Multilevel Structured Additive Regression

Lang, Stefan[1] Umlauf, Nikolaus[2] Wechselberger, Peter[3]
Kneib, Thomas[4]

[1 2 3]University of Innsbruck, Austria
[4]Carl von Ossietzky Universität Oldenburg, Germany

March 2010

1. Example of multilevel/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 multilevel/hierarchical data structures

Hedonic regression data for house prices in Austria

Variable of primary interest

house price or log house price

Covariates

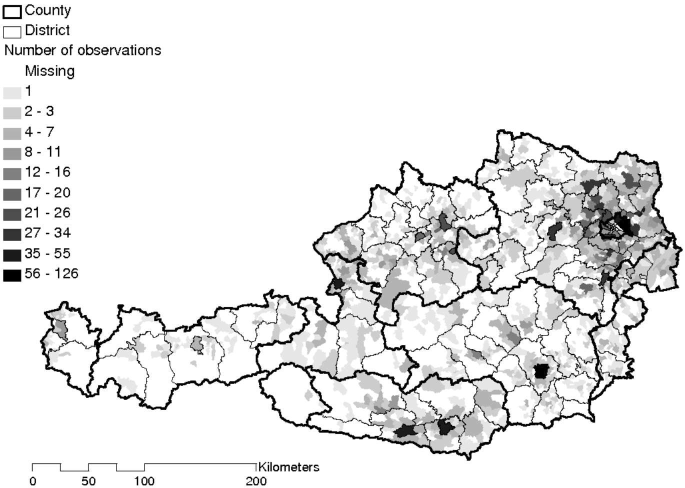
- Structural (physical) characteristics, like floor space area, constructional condition, age etc., and
- neighborhood (locational) characteristics, often on various levels of aggregation, like the proximity to places of work, the social composition of the neighborhood etc.

Four-level hierarchical model

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

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

This is an example of *multilevel/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
- $E(\mathbf{y}|\mathbf{x}, \mathbf{v}) = h(\boldsymbol{\eta})$ with structured additive predictor

$$\boldsymbol{\eta} \quad = \quad f_1(\mathbf{x}_1) + \ldots + f_p(\mathbf{x}_p) + \mathbf{v}\gamma$$

  In the following we only consider additive models with

$$\mathbf{y} = \boldsymbol{\eta} + \boldsymbol{\varepsilon} \qquad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{W}^{-1})$$

  - $\mathbf{v}\gamma$ parametric part of the predictor
  - $\mathbf{x}_j$ continuous covariate, time scale, location or unit-or cluster index
  - $\mathbf{x}_j$ 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

$f_j(\mathbf{x}_j) = f(\mathbf{x})$	$\mathbf{x}_j = \mathbf{x}$	nonlinear effect of $\mathbf{x}$
$f_j(\mathbf{x}_j) = f_{spat}(\mathbf{s})$	$\mathbf{x}_j = \mathbf{s}$	spatial effect of location variable $\mathbf{s} = (1, 2, \dots, S)'$
$f_j(\mathbf{x}_j) = diag(\mathbf{x}_2) f(\mathbf{x}_1)$	$\mathbf{x}_j = (\mathbf{x}_1, \mathbf{x}_2)$	interaction effect between $\mathbf{x}_1$ and $\mathbf{x}_2$
$f_j(\mathbf{x}_j) = f_{1\|2}(\mathbf{x}_1, \mathbf{x}_2)$	$\mathbf{x}_j = (\mathbf{x}_1, \mathbf{x}_2)$	nonlinear interaction between $\mathbf{x}_1$ and $\mathbf{x}_2$
$f_j(\mathbf{x}_j) = \mathbf{Z}\boldsymbol{\beta}$	$\mathbf{x}_j = (\mathbf{u}, \mathbf{x})$	individual specific random effect with design matrix $\mathbf{Z}$ of covariate $\mathbf{u} = (1, 2, \dots, U)'$ and/or possible $\mathbf{x}$

General form

- Vector of function evaluations can be written as:

$$\mathbf{f}_j = \mathbf{Z}_j \boldsymbol{\beta}_j = f_j(\mathbf{x}_j)$$

  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:

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

General form

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

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

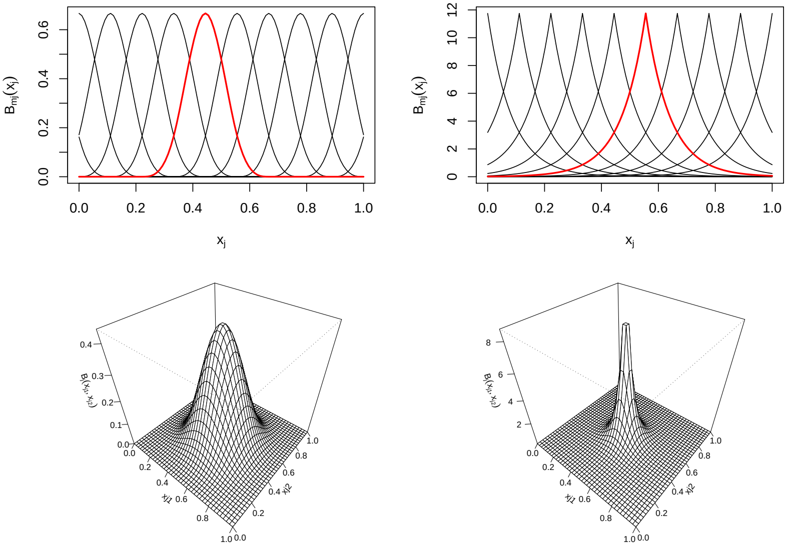
  τ_j^2 variance parameter, governs the smoothness of f_j, relation to frequentists by $\lambda_j = \sigma^2/\tau_j^2$
- $\mathbf{A}\boldsymbol{\beta}_j = \mathbf{0}$ is an identifiability constraint, e.g. $\mathbf{A} = (1, \dots, 1)'$ such that the $\boldsymbol{\beta}$'s sum up to zero
- Structure of $\mathbf{K}_j$ also depends on the type of covariates and on assumptions about smoothness of $\mathbf{f}_j$

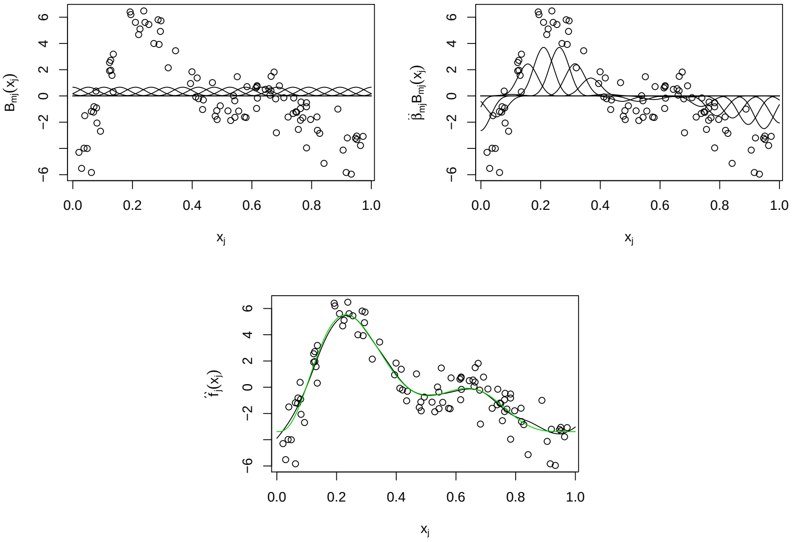
General form

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

$$\mathbf{f}_j = \sum_{m=1}^{M_j} \beta_{mj} B_{mj}(\mathbf{x}_j)$$

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


B_mj(x_j)
0.0
0.2
0.4
0.6
0.0 0.2 0.4 0.6 0.8 1.0
x_j
B_mj(x_j)
0 2 4 6 8 10 12
0.0 0.2 0.4 0.6 0.8 1.0
x_j
B_l(x_j1, x_j2)
0.0 0.1 0.2 0.3 0.4
xj1
xj2
B_l(x_j1, x_j2)
2 4 6 8
xj1
xj2



B_mj(x_j)
β̂_mj B_mj(x_j)
f̂_j(x_j)
x_j
x_j
x_j
−6
−2
0
2
4
6
0.0
0.2
0.4
0.6
0.8
1.0

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

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

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 $x_{j(1_j)}, \dots, x_{j(n_j)}$ in $\mathbf{Z}_j$ is much smaller than the total number n of observations, i.e. $n_j \ll n$.

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 $\mathbf{Q}\mathbf{S}\mathbf{Q}' = \mathbf{R}^{-1}\mathbf{K}(\mathbf{R}')^{-1}$,
$\mathbf{S} = diag(s_1, \ldots, s_M)$: 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
$\boldsymbol{\beta}'\mathbf{K}\boldsymbol{\beta} = \tilde{\boldsymbol{\beta}}'\mathbf{Q}'(\mathbf{R}')^{-1}\mathbf{K}(\mathbf{R}')^{-1}\mathbf{Q}\tilde{\boldsymbol{\beta}} = \tilde{\boldsymbol{\beta}}'\mathbf{S}\tilde{\boldsymbol{\beta}}$

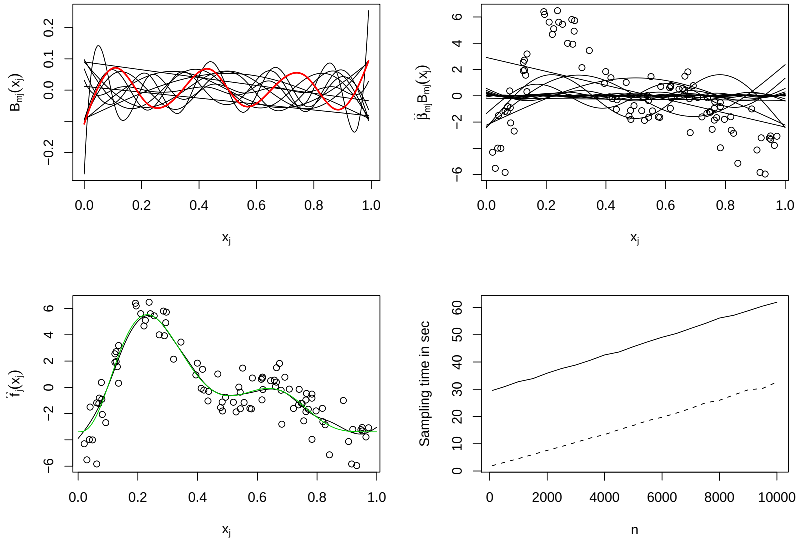
Mean and precision matrix are now given by

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

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

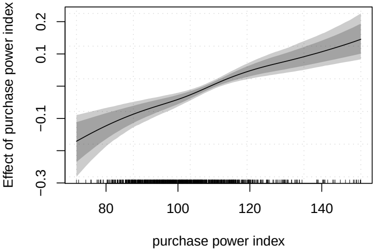
   }

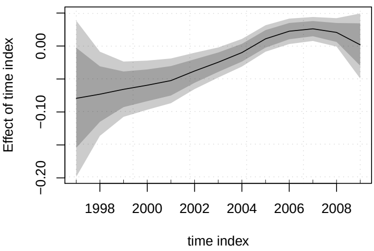
2. $\tilde{\gamma}^{(t+1))} | \cdot \sim N\left(\boldsymbol{\mu}_{\tilde{\gamma}}^{(t)}, \boldsymbol{\Sigma}_{\tilde{\gamma}}^{(t)}\right)$
3. update $\boldsymbol{\eta}$
4. $\sigma^{2^{(t+1)}} | \cdot \sim IG\left(a + \frac{n}{2}, b + \frac{1}{2}(\mathbf{y} - \boldsymbol{\eta}^{(t+1)})'(\mathbf{y} - \boldsymbol{\eta}^{(t+1)})\right)$

}

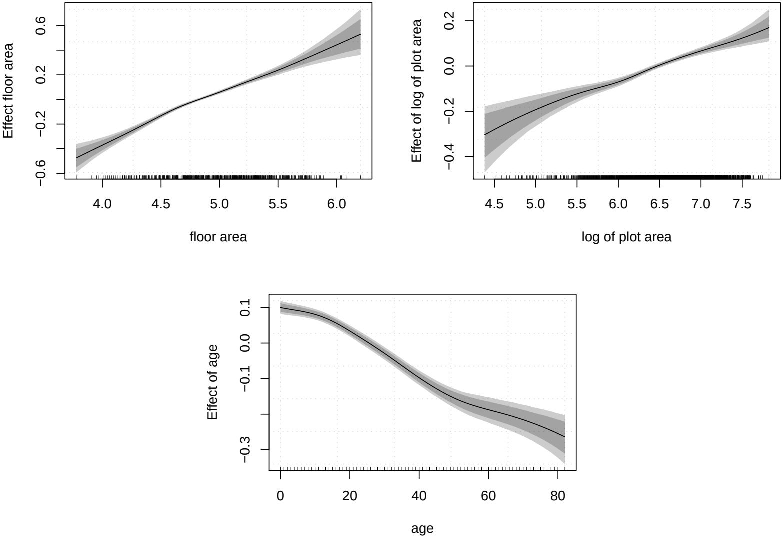
Results: Hedonic regression data for house prices

Structural continuous covariates

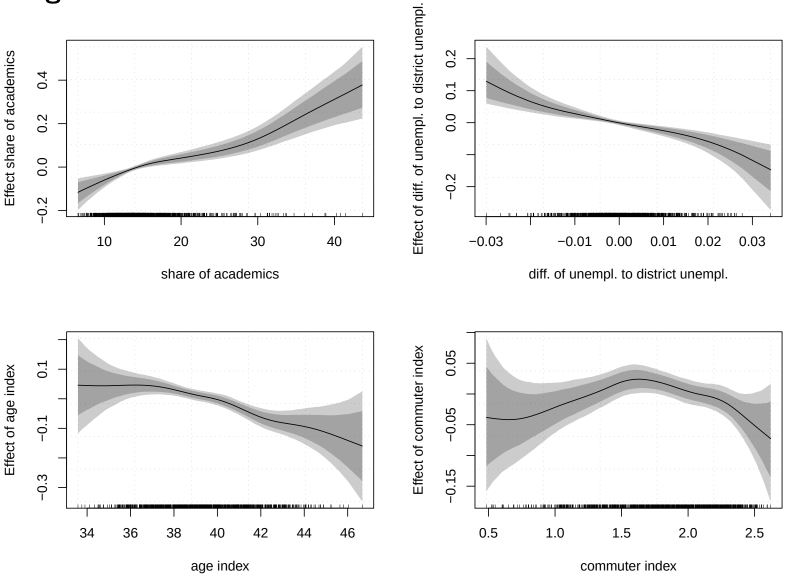




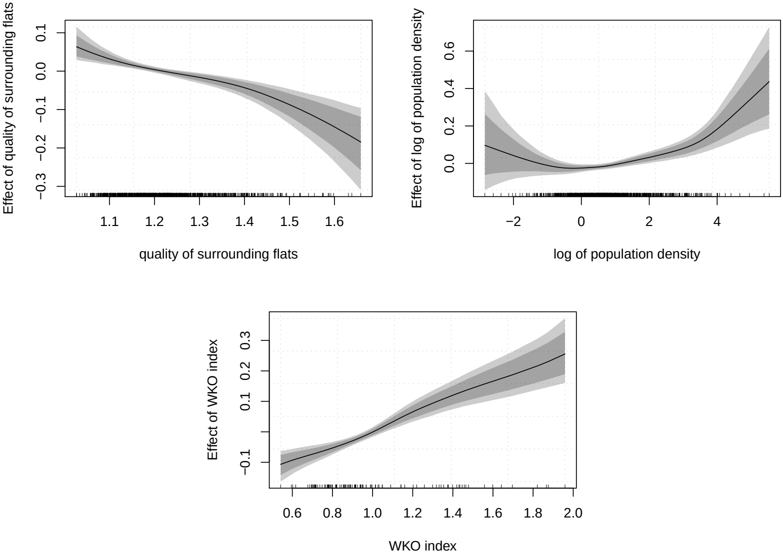
Structural continuous covariates



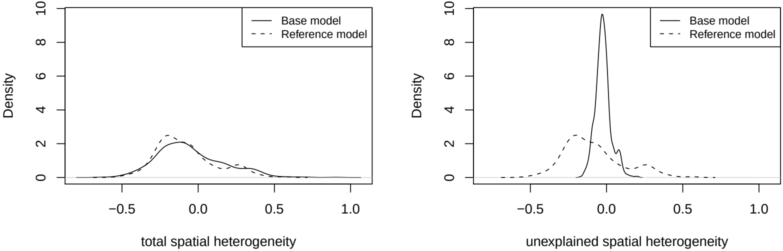
Neighborhood effects



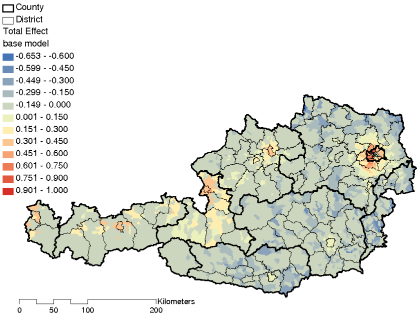
Neighborhood effects

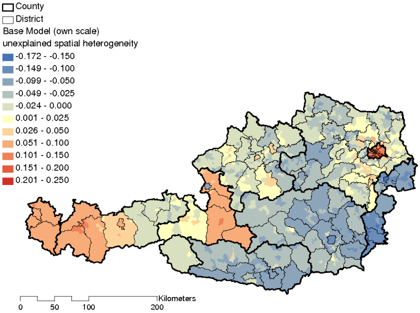


Neighborhood effects



Results: Hedonic regression data for house prices





Thank you!!!