Modeling Real Estate Data using Semiparametric Quantile Regression

Alexander Razen
Department of Statistics
University of Innsbruck

September 9th, 2011

Overview

1. Application: Hedonic regression data for real estate prices
2. Quantile regression models
3. Bayesian inference
4. Results

Hedonic regression data for house prices in Austria

Variable of primary interest

House price per square meter

Covariates

- Structural characteristics, like the floor space area, the plot area, the age, the equipment etc.
- Locational characteristics at different levels, like the buying power index (municipal), the share of academics (municipal), the real estate price index (district), etc.

Hedonic regression data for house prices in Austria

Hierarchical semiparametric model

$$
\begin{aligned}
p_{qm} &= f_{1,1}(\textit{municipal}) + f_{1,2}(\textit{area}) + \ldots + \\
&\quad f_{1,l}(\textit{age}) + \boldsymbol{X}\beta + \epsilon_1 \\
f_{1,1}(\textit{municipal}) &= f_{2,1}(\textit{district}) + f_{2,2}(\textit{buying power}) + \ldots + \\
&\quad + f_{2,m}(\textit{academics}) + \epsilon_2 \\
f_{2,1}(\textit{district}) &= f_{3,1}(\textit{state}) + f_{3,2}(\textit{real estate index}) + \\
&\quad + g(\textit{dist}) + \epsilon_3 \\
f_{3,1}(\textit{state}) &= \textit{const} + \epsilon_4
\end{aligned}
$$

The term $\boldsymbol{X}\beta$ contains the linear effects.

The functions f_i are possibly nonlinear functions of the covariates.

The function g describes a spatial district effect.

Hedonic regression data for house prices in Austria

Goal

Determining the conditional quantiles of the distribution of the house prices

Approaches

- Mean regression based on a normal distribution assumption
- Quantile regression

Overview

1. Application: Hedonic regression data for real estate prices
2. **Quantile regression models**
3. Bayesian inference
4. Results

Linear quantile regression

Linear model

Given: Observations $y_i, x_{i1}, \ldots, x_{ip}$ for $i = 1, \ldots, n$ from the model

$$\boldsymbol{y} = \boldsymbol{X}\beta + \boldsymbol{\epsilon}.$$

Assumptions for a particular quantile φ:

$$\epsilon_i \overset{iid}{\sim} F$$

$$Q_\varphi(\epsilon_i) := F^{-1}(\varphi) = 0$$

Then:

$$Q_\varphi(\boldsymbol{y}) = \boldsymbol{X}\beta$$

Linear quantile regression

Loss function

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

Empirical loss of an estimation $\hat{y}$:

$$\sum_{i=1}^{n} \rho_\varphi(y_i - \hat{y})$$

Linear quantile regression

Regression quantile

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{\arg\min} \left\{ \sum_{i=1}^{n} \rho_\varphi (y_i - \boldsymbol{x}_i' \boldsymbol{\beta}) \right\}$$

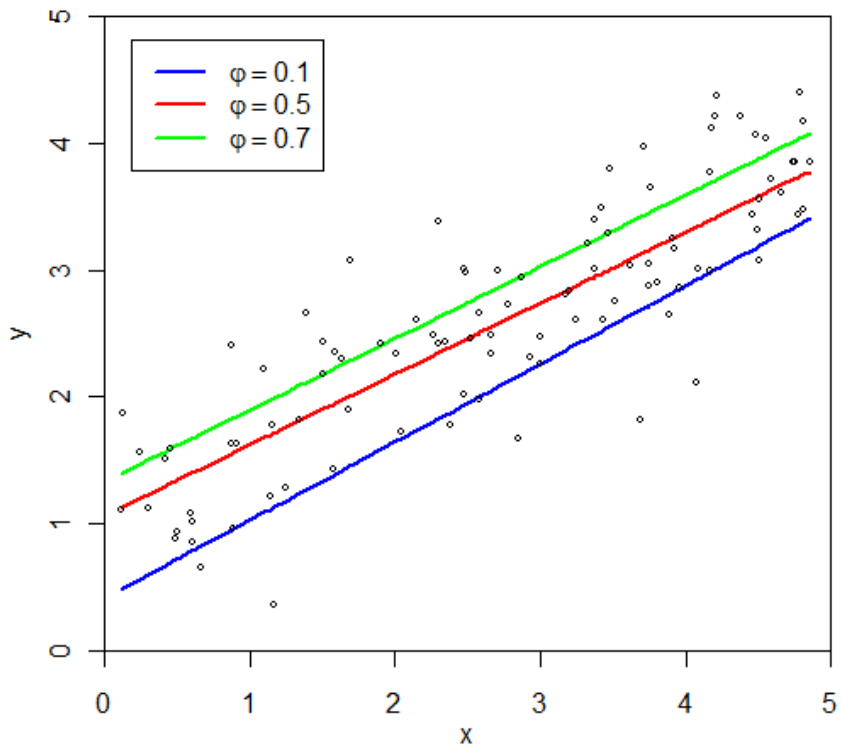
Estimation of the conditional quantile:

$$\widehat{Q_\varphi(\boldsymbol{y})} = \boldsymbol{X} \hat{\boldsymbol{\beta}}$$

Linear quantile regression

Example

$$y = 1 + \frac{3}{5}x + \epsilon, \quad \epsilon \sim \mathcal{N}(0, 0.5)$$



Nonlinear and spatial quantile regression

Nonlinear or spatial model:

$$\boldsymbol{y} = f(\boldsymbol{z}) + \boldsymbol{\epsilon} = \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon}$$

Smooth effects: Penalize differences between the coefficients of adjacent B-splines or the coefficients of neighbouring regions, respectively.

Penalized optimization problem

$$\hat{\gamma} = \underset{\gamma \in \mathbb{R}^d}{\arg\min} \left\{ \sum_{i=1}^{n} \rho_\varphi \left(y_i - \boldsymbol{z}_i' \gamma \right) + \lambda \gamma' \boldsymbol{K} \gamma \right\}$$

Semiparametric quantile regression

Semiparametric model:

$$\boldsymbol{y} = \boldsymbol{\eta} + \boldsymbol{\epsilon} = \boldsymbol{X}\boldsymbol{\beta} + f_1(\boldsymbol{z}_1) + \ldots + f_q(\boldsymbol{z}_q) + \boldsymbol{\epsilon}$$

Penalized optimization problem

$$\min_{\boldsymbol{\beta},\gamma_k} \left\{ \sum_{i=1}^{n} \rho_\varphi \left(y_i - \eta_i\right) + \sum_{j=1}^{q} \lambda_j \gamma_j' \boldsymbol{K}_j \gamma_j \right\}$$

Overview

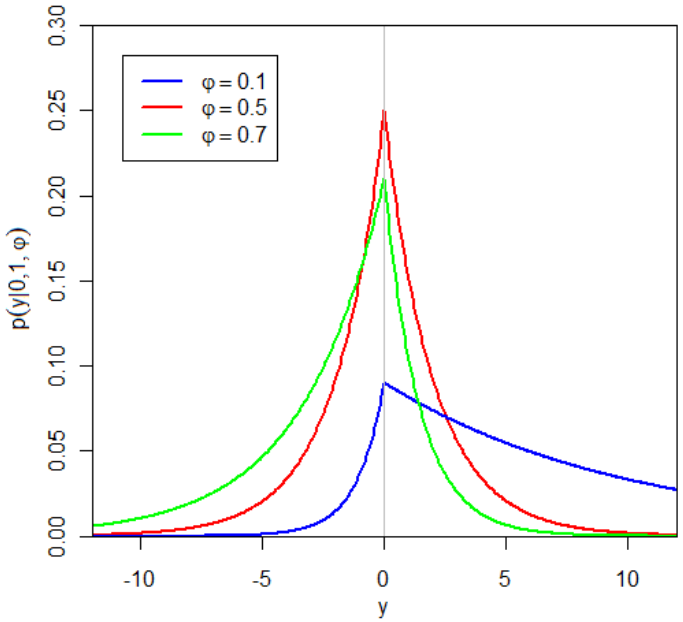
1. Application: Hedonic regression data for real estate prices
2. Quantile regression models
3. **Bayesian inference**
4. Results

Bayesian Inference

Asymmetric Laplace distribution

Density function:

$$p(y|\mu,\sigma^2,\varphi) = \frac{\varphi(1-\varphi)}{\sigma^2}\exp\left(-\frac{1}{\sigma^2}\rho_\varphi(y-\mu)\right)$$



Bayesian Inference

Assumption:

$$y_i \sim \text{ALD}(\eta_i, \sigma^2, \varphi)$$

Joint likelihood:

$$p(\boldsymbol{y}|\boldsymbol{\eta}, \sigma^2, \varphi) \propto \frac{1}{(\sigma^2)^n} \exp\left(-\frac{1}{\sigma^2} \sum_{i=1}^{n} \rho_\varphi(y_i - \eta_i)\right)$$

Maximizing this likelihood is equivalent to minimizing the former loss function in the linear case.

Bayesian Inference

Priors for nonlinear or spatial effects:

$$p\left(\gamma_j|\tau_j^2\right) \propto \frac{1}{\left(\tau_j^2\right)^{\frac{\text{rk}(\boldsymbol{K}_j)}{2}}} \exp\left(-\frac{1}{2\tau_j^2}\gamma_j'\boldsymbol{K}_j\gamma_j\right)$$

τ_j^2 variance parameter, governs the smoothness of the respective function.

Bayesian Inference

Representation of an asymmetric Laplace distribution:

$$Y \stackrel{D}{=} \eta + \frac{1-2\varphi}{\varphi(1-\varphi)} V + W \sqrt{\frac{2}{\sigma^2 \varphi(1-\varphi)} V}$$

V, W independent random variables with exponential and normal distributions respectively:

$$p\left(v|\sigma^2\right) = \sigma^2 \exp\left(\sigma^2 v\right) \quad \text{and} \quad W \sim \mathcal{N}(0,1)$$

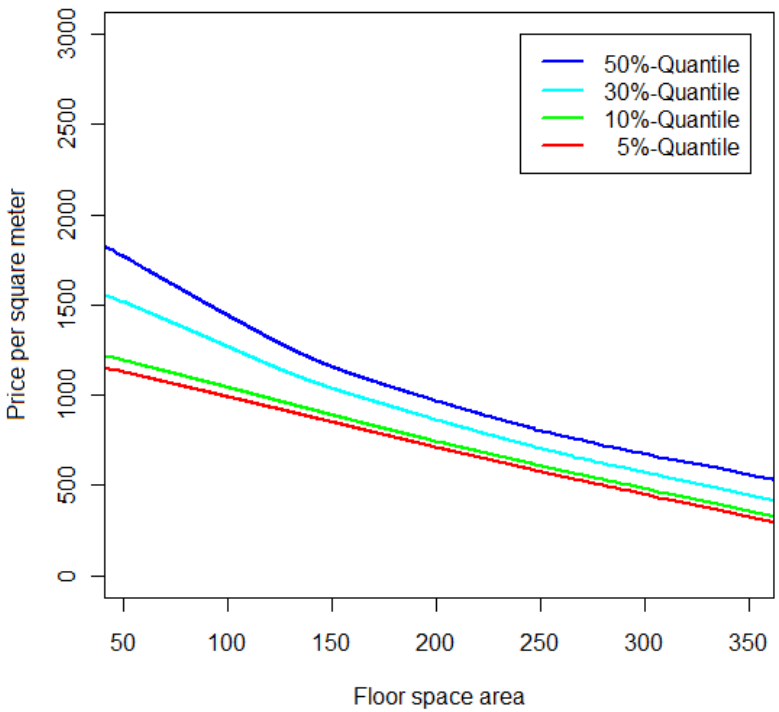
Important feature for MCMC-inference.

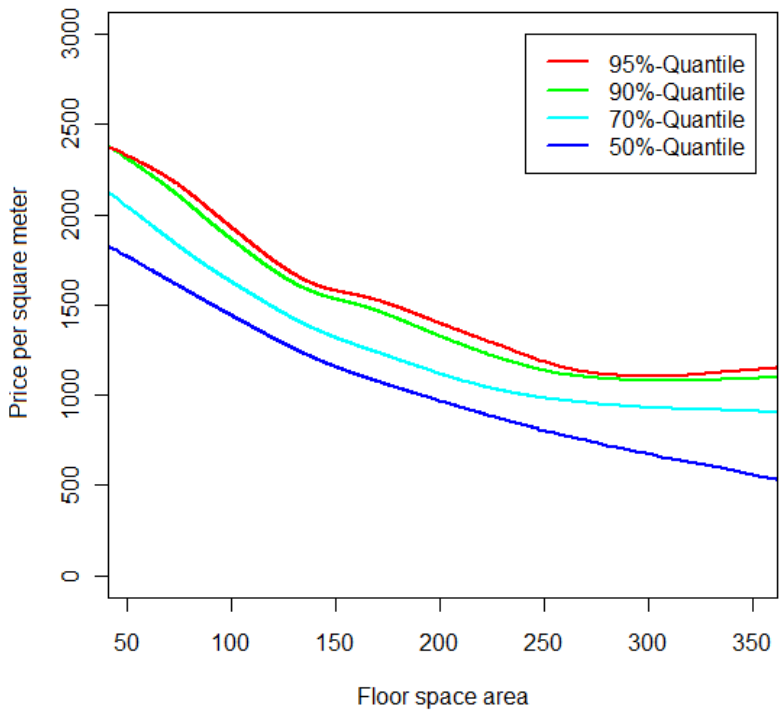
Overview

1. Application: Hedonic regression data for real estate prices
2. Quantile regression models
3. Bayesian inference
4. **Results**

Results

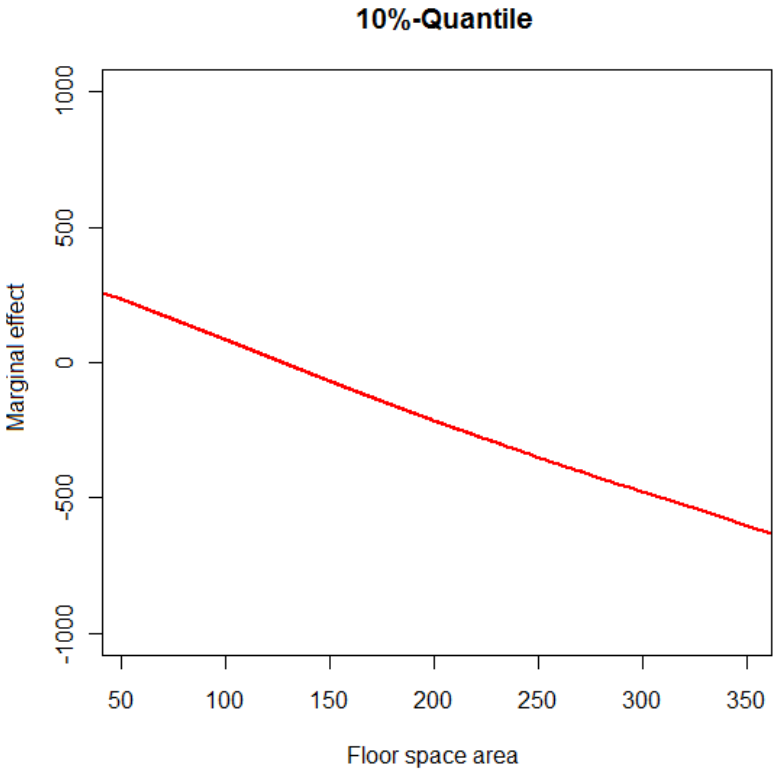
Floor space area:

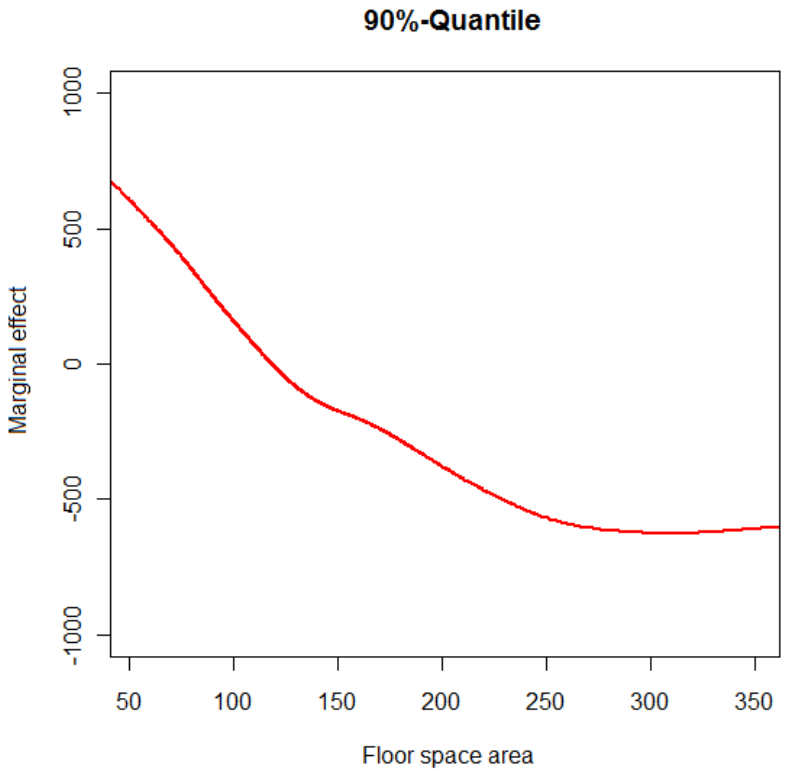




Results

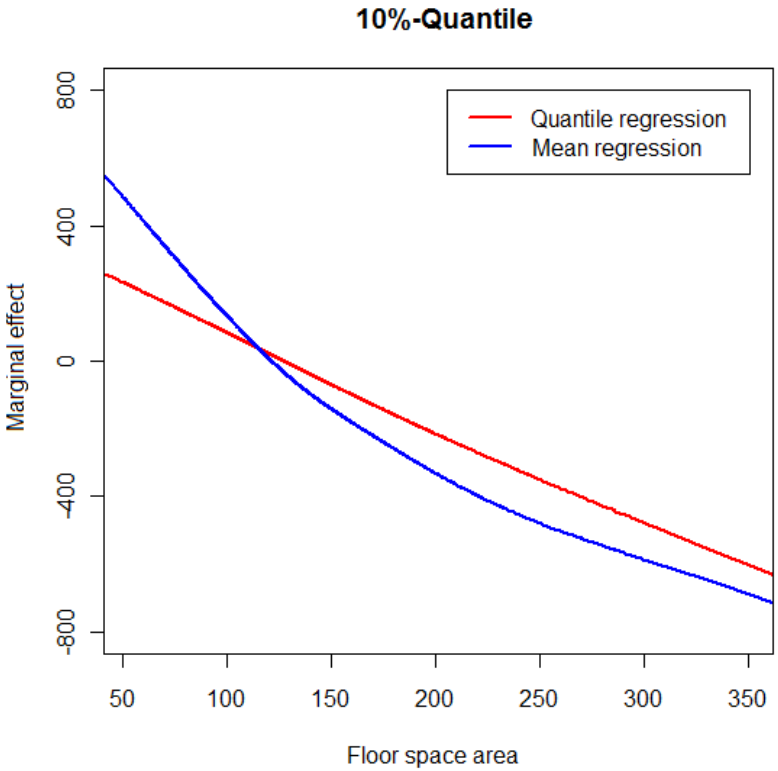
Floor space area:

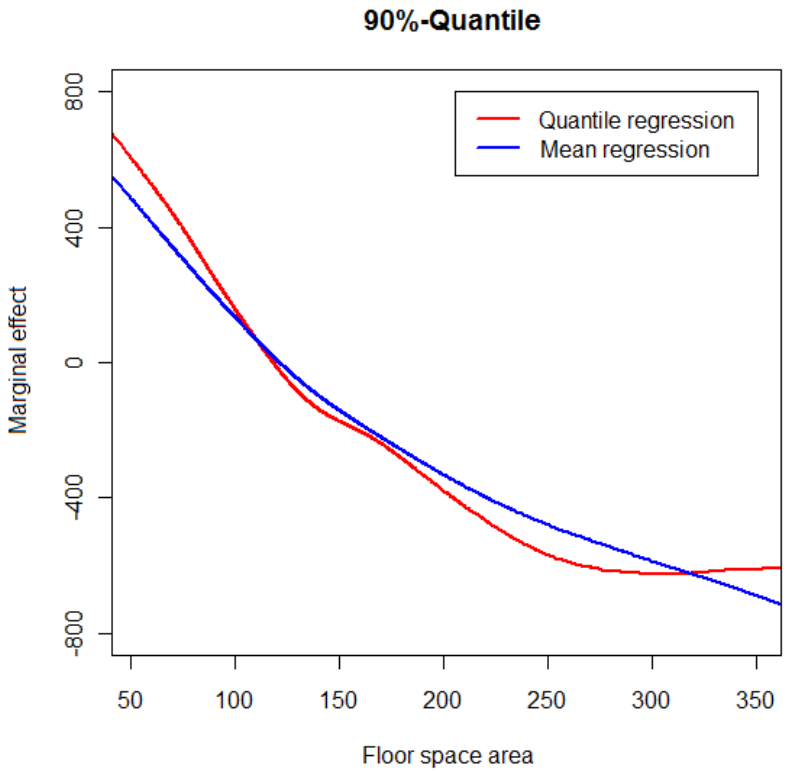




Results

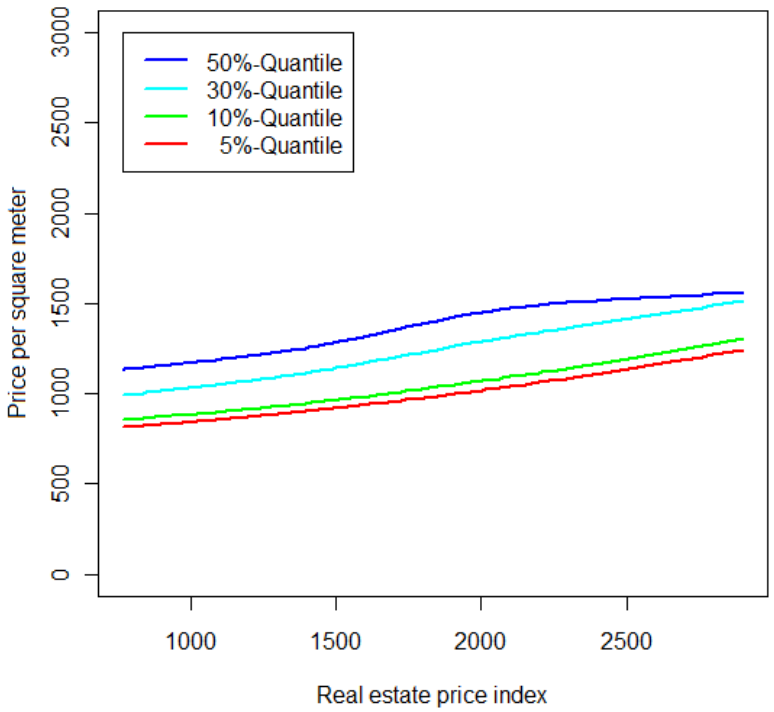
Floor space area:

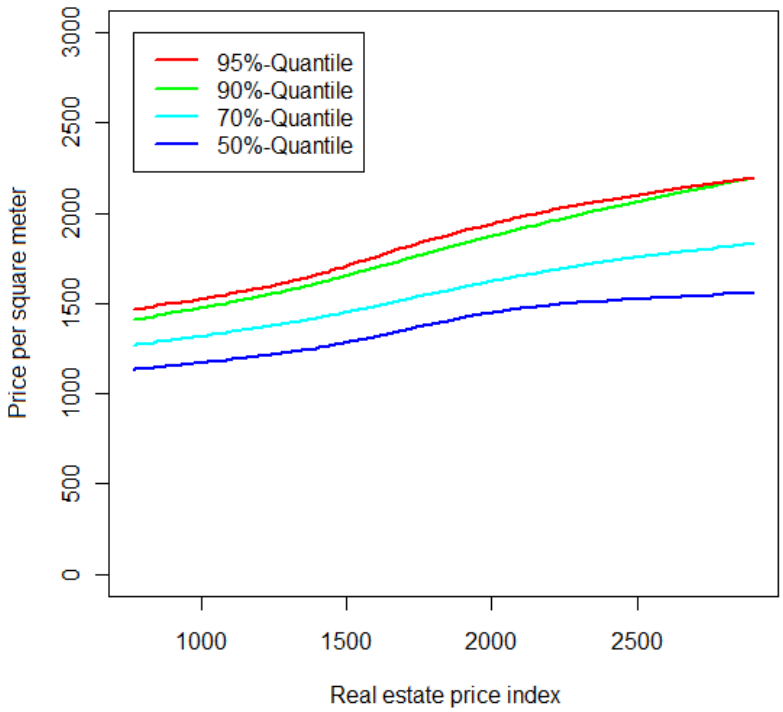




Results

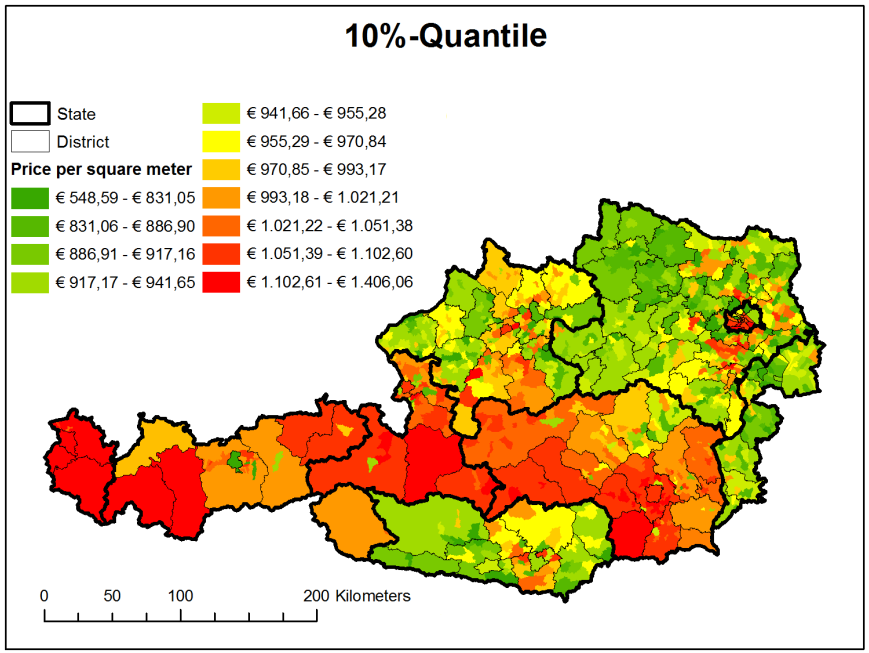
Real estate price index:

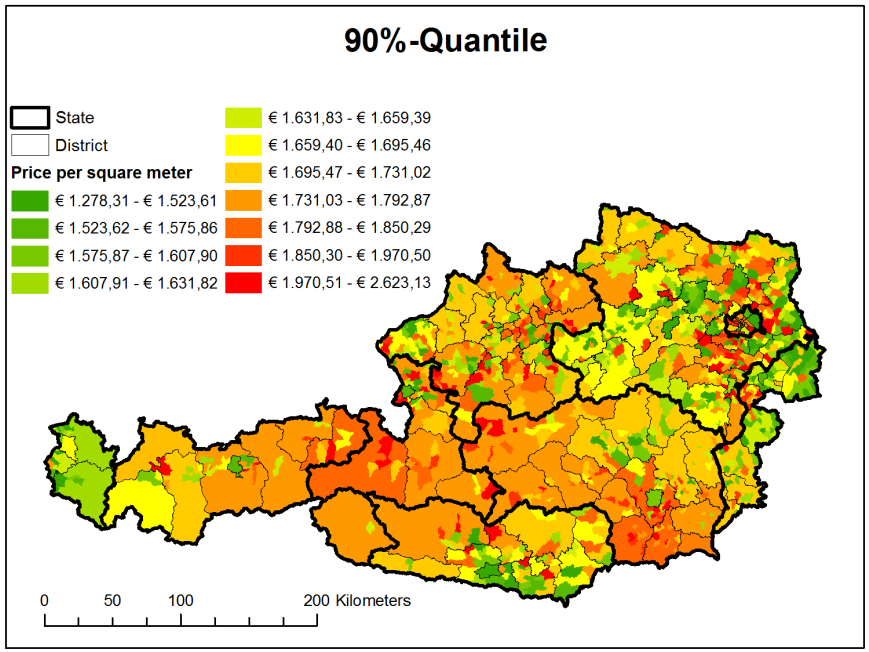




Results

Unexplained spatial effects:





Results

Model selection:

Quantile	Method	G1	...	G5	∅
10%	Quantile reg.	61.32	...	61.74	61.11
	Mean reg.	67.63	...	67.68	66.62
30%	Quantile reg.	124.10	...	125.61	125.42
	Mean reg.	125.15	...	128.41	127.71
50%	Quantile reg.	143.15	...	149.81	147.99
	Mean reg.	143.74	...	149.50	148.32
70%	Quantile reg.	129.89	...	136.01	134.63
	Mean reg.	132.45	...	135.68	136.46
90%	Quantile reg.	75.75	...	72.67	76.18
	Mean reg.	74.88	...	73.75	77.07

Conclusion

- Efficient Bayesian inference based on the ALD
- Individual marginal effects for each quantile
- Superior to mean regression

Thank you!