Multilevel Structured Additive Regression

Stefan Lang

Department of Statistics, University of Innsbruck

November 2012

Overview

1. Hedonic regression data for house prices
2. Structured additive regression models
3. Hierarchical structured additive and multiplicative regression models
4. Application: hedonic regression data for house prices
5. Application: store level scanner data

Hedonic regression data for house prices in Austria

Variable of primary interest

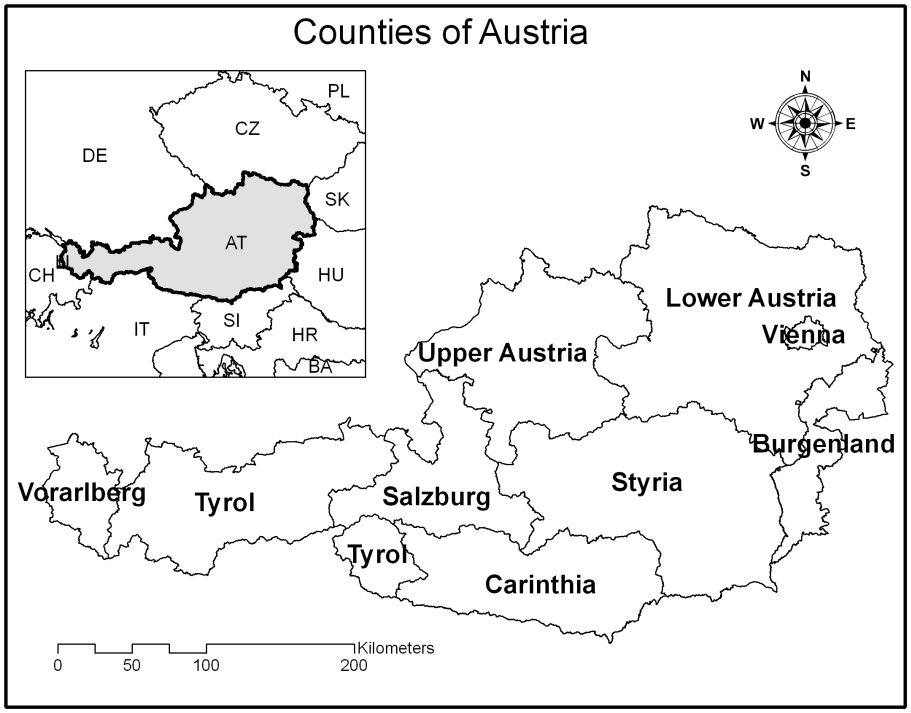
house price or log house price

Covariates

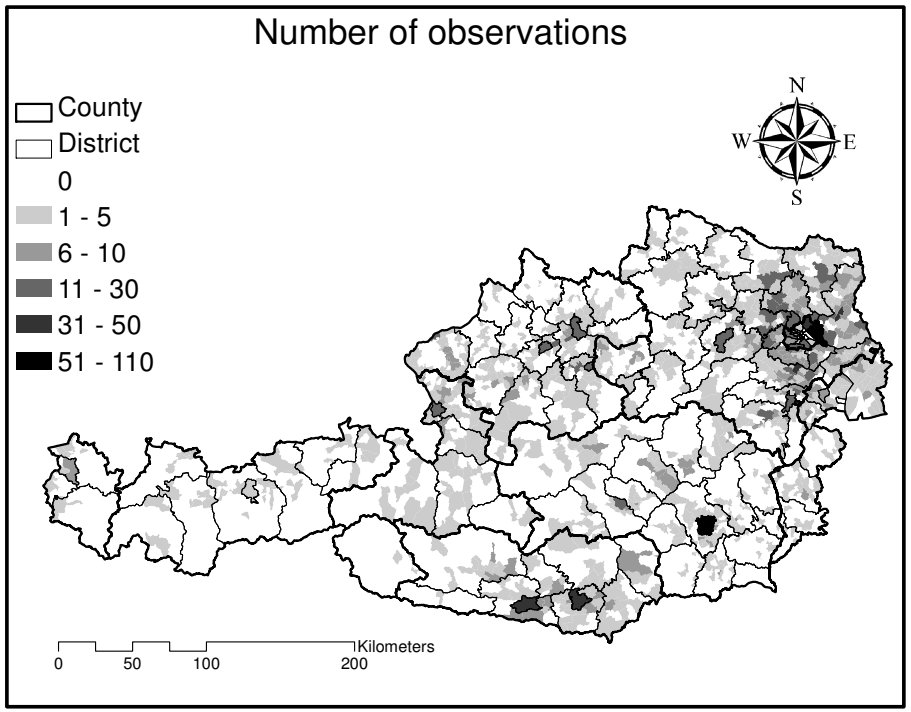
- Structural (physical) characteristics, like floor space area, constructional condition, age etc., and
- neighborhood (locational) characteristics, often on various levels of aggregation, like the proximity to places of work, the social composition of the neighborhood etc.

Hedonic regression data for house prices in Austria

[a]



[b]



Hedonic regression data for house prices in Austria

Four-level hierarchical model

$$
\begin{aligned}
\text{level 1:} \quad & lnp && = f_{1,1}(area) + \cdots + f_{1,q}(age) + \mathbf{X}\boldsymbol{\beta} + f_{municipal}(s_1) + \boldsymbol{\varepsilon}_1 \\
\text{level 2:} \quad & f_{municipal}(s_1) && = f_{2,1}(purchase\ power) + \cdots + f_{2,l}(level\ of\ education) \\
& && \quad + f_{district}(s_2) + \boldsymbol{\varepsilon}_2 \\
\text{level 3:} \quad & f_{district}(s_2) && = f_{3,1}(unemployment\ rate) + f_{county}(s_3) + \boldsymbol{\varepsilon}_3 \\
\text{level 4:} \quad & f_{county}(s_3) && = \boldsymbol{\varepsilon}_4,
\end{aligned}
$$

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

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

Structured additive regression models

- Distributional and structural assumptions, given covariates and parameters, are based on Generalized Linear Models.

- $E(y_i \mid \mathbf{z}_i, \mathbf{x}_i) = h(\eta)$ with structured-additive predictor

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

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

Overview: modeling the functions f_j

- nonlinear functions of continuous covariates
  Random Walks, see Fahrmeir, Lang (2001a, JRSS C), Fahrmeir, Lang (2001b, AISM); P-Splines, see Lang, Brezger (2004, JCGS), Brezger, Steiner (2008,JBES)
- two dimensional surface
  two dimensional P-Splines based on tensor products of one dimensional splines, see Lang, Brezger (2004, JCGS); Brezger, Lang (2006, CSDA), Belitz, Lang (2008, CSDA)
- Modelling spatial heterogeneity
  Markov-random fields, see Besag, York, Mollie (1991, AISM); two dimensional P-Splines; Gaussian random fields (Kriging), Lang et al. (2013);
- Modelling unit- or cluster specific heterogeneity
  I.i.d Gaussian random effects
- varying coefficients
  includes random slopes and geographically weighted regression

General form

- Vector of function evaluations can be written as:

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

$\mathbf{Z}$ design matrix; $\boldsymbol{\beta}$ vector of regression coefficients

- Special case varying coefficients

$$\eta = \ldots + z^{(1)} g\left(z^{(2)}\right) + \ldots, \qquad \text{i.e.} \qquad f(z^{(1)}, z^{(2)}) = z^{(1)} g\left(z^{(2)}\right)$$

$$\mathbf{f} = \mathbf{Z}\boldsymbol{\beta} = diag(z_1^{(1)}, \ldots, z_n^{(1)})\mathbf{Z}^{(2)}\boldsymbol{\beta} \qquad \mathbf{Z} = diag(z_1^{(1)}, \ldots, z_n^{(1)})\mathbf{Z}^{(2)}$$

- Penalized least squares:

$$PLS(\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_q) = \sum_{i=1}^{n}(y_i - \eta_i)^2 + \lambda_1 \boldsymbol{\beta}_1' \mathbf{K}_1 \boldsymbol{\beta}_1 + \ldots + \lambda_q \boldsymbol{\beta}_q' \mathbf{K}_q \boldsymbol{\beta}_q$$

General form

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

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

  τ^2 variance parameter, governs the smoothness of f.

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

- Structure of $\mathbf{Z}$ and $\mathbf{K}$ depends on the type of the covariates and on assumptions about smoothness of f_j. $\mathbf{Z}'\mathbf{Z}$ and $\mathbf{K}$ are most often band or sparse matrices.

Bayesian P-splines

- Idea: define a prior for the regression coefficients that acts as a penalty, e.g.

$$\boldsymbol{\beta} \,|\, \tau^2 \sim N(\mathbf{0}, \tau^2 \mathbf{I})$$

- Variance parameter τ^2 is the analogue to the smoothing parameter.

- Define a hyperprior for τ^2, e.g. $\tau^2 \sim IG(a, b)$, to be able to estimate the regression coefficients and the amount of smoothness simultaneously.

Hierarchical formulation and MCMC inference

For simplicity restrict the presentation to a two level hierarchical Gaussian model with one level-2 equation for the regression coefficients of the first term $\mathbf{Z}_1\boldsymbol{\beta}_1$.

- First stage

$$\mathbf{y} = \mathbf{Z}_1\boldsymbol{\beta}_1 + \ldots + \mathbf{Z}_q\boldsymbol{\beta}_q + \mathbf{X}\boldsymbol{\gamma} + \boldsymbol{\varepsilon}, \quad \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{W}^{-1})$$

- Second stage

$$\boldsymbol{\beta}_1 = \boldsymbol{\eta}_1 + \boldsymbol{\varepsilon}_1 = \mathbf{Z}_{11}\boldsymbol{\beta}_{11} + \ldots + \mathbf{Z}_{1q_1}\boldsymbol{\beta}_{1q_1} + \mathbf{X}_1\boldsymbol{\gamma}_1 + \boldsymbol{\varepsilon}_1,$$

and

$$\boldsymbol{\beta}_j \,|\, \tau_j^2 \sim N(\mathbf{0}, \tau_j^2\mathbf{K}^-) \quad \text{with} \quad \mathbf{A}_j\boldsymbol{\beta}_j = \mathbf{0} \qquad j = 2, \ldots, q$$

and

$$\boldsymbol{\beta}_{1l} \,|\, \tau_{1l}^2 \sim N(\mathbf{0}, \tau_{1l}^2\mathbf{K}^-) \quad \text{with} \quad \mathbf{A}_{1l}\boldsymbol{\beta}_{1l} = \mathbf{0} \qquad l = 1, \ldots, q_1$$

Hierarchical formulation and MCMC inference

Full conditionals for the regression coefficients at different levels are multivariate Gaussian. Posterior precision $\boldsymbol{\Sigma}^{-1}$ and mean $\boldsymbol{\mu}$ given by

$\boldsymbol{\beta}_1$ with compound prior

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{\sigma^2}\left(\mathbf{Z}_1'\mathbf{W}\mathbf{Z}_1 + \frac{\sigma^2}{\tau_1^2}\mathbf{I}\right) \qquad \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} = \frac{1}{\sigma^2}\mathbf{Z}_1'\mathbf{W}\mathbf{r}_1 + \frac{1}{\tau_1^2}\boldsymbol{\eta}_1$$

$\boldsymbol{\beta}_j$ in level-1 equation

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{\sigma^2}\left(\mathbf{Z}_j'\mathbf{W}\mathbf{Z}_j + \frac{\sigma^2}{\tau_j^2}\mathbf{K}_j\right) \qquad \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} = \frac{1}{\sigma^2}\mathbf{Z}_j'\mathbf{W}\,\mathbf{r}_j$$

$\boldsymbol{\beta}_{1l}$ in level-2 equation

$$\boldsymbol{\Sigma}^{-1} = \frac{1}{\tau_1^2}\left(\mathbf{Z}_{1l}'\mathbf{Z}_{1l} + \frac{\tau_1^2}{\tau_{1l}^2}\mathbf{K}_{1l}\right) \qquad \boldsymbol{\Sigma}^{-1}\boldsymbol{\mu} = \frac{1}{\tau_1^2}\mathbf{Z}_{1l}'\,\mathbf{r}_{1l}$$

Hierarchical formulation and MCMC inference

Properties

- *Reduced complexity in the second or third stage of the hierarchy:*
  - Number of "observations" in the level-2 equation 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 $z_{(1)}, \ldots, z_{(m)}$ in $\mathbf{Z}$ is much smaller than the total number n of observations, i.e. $m \ll n$.

Some details

Efficient computation of $\mathbf{Z}'\mathbf{W}\mathbf{Z}$ and $\mathbf{Z}'\mathbf{W}\,\mathbf{r}$

- Describe computation for a varying coefficient term

$$f(z) = f\left(z^{(1)}, z^{(2)}\right) = z^{(1)}\, g(z^{(2)})$$

with design matrix

$$\mathbf{Z} = \text{diag}\left(z_1^{(1)}, \ldots, z_n^{(1)}\right)\, \mathbf{Z}^{(2)} = \mathbf{D}\mathbf{Z}^{(2)}$$

where $\mathbf{D} = \text{diag}\left(z_1^{(1)}, \ldots, z_n^{(1)}\right)$.

- Computation for a pure additive term, i.e. $\mathbf{D} = \mathbf{I}$, arises as a special case.

Some details

- Denote by $z_{(1)}^{(2)} < z_{(2)}^{(2)} < \cdots < z_{(m)}^{(2)}$ the m ordered different observations of $z^{(2)}$.

- Compute the index vector $\boldsymbol{ind}$ with elements $\boldsymbol{ind}[i] \in \{1, \ldots, m\}$ denoting the category of the i-th observation, i.e. if $z_i^{(2)} = z_{(j)}^{(2)}$ then $\boldsymbol{ind}[i] = j$.

- 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 $z_{(1)}^{(2)}, \ldots, z_{(m)}^{(2)}$, i.e. $\tilde{\mathbf{Z}}[s,k] = B_k\left(z_{(s)}^{(2)}\right)$, $s = 1, \ldots, m$, $k = 1, \ldots, K$,
  - $\mathbf{P}$ is a $n \times m$ permutation matrix, which reverts the sorting, i.e. $\mathbf{P}[i,s] = I(\boldsymbol{ind}(i) = s)$.

- For the vector of function evaluations we obtain $\mathbf{f} = \mathbf{Z}\boldsymbol{\beta} = \mathbf{D}\mathbf{P}\tilde{\mathbf{Z}}\boldsymbol{\beta}$.

Some details

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

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

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

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

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 $\left((z_i^{(1)}\right)^2 w_i$ to $\tilde{w}_{\boldsymbol{ind}[i]}$.

Some details

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

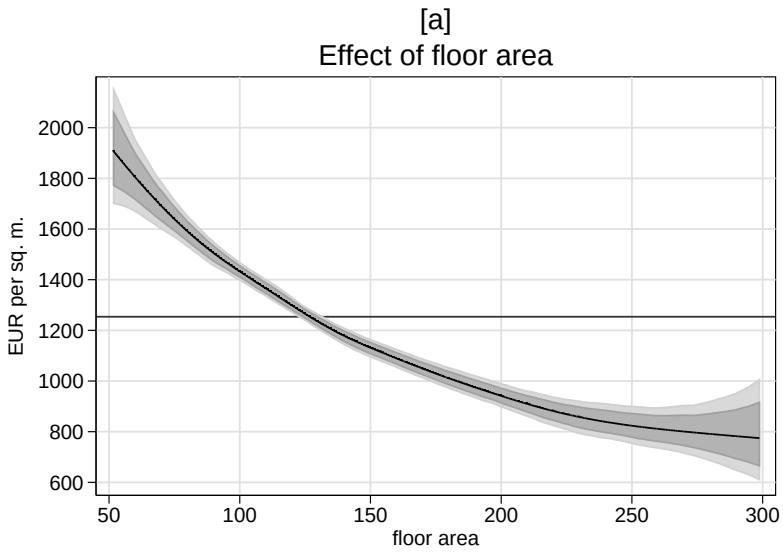
$$\tilde{r}_s = \sum_{i\,:\,\boldsymbol{ind}[i]=s} z_i^{(1)}\, w_i\, r_i.$$

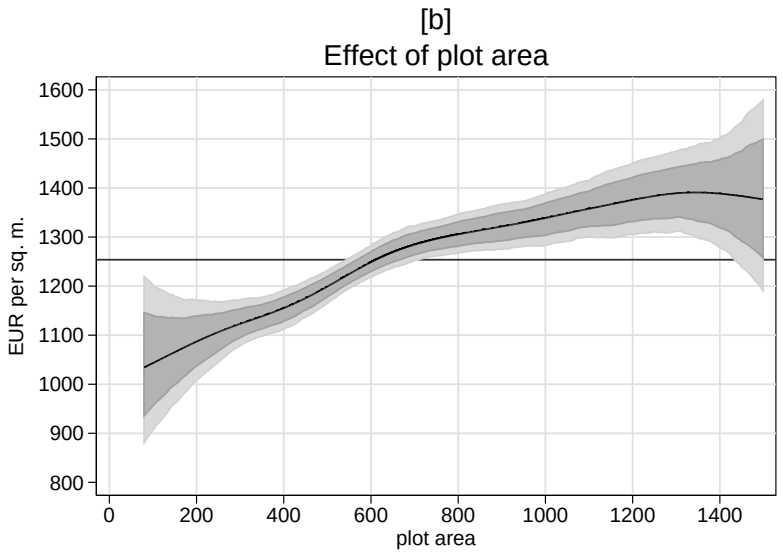
The $\tilde{r}_s$ are computed by first initializing $\tilde{r}_s = 0$ followed by the loop: For $i = 1, \ldots, n$ add $z_i^{(1)} w_i\, r_i$ to $\tilde{r}_{ind(i)}$.

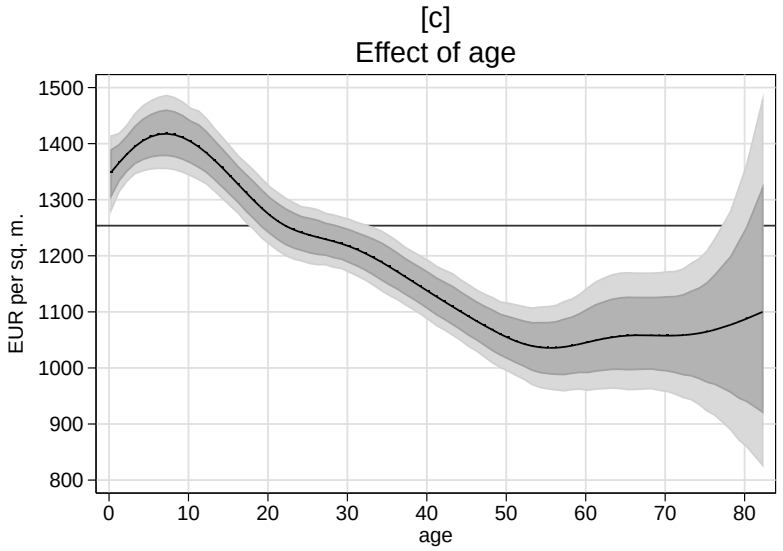
Application hedonic regression data for house prices in Austria

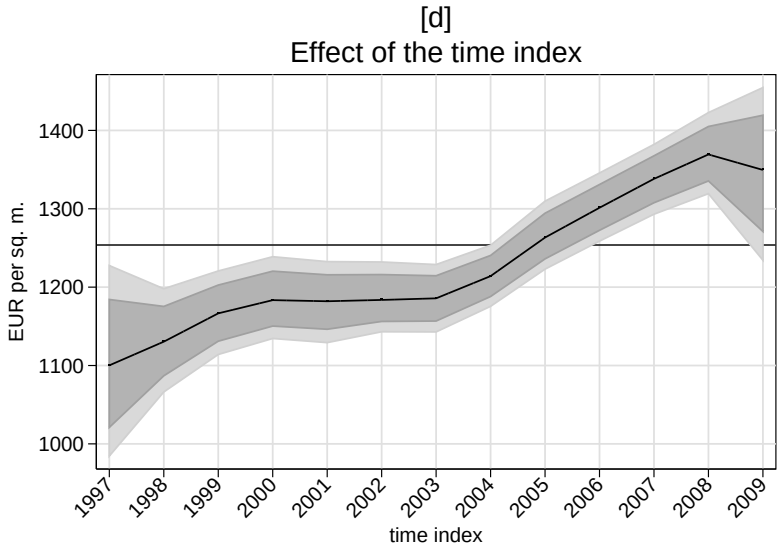
Four-level hierarchical model

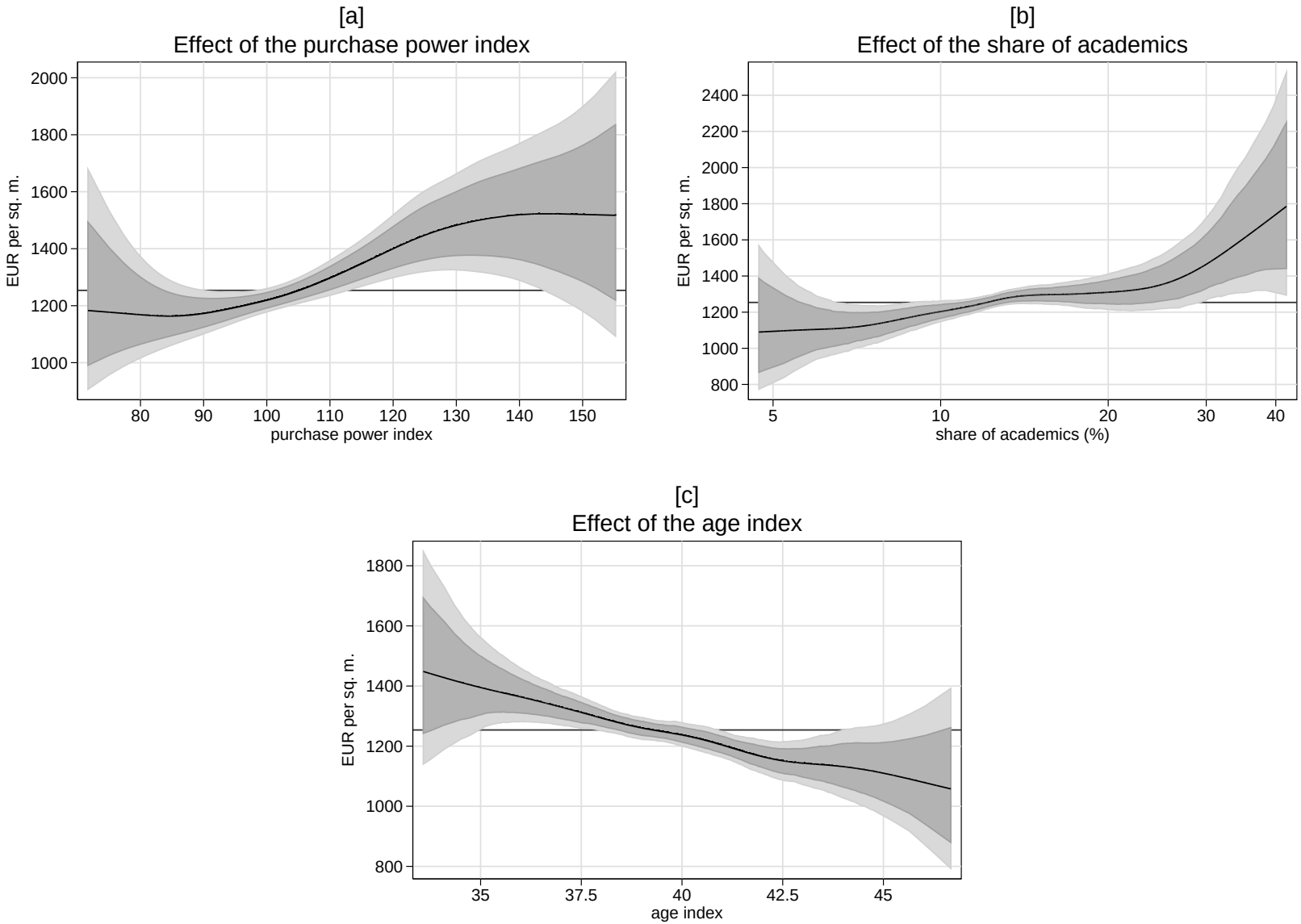
$$
\begin{aligned}
\text{level 1:} \quad & lnp && = f_{1,1}(area) + \cdots + f_{1,q}(age) + \mathbf{X}\boldsymbol{\beta} + f_{municipal}(s_1) + \boldsymbol{\varepsilon}_1 \\
\text{level 2:} \quad & f_{municipal}(s_1) && = f_{2,1}(purchase\ power) + \cdots + f_{2,l}(level\ of\ education) \\
& && \quad + f_{district}(s_2) + \boldsymbol{\varepsilon}_2 \\
\text{level 3:} \quad & f_{district}(s_2) && = f_{3,1}(unemployment\ rate) + f_{county}(s_3) + \boldsymbol{\varepsilon}_3 \\
\text{level 4:} \quad & f_{county}(s_3) && = \boldsymbol{\varepsilon}_4,
\end{aligned}
$$

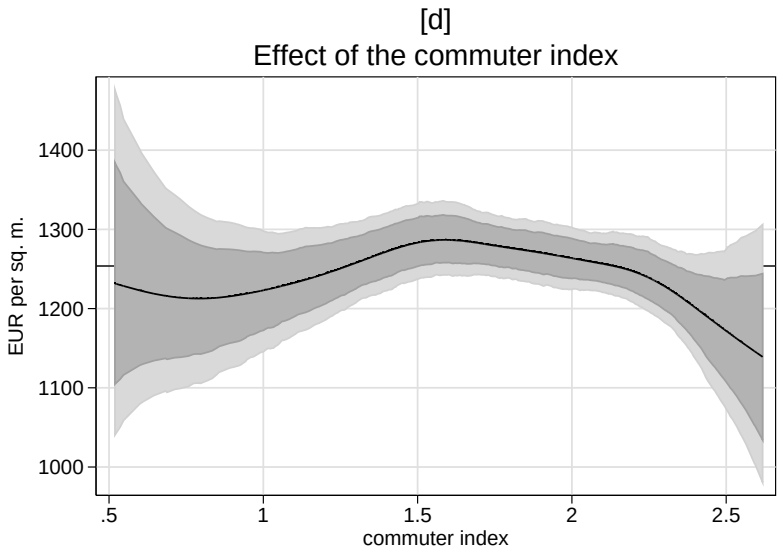

[a]
Effect of floor area
EUR per sq. m.
2000
1800
1600
1400
1200
1000
800
600
50
100
150
200
250
300
floor area

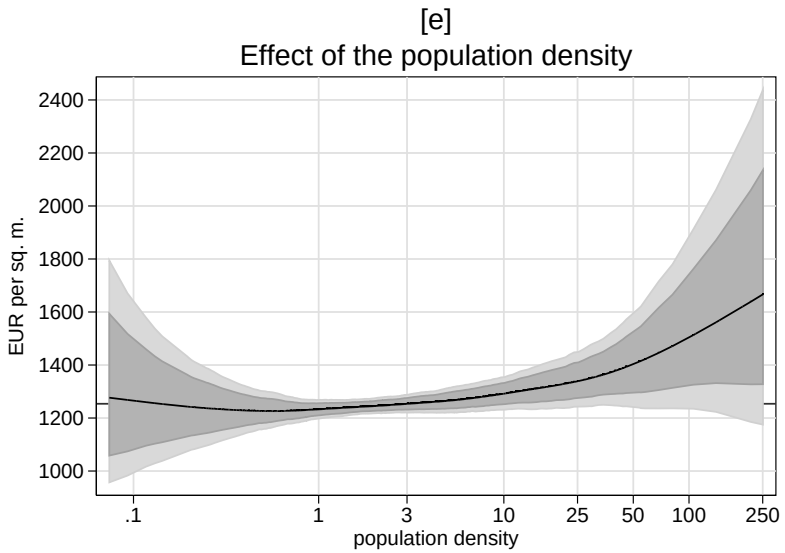

[b]
Effect of plot area
EUR per sq. m.
1600
1500
1400
1300
1200
1100
1000
900
800
0
200
400
600
800
1000
1200
1400
plot area

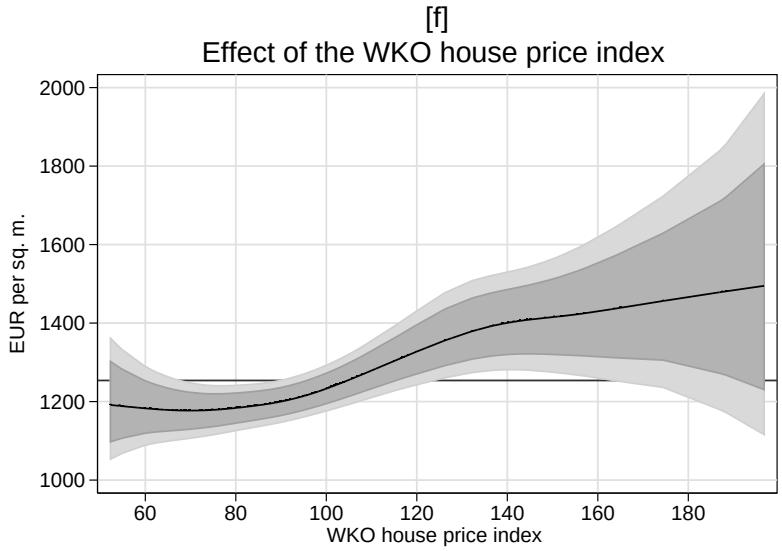

[c]
Effect of age
EUR per sq. m.
1500
1400
1300
1200
1100
1000
900
800
0
10
20
30
40
50
60
70
80
age


[d]
Effect of the time index
EUR per sq. m.
1400
1300
1200
1100
1000
1997
1998
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
time index

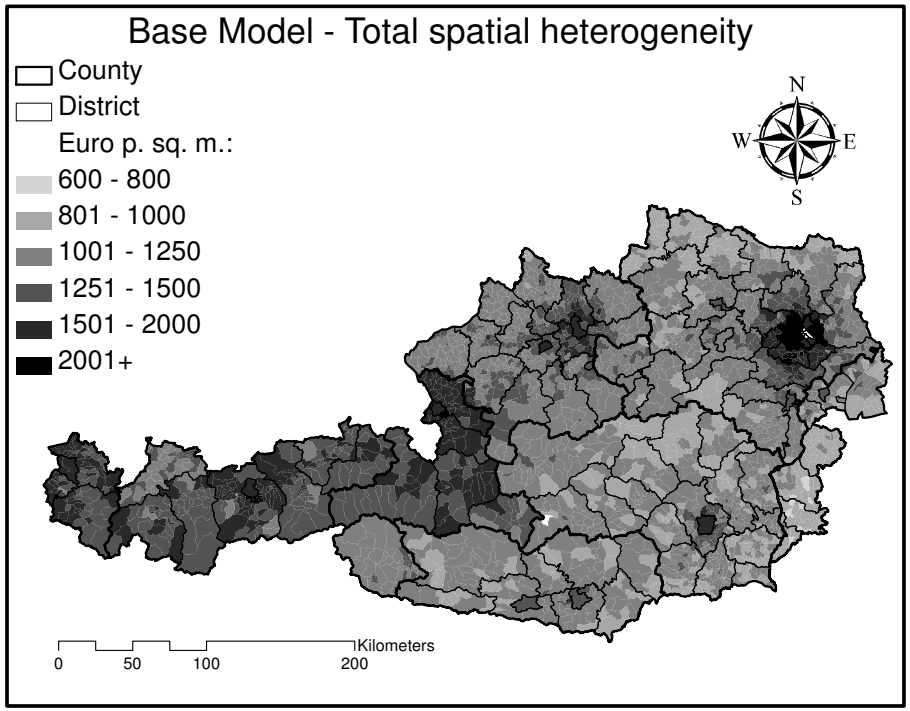

[a]
Effect of the purchase power index
EUR per sq. m.
2000
1800
1600
1400
1200
1000
80
90
100
110
120
130
140
150
purchase power index
[b]
Effect of the share of academics
EUR per sq. m.
2400
2200
2000
1800
1600
1400
1200
1000
800
5
10
20
30
40
share of academics (%)
[c]
Effect of the age index
EUR per sq. m.
1800
1600
1400
1200
1000
800
35
37.5
40
42.5
45
age index


[d]
Effect of the commuter index
EUR per sq. m.
1000
1100
1200
1300
1400
.5
1
1.5
2
2.5
commuter index

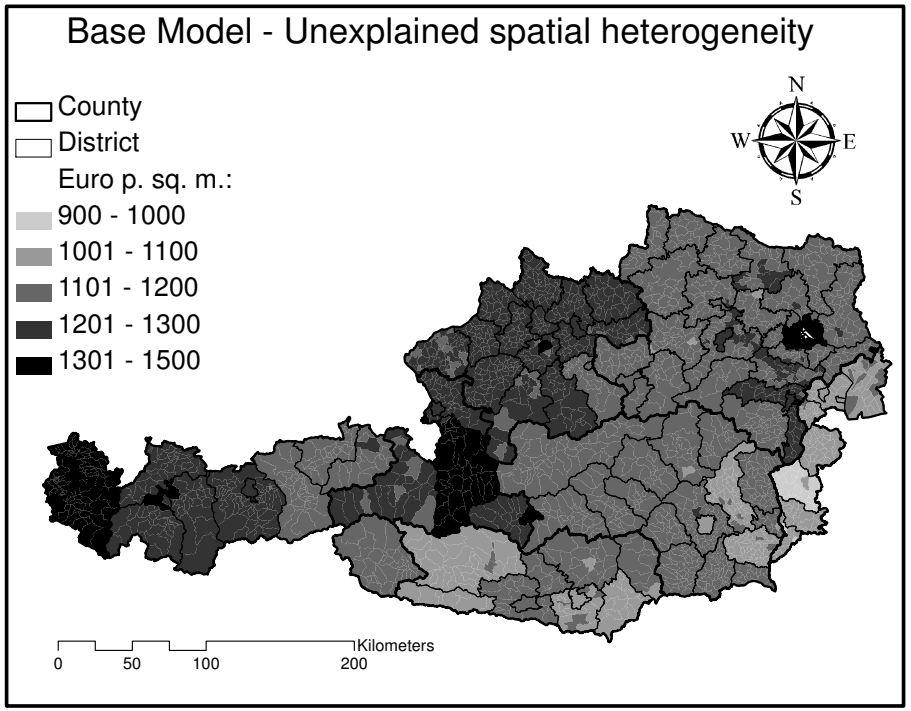

[e]
Effect of the population density
EUR per sq. m.
1000
1200
1400
1600
1800
2000
2200
2400
.1
1
3
10
25
50
100
250
population density

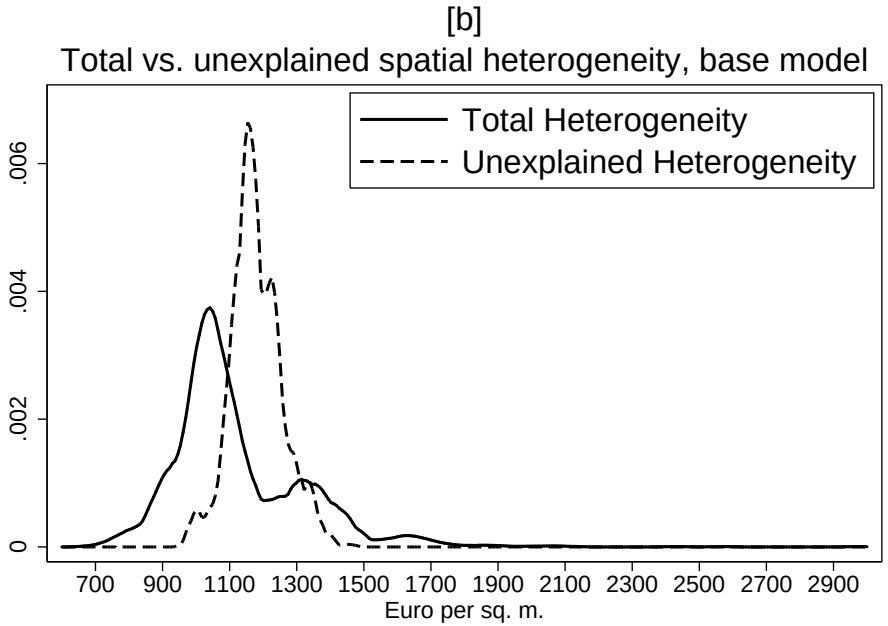

[f]
Effect of the WKO house price index
EUR per sq. m.
1000
1200
1400
1600
1800
2000
60
80
100
120
140
160
180
WKO house price index

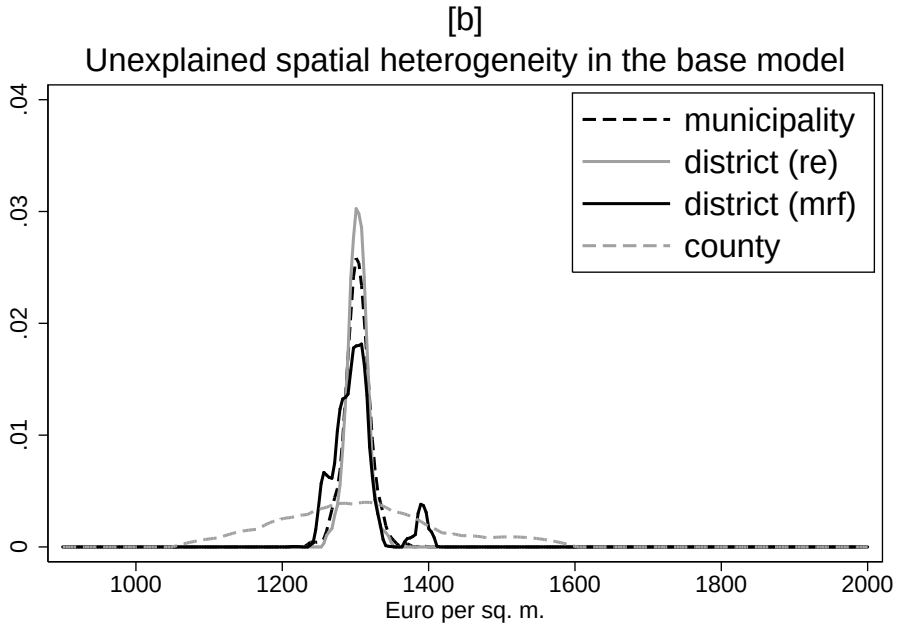
[b]


Base Model - Total spatial heterogeneity
County
District
Euro p. sq. m.:
600 - 800
801 - 1000
1001 - 1250
1251 - 1500
1501 - 2000
2001+
N
W
E
S
Kilometers
0
50
100
200

[c]


Base Model - Unexplained spatial heterogeneity
County
District
Euro p. sq. m.:
900 - 1000
1001 - 1100
1101 - 1200
1201 - 1300
1301 - 1500
N
W
E
S
Kilometers
0
50
100
200

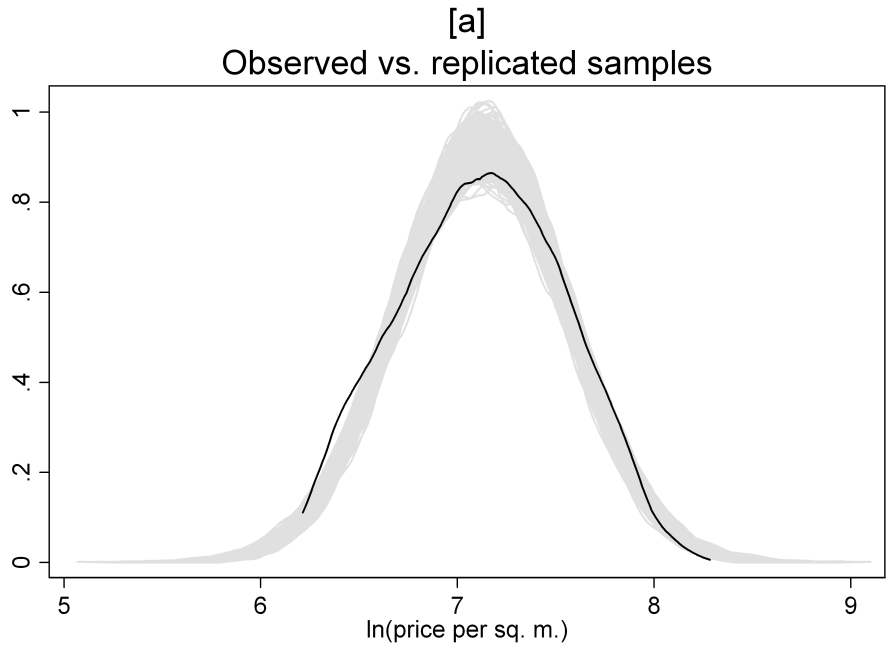

[b]
Total vs. unexplained spatial heterogeneity, base model
Total Heterogeneity
Unexplained Heterogeneity
0
.002
.004
.006
700 900 1100 1300 1500 1700 1900 2100 2300 2500 2700 2900
Euro per sq. m.

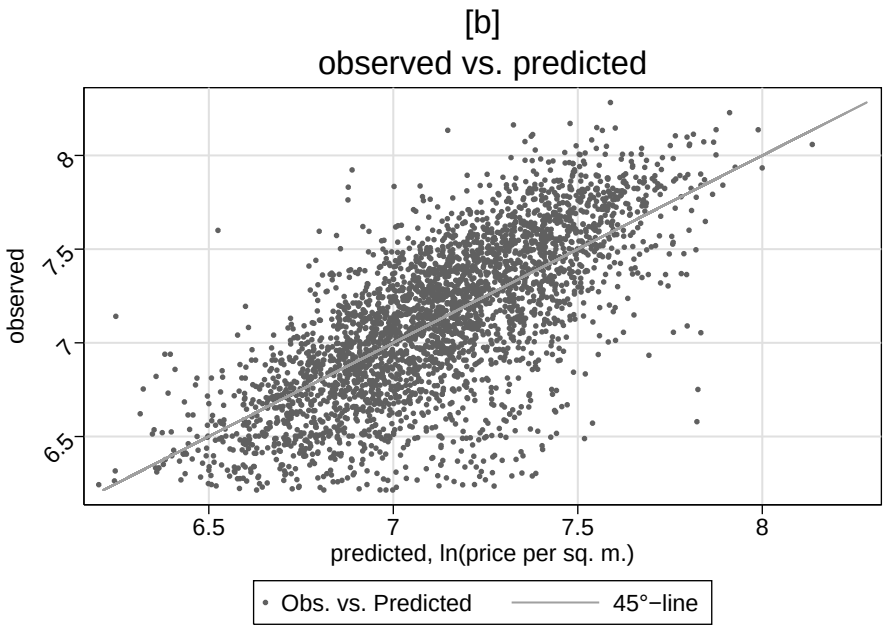

[b]
Unexplained spatial heterogeneity in the base model
municipality
district (re)
district (mrf)
county
0
.01
.02
.03
.04
1000 1200 1400 1600 1800 2000
Euro per sq. m.

Model diagnosis

- Systematic differences between the data and the estimated model can be detected with the aid of posterior predictive checks.

- Compare the empirical distribution of logged house prices per sq. m. with the simulated posterior distributions of house prices obtained from our base model.





[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.

	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

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

Model diagnosis

- The predictive checks indicate some misspecification as the simulated responses are sampled in a wider range and are 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.
- The age effect is not in line with our expectations for buildings of an age of less than three years.
- Inspection of the "problematic observations" shows that the corresponding houses are mostly in a group with age less than three years.

Improving models on the basis of model diagnosis

- *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.

  The deviance and with it the DIC decreases dramatically for all model specifications.

- *Include interactions with age*

Application store-level scanner data

Variable of primary interest

Q_{st} weekly unit sales in store s and week t

Covariates

- own price of brand ($price$)
- prices of competing brands ($price\_national$, $price\_dominicks$, $pprice\_premium$)
- Store characteristics (e.g. share of women working full-time, share of retirees, driving time to the nearest supermarket, etc.)

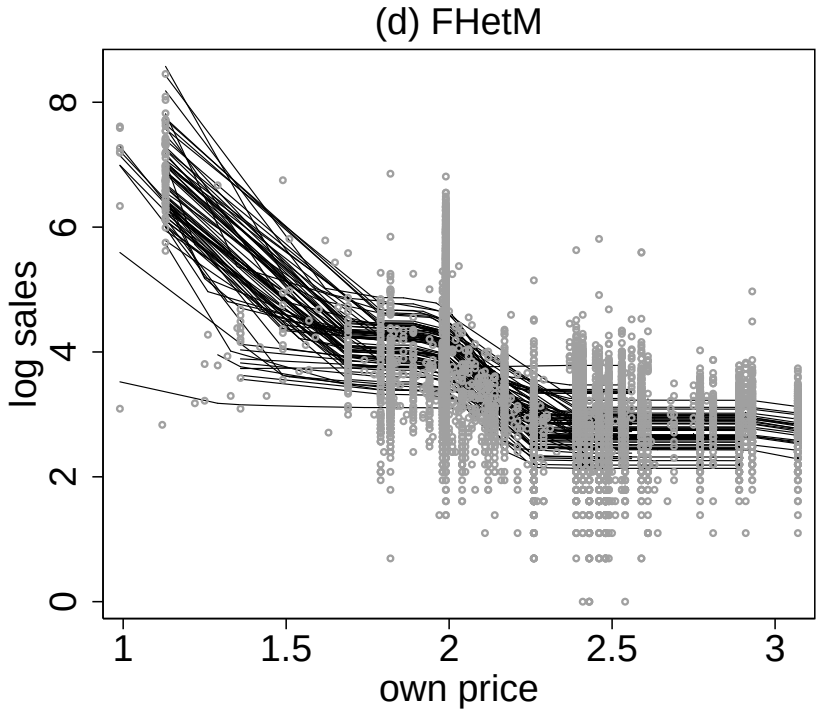
Application: store-level scanner data

Model

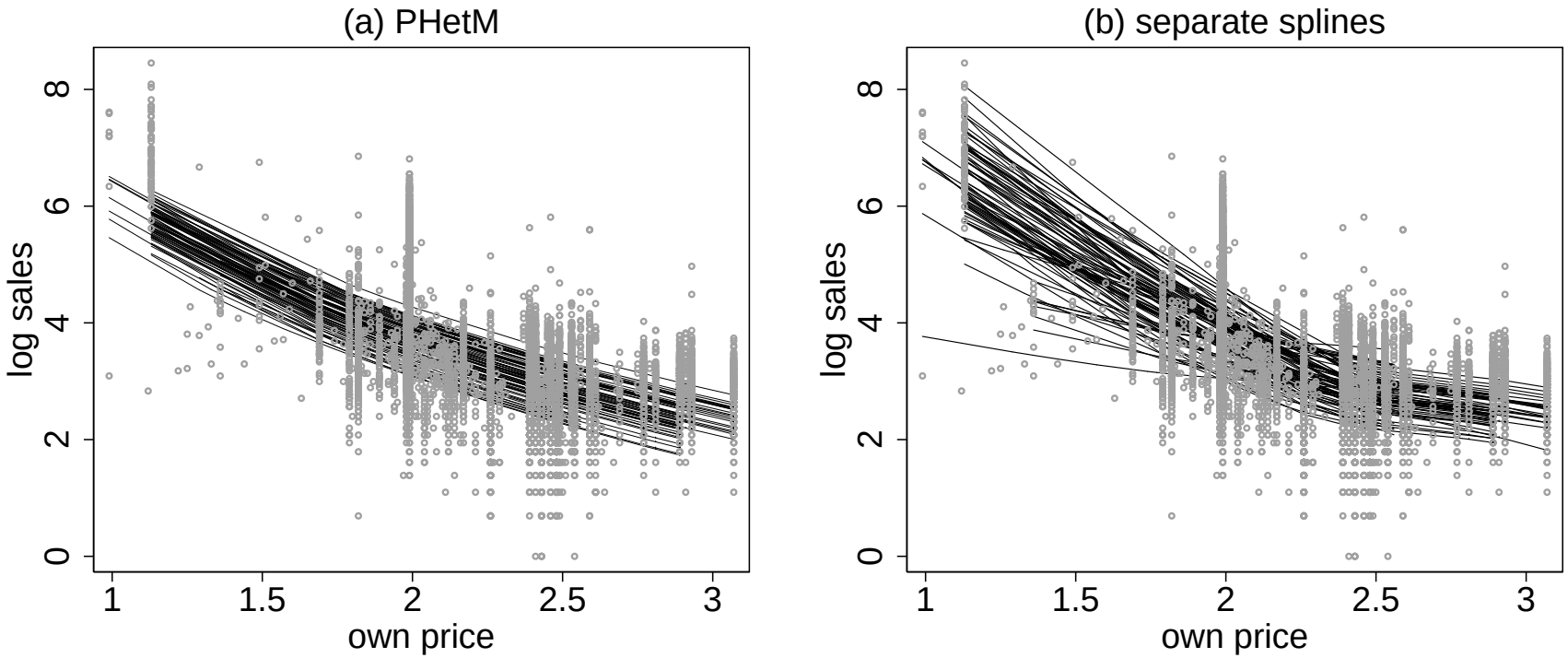
$$
\begin{aligned}
ln\, Q_{st} \quad = \quad & f_0(m_t) + (1+\alpha_{s1})\, f_1\left(price_{st}\right) + (1+\alpha_{s2})\, f_2\left(price\_premium_{st}\right) \\
& + (1+\alpha_{s3})\, f_3\left(price\_national_{st}\right) + (1+\alpha_{s4})\, f_4\left(price\_dominicks_{st}\right) \\
& + \mathbf{x}'_{st}\boldsymbol{\gamma}_s + \varepsilon_{st},
\end{aligned}
$$

$$
\alpha_{sj} = f_{j1}(v_{s1}) + \ldots + f_{j11}(v_{s11}) + u_{sj}, \quad s = 1, \ldots, 81, \qquad j = 1, \ldots, 4,
$$

Application: store-level scanner data

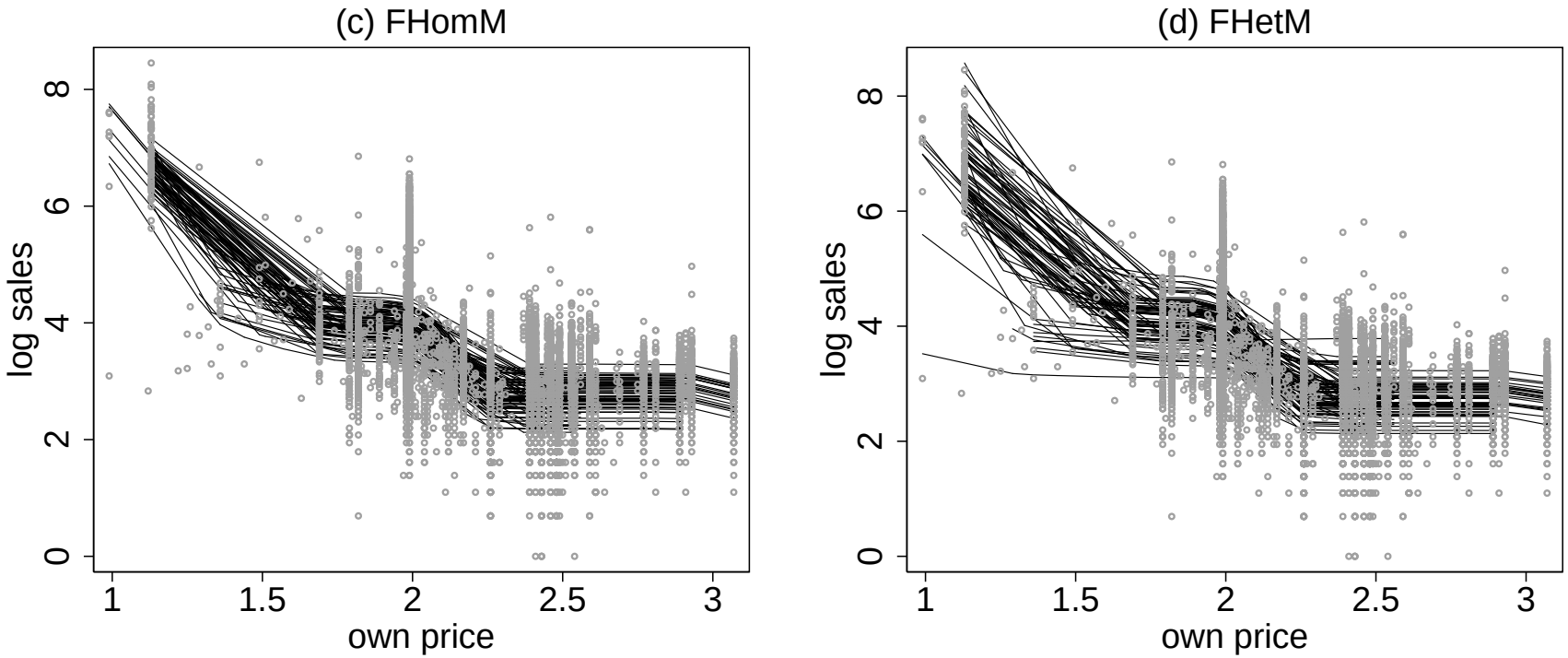


Application: store-level scanner data



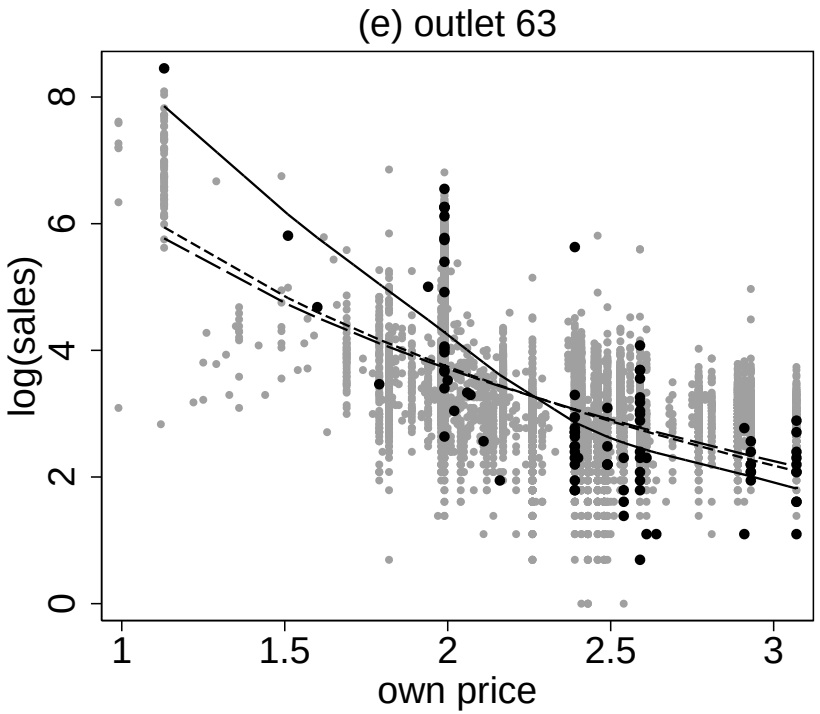
PHetM corresponds to usual parametric random effects model

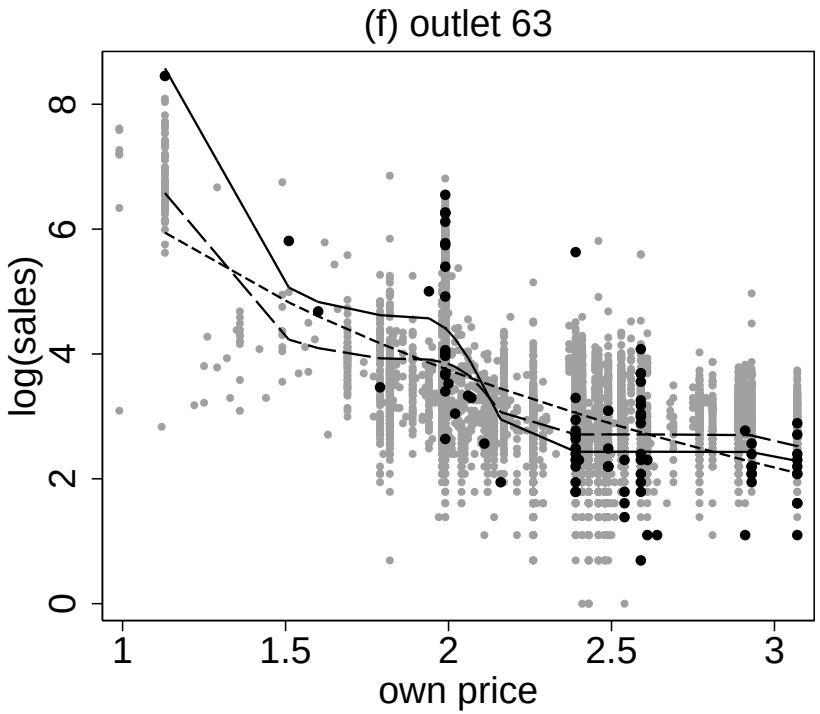
Application: store-level scanner data



FHomM corresponds to usual nonlinear model without random effects

Application: store-level scanner data





Some references

Review, teaching

- Fahrmeir, L. and Kneib, T. and Lang, S. (2012): Bayesian multilevel models. *The SAGE handbook of multilevel Modeling* edited by Scott, M.A. and Simonoff, J.S. and Marx, B.D.
- Fahrmeir, L. and Kneib, T. and Lang, S. (2009): *Regression: Modelle, Methoden und Anwendungen.* Springer, Berlin. (2nd edition)
- Fahrmeir, L. and Kneib, T. and Lang, S. and Marx, B. (2012) *Regression: Models, Methods and Applications.* Springer, New York, forthcoming.

Statistical methodology and software

- Lang, S. and Brezger, A. (2004): Bayesian P-Splines. *Journal of Computational and Graphical Statistics*, 13, 183–212.
- Brezger, A. and Lang, S. (2006): Generalized additive regression based on Bayesian P-splines. *Computational Statistics and Data Analysis*, 50, 967–991.
- Lang, S. and Umlauf, N. and Wechselberger, P. and Harttgen, K. and Kneib, T. (2013): Multilevel Structured Additive Regression. *Statistics and Computing, to appear*.

- Lang, S. and Steiner, W. and Wechselberger, P. (2012): Simultaneously Accommodating Heterogeneity and Functional Flexibility in Store Sales Models, Revised for *Marketing Science*.
- Umlauf, N. and Kneib, T. and Lang, S. and Zeileis, A. (2012): Structured Additive Regression Models: An R Interface to BayesX. *Journal of Statistical Software, conditionally accepted*.
- Waldmann, E. and Kneib, T. and Lang, S. and Yue, Y. (2012): Bayesian Semiparametric Additive Quantile Regression, *In Revision for Statistical Modelling*,

Real estate modeling

- Brunauer, W. A. and Lang, S. and Wechselberger, P. and Bienert, S. (2010): Additive Hedonic Regression Models with Spatial Scaling Factors: An Application for Rents in Vienna. *Journal of Real Estate Finance and Economics*, 40,390–410.
- Brunauer, W. A. and Lang, S. and Umlauf, N. (2012): Modeling House Prices using Multilevel Structured Additive Regression. To appear in *Statistical Modelling*.
- Brunauer, W. and Keiler, S. and Lang, S. (2011): Cost drivers of operation charges and variation over time: An analysis based on semiparametric SUR models. *Energy and Buildings*, 43, 1189-1199.