI versus Exports. Substitutes or Complements? A Three Nation Model and Empirical Evidence.

2007-27 **Peter Wechselberger, Stefan Lang and Winfried J. Steiner:** Additive models with random scaling factors: applications to modeling price response functions.

2007-26 **Matthias Sutter:** Deception through telling the truth?! Experimental evidence from individuals and teams*. Revised version published in The Economic Journal, Vol. 119 (2009), 47-60.*

2007-25 **Andrea M. Leiter, Harald Oberhofer and Paul A. Raschky:** Productive disasters? Evidence from European firm level data. *Revised version forthcoming in Environmental and Resource Economics.*

2007-24 **Jesus Crespo Cuaresma:** Forecasting euro exchange rates: How much does model averaging help?

2007-23 **Matthias Sutter, Martin Kocher and Sabine Strauß:** Individuals and teams in UMTS-license auctions. *Revised version with new title "Individuals and teams in auctions" published in Oxford Economic Papers, Vol. 61 (2009): 380-394***).**

2007-22 **Jesus Crespo Cuaresma, Adusei Jumah and Sohbet Karbuz:** Modelling and Forecasting Oil Prices: The Role of Asymmetric Cycles. *Revised version accepted for publication in The Energy Journal.*

2007-21 **Uwe Dulleck and Rudolf Kerschbamer:** Experts vs. discounters: Consumer free riding and experts withholding advice in markets for credence goods. *Revised version published in International Journal of Industrial Organization*, *Vol. 27, Issue 1 (2009): 15-23.*

2007-20 **Christiane Schwieren and Matthias Sutter:** Trust in cooperation or ability? An experimental study on gender differences*. Revised version published in Economics Letters, Vol. 99 (2008): 494-497.*

2007-19 **Matthias Sutter and Christina Strassmair:** Communication, cooperation and collusion in team tournaments – An experimental study. *Revised version published in: Games and Economic Behavior, Vol.66 (2009), 506-525.*

2007-18 **Michael Hanke, Jürgen Huber, Michael Kirchler and Matthias Sutter:** The economic consequences of a Tobin-tax – An experimental analysis. *Revised version forthcoming in Journal of Economic Behavior and Organization.*

2007-17 **Michael Pfaffermayr:** Conditional beta- and sigma-convergence in space: A maximum likelihood approach. *Revised version forthcoming in Regional Science and Urban Economics.*

2007-16 **Anita Gantner:** Bargaining, search, and outside options. *Published in: Games and Economic Behavior, Vol. 62 (2008), pp. 417-435.*

2007-15 **Sergio Currarini and Francesco Feri:** Bilateral information sharing in oligopoly.

2007-14 **Francesco Feri:** Network formation with endogenous decay.

2007-13 **James B. Davies, Martin Kocher and Matthias Sutter:** Economics research in Canada: A long-run assessment of journal publications. *Revised version published in: Canadian Journal of Economics, Vol. 41 (2008), 22-45.*

2007-12 **Wolfgang Luhan, Martin Kocher and Matthias Sutter:** Group polarization in the team dictator game reconsidered. *Revised version published in: Experimental Economics, Vol. 12 (2009), 26-41.*

2007-11 **Onno Hoffmeister and Reimund Schwarze:** The winding road to industrial safety. Evidence on the effects of environmental liability on accident prevention in Germany.

2007-10 **Jesus Crespo Cuaresma and Tomas Slacik:** An “almost-too-late” warning mechanism for currency crises. *(Revised version accepted for publication in Economics of Transition***)**

2007-09 **Jesus Crespo Cuaresma, Neil Foster and Johann Scharler:** Barriers to technology adoption, international R&D spillovers and growth.

2007-08 **Andreas Brezger and Stefan Lang:** Simultaneous probability statements for Bayesian P-splines.

2007-07 **Georg Meran and Reimund Schwarze:** Can minimum prices assure the quality of professional services? (*Accepted for publication in European Journal of Law and Economics*)

2007-06 **Michal Brzoza-Brzezina and Jesus Crespo Cuaresma:** Mr. Wicksell and the global economy: What drives real interest rates?.

2007-05 **Paul Raschky:** Estimating the effects of risk transfer mechanisms against floods in Europe and U.S.A.: A dynamic panel approach.

2007-04 **Paul Raschky and Hannelore Weck-Hannemann:** Charity hazard - A real hazard to natural disaster insurance. *Revised version forthcoming in: Environmental Hazards.*

2007-03 **Paul Raschky:** The overprotective parent - Bureaucratic agencies and natural hazard management.

2007-02 **Martin Kocher, Todd Cherry, Stephan Kroll, Robert J. Netzer and Matthias Sutter:** Conditional cooperation on three continents. *Revised version published in: Economics Letters, Vol. 101 (2008): 175-178.*

2007-01 **Martin Kocher, Matthias Sutter and Florian Wakolbinger:** The impact of naïve advice and observational learning in beauty-contest games.

University of Innsbruck

Working Papers in Economics and Statistics

2010-19

Wolfgang Brunauer, Stefan Lang and Nikolaus Umlauf

Modeling House Prices using Multilevel Structured Additive Regression

Abstract

This paper analyzes house price data belonging to three hierarchical levels of spatial units. House selling prices with associated individual attributes (the elementary level-1) are grouped within municipalities (level-2), which form districts (level-3), which are themselves nested in counties (level-4). Additionally to individual attributes, explanatory covariates with possibly nonlinear effects are available on two of these spatial resolutions. We apply a multilevel version of structured additive regression (STAR) models to regress house prices on individual attributes and locational neighborhood characteristics in a four level hierarchical model. In multilevel STAR models the regression coefficients of a particular nonlinear term may themselves obey a regression model with structured additive predictor. The framework thus allows to incorporate nonlinear covariate effects and time trends, smooth spatial effects and complex interactions at every level of the hierarchy of the multilevel model. Moreover we are able to decompose the spatial heterogeneity effect and investigate its magnitude at different spatial resolutions allowing for improved predictive quality even in the case of unobserved spatial units. Statistical inference is fully Bayesian and based on highly efficient Markov chain Monte Carlo simulation techniques that take advantage of the hierarchical structure in the data.

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