Austrian Journal of Statistics
2025, Volume 54, 44–68.
http://www.ajs.or.at/
doi:10.17713/ajs.v54i2.1904

A Parsimonious Hedonic Distributional Regression Model for Large Data with Heterogeneous Covariate Effects

Julian Granna
University of Innsbruck
Innsbruck, Austria

Wolfgang Brunauer
DataScience Service GmbH
Vienna, Austria

Stefan Lang
University of Innsbruck
Innsbruck, Austria

Nikolaus Umlauf
University of Innsbruck
Innsbruck, Austria

Abstract

Modeling real estate prices in the context of hedonic models often involves fitting a Generalized Additive Model, where only the mean of a (lognormal) distribution is regressed on a set of variables without taking other parameters of the distribution into account. Thus far, the application of regression models that model the full conditional distribution of the prices, has been infeasible for large data sets, even on powerful machines. Moreover, accounting for heterogeneity of effects regarding time and locale, is often achieved by naive stratification of the data rather than on a model basis.

A novel batchwise backfitting algorithm is applied in the context of a structured additive distributional regression model, which enables us to efficiently model all distributional parameters of the price distribution. Using a large German dataset of apartment asking prices with over one million observations, we employ a model-based clustering algorithm to capture the heterogeneity of covariate effects on the parameters with respect to dwelling locale. We thus identify clusters that are homogeneous with respect to the influence of dwelling locale on price. A boosting type algorithm of the batchwise backfitting algorithm is then used to automatically determine the variables relevant for modelling the location and scale parameters in each regional cluster. This allows for a different influence of variables on the distribution of prices depending on the locale and price segment of the dwelling.

Keywords: generalized additive models for location scale and shape, hedonic regression, regional cluster, real estate prices.

1. Introduction

Rosen (1974) interprets hedonic prices as the sum of implicit prices of attributes of a good. These directly unobservable prices, hence often referred to as "shadow prices", are estimated

employing regression models. Hedonic price models are of major importance, e.g. for the compilation of unbiased hedonic house price indices, as discussed in Granna, Brunauer, and Lang (2022). There exists a vast literature that extents the basic linear regression model discussing model assumptions and improving predictive accuracy.
The use of generalised additive models (GAM; Hastie and Tibshirani 2017; Wood 2017) allows fitting smooth effects without making restrictive assumptions about the corresponding functional relationship between the price and dwelling attributes. Applications include those of Waltl (2016), Brunauer, Lang, and Feilmayr (2013) or Hill and Scholz (2018). The latter model geospatial dependence of prices by estimating a spline surface of the longitude and latitude of houses. Razen and Lang (2020) fit smooth covariate effects on the price by employing penalised splines.

Another widely discussed issue in the literature is possible heterogeneity of effects on prices with respect to a) the locale of a dwelling and b) its price segment. The former stems from the recognition that diverse local conditions lead to varying price effects, as documented in numerous works such as Straszheim (1975) or Malpezzi, Ozanne, and Thibodeau (1980). There exists a large body of literature that segments global markets into smaller submarkets. Some authors explain the formation of such submarkets by market imperfections, e.g. Can (1992) or Goodman and Thibodeau (2003). In functioning markets, they would expect prices to equalise across regions to eliminate arbitrage opportunities. Other authors, such as Nesheim (2002) or Ekeland, Heckman, and Nesheim (2004), see the formation of submarkets as an integral component of the price-setting mechanism in functioning markets.
In terms of model methodology, heterogeneous effects on prices is often modeled using some form of clustering. There exists a wide spectrum of techniques aimed at identifying such homogeneous clusters. A popular strategy is k-means clustering. Examples include Abraham, Goetzmann, and Wachter (1994), Bourassa, Hamelink, Hoesli, and MacGregor (1999) and Tomal (2021). The latter use k-means clustering to identify county housing markets in Poland. Abraham *et al.* (1994) identify three groups of homogeneous markets in the US: West Coast, East Coast and Central US. Bourassa *et al.* (1999) combine principal component analysis (PCA) with k-means clustering to identify housing submarkets in Australia. Although the choice of clusters in these examples is data-driven, the clustering algorithm is not related to the actual hedonic modelling and the modelling itself is carried out separately. In contrast, examples of model-based clustering include Day, Bateman, and Lake (2004), who fit a finite mixture model to a dataset of house and bungalow sales from Birmingham, UK, to identify observations that are "close" to each other, although not in a spatial sense. McMillen and Redfearn (2010) apply nonparametric locally weighted regression (LWR) to a dataset in Chicago, USA, allowing for spatially varying coefficients. In this way, they investigate how house sales prices vary regarding their proximity to a rapid transit line. Still, all of these applications do not consider the full distribution of prices, but rather a single moment. Thus, these attempts fail in identifying varying effects of the covariates on the price depending on the underlying price segment.

Regression models that consider only a single moment of the distribution, e.g. the mean, fail to achieve this objective. Hence arises the need for models that take into account the entire distribution of prices, i.e. distributional or quantile regression. However, there are much fewer attempts to simultaneously account for spatial heterogeneity and the full conditional distribution of prices. Zietz, Zietz, and Sirmans (2008) apply a quantile regression approach (Koenker and Bassett 1978) on data investigating home sales from the Orem/Provo area in Utah, USA. Considering 1,366 sales between 1999 and 2000, they find that housing characteristics are not equally valued across the price distribution. Waltl (2019) also adopts a quantile regression approach and examines the differential influence of characteristics in three different price segments, i.e. quantiles, and three regional clusters for 565,587 houses in Sydney, Australia. She finds differing price developments both in urban versus suburban areas and in low-cost versus high-cost dwellings. Razen and Lang (2020) apply distributional regression

with cluster-specific heterogeneity using random scaling factors (Wechselberger, Lang, and Steiner 2016), assuming a homogeneous functional form between clusters but heterogeneity in their scaling. Using data on nearly 100,000 single-family homes in Germany, they find spatially diverse effects, especially with respect to differences between former East and West Germany.
Regarding the use of quantile versus distributional regression, Razen, Brunauer, Klein, Kneib, Lang, and Umlauf (2014) provide a comparison of the two and find that distributional regression is superior to quantile regression in the real estate context.
However, none of these approaches form the clusters on a model basis. Furthermore, while they allow for the varying influence of dwelling characteristics, they do not take into account the actual *relevance* of the covariates in each cluster.

Overall, three main problems still persist in the context of hedonic models of real estate data:

1. Many applications still do not consider the full conditional distribution of prices, thus neglecting heterogeneous effects depending on the price segment. In the context of quantile and distributional regression, there exists no model-based clustering mechanism for identification of locally, spatially coherent, homogeneous submarkets.

2. One possible reason for the lack of use of distributional models within the scope of real estate data, is that fitting distributional models has not been feasible so far because of the computational complexity. This is especially true for (very) large data sets.

3. Applications to date have completely ignored variable selection with the locally identified submarkets. Potential variation in the *relevance*, in addition to the magnitude, of covariate effects is a facet that has been completely ignored in the literature.

In addressing all of the aforementioned challenges, the contribution of this paper can be distilled into the following aspects:

- We propose a model-based clustering algorithm that partitions the data into more homogeneous, spatially coherent regions. We use a tree-based partitioning algorithm where each leaf in the regression tree is associated with a distributional regression model. This achieves optimal clustering such that the global negative log-likelihood is minimised.

- By using a boosting type algorithm in each cluster, we are able to *automatically* identify (ir)relevant covariates responsible for price formation. Thereby, we not only account for the varying influence of dwelling characteristics on the price, but also take their relevance into account.

- We apply a novel *batchwise backfitting* algorithm (Umlauf, Seiler, Wetscher, Thorsten Simon, and Klein 2024) that is able to fit distributional regression models efficiently even for very large data on a conventional machine. Our dataset of more than 1.2 million apartments in Germany is, to our knowledge, the largest dataset ever used in the context of distributional hedonic real estate models.

Thus, our proposed method combines model-based clustering, which accounts for spatial heterogeneity, and distributional regression, which sheds light on the entire distribution of prices rather than a single moment.

The rest of the paper is structured as follows. In section 2 we present our general estimation strategy and model design. In addition, we detail the methodological ideas behind the applied structured additive distributional regression framework and the applied model-based clustering algorithm. In section 3, we apply the introduced methodology to a dataset

of over 1.2 million apartment prices in Germany and discuss our findings regarding effect heterogeneity. Finally, we conclude our work in section 4.

2. Methodology

In this section, we begin with a description of our estimation strategy for identifying heterogeneous effects across locale and price segments. We then provide details about the utilised methods in the context of hedonic real estate modelling.
Our estimation strategy can be verbally summarised as follows.

1. We identify spatially homogeneous clusters of apartment rents by employing a model-based recursive partitioning algorithm. The algorithm is a tree-based method similar to a regression tree, where each leaf in the tree is not associated with a simple average, but instead, a fully specified structured additive distributional regression model. For clustering, we use a model containing all available variables. We choose the longitude and latitude of the centroids of the counties, which are similar to counties, as partitioning variables.

2. Following the identification of clusters, we perform variable selection in all of the clusters by applying a boosting-type variant of the computationally highly efficient batchwise-backfitting algorithm.

3. Finally, after identification of relevant variables in the clusters, we fit a final distributional regression model containing all relevant covariates.

2.1. Hedonic structured additive distributional regression

Early applications of hedonic regression go back to Haas (1922) and Wallace (1926), both of whom relate farmland prices to land attributes. Later applications include those of Court (1939), who constructed hedonic price indexes for automobiles. Rosen (1974) is the first to develop a theory of hedonic pricing in the context of housing, the idea being to decompose housing prices into the sum of the prices at which buyers value the characteristics of the dwellings. Thus, as stated by Sopranzetti (2015), the simplest hedonic model is given by

$$p_i = \sum_{j=1}^{J} \beta_j x_{ij} + \epsilon_i, \tag{1}$$

where p_i is the price (per square metre) of the dwelling is regressed on a set of J associated house characteristics x_{ij} with the corresponding coefficients, β_j. To deal with nonlinearities in the pricing structure as well as favourable distributional properties, a semilog form of equation (1) is often preferred. For notational simplicity, we refer to the log price per square metre as p_i.
To allow for smooth effects of the covariates that do not assume a specific relationship between housing attributes and the price, this model is often extended using generalised additive models (GAM; Hastie and Tibshirani 2017), such that

$$p_i = \sum_{j=1}^{J} \beta_j x_{ij} + \sum_{l=1}^{L} f_l(z_{il}) + \epsilon_i,$$

where f_l is a function that smoothly relates a covariate z_l to the dependent variable. This smooth function is usually modelled by basis, P(enalised) or thin-plate splines. Models incorporating splines as smooth functions are of increasing popularity in the literature. Brunauer *et al.* (2013) fit P-splines to model the effects of the metric covariates used in their analysis. Schäfer and Hirsch (2017) compare the performance of ordinary least squares (OLS) regression

with that of a GAM model on data from Berlin, Germany. They find that GAMs outperform OLS regression models and better capture effects on price related to spatial attributes of a dwelling, such as distance to amenities.

The hedonic approaches described thus far only model the mean of the prices' distribution by regressing it on a set of covariates. Quantile and distributional regression consider either multiple quantiles (quantile regression) or the full conditional distribution of prices. Razen *et al.* (2014) carry out a comparison of both approaches in the context of hedonic real estate regression and find distributional regression to be favourable.
Structured additive distributional models are employed to capture the relationship between a response variable and covariates while accounting for all distributional parameters of the response distribution. Using distributional regression, we are able to account for heterogeneous effects with regard to the price segment of the dwelling. By modeling all parameters of the distribution, we are able to identify varying influences on the price for varying quantiles of the price. Only considering the mean (or any single moment of a distribution) provides only very restricted insights on the covariate effects and completely disregards varying influence depending on the price segment. Zietz *et al.* (2008) find that buyers of high-priced dwellings in Utah, USA, value dwelling characteristics differently from buyers of lower-priced dwellings. Waltl (2019) also examines data from Sydney with a particular focus on differential effects across price segments and finds varying influences of features across price segments. She then constructs indices and shows differences in price trends between high- and low-cost dwellings.
The following notation follows that of Umlauf, Klein, and Zeileis (2018). In the context of distributional regression, the distribution of the response variable can be defined very broadly to follow any desired distribution. In our application, the response is the log price per square metre, p, and is defined, given a set of covariates $x = x_1, \ldots, x_J$, and $z = z_1, \ldots, z_L$ as

$$p|x, z \sim \mathcal{D}_p \left(\theta_1(x,z) = h_1^{-1}\left(\eta_1(x,z)\right), \theta_2(x,z) = h_2^{-1}(\eta_2(x,z)), \ldots, \theta_K(x,z) = h_K^{-1}(\eta_K(x,z))\right).$$

$\mathcal{D}_p$ represents an arbitrary parametric distribution of the log price p with K parameters θ_k, where each $\theta_k(x,z)$ is a function of x and z, i.e., the covariates. h_k^{-1} denote inverse link functions that relate the additive predictors η_k to the parameters of the distribution. η_k, representing the additive predictor for the k-th parameter $\theta_k(x,z)$, is then specified as

$$\eta_k = \beta_{1k}x_1 + \ldots + \beta_{J_k k}x_{J_k} + f_{1k}(z_1) + \ldots + f_{L_k k}(z_{L_k}),$$

where f_{lk} represents one of L_k functions of the predictor variables z_L. Each function captures an unspecified relationship between a predictor and the k-th parameter. They could be specified as, e.g., B-spline basis functions or, as in our application, as thin plate splines. β_{jk} corresponds to one of J_k coefficients capturing a linear effect of covariate x. Note that each predictor η_k is not required to contain the same set of predictor variables x and z. For each predictor, we actually allow different predictor variables based on automatic variable selection. However, we do not include this in the formula to keep the notation simple.

In our application, we assume a log-normal distribution for the prices for apartment prices given by

$$p \sim \text{logNO}(\eta_1 = \mu,\, \eta_2 = \log(\sigma)),$$

where μ is the mean and σ the corresponding standard deviation. In the course of our analysis, we modelled the price using a variety of distributions. In general, we found very little difference between the fitted distributions. In our view, the log-normal distribution offers a good marriage of high predictive accuracy and simplicity. η_1 is the predictor for the μ-parameter of the distribution with identity link. The additive predictor for the σ-parameter is given by η_2. We use a log link to ensure strictly positive values for σ.
To account for spatial heterogeneity, we further allow for varying effects of covariates based on model-based cluster membership. For each of the clusters identified using the model-based

rescursive partitioning algorithm (see section 2.2), we further perform automatic variable selection using a boosting variant of the batchwise backfitting algorithm (see section 2.2). In this way, we also allow covariates to be included only in those regional clusters where they are relevant to price formation.

2.2. Strategy for model choice and variable selection

Model-based identification of clusters

There are few examples in the literature, where spatial heterogeneity of effects is modelled in the context of distributional or quantile regression using real estate data. Waltl (2019) employs a quantile regression approach and considers varying effects of characteristics in three different price segments and three regional clusters (inner city, metropolitan area, and suburban districts) for houses in Sydney, Australia. She finds differing price developments both in urban versus suburban areas and in low-cost versus high-cost dwellings. Razen and Lang (2020) apply distributional regression with cluster-specific heterogeneity using random scaling factors (Wechselberger *et al.* 2016), assuming a homogeneous functional form between clusters but heterogeneity in their scaling. Using data on single-family homes in Germany, they find different effects across clusters, especially with respect to clusters in former East and West Germany. However, the choice of clusters is based on administrative boundaries rather than on a model. To date, there has been no application in the residential real estate context that combines distributional regression with a model-based clustering approach.

We achieve the identification of locally homogeneous clusters by adopting the model-based recursive partitioning algorithm as introduced by Zeileis, Hothorn, and Hornik (2008). Like generalised additive models, model-based recursive partitioning is a supervised statistical learning technique. The resulting model is similar to a standard regression tree, where each terminal node, i.e. leaf, is associated with a model (in our case a distributional regression model) rather than a simple average. Since this algorithm is a tree-based method, we introduce its idea by embedding it in a regression tree context. Our structure and notation follow the work of Hastie, Tibshirani, and Friedman (2009).

In general terms, regression trees are tree-based models fitted to a metric dependent variable. They present a relatively simple yet powerful tool. The feature space is partitioned into rectangles called nodes. In each of these nodes, a simple mean of the corresponding data is fitted. The rationale behind the concept is to identify split points within the explanatory variables that divide the characteristic space into two regions. The data are split until some stopping criterion, such as a pre-specified minimum number of observations in a node, is met. The final model can be visualised as a tree depicting the regions into which the data is split. Formally, a dependent variable, the log price per square metre, p, is considered along with L explanatory variables for n observations. For simplicity, we refer to all explanatory variables as z, including the linearly modelled covariates as x. The algorithm is designed to identify splitting variables and splitting points. The regression tree model is then expressed as

$$g(z) = \sum_{m=1}^{M} c_m I(z \in R_m),$$

which represents the response as a constant c_m and a partition with M regions $R_1, R_2, \ldots, R_m$ is created. If we then set the minimisation of the sum of squares $\sum(p_i - f(z_i))^2$, we get the optimal $\hat{c}_m$ as

$$\hat{c}_m = ave(p_i \mid z_i \in R_m),$$

which is simply the average of p_i in R_m. Since computing an optimal partition in terms of the sum of squares is usually infeasible, a greedy algorithm is adopted: First, a splitting variable

l is defined for which the characteristic space is split at point s such that

$$R_1(l,s) = \{z \mid z_l \leq s\} \text{ and } R_2(l,s) = \{z \mid z_l > s\}$$

are obtained. Finally, the splitting variable l and split point s are received by solving

$$\min_{l,s} \left[\min_{c_1} \sum_{z_i \in R_1(l,s)} (p_i - c_1)^2 + \min_{c_2} \sum_{z_i \in R_2(l,s)} (p_i - c_2)^2 \right],$$

where

$$\hat{c}_1 = ave(p_i \mid z_i \in R_1(l,s)) \text{ and } \hat{c}_2 = ave(p_i \mid z_i \in R_2(l,s))$$

solve the inner minimisation. In this way, the optimal pair (l, s) is obtained and the procedure is typically repeated until some minimum terminal node size is reached. The tree can then be pruned to avoid overfitting (Hastie *et al.* 2009).

We now extend the concept of regression trees by fitting a structured additive distributional regression model in each partition instead of the simple mean. This methodology, called *model-based recursive partitioning*, was introduced by Zeileis *et al.* (2008), whose notation we adapt to briefly introduce the method. Model-based recursive partitioning is an integration of parametric models into regression trees:
Suppose a global parametric model $\mathcal{M}(p, \theta)$ is given with observed log prices per square metre p and parameter vector θ. The model is then estimated by minimization of the objective function $\Psi(p, \theta)$, which is the negative log-likelihood, resulting into

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}} \sum_{i=1}^{n} \Psi(p_i, \theta), \tag{2}$$

where $\hat{\theta}$ is the parameter estimate given n observed prices $p_i (i = 1, \ldots, n)$. Then, instead of a global model $\mathcal{M}$, the characteristic space is divided into regions, or partitions, $R_1, R_2, \ldots, R_m$. Thus, each cell R_m holds a model $\mathcal{M}_m(p, \theta_m)$ corresponding to a cell-specific parameter θ_m yielding a globally segmented model $\mathcal{M}_M(p, \{\theta_m\})$. $\{\theta_m\}_{m=1,\ldots,M}$ thereby corresponds to the full combined parameter. Equation (2) formulated over all regions can then be written as the optimization problem

$$\sum_{m=1}^{M} \sum_{i \in I_m} \Psi(p_i, \theta_m) \to \min, \tag{3}$$

over all partitions $\{R_m\}$ with the indexes $I_m, m = 1, \ldots, M$. Equation (3) corresponds to a single model corresponding to each terminal node in a tree. The fitting of a model-based recursive partitioning model can then be summarised in the following algorithm:

1. In a possible node, fit the model with $\hat{\theta}$ to all corresponding observations by minimising the objective function Ψ, in our case the negative log-likelihood.
2. Calculate the split point s that locally minimises Ψ.
3. Split the current node into a set of child nodes and repeat the previous steps.
4. Grow a large tree until a defined minimum of observations is reached in each node. Then postprune the tree using BIC for regularisation.

Note: The original algorithm includes an optional fluctuation test in the second step, to evaluate whether the parameter estimates are stable with respect to any order in the partitioning variables j. However, we choose not to apply the fluctuation test and instead grow and postprune a very large tree. For a more detailed description of the steps, see Zeileis *et al.* (2008). The above algorithm is implemented in R (R Core Team 2023) in the `partykit` package by Hothorn and Zeileis (2015), which includes the `mob()` function.

Efficient estimation of models with automated variable selection

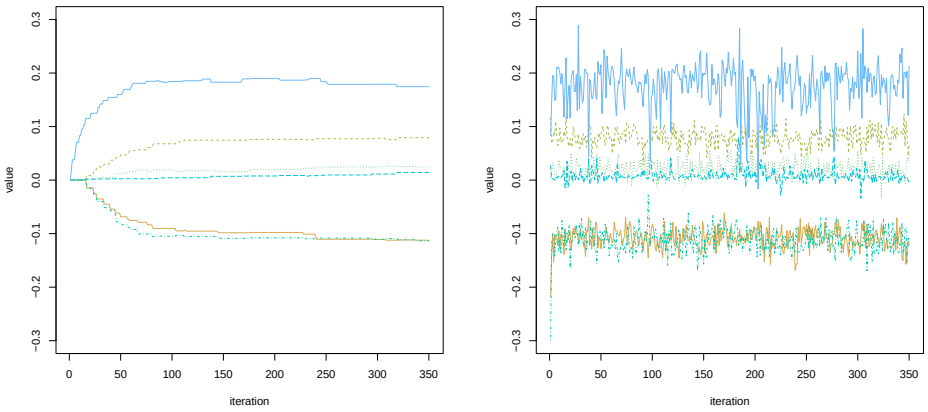
Until now, one of the main drawbacks of distributional regression models has been the computational complexity of their estimation. Maximising the (penalised) likelihood of the model is achieved by a backfitting algorithm introduced by Rigby and Stasinopoulos (2005), based on Marx (1996)'s iteratively reweighted (penalised) least squares (IRPLS), which incorporates first- and second-order information of the penalised likelihood. For a more detailed description of the algorithm, see Umlauf *et al.* (2024) or Umlauf *et al.* (2018).
Although estimation of the model is feasible for smaller datasets, estimation using the standard backfitting algorithm based on IRPLS is not feasible for large datasets such as those used in our analysis. This is especially true since the identification of clusters, as introduced in section 2.2, requires the fitting of hundreds of models to identify the optimal split point.
Batchwise backfitting (Umlauf *et al.* 2024) is based on the idea of updating model coefficients based on a random batch (or batches) of data. Given a randomly drawn subset $[i] \subseteq \{1, \ldots, n\}$, the model coefficients are updated in step $[t+1]$ with

$$\beta_{jk}^{[t+1]} = (1 - \nu) \cdot \beta_{jk}^{[t]} + \nu \cdot \beta_{[i],jk} \tag{4}$$

where ν is the step length control parameter, or learning rate, at which the coefficient $\beta_{jk}^{[t]}$ is updated to $\beta_{jk}^{[t+1]}$. For each iteration of the batchwise backfitting algorithm, (4) is run on a random batch. The batch size is considerably smaller than the full sample. This approach has two main advantages towards fitting the model on the full data:

- The estimation of the distributional regression model is feasible even for very large data sets like the one we have here. For the clustering algorithm explained in section 2.2, the model has to be fitted hundreds of times to compare improvements in the objective function, i.e. the negative log-likelihood. This would be impractical if the model were fitted to the full data. In this way, the estimation of the models would easily be computable on a conventional laptop within a day's computing time.
- Similar to stochastic gradient descent, batchwise backfitting carries less risk of getting stuck in a local minimum of the objective function.

When choosing the optimal batch size, there is a trade-off between stability of estimates (larger batch size) and faster computation time (smaller batch size), which we discuss in more detail in the context of our application.
A central contribution of this paper is to include the automated choice of *relevant* variables for modelling distribution parameters. For variable selection, (4) is altered such that not all model terms are updated, but only the term with the highest improvement in some information criterion, e.g. out-of-sample log-likelihood or AIC. This corresponds to a boosting-type algorithm. In our application, we only update the model term that maximises the reduction in AIC.
The estimation of the final model after variable selection is achieved by applying a resampling variant of the algorithm: We set the step size parameter $\nu = 1$ so that the updated coefficient $\beta_{jk}^{[t]}$ consists only of the new estimate $\beta_{[i],jk}$ based on the batch $[i]$. In this way, we obtain samples of a distribution of β_{jk} and obtain $\hat{\beta}$ by taking the mean (or median) of the coefficient paths after a burn-in period.
In Figure 1, we demonstrate the functionality of both approaches for a model fit on the pooled data with a batch size of $10,000$ and 350 iterations. The left panel in Figure 1 corresponds to the boosting-type flavor of the algorithm. The stepwise coefficient updates lead to the shown coefficient paths that converge towards their final estimated value. The resampling flavor is shown in the right panel in Figure 1. After a, in this application rather short, burn-in phase, each update resembles a sample from the coefficient's distribution. Because, $\nu = 1$, the updated coefficient is solely the estimate based on the current batch and the estimated values of the coefficient fluctuate around a steady state.



(a) Boosting (b) Resampling

Figure 1: Selection of paths for six coefficients a global model using batchwise backfitting with boosting and resampling variants of the algorithm

A considerable advantage of the algorithm is that, apart from specifying the model, we only have to specify the step length parameter ν, the number of batches and the batch size.
For the step length, simulations show that $\nu = 0.1$ represents a reasonable balance between convergence and numerical stability.
Regarding batch size and number of batches, no universally appropriate size exists. Increasing the batch size increases the stability of the estimates, but also increases the computation time of the models. As advised by Umlauf *et al.* (2024), we estimate intercept only models and increase the batch size until the coefficient paths become stationary. We also consider the convergence of the log-likelihood contribution plots of the full model, which track the contribution path of the corresponding variable to the (log-) likelihood.
With respect to the number of batches, a high number of batches ensures convergence of the algorithm at the cost of increased computational time. For both the boosting and resampling variants of the algorithm, we find that a batch size of $5,000$ and a number of batches of 350 is a good choice. Figure 2 depicts the contribution to the log-likelihood of corresponding variables. The plot shows the convergence of the paths from about iteration 200.

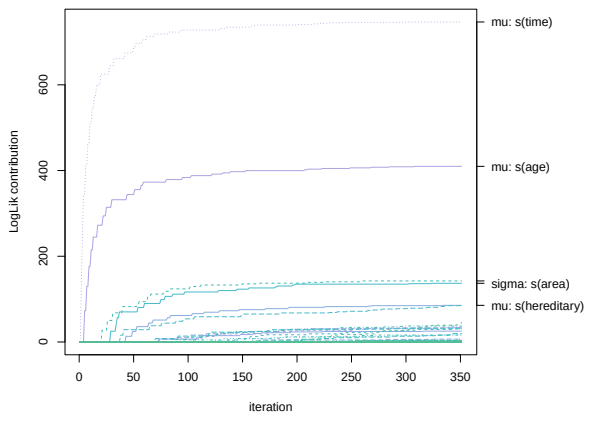


Figure 2: Log-likelihood contribution paths for model terms

The described methodology along with other functionalities is implemented in the R package

`bamlss` (Umlauf, Klein, Simon, and Zeileis 2021). We apply the method using the included `bamlss()`-function. For the boosting variant, we apply `opt_bbfit` as an optimizer, and for resampling, we use `opt_bbfitp`. For details on the estimation see also Umlauf *et al.* (2024), as well as the vignettes on the official website `http://bamlss.org/`.

3. Case study

3.1. Data

The data set we exploit for our application, is provided by DataScience Service GmbH, located in Vienna, Austria. The sample contains (log) asking prices for 1,235,002 observed dwellings comprising 40 variables distributed over 400 *Landkreise* (which we will refer to as *counties*) in Germany. Asking prices generally come with advantages and disadvantages. Major advantages lie in a larger sample size and earlier availability in contrast to transaction prices. This implies smaller standard errors in the predicted prices and price indices next to larger variability in the explanatory variables. The major disadvantage is a potential upward bias of the asking prices as the last offer price is usually greater than transaction prices. In our work, we disregard this upward bias of the prices.
Figure 3 illustrates the spatial distribution of observations across the counties in Germany. The density of observations is higher in rural areas and lower in urban areas. The maximum number of observations is in county 11000, which is the Berlin, where we observe 90,907 units. County 12070, Prignitz in the state of Brandenburg, contains only 83 observations, representing the global minimum in our data.

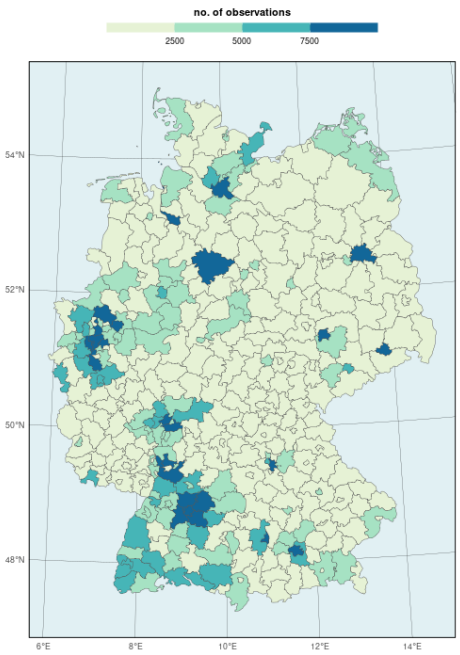


Figure 3: Number of observations over counties in Germany

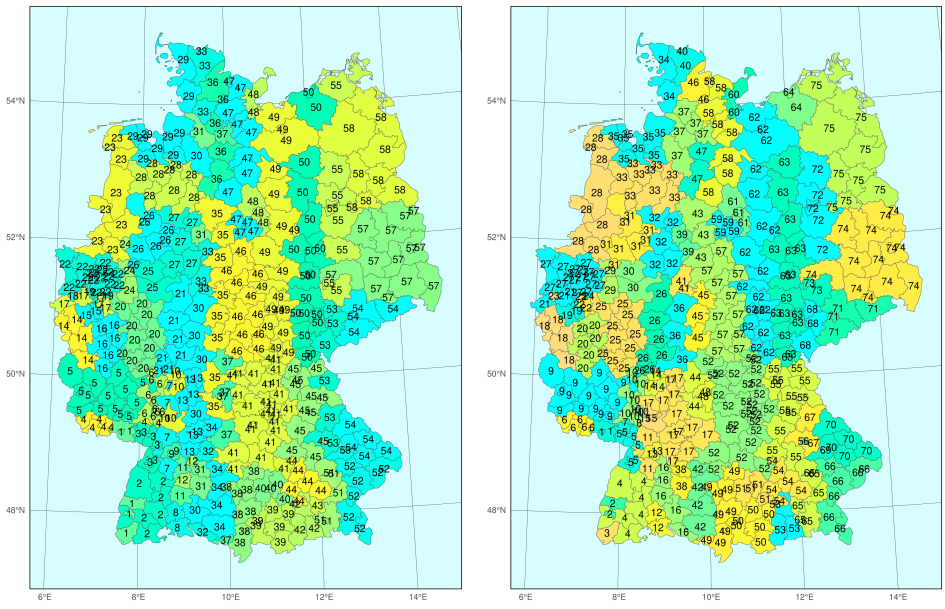
For our analysis, we consider a time horizon from the first quarter in 2016 to the third quarter in 2022. We offer a comprehensive list of variables considered in our analysis in Tables 2 and 3. For continuous covariates, we present the (arithmetic) mean, standard deviation, minimum, and maximum values. For discrete variables, we provide the relative frequencies of their respective values.

3.2. Identified housing clusters

As explained in more detail in the algorithm outlined in section 2, we fit a tree with a structured additive distributional model in each leaf to the data. Because the algorithm does not take into account the neighborhood structure of the counties, it does not necessarily yield spatially contiguous clusters. Thus, we postprocess the obtained clusters to ensure that all counties within each cluster are spatially contiguous:

1. Identify all clusters that contain spatially non-contiguous counties and split them into sub-clusters, such that all counties within each sub-cluster are spatially contiguous.
2. For each sub-cluster, identify all neighboring clusters.
3. Assign the sub-cluster to each neighbor and compute the corresponding prediction accuracy, measured as the out-of-sample mean squared error (MSE). Additionally, compute the out-of-sample MSE when the sub-cluster becomes its own cluster.
4. Identify the cluster for which the union with the sub-cluster yields the minimum MSE, and assign the sub-cluster to the corresponding cluster.

Figure 4 provides an overview of the obtained clusters. The first panel in Figure 4 displays the clusters as obtained from the model-based recursive partitioning algorithm. While most counties within the clusters are contiguous, some, like cluster 36, are not. The right panel in Figure 4 shows the clusters obtained after applying the postprocessing algorithm described above. The number of clusters increases from 58 to 75. Regarding predictive accuracy, the postprocessing of clusters improves the out-of-sample MSE from 0.122 to 0.113, which corresponds to a reduction of roughly 7.4%



(a) Preprocessed clusters (b) Postprocessed clusters

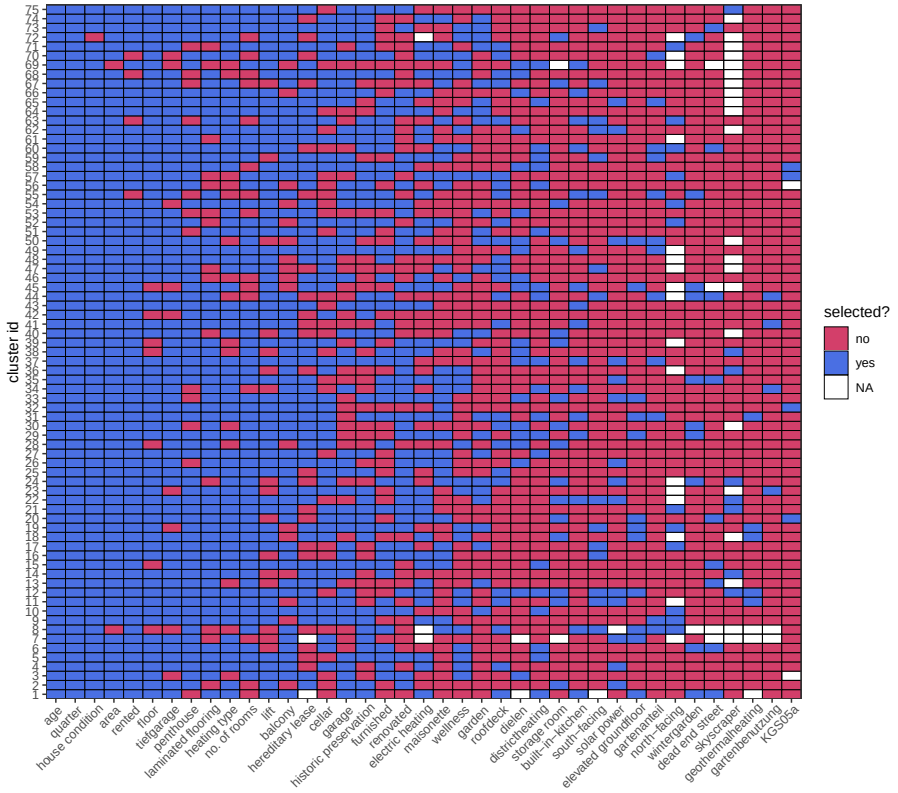
Figure 4: Resulting clusters from Model-Based Recursive Partitioning and postprocessing. Colours and numbers refer to obtained clusters.

3.3. Heterogeneity in effects across locale and price segment

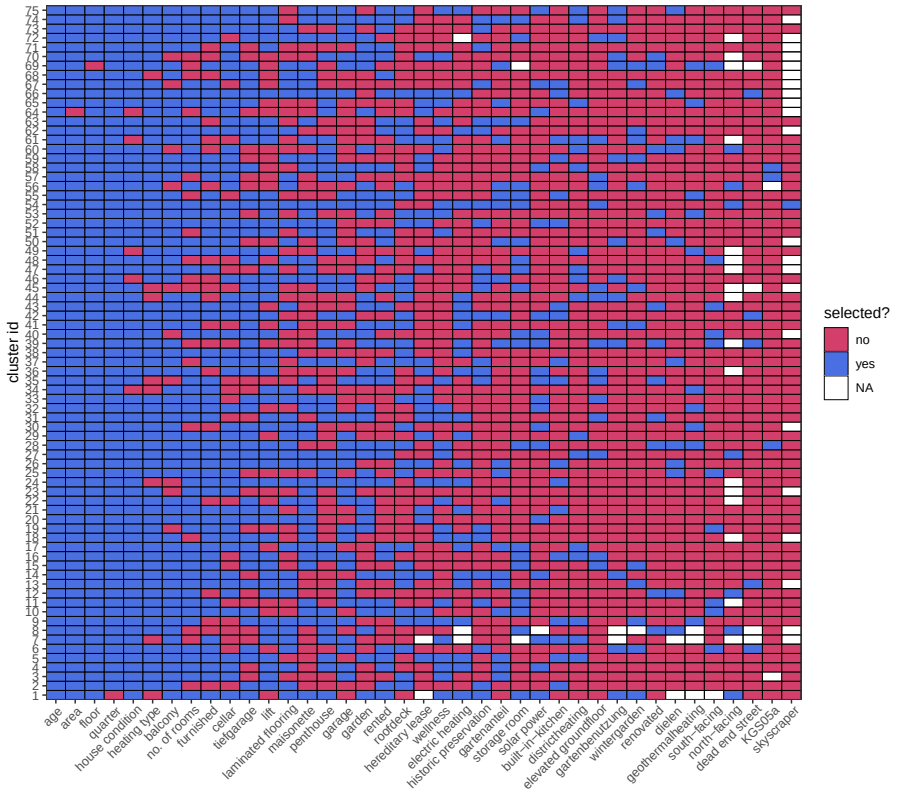
We account for spatial heterogeneity in covariate effects on μ and σ parameters of the lognormally distributed prices by fitting a distributional regression model in each identified cluster. A boosting flavor of the estimation procedure allows automated variable selection in each cluster and distributional parameter. Figure 5 indicates whether variables have been selected

for the μ and σ terms. The graph allows us to assess how homogenous log-prices are in general within each cluster. If a cluster contains homogenous flats, we expect in general very little variables to be included for modeling prices within each cluster. Moreover, we can assess, whether the same variables are (not) relevant within each cluster. If some variables are relevant to determination of prices only in specific clusters, we are able to identify the concerning clusters in this graph.
For some clusters, not all underlying categories of the corresponding variable are observed. Thus, variable selection is not possible for these variables in the concerned clusters (indicated with an "NA"). Both panels consist of a grid of 39 variables across 75 clusters, resulting in a total of $2,925$ fields in each graph.
The variables chosen for the μ term of the distribution are illustrated in the upper panel in Figure 5. Out of all the $2,925$ selection choices, $1,460$ indicate that a variable has been selected, while $1,403$ indicate that a variable has not been selected. Variables *age* and *quarter* are chosen for all clusters, while the continuous variable *area* is included in all clusters except for two. These results are in line with our expectations, as one would typically anticipate age, quarter, and area to consistently influence prices. Conversely, variables such as *skyscraper*, *use of garden*, and *geothermal heating* are infrequently selected for inclusion in the model.
The infrequent selection of the *skyscraper* variable highlights the benefit of automated variable selection within the model framework. In many regions, incorporating this variable would be impractical due to the limited presence of skyscrapers in Germany, which are concentrated in a few specific regions. Additionally, the inclination of residents to live in skyscrapers may differ from city to city based on the area's characteristics. Manually selecting such variables would be arduous, whereas automated selection based on the model streamlines the process and provides meaningful criteria for including or excluding variables in each cluster. Our model indicates that the *skyscraper* variable is relevant only in close proximity to large cities such as Frankfurt, Düsseldorf, or Cologne.
The lower panel in Figure 5 displays the variables (not) chosen for the σ term of the distribution. Out of all $2,925$ selection choices, $1,322$ fields indicate that the corresponding variable has been selected in the respective cluster. In $1,542$ cases, the variable of interest is not selected in the given cluster. Overall, the number of relevant variables for the σ term is lower compared to the μ term. Similarly to the variable selection in the μ term, the continuous variables *age*, *area*, and *quarter* are consistently chosen to be included in the model, with few exceptions. While there are some variations between the panels, the overall insight is that both the location and scale of the distribution are influenced by a substantial number of variables, and the selection of relevant variables in each cluster exhibits heterogeneity.
The figures indicate that identified clusters are very homogeneous with respect to the influence of dwelling locale on the price. The *county* variable *KGS05a*, which gives the region ID, is only selected to be included into both the μ and σ parameter of the model in three out of 75 clusters. This shows that the identified clusters are quite homogeneous regarding the influence of object locale.

To gain an impression of the heterogeneity of the metric covariates' effects, we depict the marginal effects at median for variables *age*, *area*, and *time* in Figure 6. For better interpretability, we show all effects on a linear scale instead of a log-scale. To correct for bias, we apply the bias correction as described by Greene (2017). For all effects, we see that the effects' levels differ substantially between the clusters: Specifically, clusters 53 and 54, which pertain to counties surrounding the greater Munich metropolitan area in Bavaria, exhibit the highest curves in all panels. Conversely, the lowest curve, representing cluster ID 68, corresponds to a rural area south of Leipzig in the state of Saxony.
Regarding the effect of *age* on flat prices, some spatial heterogeneity is present. In many clusters, prices per square meter decline monotonously with increasing age of the dwelling. However, in other areas, both new buildings and buildings over 60 years old sell for higher prices than middle-aged buildings. This difference in effects could be attributed to varying quality levels of older buildings in different clusters.



(a) μ-term



(b) σ-term

Figure 5: Selected variables for μ and σ terms. Variables are ordered from left to right from the most frequently selected variable to the least.

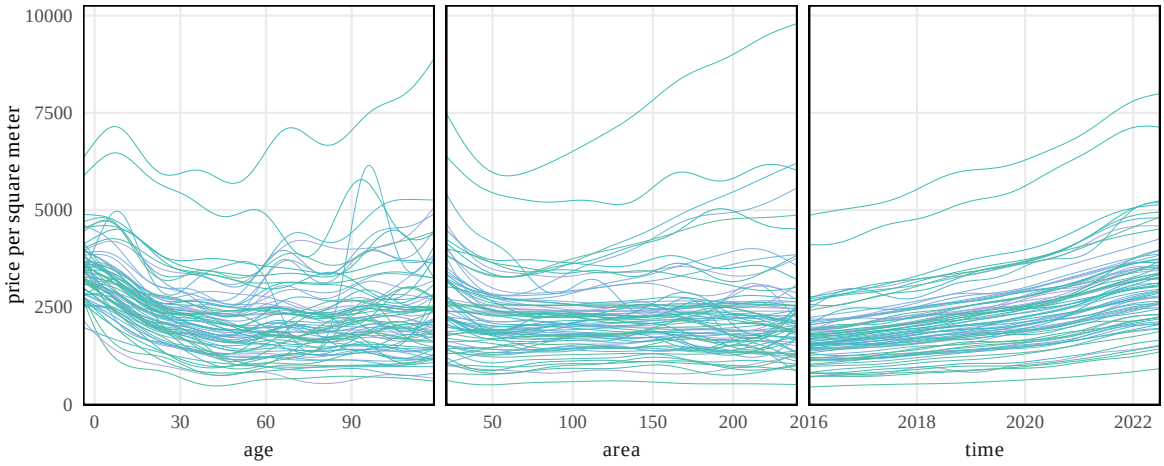


Figure 6: Marginal effects at median values for continuous variables age, area, and quarter for each of the 75 clusters

For *area*, there exists heterogeneity with respect to the effects as well. In most clusters, prices per square meter decrease monotonously with increasing area, which can be seen as a bulk discount for larger flats. However, in certain regions, especially those with higher overall price levels, prices increase again after surpassing an area threshold of about $60m^2$.

For all clusters, prices per square meter rise over the domain of time as the data does not cover more recent time periods. Similar to *age* and *area* effect plots, there is some heterogeneity of effects present. The slope of the functions varies between clusters, with higher priced regions exhibiting steeper slopes compared to regions with lower overall price levels.
In general, the relationships between metric covariates and prices show heterogeneity across different locales of flats. Particularly, flats located in areas with higher overall price levels show differing effects compared to those flats located in lower-priced regions.

To further assess the heterogeneity of effects alongside the improvement in the fit of our model, we report the out-of-sample prediction accuracy in Table 1. To provide a more complete overview of the distribution of (squared) prediction errors, we further report the median squared error (MedianSE) alongside the mean squared error (MSE). The table illustrates the importance of allowing for heterogeneous effects across regions in the context of hedonic house price models: Clustering alone, without automated variable selection, reduces the MSE by about 56% and the MedianSE by 61%. When the relevance of the effects within each cluster is taken into account by applying automated variable selection based on boosting, the MSE falls by a further 5% and the MedianSE by a further 6%. Hence, clustering, and thus accounting for heterogeneous effects, is by far the most relevant step, but the reduction in prediction error by introducing variable selection is also of relevant magnitude.

Table 1: Out-of-sample prediction accuracy of different models. The first line reports the mean squared error (MSE) of the employed models, the second line gives the corresponding median squared error (MedianSE).

	Global Model	With Clustering (no variable selection)	With Clustering
MSE	0.2670	0.1183 (-56%)	0.1130 (-5%)
MedianSE	0.0947	0.0372 (-61%)	0.0351 (-6%)

Figure 7 shows the marginal effect plots for the μ parameter at median values for the county variable. The panels reveal the heterogeneity of the effect both in terms of locale and price segment. For orientation, the maps also mark cities hosting more than 500,000 inhabitants.

As far as the overall price level is concerned, two main findings emerge. Firstly, prices per square metre are elevated in urban areas around large cities. This in particular applies to Hamburg in the north, the German capital Berlin and Munich. The high-priced region around Munich bleeds down into the south, likely reflecting homeowners' preferences to live closer to the Alps in the very south of Germany. Second, former East Germany is associated with lower prices per square metre than areas in former West Germany, with the exception of areas around Berlin.
In terms of the effect of dwelling locale across price segments, the overall pattern of price levels is similar across all price segments. The patterns of high and low priced regions identified above hold for all quantiles plotted. However, the variability of the county effect is different. For the 10% price quantile, the map features both very bright and very dark areas. On the other hand, the panel for the 90% quantile is dominated by darker shaded areas. The difference in price levels between high and low price counties is therefore smaller than in the lower price segments.

Provided in the Appendix, Figure 9 illustrates the influence of automatic variable selection implemented in boosting flavour of the batchwise backfitting algorithm. The left panel corresponds to a model in which the county variable is consistently included in all models for all clusters, neglecting the relevance of its influence within each cluster as suggested by the automatic variable selection. In contrast, the right panel represents the model where boosting determines whether the county variable is included in the models, allowing for a data-driven selection process. As previously established, the boosting flavour model only selects the county variable to be included in the μ term in four clusters. In all other clusters, the county variable is chosen not to be included (or, in two cases, the cluster consisted of only a single county). The differences between the two types of model are particularly evident in the northern and north-eastern parts of the country. In Figure 9, prices in the corresponding urban regions are more distinct from their surroundings. Conversely, in Figure 7 the former East Germany is more clearly differentiated in terms of its price level from the former West Germany.
Regarding out-of-sample prediction accuracy, always including the county variable in the regional models leads to an increase in MSE of approximately 4% and an increase in MedianSE of 6%. Hence, the model-based choice of including the county variable in the models accounts for most of the reduction in prediction error for all variables. This improvement in prediction accuracy cannot be solely attributed to low observation counts in the corresponding areas, as clusters also include observation-dense areas. Additionally, with 1.2 million observations in total, most counties contain a sufficient amount of data. These findings indicate that model-based recursive partitioning based on distributional regression models is capable of identifying homogeneous clusters, particularly concerning the influence of objects' locale on prices. This is in line with our expectations, as one would expect that a clustering algorithm based on locale is likely to identify clusters that are homogeneous in terms of dwelling locale. This result underlines the usefulness of our approach.

Figure 8 sheds further light on the homogeneity of the influence of covariates across price segments. The panels correspond to clusters 4, 15, 35, 54 and 74 or regions in or near Freiburg (Breisgau), Heidelberg, Bremen, Munich and Berlin (in that order). However, cluster 74 corresponds rather to an area forming a triangle between Berlin, Dresden and Leipzig in former East Germany. Plotting seven quantiles provides more insight into the entire distribution of prices. Similar to Figure 6, the presence of heterogeneity of effects across dwelling locale is visible, which manifests itself in diverse functional relationships between clusters. In some clusters, the effect of age on price is monotonically decreasing, as in clusters 4 and 35. In others, the relationship between the two is U-shaped. Thus, in some clusters, buyers strictly prefer newer buildings to older ones, while in others they prefer either new or historic buildings (clusters 15 and 74 for higher quantiles).
Heterogeneity across price segments is less relevant in some clusters, such as cluster 4, than in others, such as cluster 15. Regarding the effect of age on μ for the latter, the influence of

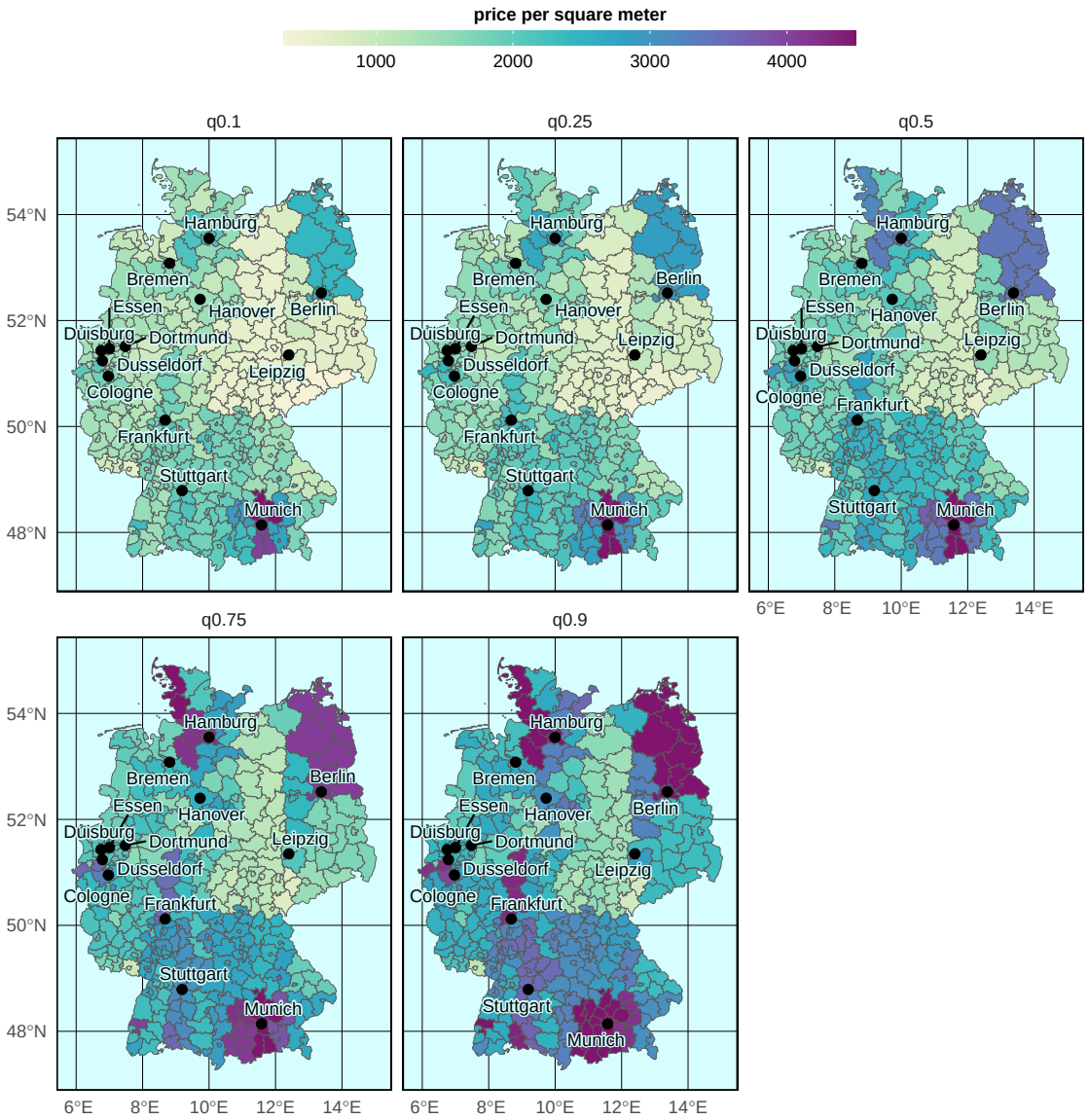


Figure 7: Marginal effects on the μ parameter at the county median (KGS) for each of the 75 clusters. County variable (not) included on the basis of automatic variable selection. Panel titles refer to the quantile of the price segment. Thus 'q0.1' refers to the county effect on μ for the 10% quantile of prices per square metre. Darker colours refer to higher prices and vice versa.

age differs between flat prices in the lower and higher segments. For the 2.5% quantile, the age effect is (almost) monotonically decreasing. In contrast, for high-priced dwellings in the 97.5% quantile, the relationship between age and price is U-shaped, indicating a higher total price for expensive historic buildings than for less expensive historic dwellings.
Regarding time, a similar pattern is apparent. In clusters 4 and 35, i.e. regions around Freiburg (Breisgau) and Bremen, the price trend appears homogeneous across all price segments. In other clusters, such as 54 and 74, referring to Munich and the Berlin - Dresden - Leipzip triangle, price trends are steeper in higher quantiles than in lower. Although prices have increased in all price segments, high priced apartments are hence subject to stronger price appraisals compared to low priced apartments.
For the sake of clarity, we only include graphs for 5 of the 75 clusters. For completeness, and also for a more comprehensive overview of all clusters, we provide an animated graph online containing the marginal effect curves for the seven quantiles for all 75 clusters.[1]

Overall, we find that accounting for heterogeneity across price segments is relevant. However, our results also suggest that heterogeneity across object locale plays a greater role. This is reflected in the fact that fewer variables, less than half the total number, are selected in the σ parameter of the model than in the μ parameter. We additionally fitted a model assuming homoscedasticity, i.e. fitting σ only to an intercept term, and compared it to our final model with variables included on the basis of boosting. We then compared the two models using the continuous ranked probability score (CRPS), which is a generalisation of the mean absolute error (MAE). The CRPS metric indicates the discrepancy between the predicted cumulative distribution function (CDF) and the observed value. It therefore provides a more comprehensive picture of the quality of the predicted distribution. The model assuming homoscedasticity has a CRPS of circa 0.1791, while our final model accounting for heteroscedasticity has a score of around 0.1780 (lower is better). This corresponds to a reduction of about 0.6%. Thus, modelling the σ parameter, and thus accounting for the heterogeneity of effects across price segments, slightly improves the fit to the data, but is far less relevant in terms of model fit compared to the clustering of observations.

4. Discussion

Accounting for heterogeneous effects of housing attributes on prices, both with respect to the position of dwellings and the price segment, is of great but still growing interest in the literature. Clustering techniques are often not model-based and unrelated to the underlying hedonic model. Although some authors take into account the varying impact of dwelling characteristics, variable selection is usually not carried out in each cluster, implicitly assuming that the same set of variables is relevant in each region. Furthermore, distributional regression was not feasible for very large datasets due to computational complexity.

In this paper, we apply a model-based clustering approach in combination with a novel batch-wise backfitting algorithm on a large dataset to account for both varying influence of housing characteristics based on the dwellings' price segment and locale. Further, we allow not only for varying influence of covariates, but also allow for variables to be automatically included in regions, where they are relevant and excluded in regions, where they have no influence on the price. Thus, our model is not only less biased, but also parsimonious.

We identify spatially coherent clusters that are homogeneous especially with respect to the influence of dwelling locale on the price. The model-based clustering algorithm reduces the out-of-sample MSE, and thereby the bias, by 56% compared with a model assuming homogeneous effects of housing positioning. We find differing functional forms in the relationships between continuous covariates and the price, especially regarding the influence of time, which

[1] https://jgranna.github.io/quantile_plots

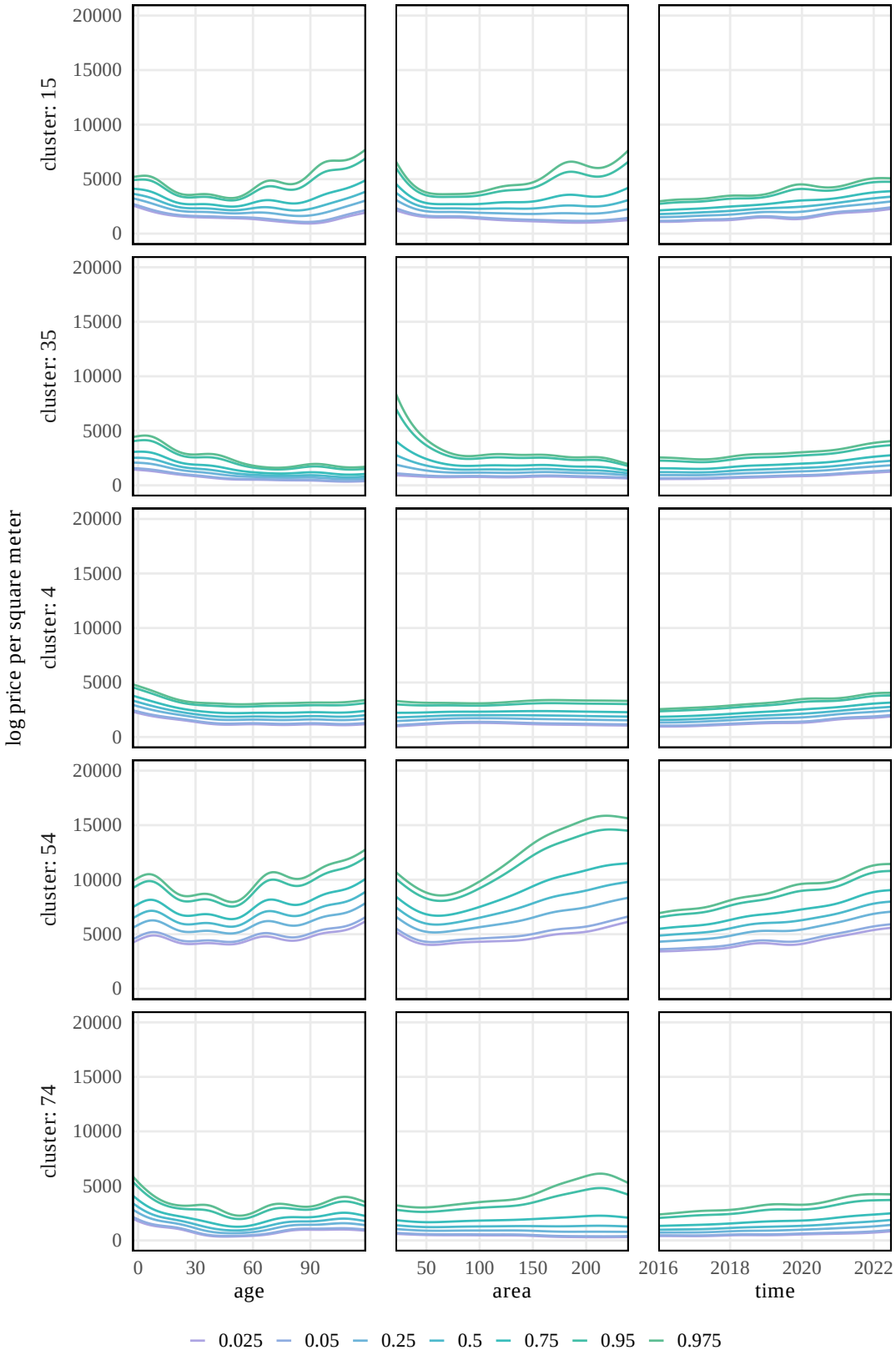


Figure 8: Marginal effects on μ at median values for continuous variables age, area, and time for 5 selected clusters across varying price segments, i.e. quantiles. The rows correspond to the indicated cluster. Columns refer to the indicated covariate effect.

indicates diverse time trends in the clusters. Accounting for heteroskedasticity and thus allowing for differential effects across price segments is also relevant, but not to the same extent as accounting for spatial heterogeneity. Model fit, as measured by the CRPS, is improved by modelling the σ parameter, but by far not to the same magnitude. Still, we find diverse effects of covariates depending on the price quantile in some regional clusters. Our results show a substantial variation in the variables that are automatically selected into the cluster models. Some variables, like *area* and *age* are always selected to be included, other variables are very rarely selected to be included. With regard to improvement of prediction accuracy, automated variable selection improves MSE by another 5%.

A possible extension of our work could be to assess whether the covariate effects are indeed subject to functional heterogeneity, as assumed in our analysis. An alternative approach could be to assume functional homogeneity across clusters and heterogeneity across clusters only with respect to the scaling of the effects, as suggested by Wechselberger *et al.* (2016) and investigated in Chapter 3. A useful extension of the framework used in this paper could be the inclusion of such random scaling factors in the context of batchwise backfitting. However, this remains a matter for future research.

References

Abraham JM, Goetzmann WN, Wachter SM (1994). "Homogeneous Groupings of Metropolitan Housing Markets." *Journal of Housing Economics*, **3**, 186–206. ISSN 10511377. doi:10.1006/jhec.1994.1008.

Bourassa SC, Hamelink F, Hoesli M, MacGregor BD (1999). "Defining Housing Submarkets." *Journal of Housing Economics*, **8**, 160–183. ISSN 10511377. doi:10.1006/jhec.1999.0246.

Brunauer WA, Lang S, Feilmayr W (2013). "Hybrid Multilevel STAR Models for Hedonic House Prices." *Jahrbuch für Regionalwissenschaft*, **33**, 151–172. ISSN 0173-7600. doi:10.1007/s10037-013-0074-9.

Can A (1992). "Specification and Estimation of Hedonic Housing Price Models." *Regional Science and Urban Economics*, **22**, 453–474. ISSN 01660462. doi:10.1016/0166-0462(92)90039-4.

Court AT (1939). *Hedonic Price Indexes with Automotive Examples*, pp. 98–119. General Motors.

Day B, Bateman I, Lake I (2004). "Nonlinearity in Hedonic House Price Equations: An Estimation Strategy Using Model-Based Clustering." CSERGE Working Paper EDM, No. 04-02.

Ekeland I, Heckman JJ, Nesheim L (2004). "Identification and Estimation of Hedonic Models." *Journal of Political Economy*, **112**, S60–S109. ISSN 0022-3808. doi:10.1086/379947.

Goodman AC, Thibodeau TG (2003). "Housing Market Segmentation and Hedonic Prediction Accuracy." *Journal of Housing Economics*, **12**, 181–201. ISSN 10511377. doi:10.1016/S1051-1377(03)00031-7.

Granna J, Brunauer W, Lang S (2022). "Proposing a Global Model to Manage the Bias-variance Tradeoff in the Context of Hedonic House Price Models." Working Papers in Economics and Statistics, No. 2022-12, Faculty of Economics and Statistics, Universität Innsbruck. URL https://EconPapers.repec.org/RePEc:inn:wpaper:2022-12.

Greene WH (2017). *Econometric Analysis, 8th edition*. Pearson. ISBN 9781292231136.

Haas G (1922). *Sale Prices as a Basis for Farmland Appraisal*. Technical Bulletin. University Farm. ISBN 9781248525999.

Hastie T, Tibshirani R, Friedman J (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer. doi:10.1007/978-0-387-84858-7.

Hastie TJ, Tibshirani RJ (2017). *Generalized Additive Models*. Routledge. ISBN 9780203753781. doi:10.1201/9780203753781.

Hill RJ, Scholz M (2018). "Can Geospatial Data Improve House Price Indexes? A Hedonic Imputation Approach with Splines." *Review of Income and Wealth*, **64**, 737–756. ISSN 00346586. doi:10.1111/roiw.12303.

Hothorn T, Zeileis A (2015). "partykit: A Modular Toolkit for Recursive Partytioning in R." *Journal of Machine Learning Research*, **16**, 3905–3909.

Koenker R, Bassett G (1978). "Regression Quantiles." *Econometrica*, **46**, 33. ISSN 00129682. doi:10.2307/1913643.

Malpezzi S, Ozanne L, Thibodeau T (1980). *Characteristic Prices of Housing in Fifty-Nine Metropolitan Areas*. Urban Institute.

Marx BD (1996). "Iteratively Reweighted Partial Least Squares Estimation for Generalized Linear Regression." *Technometrics*, **38**, 374. ISSN 00401706. doi:10.2307/1271308.

McMillen DP, Redfearn CL (2010). "Estimation and Hypothesis Testing for Nonparametric Hedonic House Price Functions." *Journal of Regional Science*, **50**, 712–733. ISSN 00224146. doi:10.1111/j.1467-9787.2010.00664.x.

Nesheim L (2002). "Equilibrium Sorting of Heterogeneous Consumers across Locations: Theory and Empirical Implications." cemmap working paper, No. CWP08/02. doi:10.1920/wp.cem.2002.0802.

R Core Team (2023). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org.

Razen A, Brunauer W, Klein N, Kneib T, Lang S, Umlauf N (2014). "Statistical Risk Analysis for Real Estate Collateral Valuation Using Bayesian Distributional and Quantile Regression." Working Papers in Economics and Statistics, No. 2014-12.

Razen A, Lang S (2020). "Random Scaling Factors in Bayesian Distributional Regression Models with an Application to Real Estate Data." *Statistical Modelling*, **20**, 347–368. ISSN 1471-082X. doi:10.1177/1471082X18823099.

Rigby RA, Stasinopoulos DM (2005). "Generalized Additive Models for Location, Scale and Shape." *Journal of the Royal Statistical Society Series C: Applied Statistics*, **54**, 507–554. ISSN 0035-9254. doi:10.1111/j.1467-9876.2005.00510.x.

Rosen S (1974). "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition." *Journal of Political Economy*, **82**, 34–55. ISSN 0022-3808. doi:10.1086/260169.

Schäfer P, Hirsch J (2017). "Do Urban Tourism Hotspots Affect Berlin Housing Rents?" *International Journal of Housing Markets and Analysis*, **10**, 231–255. ISSN 1753-8270. doi:10.1108/IJHMA-05-2016-0031.

Sopranzetti BJ (2015). *Hedonic Regression Models*, pp. 2119–2134. Springer New York. doi:10.1007/978-1-4614-7750-1_78.

Straszheim MR (1975). *An Econometric Analysis of the Urban Housing Market*. NBER. ISBN 0870145126.

Tomal M (2021). "Housing Market Heterogeneity and Cluster Formation: Evidence from Poland." *International Journal of Housing Markets and Analysis*, **14**, 1166–1185. ISSN 1753-8270. doi:10.1108/IJHMA-09-2020-0114.

Umlauf N, Klein N, Simon T, Zeileis A (2021). "bamlss : A Lego Toolbox for Flexible Bayesian Regression (and Beyond)." *Journal of Statistical Software*, **100**. ISSN 1548-7660. doi:10.18637/jss.v100.i04.

Umlauf N, Klein N, Zeileis A (2018). "BAMLSS: Bayesian Additive Models for Location, Scale, and Shape (and Beyond)." *Journal of Computational and Graphical Statistics*, **27**, 612–627. ISSN 1061-8600. doi:10.1080/10618600.2017.1407325.

Umlauf N, Seiler J, Wetscher M, Thorsten Simon SL, Klein N (2024). "Scalable Estimation for Structured Additive Distributional Regression." *Journal of Computational and Graphical Statistics*, pp. 1–23. doi:10.1080/10618600.2024.2388604.

Wallace HA (1926). "Comparative Farm-Land Values in Iowa." *The Journal of Land & Public Utility Economics*, **2**, 385. ISSN 15489000. doi:10.2307/3138610.

Waltl SR (2016). "A Hedonic House Price Index in Continuous Time." *International Journal of Housing Markets and Analysis*, **9**, 648–670. ISSN 1753-8270. doi:10.1108/IJHMA-10-2015-0066.

Waltl SR (2019). "Variation across Price Segments and Locations: A Comprehensive Quantile Regression Analysis of the Sydney Housing Market." *Real Estate Economics*, **47**, 723–756. ISSN 1080-8620. doi:10.1111/1540-6229.12177.

Wechselberger P, Lang S, Steiner WJ (2016). "Additive Models with Random Scaling Factors: Applications to Modeling Price Response Functions." *Austrian Journal of Statistics*, **37**, 255–270. doi:10.17713/ajs.v37i3&4.307.

Wood SN (2017). *Generalized Additive Models: An Introduction with R*. Chapman and Hall/CRC. doi:10.1201/9781315370279.

Zeileis A, Hothorn T, Hornik K (2008). "Model-Based Recursive Partitioning." *Journal of Computational and Graphical Statistics*, **17**, 492–514. ISSN 1061-8600. doi:10.1198/106186008X319331.

Zietz J, Zietz EN, Sirmans GS (2008). "Determinants of House Prices: A Quantile Regression Approach." *The Journal of Real Estate Finance and Economics*, **37**, 317–333. ISSN 0895-5638. doi:10.1007/s11146-007-9053-7.

A. Descriptive tables

Table 2: Summary statistics of variables considered for modeling. Arithmetic mean is provided for metric covariates, relative frequency for discrete variables (Part 1).

variable	description	mean / rel. frequency	std. deviation	min	max
ppqm	Price per square meter	2.070	0.080	1.360	2.420
log.ppqm	log(Price per square meter)	7.930	0.600	3.910	11.260
age	Age of flat in years	37.680	30.230	-4.000	119.000
area	Area of flat in square meters	80.080	32.380	21.000	240.000
quarter	Quarter of last offer	2019.200	2.010	2016.000	2022.500
KGS	county ID				
balcony	Whether object has balcony				
	0 = no	0.435			
	1 = yes	0.565			
built.in.kitchen	Whether object has built-in kitchen				
	0 = no	0.993			
	1 = yes	0.007			
cellar	Whether object has a cellar				
	0 = no	0.461			
	1 = yes	0.539			
condition	condition of object				
	1 = other	0.634			
	2 = first-time occupancy	0.161			
	3 = renovated	0.173			
	4 = in need of renovation	0.031			
cul.de.sac	Whether object is located in cul de sac				
	0 = no	0.998			
	1 = yes	0.002			
district.heating	Whether object features district heating				
	0 = no	0.946			
	1 = yes	0.054			
electric.heating	Whether object has electric heating				
	0 = no	0.997			
	1 = yes	0.003			
elevated.ground.floor	Whether object is on elevated ground floor				
	0 = no	0.994			
	1 = yes	0.006			
floor.boards	Whether object has floorboards				
	0 = no	0.995			
	1 = yes	0.005			
floor.heating	Whether object has floor heating				
	0 = no	0.909			
	1 = yes	0.091			
furnished	Whether object is furnished				
	0 = no	0.967			
	1 = yes	0.033			
garage	Whether object has a garage				
	0 = no	0.882			
	1 = yes	0.118			
garden	Whether object features a garden				
	0 = no	0.924			
	1 = yes	0.076			
garden.share	Whether object inlcudes share of a garden				
	0 = no	0.990			
	1 = yes	0.010			
garden.use	Whether object inlcudes a shared garden				
	0 = no	0.997			
	1 = yes	0.003			
geothermal.heating	Whether object features geothermal heating				
	0 = no	0.998			
	1 = yes	0.002			

Table 3: Summary statistics of variables considered for modeling. Arithmetic mean is provided for metric covariates, relative frequency for discrete variables (Part 2).

variable	description	mean / rel. frequency	std. deviation	min	max
heating.type	Type of heating in dwelling				
	1 = other	0.079			
	2 = unknown	0.661			
	3 = district_heating	0.052			
	4 = gas	0.189			
	5 = oil	0.019			
hereditary.lease	Whether dwelling is hereditary lease type				
	0 = no	0.984			
	1 = yes	0.016			
historic.preservation	Whether dwelling is historically preserved				
	0 = no	0.978			
	1 = yes	0.022			
laminated.flooring	Whether object features laminated flooring				
	0 = no	0.954			
	1 = yes	0.046			
lift	Whether object features a lift				
	0 = no	0.710			
	1 = yes	0.290			
maisonette	Whether object has multiple floors				
	0 = no	0.944			
	1 = yes	0.056			
no..of.rooms	Number of rooms in the flat				
	1	0.092			
	2	0.310			
	3	0.389			
	4	0.160			
	> 5	0.049			
north.facing	Whether object is facing north				
	0 = no	0.999			
	1 = yes	0.001			
refurbished.flat	Whether object is refurbished				
	0 = no	0.809			
	1 = yes	0.191			
rented	Whether object is rented				
	0 = no	0.719			
	1 = yes	0.281			
skyscraper	Whether object is located in skyscraper				
	0 = no	0.999			
	1 = yes	0.001			
solar.power	Whether object features solar power				
	0 = no	0.997			
	1 = yes	0.003			
south.facing	Whether object is facing south				
	0 = no	0.997			
	1 = yes	0.003			
storage.room	Whether object has a storage room				
	0 = no	0.993			
	1 = yes	0.007			
underground.parking	Whether object has underground parking				
	0 = no	0.871			
	1 = yes	0.129			
wellness	Whether object has a wellness area				
	0 = no	0.996			
	1 = yes	0.004			
winter.garden	Whether object has a winter garden				
	0 = no	0.994			
	1 = yes	0.006			

B. Estimated effects of counties: comparison between models

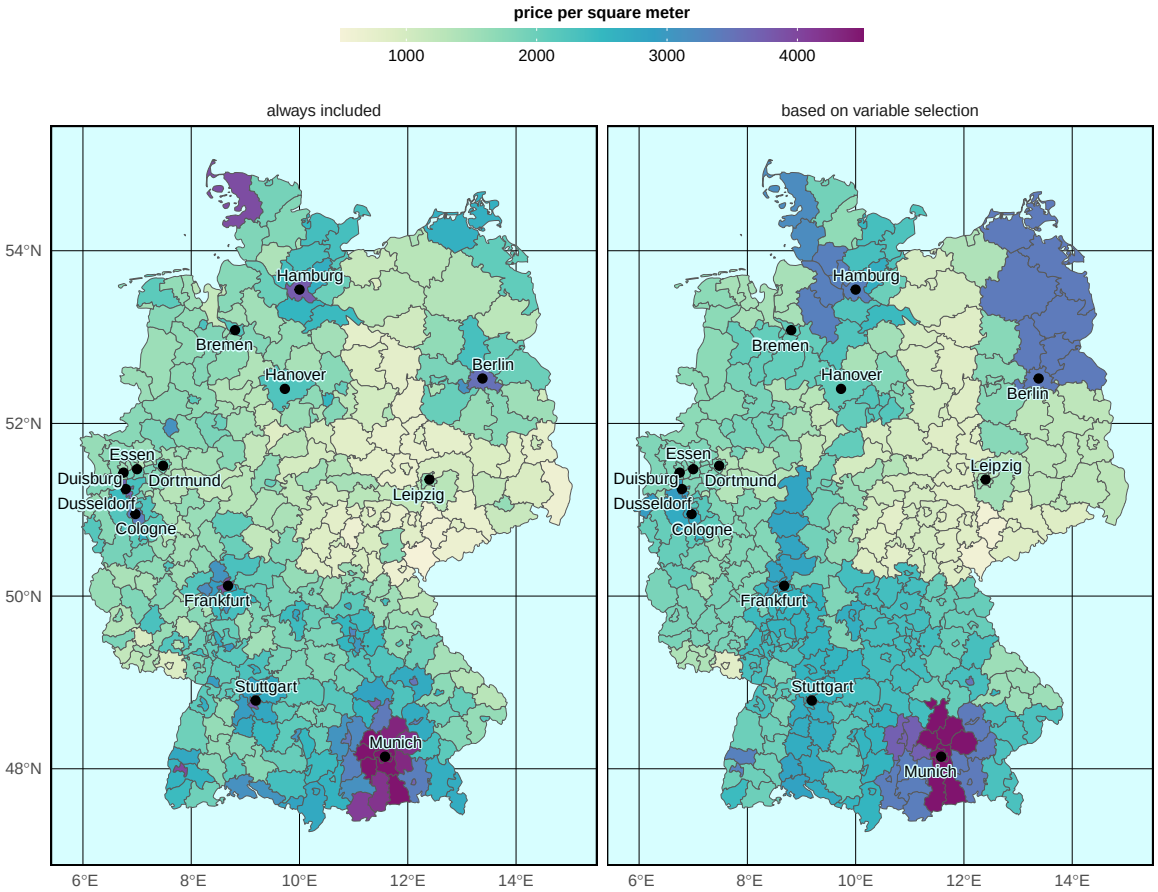


Figure 9: Marginal effects (50% quantile) on the μ parameter at county median values (KGS) for each of the 75 clusters. Panel titles refer to whether the county variable is always included (left panel), or whether the inclusion of the county variable in each cluster is based on the boosting flavor impletented in the batchwise backfitting algorithm.

Affiliation:

Stefan Lang
Department of Statistics
Universitätsstraße 15
6020 Innsbruck, Austria
E-mail: stefan.lang@uibk.ac.at
URL: https://www.uibk.ac.at/statistics/personal/lang/

Austrian Journal of Statistics
published by the Austrian Society of Statistics
Volume 54
2025

http://www.ajs.or.at/
http://www.osg.or.at/

Submitted: 2024-03-11
Accepted: 2024-09-06