Hierarchical Structured Additive Regression

Lang, Stefan[1] Umlauf, Nikolaus[2] Brunauer, Wolfgang[3]

[1 2]University of Innsbruck, Austria
[3]Immobilien Rating GmbH, Austria

July 2010

1. Example of hierarchical data structures

2. Structured additive regression models

3. Hierarchical formulation and MCMC inference

4. Alternative sampling scheme based on transformed parametrization

5. Results: Hedonic regression data for house prices

Example of hierarchical data structures

Hedonic regression data for house prices in Austria

Variable of primary interest

house price or log house price

Covariates

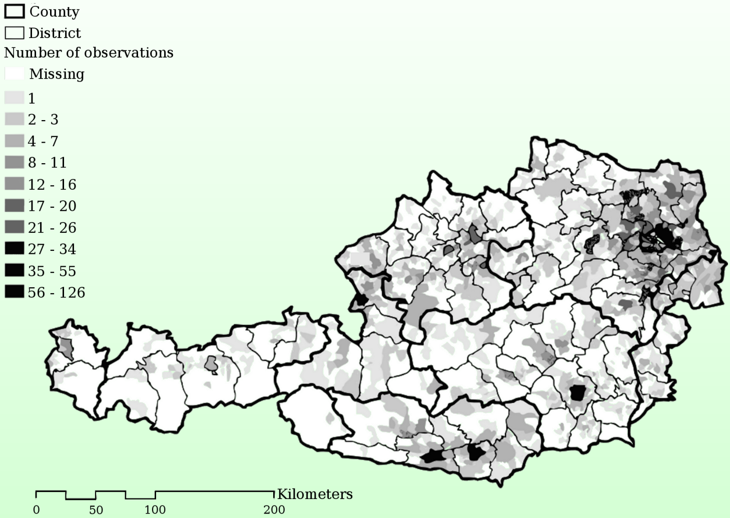
- Structural (physical) characteristics, like floor space area, constructional condition, age etc., and
- neighborhood (locational) characteristics, often on various levels of aggregation, like the proximity to places of work, the social composition of the neighborhood etc.

Four-level hierarchical model

$$
\begin{array}{llcl}
\text{level 1:} & \texttt{lnp} & = & f_1(\texttt{area}) + \cdots + f_q(\texttt{age}) + \mathbf{v}\boldsymbol{\gamma} + f_{\texttt{municipal}}(\texttt{s}_1) + \boldsymbol{\varepsilon}_1 \\
\text{level 2:} & f_{\texttt{municipal}}(s_1) & = & f_{1_1}(\texttt{purchase power}) + \cdots + f_{p_1}(\texttt{level of education}) \\
 & & & + f_{\texttt{district}}(\texttt{s}_2) + \boldsymbol{\varepsilon}_2 \\
\text{level 3:} & f_{\texttt{district}}(\texttt{s}_2) & = & f_{1_2}(\texttt{unemployment rate}) + f_{\texttt{county}}(\texttt{s}_3) + \boldsymbol{\varepsilon}_3 \\
\text{level 4:} & f_{\texttt{county}}(\texttt{s}_3) & = & \boldsymbol{\varepsilon}_4
\end{array}
$$

The f's are possibly nonlinear functions of the covariates.

This is an example of *hierarchical structured additive regression models*.


County
District
Number of observations
Missing
1
2 - 3
4 - 7
8 - 11
12 - 16
17 - 20
21 - 26
27 - 34
35 - 55
56 - 126
Kilometers
0
50
100
200

Structured additive regression models

- Distributional and structural assumptions, given covariates and parameters, are based on Generalized Linear Models
- $E(y|\mathbf{x}, \mathbf{v}) = h(\eta)$ with structured additive predictor

$$\eta = f_1(x_1) + \ldots + f_p(x_p) + \mathbf{v}'\boldsymbol{\gamma}$$

  In the following we only consider additive models with

$$y = \eta + \varepsilon \qquad \varepsilon \sim N(0, \sigma^2)$$

  - $\mathbf{v}'\boldsymbol{\gamma}$ parametric part of the predictor
  - x_j continuous covariate, time scale, location or unit-or cluster index
  - x_j may be two (even higher) dimensional for modeling interactions
  - f_j one-/two (even higher) dimensional, not necessarily continuous functions

Overview: Modeling the functions f_j

$f_j(x_j) = f(x)$	$x_j = x$	nonlinear effect of x
$f_j(x_j) = f_{spat}(s)$	$x_j = s$	spatial effect of location variable $\mathbf{s} = (1, 2, \ldots, S)'$
$f_j(x_j) = x_2\, f(x_1)$	$x_j = (x_1, x_2)$	interaction effect between x_1 and x_2
$f_j(x_j) = f_{1\|2}(x_1, x_2)$	$x_j = (x_1, x_2)$	nonlinear interaction between x_1 and x_2
$f_j(x_j) = \beta_i u$	$x_j = (u, i)$	individual specific random effect with $\mathbf{u} = (1, 2, \ldots, U)'$

General form

- Vector of function evaluations $\mathbf{f}_j = (f_{1j}, \ldots, f_{nj})'$ can be written as:

$$\mathbf{f}_j = \mathbf{Z}_j \boldsymbol{\beta}_j$$

  with $\mathbf{Z}_j$ as the design matrix, where $\boldsymbol{\beta}_j$ are unknown regression coefficients

- Form of $\mathbf{Z}_j$ only depends on the functional type chosen

- Penalized least squares:

$$\text{PLS}(\boldsymbol{\beta}, \boldsymbol{\gamma}) = ||\mathbf{y} - \boldsymbol{\eta}||^2 + \lambda_1 \boldsymbol{\beta}_1' \mathbf{K}_1 \boldsymbol{\beta}_1 + \ldots + \lambda_p \boldsymbol{\beta}_p' \mathbf{K}_p \boldsymbol{\beta}_p$$

General form

- Prior for $\boldsymbol{\beta}$ in the corresponding Bayesian approach

$$p(\boldsymbol{\beta}_j|\tau_j^2) \propto \left(\frac{1}{2\pi\tau_j^2}\right)^{rk(\mathbf{K}_j)/2} exp\left(-\frac{1}{2\tau_j^2}\boldsymbol{\beta}_j'\mathbf{K}_j\boldsymbol{\beta}_j\right) I(\mathbf{A}\boldsymbol{\beta}_j = \mathbf{0})$$

  τ_j^2 variance parameter, governs the smoothness of f_j, relation to frequentists by $\lambda_j = \sigma^2/\tau_j^2$

- $\mathbf{A}\boldsymbol{\beta}_j = \mathbf{0}$ is an identifiability constraint, e.g. $\mathbf{A} = (1, \ldots, 1)'$ such that the $\boldsymbol{\beta}$'s sum up to zero

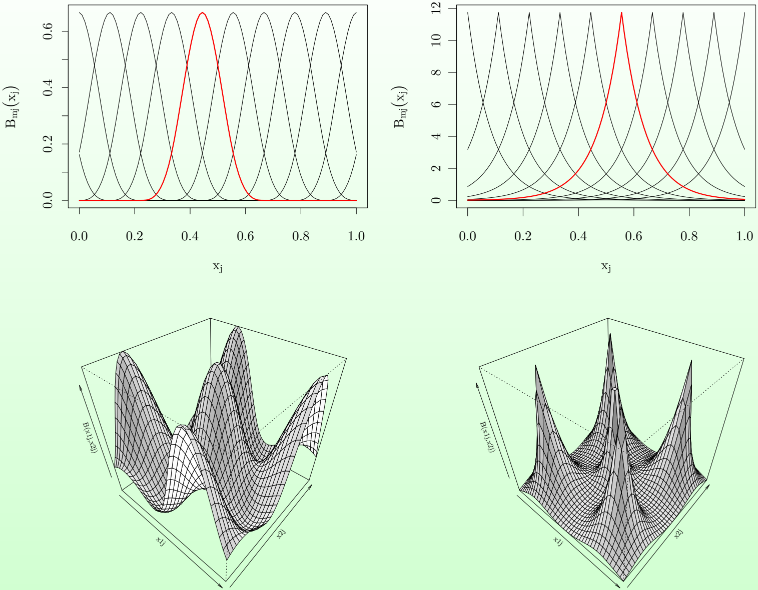
- Structure of $\mathbf{K}_j$ also depends on the type of covariates and on assumptions about smoothness of $\mathbf{f}_j$

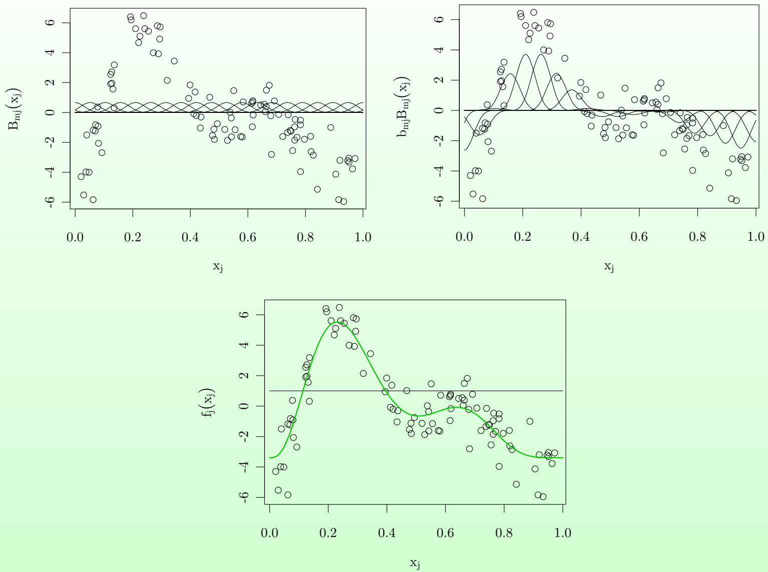
General form

- Basis functions $B_{mj}(\cdot)$ in

$$f_j(x_j) = \sum_{m=1}^{M_j} \beta_{mj} B_{mj}(x_j)$$

  may include e.g. a polynomial, B-spline, Matérn basis (one or more dimensional), etc.


0.0
0.2
0.4
0.6
$B_{mj}(x_j)$
0.0 0.2 0.4 0.6 0.8 1.0
x_j
0 2 4 6 8 10 12
$B_{mj}(x_j)$
0.0 0.2 0.4 0.6 0.8 1.0
x_j
B(x1j,x2j)
x1j
x2j
B(x1j,x2j)
x1j
x2j



$B_{mj}(x_j)$
$b_{mj}B_{mj}(x_j)$
$f_j(x_j)$
x_j

Hierarchical formulation and MCMC inference

Multilevel/Hierarchical structured additive model with k hierarchies within a first stage term $\mathbf{Z}_j\boldsymbol{\beta}_j$ may be written as

$$
\begin{aligned}
\mathbf{y} &= \mathbf{Z}_1\boldsymbol{\beta}_1 + \ldots + \mathbf{Z}_p\boldsymbol{\beta}_p + \mathbf{v}\boldsymbol{\gamma} + \boldsymbol{\varepsilon} \\
\boldsymbol{\beta}_j &= \mathbf{Z}_{j1_1}\boldsymbol{\beta}_{j1_1} + \ldots + \mathbf{Z}_{jp_1}\boldsymbol{\beta}_{jp_1} + \mathbf{v}_j\boldsymbol{\gamma}_j + \mathbf{u}_j \\
&\;\vdots \\
\boldsymbol{\beta}_{j,j_1,\ldots,j_k} &= \mathbf{z}_{j,j_1,\ldots,j_k}\boldsymbol{\beta}_{j,j_1,\ldots,j_k} + \ldots + \mathbf{z}_{j,j_1,\ldots,j_k}\boldsymbol{\beta}_{j,j_1,\ldots,j_k} + \mathbf{v}_{j,j_1,\ldots,j_k}\boldsymbol{\gamma}_{j,j_1,\ldots,j_k} + \mathbf{u}_{j,j_1,\ldots,j_k} \\
\boldsymbol{\beta}_{j,j_1,\ldots,j_k} &= \boldsymbol{\eta}_{j,j_1,\ldots,j_k} + \mathbf{u}_{j,j_1,\ldots,j_k}
\end{aligned}
$$

with $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{W}^{-1})$ and $\mathbf{u}_{j,j_1,\ldots,j_k} \sim N(\mathbf{0}, \tau^2_{j,j_1,\ldots,j_k}\mathbf{K}^{-1}_{j,j_1,\ldots,j_k})$

The full conditionals for the regression coefficients are multivariate Gaussian. Starting from a first level view, the precision matrix $\boldsymbol{\Sigma}_{\boldsymbol{\beta}_j}$ and mean $\boldsymbol{\mu}_{\boldsymbol{\beta}_j}$ are given by

$$\begin{aligned}\boldsymbol{\Sigma}_{\boldsymbol{\beta}_j} &= \sigma^2 \left(\mathbf{Z}_j' \mathbf{W} \mathbf{Z}_j + \frac{\sigma^2}{\tau_j^2} \mathbf{K}_j \right)^{-1} \\ \boldsymbol{\mu}_{\boldsymbol{\beta}_j} &= \boldsymbol{\Sigma}_{\boldsymbol{\beta}_j} \left(\frac{1}{\sigma^2} \mathbf{Z}_j' \mathbf{W} \mathbf{r} + \frac{1}{\tau_j^2} \boldsymbol{\eta}_{\boldsymbol{\beta}_j} \right)\end{aligned}$$

and for the higher levels

$$\begin{aligned}\boldsymbol{\Sigma}_{\boldsymbol{\beta}_{j,j_1,\ldots,j_k}} &= \tau^2_{j,j_1,\ldots,j_{k-1}} \left(\mathbf{Z}'_{j,j_1,\ldots,j_k} \mathbf{Z}_{j,j_1,\ldots,j_k} + \frac{\tau^2_{j,j_1,\ldots,j_{k-1}}}{\tau^2_{j,j_1,\ldots,j_k}} \mathbf{K}_{j,j_1,\ldots,j_k} \right)^{-1} \\ \boldsymbol{\mu}_{\boldsymbol{\beta}_{j,j_1,\ldots,j_k}} &= \boldsymbol{\Sigma}_{\boldsymbol{\beta}_{j,j_1,\ldots,j_k}} \left(\frac{1}{\tau^2_{j,j_1,\ldots,j_{k-1}}} \mathbf{Z}'_{j,j_1,\ldots,j_k} \mathbf{r} + \frac{1}{\tau^2_{j,j_1,\ldots,j_k}} \boldsymbol{\eta}_{\boldsymbol{\beta}_{j,j_1,\ldots,j_k}} \right)\end{aligned}$$

Properties

- *Reduced complexity in higher stages of the hierarchy:*
  - Number of "observations" in the higher levels is much less than the actual number of observations n.
  - Full conditionals for regression coefficients are Gaussian regardless of the response distribution in the first level of the hierarchy.
- *Sparsity*
  Design matrices and posterior precision matrices are typically sparse (after reordering of parameters).
- *Number of different observations smaller than sample size*
  Typically the number of different observations $x_{j(1_j)}, \ldots, x_{j(n_j)}$ in $\mathbf{Z}_j$ is much smaller than the total number n of observations, i.e. $n_j \ll n$.

- Denote by $z^{(2)}_{(1)} < z^{(2)}_{(2)} < \cdots < z^{(2)}_{(m)}$ the m ordered different observations of $z^{(2)}$.
- Compute the index vector **ind** with elements $\mathbf{ind}[i] \in \{1, \ldots, m\}$ denoting the category of the i-th observation, i.e. if $z^{(2)}_i = z^{(2)}_{(j)}$ then $\mathbf{ind}[i] = j$.
- Decompose the design matrix in $\mathbf{Z} = \mathbf{D}\mathbf{P}\tilde{\mathbf{Z}}$ where
  - $\tilde{\mathbf{Z}}$ is the $m \times K$ reduced design matrix for the different and sorted observations $z^{(2)}_{(1)}, \ldots, z^{(2)}_{(m)}$, i.e. $\tilde{\mathbf{Z}}[s,k] = B_k\left(z^{(2)}_{(s)}\right)$, $s = 1, \ldots, m$, $k = 1, \ldots, K$,
  - $\mathbf{P}$ is a $n \times m$ permutation matrix, which reverts the sorting, i.e. $\mathbf{P}[i,s] = I(\mathbf{ind}(i) = s)$.
- For the vector of function evaluations we obtain $\mathbf{f} = \mathbf{Z}\boldsymbol{\beta} = \mathbf{D}\mathbf{P}\tilde{\mathbf{Z}}\boldsymbol{\beta}$.

We get

$$\mathbf{Z}'\mathbf{W}\mathbf{Z} = \tilde{\mathbf{Z}}'\mathbf{P}'\mathbf{D}'\mathbf{W}\mathbf{D}\mathbf{P}\tilde{\mathbf{Z}} = \tilde{\mathbf{Z}}'\tilde{\mathbf{W}}\tilde{\mathbf{Z}},$$

where

$$\tilde{\mathbf{W}} = \mathbf{P}'\mathbf{D}'\mathbf{W}\mathbf{D}\mathbf{P} = \mathrm{diag}(\tilde{w}_1, \ldots, \tilde{w}_m)$$

and the "reduced" weights $\tilde{w}_s$, are given by

$$\tilde{w}_s = \sum_{i\,:\,\boldsymbol{ind}[i]=s} \left(\left(z_i^{(1)} \right)^2 w_i. \right.$$

The weights $\tilde{w}_s$ can be computed by first initializing $\tilde{w}_s = 0$ followed by a simple loop: For $i = 1, \ldots, n$ add $\left(\left(z_i^{(1)} \right)^2 w_i \right.$ to $\tilde{w}_{\boldsymbol{ind}[i]}$.

For $\mathbf{Z}'\mathbf{W}\mathbf{r}$ we obtain

$$\mathbf{Z}'\mathbf{W}\mathbf{r} = \tilde{\mathbf{Z}}'\mathbf{P}'\mathbf{D}'\mathbf{W}\mathbf{r} = \tilde{\mathbf{Z}}'\tilde{\mathbf{r}},$$

where the $m \times 1$ vector $\tilde{\mathbf{r}} = (\tilde{r}_1, \ldots, \tilde{r}_m)'$ of "reduced" partial residuals is given by

$$\tilde{r}_s = \sum_{i\,:\,\boldsymbol{ind}[i]=s} z_i^{(1)}\, w_i\, r_i.$$

The $\tilde{r}_s$ are computed by first initializing $\tilde{r}_s = 0$ followed by the loop: For $i = 1, \ldots, n$ add $z_i^{(1)} w_i\, r_i$ to $\tilde{r}_{ind(i)}$.

Alternative sampling scheme based on transformed parametrization

(i.) Cholesky decomposition $\mathbf{R}\mathbf{R}'$ of $\mathbf{Z}'\mathbf{W}\mathbf{Z}$

(ii.) Singular value decomposition $\mathbf{Q}\mathbf{S}\mathbf{Q}' = \mathbf{R}^{-1}\mathbf{K}(\mathbf{R}')^{-1}$,
$\mathbf{S} = diag(s_1, \dots, s_M)$: Eigenvalues of $(\mathbf{R}')^{-1}\mathbf{K}(\mathbf{R}')^{-1}$
$\mathbf{Q}$: Orthogonalmatrix

(iii.) Then set transformed design matrix $\tilde{\mathbf{Z}} = \mathbf{Z}(\mathbf{R}')^{-1}\mathbf{Q}$ such that
$\mathbf{f} = \mathbf{Z}\boldsymbol{\beta} = \tilde{\mathbf{Z}}\tilde{\boldsymbol{\beta}}$ $(\boldsymbol{\beta} = (\mathbf{R}')^{-1}\mathbf{Q}\tilde{\boldsymbol{\beta}})$

(iv.) and the resulting penalty is now given by
$\boldsymbol{\beta}'\mathbf{K}\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}}'\mathbf{Q}'(\mathbf{R}')^{-1}\mathbf{K}(\mathbf{R}')^{-1}\mathbf{Q}\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}'\mathbf{S}\tilde{\boldsymbol{\beta}}$

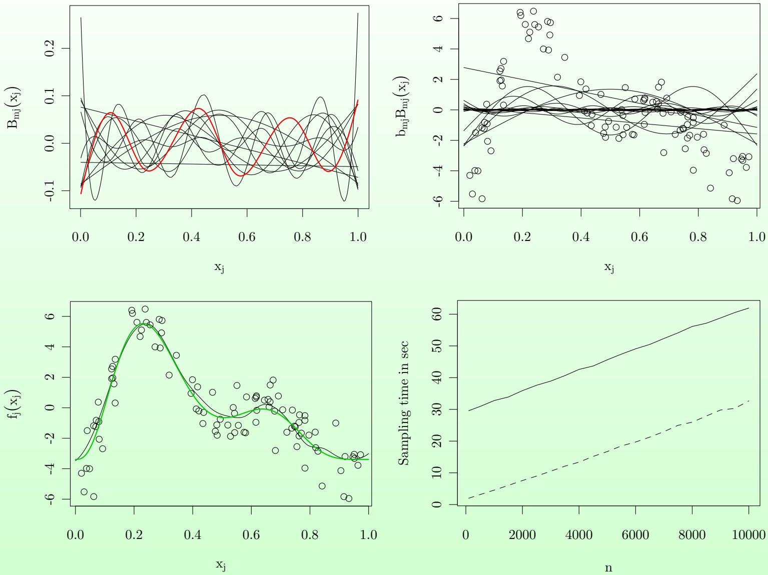
Mean and precision matrix are now given by

$$\mu_{\tilde{\beta}_{mj}} = \frac{1}{1+\lambda_j s_{mj}} \cdot u_{mj} \qquad m = 1, \ldots, M_j$$

where $\lambda_j = \sigma^2/\tau_j^2$ and u_{mj} is the m-th element of the vector $\mathbf{u}_j = \tilde{\mathbf{Z}}_j \mathbf{W} (\mathbf{y} - \boldsymbol{\eta} + \mathbf{f}_j)$, and entries of the corresponding diagonal precision matrix

$$\boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}_j}[m,m] = \frac{\sigma^2}{1+\lambda_j s_{mj}} \qquad m = 1, \ldots, M_j$$

Alternative sampling scheme based on transformed parametrization



MCMC sampling scheme

for $t = 1, \ldots, T$ {

1. for $j = 1, \ldots, p$ {

   1.1 $\tilde{\boldsymbol{\beta}}_j^{(t+1)} | \cdot \sim N\left(\mu_{\tilde{\beta}_j}^{(t)}, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\beta}}_j}^{(t)}\right)$

   1.2 if level within $\tilde{\boldsymbol{\beta}}_j$ set $\mathbf{y}^* = \tilde{\boldsymbol{\beta}}_j^{(t+1)}$ and repeat steps 1-4

   1.3 $\tau_j^{2\,(t+1)} | \cdot \sim IG\left(a + \frac{rk(\mathbf{K}_j)}{2}, b + \frac{1}{2}\tilde{\boldsymbol{\beta}}_j'^{\,(t+1)} \mathbf{K}_j \tilde{\boldsymbol{\beta}}_j^{(t+1)}\right)$

   1.4 update $\boldsymbol{\eta}$

   }

2. $\tilde{\boldsymbol{\gamma}}^{(t+1))} | \cdot \sim N\left(\boldsymbol{\mu}_{\tilde{\boldsymbol{\gamma}}}^{(t)}, \boldsymbol{\Sigma}_{\tilde{\boldsymbol{\gamma}}}^{(t)}\right)$

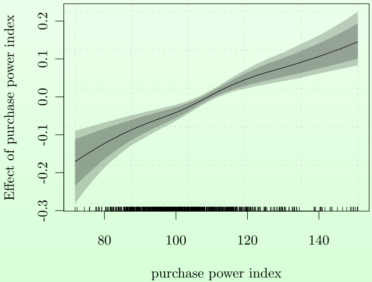
3. update $\boldsymbol{\eta}$

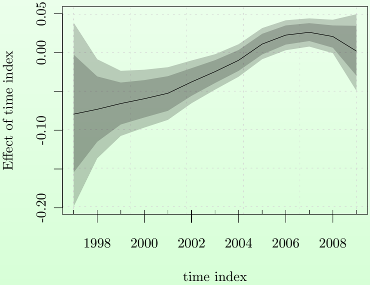
4. $\sigma^{2\,(t+1)} | \cdot \sim IG\left(a + \frac{n}{2}, b + \frac{1}{2}(\mathbf{y} - \boldsymbol{\eta}^{(t+1)})'(\mathbf{y} - \boldsymbol{\eta}^{(t+1)})\right)$

}

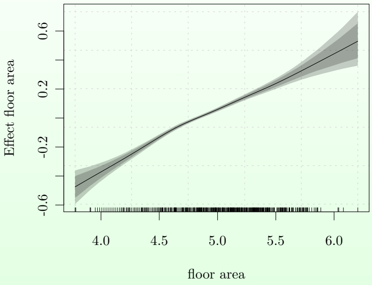
Results: Hedonic regression data for house prices

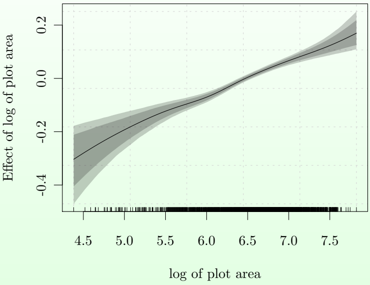
Structural continuous covariates

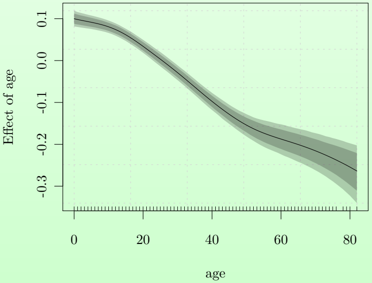




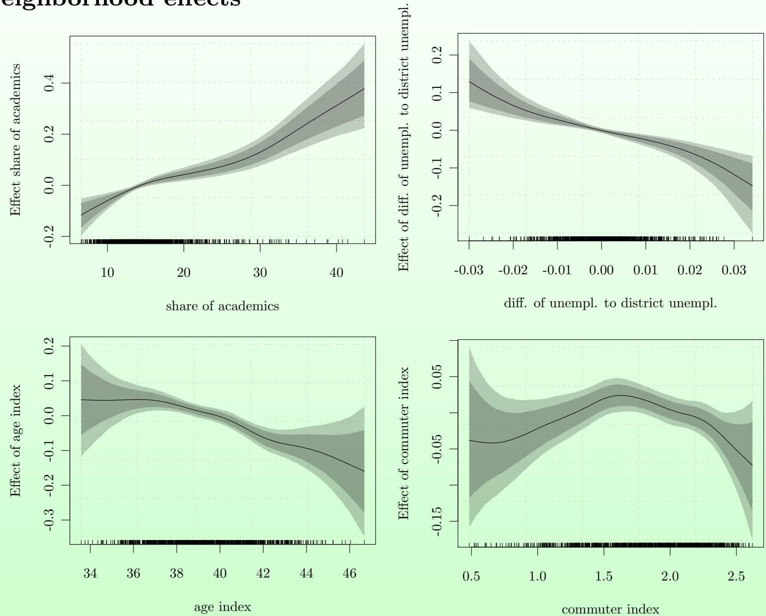
Structural continuous covariates



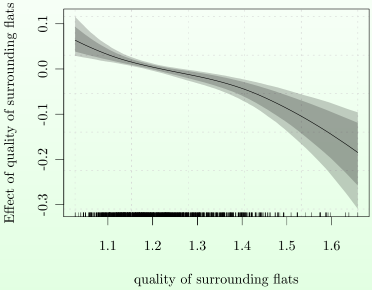


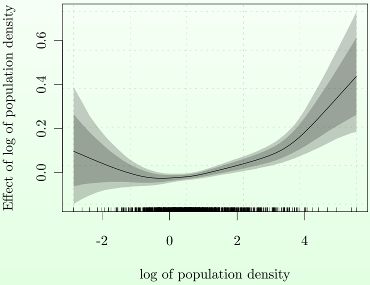


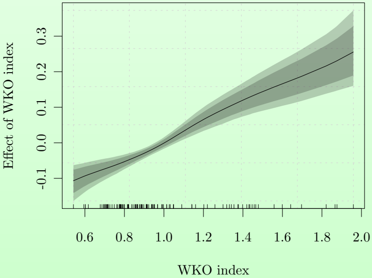
Neighborhood effects



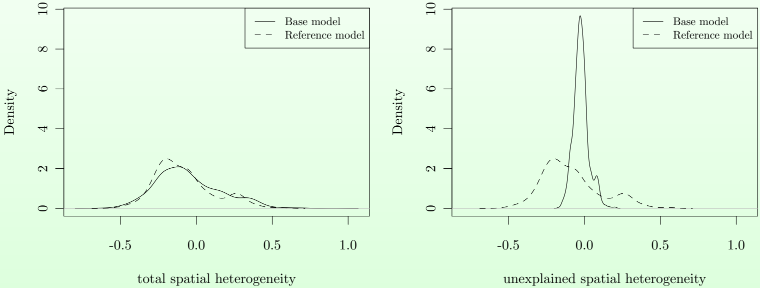
Neighborhood effects



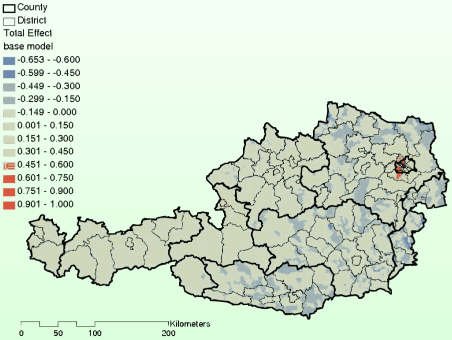


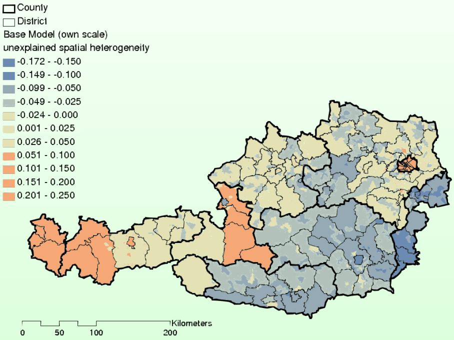


Neighborhood effects



Results: Hedonic regression data for house prices





Thank you!!!

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

Brunauer, W., Lang, S. and Umlauf, N. (2010). Modeling House Prices using Multilevel Structured Additive Regression. Technical report, University of Innsbruck.

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

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

Lang, S., Umlauf, N., Kneib, T., Hartgen, K. and Wechselberger, P. (2010): Multilevel Generalized Structured Additive Regression. Technical report, University of Innsbruck