Managing the bias-variance tradeoff in the context of house price prediction and hedonic indices - An application for German housing data

February 25, 2021

Abstract

We a) compute a Time Dummy Method index based on a Generalized Additive Model allowing for smooth effects of the metric covariates on the price utilizing the pooled data set. We b) construct an Imputation Approach model, where we fit a regression model separately for each year. Our work aims at constructing a global model that captures relevant interactions of the covariates with time, where time intervals are selected on a model basis. We therefore c) fit a model-based recursive partitioning tree to partition the time span and to account for parameter instability. We d) fit a global model, in which we interact the covariates with the time periods obtained from the recursive partitioning tree. We analyze the respective performance and choose the optimal model with respect to out-of-sample prediction accuracy.

We find that parameter instability over time plays a role as the Imputation Approach outperforms the Time Dummy Model. However by choosing model-based interactions with time, we are able to reduce both model complexity and out-of-sample prediction error. We find the interaction between location and time appears to be the most important.

Our work provides a model-based approach to account for parameter instability over time in the context of hedonic price models and index construction. We are able to reduce bias compared with standard Time Dummy Method indices, and receive less volatile results compared with the typical Imputation Approach. Further, our assessment no longer naively selects time periods that are interacted with the other explaining variables. We expect these improvements to be useful especially for smaller, e. g. regional data sets.

1 Introduction

Prices of residential property, according to de Haan and Diewert (2011), play a major role as both a macroeconomic indicator of economic activity and asset wealth as well as in monitoring risk exposure and hence financial stability. Thus, it is of great importance to assess prices of national real estate properties and their development over time.

The main challenge in the computation of house price indices lies in controlling for the dwellings' varying characteristics and locations. Following ILO et al. (2004), hedonic indices have become the gold standard for this purpose. Hedonic indices are characterized by expressing house prices as a function of characteristics within the framework of a regression model. Thus, the obtained indexes show price evolutions controlled for variation in the underlying characteristics. Potential problems include an omitted variable bias next to the usually unknown functional relationship between the house price and its regressors. Thus, in many applications, it has been shown to be advantageous to utilize more flexible non-parametric estimation techniques. A common way to incorporate these, is the construction of Generalized Additive Models and within its framework, the use of splines. Applications of such methodologies include Waltl (2016) and Brunauer, Lang, and Feilmayr (2013), who compute hedonic indices utilizing penalized splines within flexible Generalized Additive Models or Hill and Scholz (2018), who employ a spline surface to capture geospatial effects.

Within the context of hedonic regression, the Time Dummy Method, next to the Imputation Approach, are the most relevant approaches. They are both characterized as hedonic indexes and their differences arise usually from changes in average characteristics. When utilizing the Time Dummy Method, a model is usually fit to the pooled data set comprising all periods. In this way, it is straightforward to obtain the price index simply as the (exponential) coefficients of the time dummies. The Imputation Approach is more flexible as it does not rely on fitting the model to the pooled data. In practice, separate models are usually fit to each time period, which relaxes the constant parameter assumption over time, which is a restrictive assumption within the framework of Time Dummy indexes. This setting represents a typical bias-variance tradeoff: Generally, increased complexity of a model results in a decreased bias at the cost of inflated variance. Assuming parameter stability over time could be inappropriate, but modeling each time period separately possibly poses an extreme methodology as well. We propose the application of model-based recursive partitioning to stratify the data into time partitions in contrast to a naive stratification strategy within the context of Imputation Approach indices. On the basis of the obtained model-based partitions, we fit a global model incorporating interaction terms between the covariates and the time regions obtained from the recursive partitioning algorithm. This approach enables us to allow for possible parameter stability over time and to identify (ir)relevant interaction terms using model choice criteria. We are able to reduce bias, while minimizing variation in the predictions.

We contribute to the discussion of the bias variance tradeoff by analyzing a large sample

of semi-detached and single family houses in Germany from 2005 to 2019.
Our work is structured as follows. This introduction is followed by a brief presentation of the concept of hedonic price indices in section 2, where we describe the computation of Time Dummy and Imputation Approach indices within the context of a brief discussion of the laid-out concepts. In the following section 3, we briefly describe the methodologies of Generalized Additive Models and model-based recursive partitioning of which application our work heavily relies on. Our empirical analysis in section 4 involves, following a description of the involved data, a) the computation of a Time Dummy Method model based on a Generalized Additive Model allowing for smooth effects of the metric covariates on the price utilizing the pooled data set. We then b) construct an Imputation Approach model, where we fit a regression model separately for each time period and c) fit a model-based recursive partitioning tree to partition the time span into a set of regions. We d) fit a global model, in which we interact the covariates with the time regions obtained from the recursive partitioning tree. We analyze the respective performance and choose the optimal model with respect to out-of-sample prediction accuracy. Finally, e) we construct hedonic indices on the basis of the employed models and discuss them regarding their differences and implications. In section 5 we conclude with a discussion of our results.

2 Hedonic Price Indices

In accordance with de Haan and Diewert (2011), hedonic regression methods serve the purpose of constructing quality-adjusted price indices. At the basis of this lies the assumption that property prices depend on a set of characteristics, such as location and structure, which cannot be observed separately. Regression methods are employed in order to assess the marginal effects and ultimately, to construct indices.
Following Brunauer, Feilmayr, and Wagner (2012), the literature distinguishes two approaches: The Time Dummy Method and the Imputation Approach. Both techniques involve regressing the price of a property on its characteristics. The Time Dummy Approach usually utilizes a pooled regression comprising all time periods, while the Imputation Method often assesses the characteristics' marginal effects through a separate regression for each time period.

2.1 Time Dummy Indices

The Time Dummy Approach is, following Triplett (2004), the most frequently applied method to construct price indices. The convenience in this approach lies in its simplicity, as the index is derived directly from the regression coefficients, making its application and interpretation very straightforward. The subsequent taxonomy is notationally motivated as laid down in the Handbook on Residential Property Price Indices, commissioned by the European Union, the International Labor Association, the International Monetary Fund, the Organisation for Economic Co-operation and Development, the United Nations Economic Commission for Europe, and The World Bank, authored by de Haan and Diewert (2011). The standard Time Dummy Variable model is formulated in a semi-logarithmic form

$$ln\ p_{it} = \beta_0 + \sum_{t=1}^{T} \delta_t D_{it} + \sum_{k=1}^{K} \beta_k z_{ikt} + \epsilon_{it}, \tag{1}$$

where p_{it} is the price of property i in period t as a function of K characteristics captured by z_{ikt}. Thereby, β_0 and β_k give the intercept term and the characteristics' parameters estimated by the model, respectively. D_{it} is the time dummy variable taking the value 1, if an observation comes from period t and 0 otherwise, where a time dummy for the base period 0 is left out to prevent an identification problem. Finally, ϵ_{it} is the error-term and is considered to be white noise. The given model is estimated on the pooled data comprising all time periods. Hence, the time dummies provide a measure for the marginal effect of time on the logarithm of price. The price index P_{0t}^{TD} from period 0 to period 1 is usually derived by

$$P_{0t}^{TD} = exp(\hat{\delta}_t), \tag{2}$$

i. e. is simply given by the respective exponential of the estimated time dummy coefficients. However, since equation (2) represents a nonlinear transformation, the obtained price index is biased. Under the assumption of normally distributed errors, Kennedy (1981) proposes the unbiased estimator

$$P_{0t}^{TD*} = exp\left(\hat{\delta}_t + \frac{1}{2}V\hat{a}r(\hat{\delta}_t)\right). \tag{3}$$

Although some authors note that the actual bias is small, like de Haan (2010) or Yu and Prud'homme (2010), we include the bias correction, since it is not computationally costly.

2.2 Imputation Approach Indices

Imputation Approach Indices are, next to the Time Dummy Method, a prominent method to compute hedonic indices. As formulated by de Haan and Diewert (2011), the approach is easily motivated by regarding it from an index construction view: Prices of dwellings sold in period t can only be observed at time t, but are unknown in all other periods. In order to obtain standard price indices, these unobserved prices need to be *imputed.* Thus, price predictions for housings are obtained, whose characteristics are held fixed, while the time period is varied. In many applications, although not necessarily, this involves a hedonic regression model that is run separately for each period, see Hill and Melser (2008). In this, according to Hill (2011), lies criticism of the method, as the exploitation of interactions between the regression equations is prevented. Further, model complexity increases and other potential important interactions, regarding e. g. with location, are neglected.

Within the methodology of Imputation type indexes, single imputation and double imputation indices are distinguished. The single imputation index imputes solely missing observations, while the computation of the double imputation indices involves imputing both missing and observed prices. Hill (2011) argues that imputing both actual and unobserved prices decreases a potential omitted variable bias. Thus, henceforth, only double imputation indices are considered. To finally obtain a price index, classic price index formulae are applied. There exists a broad range of formulae in literature, which include e. g. Laspeyres, Paasche, Törnqvist, or Fisher type indexes. The most commonly applied are Laspeyres and Paasche indices, although these two approaches have some disadvantages compared to e. g. the Törnqvist index. However, since this work is not aimed at contributing to the discussion of index formulae, we restrict ourselves to the application of the Laspeyres index. For a detailed discussion of the mentioned index alternatives, see for example Balk (1995), Diewert (2007), Hill and Melser (2008), or de Haan (2010). Again, the taxonomy is (mostly) in analogy to de Haan and Diewert (2011). Verbally, for the computation of the double imputation Laspeyres index, the following steps are undertaken:

1. A model is fit separately to each time period, e.g. every quarter for a quarterly price index and every year for a yearly index, respectively.

2. To receive the Laspeyres type index, the base period characteristics are plugged into each model to obtain predictions for each period. Hence, base model characteristics

are evaluated at time t.

3. Finally, the sum of the predictions is obtained in each period and divided by the sum of predicted prices in the base period. Thus, the price evolution over time is tracked.

In terms of notation, the described methodology translates into the regression model

$$ln\ p_{it} = \beta_{0t} + \sum_{k=1}^{K} \beta_{kt} z_{ikt} + \epsilon_{it}, \tag{4}$$

which differs from equation (1) with respect to a) the missing time dummy term, and b) the subscript t for the estimated coefficients, i. e. shadow prices for characteristics z_{ikt}. The subscript is added, since there is a model for each time period.
Finally, predicted property prices for period 0 and t are $ln\ \hat{p}_{i0} = \hat{\beta}_{00} + \sum_{k=1}^{K} \hat{\beta}_{k0} z_{ik0}$ and $ln\ \hat{p}_{it} = \hat{\beta}_{0t} + \sum_{k=1}^{K} \hat{\beta}_{kt} z_{ikt}$, respectively. It is apparent that predicted prices and therefore indexes vary with differing values of z_k. To address this issue, the sample average characteristics of the base period, $\bar{z}_{k0}$, as well as the sample average characteristics of the comparison period, $\bar{z}_{kt}$, are utilized to compute the indices. the hedonic double imputation (DI) Laspeyres index is then defined as

$$P_{0t}^{HDIL} = \frac{\hat{\beta}_{0t} + \sum_{k=1}^{K} \hat{\beta}_{kt} \bar{z}_{k0}}{\hat{\beta}_{00} + \sum_{k=1}^{K} \hat{\beta}_{k0} \bar{z}_{k0}}, \tag{5}$$

where base period prices are imputed for properties corresponding to the period t sample, evaluated at base period 0 characteristics. Exponentiation to convert prices back onto a linear scale, the expression from equation (5) becomes

$$P_{0t*}^{HDIL} = \frac{exp(\hat{\beta}_{0t} + \sum_{k=1}^{K} \hat{\beta}_{kt} \bar{z}_{k0})}{exp(\hat{\beta}_{00} + \sum_{k=1}^{K} \hat{\beta}_{k0} \bar{z}_{k0})}. \tag{6}$$

However, as shown by Jensen et al. (1906), this estimate is biased as $\varphi(E[X]) \leq E[\varphi(X)]$. Analogous to the bias correction in equation 3, we thus need to correct the estimated prices for bias, when transforming them to a linear scale. Suppose the general case of computing a fitted value for an object with explanatory variables $\boldsymbol{x}_0$, the naive approach to convert a predictor $ln\ \boldsymbol{y}_0 = \boldsymbol{x}_0' b$ would be

$$\hat{y}_0 = exp(\boldsymbol{x}_0' \boldsymbol{b}).$$

Then, the bias corrected estimator is, as outlined by Greene (2018), given by

$$\hat{y}_0 = exp(\boldsymbol{x}_0' \boldsymbol{b} + s^2/2) > exp(\boldsymbol{x}_0' \boldsymbol{b}), \tag{7}$$

where s^2 is the sample variance. Finally, the double Imputation hedonic price index is obtained by combining equations (6) and (7), so that

$$P_{0t*}^{HDIL} = \frac{exp(\hat{\beta}_{0t} + \sum_{k=1}^{K} \hat{\beta}_{kt} \bar{z}_{k0} + s_t^2/2)}{exp(\hat{\beta}_{00} + \sum_{k=1}^{K} \hat{\beta}_{k0} \bar{z}_{k0} + s_0^2/2)}.$$

As pointed out by Hill (2013), the issue of biased transformation from log to linear scale often is completely disregarded or neglected as the resulting bias is usually very small. In this application, we compute the bias-corrected version, since it is not computationally costly. One great advantage of the Imputation Approach towards the Time Dummy Method is that it implicitly relaxes the assumption of constant characteristics' parameters. Thus, if the assumption is violated, this method still yields unbiased estimates. However, as stated before, this advantage is payed for with increased model complexity and it is questionable whether all covariates are instable over time, which is implicitly assumed.

Further, since the method employs regressions for all time periods separately, the characteristics coefficients, and hence the index numbers, do not vary when additional samples are added for future periods. Within the classic framework of Time Dummy Models, the adding of time periods automatically leads to an update of the whole index, which is problematic. However, this problem can be avoided by employing e. g. rolling Time Dummy indices.

3 Model Methodology

Following de Haan and Diewert (2011), the construction of any house price index is generally based on matching the prices for identical dwellings over time. However, this matching is problematic for several reasons. First, all housings are different in nature, i. e. their characteristics largely differ both in quality and location. Second, even if the same dwelling is sold in differing time periods, an exact comparison leads to biased indices as stated by Diewert (2009). Issues arise, because regarded buildings depreciate over time or when properties have been subject to substantial changes in form of additions, repairs or remodeling. Hedonic indices address these issues, as they are constructed on the basis of regression models that explain the observed prices as a function of the dwellings' characteristics. Hence, an appropriate index relies on a model that captures the relationship between the price and its regressors accurately.

In this section, I shortly present the employed model methodology. I begin by outlining the Generalized Additive Model (GAM) and the concept of nonparametric regression approach within its framework. The provided illustrations of the various methodologies primarily follow those by Fahrmeir et al. (2013). The remainder of the section briefly outlines the concept of model-based recursive partitioning.

3.1 Generalized Additive Models

Estimating the prices utilizing penalized spline regression within a Generalized Additive Model framework comes along with several advantages over more classic approaches. Linear regression seeks to capture the relationship between the target variable and the explanatory variables. It is often unclear, however, what the functional relationship between the dependent variable and a specific regressor is. While classic linear models allow a nonlinear functional relationship through a transformation of the covariates or inclusion of polynomials, the nature of the exact functional dependence often remains unclear, however. Over the years, nonparametric regression methods have become increasingly popular. The goal of nonparametric methods is to obtain a smooth function to capture the relationship between the dependent variable and its regressor.

Hastie and Tibshirani (1987) introduced the framework of Generalized Additive Models (GAM), which was later implemented in R by Wood (2001) and Wood (2007). Following Fahrmeir et al. (2013), the GAM can be regarded as an extension of the multiple linear regression model with

$$y_i = f_1(z_{i1}) + \cdots + f_q(z_{iq}) + \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i, \tag{8}$$

where the dependent variable y_i is regressed on an intercept β_0, and a set of regressors $x_{i1}, \ldots, x_{ik}$. This classic linear model is then extended by the $f_1(z_{i1}), \ldots, f_q(z_{iq})$ terms, which are nonlinear and smooth effects of the regressors. Thereby, it must hold that

$\sum_{i=1}^{n} f_1(z_{i1}) = \cdots = \sum_{i=1}^{n} f_q(z_{iq}) = 0$ to avoid an identification problem.

3.1.1 Basis splines

The idea of splines is to divide the range of regressors into several equidistant segments. The points dividing these segments are subsequently referred to as *knots*. For polynomial splines, a separate polynomial is fitted for each of the intervals. These in turn are restricted to be continuous and differentiable at the knots.

The application of polynomial splines in nonparametric regression requires a constructive representation of polynomial splines. Following Wood (2017), this means that the function f needs to be represented in a way so that it becomes a linear model. One possible way to achieve this representation is to choose basis functions. Among these, we restrict ourself to the use of basis splines, as they offer several advantages from a numerical viewpoint.

Again, the derivation of the motivation and method closely follows Fahrmeir et al. (2013). Basic references include Dierckx (1995) and de Boor (1978). The starting point is the construction of piecewise polynomials. Now, basis functions are constructed in such manner that the transitions are sufficiently smooth at the knots. A B-Spline then consists of $(l+1)$ polynomial fractures, where l is the degree of the respective spline. The fractures are put together in a way so that they are $(l-1)$-times continuously differentiable. Through a linear combination of $d = m + l - 1$ basis functions with m knots, a representation of $f(z)$ is obtained. Hence, one obtains

$$f(z) = \sum_{j=1}^{d} \gamma_j B_j(z), \tag{9}$$

where B_j are the basis functions and γ_j its corresponding coefficients. The derivation of basis function has been done by Wahba (1990) and Gu (2013), so that for B-splines of degree l ¿ 1, the basis functions are defined as

$$B_j^l(z) = \frac{z - \kappa_j}{\kappa_{j+1} - \kappa_j} B_j^{l-1}(z) + \frac{\kappa_{j+l+1} - z}{\kappa_{j+l+1} - \kappa_{j+1}} B_{j+1}^{0}(z), \tag{10}$$

where $\kappa_1, \ldots, \kappa_m$ are the inner m knots. As equation (10) has a recursive structure, an extended knot range of length $2l$, $\kappa_{1-l}, \kappa_{\kappa_{1-l+1}}, \ldots, \kappa_{m+l-1}, \kappa_{m+l}$, is required. A notation for splines of degree $l <= 1$ is omitted here, since we don't apply it. See e. g. Fahrmeir et al. (2013) for a more detailed overview.

Finally, in order to obtain and estimate a model that is linear in its parameters, equation (9) is substituted into equation (8)and written in matrix notation such that

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\gamma} + \boldsymbol{\epsilon},$$

where $\boldsymbol{y}$ is the vector of obser vations $(y_1, \ldots, y_n)'$, $\boldsymbol{\gamma}$ is a vector containing the basis functions' coefficients $(\gamma_1, \ldots, \gamma_d)'$ and $\boldsymbol{Z}$ is the $n \times d$ design matrix, which is defined for the basis splines as

$$\boldsymbol{Z} = \begin{pmatrix} B_1^l(z_1) & \cdots & B_d^l(z_1) \\ \vdots & & \vdots \\ B_1^l(z_n) & \cdots & B_d^l(z_n) \end{pmatrix}.$$

Then the least squares estimator is

$$\hat{\gamma} = \left(\boldsymbol{Z}^\top \boldsymbol{Z}\right)^{-1} \boldsymbol{Z}^\top \boldsymbol{y}.$$

The interpretation of the resulting coefficients is however not meaningful. Diagnostic plots from the predictions are more insightful.

3.1.2 Penalized splines

Given the implementation into statistical software, such as the `mgcv`-package in R, B-splines are relatively easy to construct and compute. In most applications, the main challenge lies in choosing the quantity of (equidistant) knots and find a good compromise between a good fit to the data and increasing model complexity and thus overfitting. A different approach is to use a fixed, and relatively large, number of equidistant knots (usually circa 20-40) and introduce a term into the least squares condition that penalizes complexity in the model. The first implementations of such penalties were introduced by Silverman (1985) or O'Sullivan (1986). The latter introduced the penalty term

$$\lambda \int (f''(z))^2,$$

where the smoothing parameter λ drives the penalty's influence. Hence, higher curvature in f(z) implies a higher penalty term and a smoother is favored over a wiggly fit.
Eilers and Marx (1996) translate the problem into a penalized least squares criterion

$$PLS(\lambda) = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{d} \gamma_j B_j(z_i) \right)^2 + \lambda \sum_{j=k+1}^{d} \left(\Delta^k \gamma_j \right)^2, \tag{11}$$

which puts a difference penalty on the coefficients rather than the integral over the second derivative of the fitted curve. Δ^k are the differences of k-th order and are defined recursively as

$$\begin{aligned} \Delta^1 \gamma_j &= \gamma_j - \gamma_{j-1} \\ \Delta^2 \gamma_j &= \Delta^1 \Delta^1 \gamma_j = \Delta^1 \gamma_j \Delta^1 \gamma_{j-1} = \gamma_j - 2\gamma_{j-1} + \gamma_{j-2} \\ &\vdots \\ \Delta^k \gamma_j &= \Delta^{k-1} \gamma_j - \Delta^{k-1} \gamma_{j-1}. \end{aligned}$$

In matrix notation, equation (11) can be written as

$$PLS(\lambda) = (\boldsymbol{y} - \boldsymbol{Z\gamma})'(\boldsymbol{y} - \boldsymbol{Z\gamma}) + \lambda\boldsymbol{\gamma}'\boldsymbol{K}_k\boldsymbol{\gamma},$$

where $\boldsymbol{K}_k$ is the penatly matrix for the k-th difference of Δ^k.

Finally, as written by Fahrmeir et al. (2013), the penalized least squares estimator is defined as

$$\hat{\gamma} = (\boldsymbol{Z}'\boldsymbol{Z} + \lambda\boldsymbol{K})^{-1}\boldsymbol{Z}'y.$$

The only term that differs from the B-spline least squares estimator is $\lambda\boldsymbol{K}$, which is in turn mainly driven by the smoothing parameter λ. If $\lambda = 0$, then the penalized estimator becomes the standard least squares estimator. As λ grows very large, the obtained fit becomes equivalent to a linear fit. Eilers and Marx (1996) propose the use of the Akaike information criterion (AIC) as introduced by Sakamoto, Ishiguro, and Kitagawa (1986) or the generalized cross-validation method (GCV). The latter is implemented in the context of penalized splines estimation in the `mgcv`-package in R by Wood (2007).

3.2 Model-Based Recursive Partitioning

Like generalized additive models, model-based recursive partitioning are considered techniques of supervised statistical learning. In this section, we briefly explain the utilized model-based recursive partitioning within the class of tree-based methods, and more specifically, regression trees. Our outline and notation follows the work by Hastie, Tibshirani, and Friedman (2009).

3.2.1 Regression Trees

Regression trees refer to tree-based models that are fit for a metric target variable. They pose a relatively simple, but mighty tool. In their basic form, they partition the characteristic space into rectangles and simply fit an average in each space. The main concept of regression trees is to identify split points within the covariates, at which the characteristic space is split into two regions. For each of the obtained regions, the average is computed. This procedure is repeated until there some minimum threshold of observations is reached in a node, or some other stopping criterion is met. The final graphical representation resembles a tree, which is where the name stems from.

In order to shortly illustrate the approach, a dependent variable Y is considered along with p explanatory variables for n observations. The algorithm is designed, so that it identifies splitting variables and split points. We then create a partition with M regions $R_1, R_2, \ldots, R_m$ and model the response as constant c_m in each region, so that

$$f(x) = \sum_{m=1}^{M} c_m I(x \in R_m).$$

If we then set the minimization of the sum of squares $\sum(y_i - f(x_i))^2$, we obtain the optimal $\hat{c}_m$ as

$$\hat{c}_m = ave(y_i \mid x_i \in R_m),$$

which is simply the average y_i in R_m. Since the computation of an optimal partition regarding the sum of squares numerically is usually infeasible, a greedy algorithm is utilized: First, define a splitting variable j for which the characteristic space is split at point s, so that

$$R_1(j,s) = \{X \mid X_j \leq s\} \text{ and } R_2(j,s) = \{X \mid X_j > s\}$$

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

$$\min_{j,s}\left[\min_{c_1} \sum_{x_i \in R_1(j,s)} (y_i - c_1)^2 + \min_{c_2} \sum_{x_i \in R_2(j,s)} (y_i - c_2)^2\right],$$

where

$$\hat{c}_1 = ave(y_i \mid x_i \in R_1(j,s)) \text{ and } \hat{c}_2 = ave(y_i \mid x_i \in R_2(j,s))$$

solve the inner minimization. In this way, the optimal pair (j,s) is obtained and the procedure is repeated typically until some minimum terminal node size is reached. In the subsequent, the tree may be pruned to avoid overfitting.

3.2.2 Model-Based Recursive Partitioning

The methodology of model-based recursive partitioning was introduced by Zeileis, Hothorn, and Hornik (2008), whose notation we adapt to shortly outline the method in the following. Model-based recursive partitioning represents an integration of parametric models into regression trees. Within this methodology, a tree is computed, in which every leaf is not associated with a simple average, but instead with a fitted model, e. g. a linear regression: Suppose a global parametric model $\mathcal{M}(Y,\theta)$ is given with observations Y and parameter vector θ. The model is then estimated by minimization of some objective function $\Psi(Y,\theta)$ resulting into

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}} \sum_{i=1}^{n} \Psi(Y_i, \theta), \tag{12}$$

where $\hat{\theta}$ is the parameter estimate given n observations $Y_i (i = 1, \ldots, n)$. For OLS, Ψ is simply the error sum of squares. 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 $\mathcal{M}_m(Y, \theta_m)$ corresponding to a cell-specific parameter θ_m yielding a globally segmented model $\mathcal{M}_M(Y, \{\theta_m\})$. $\{\theta_m\}_{m=1,\ldots,M}$ thereby corresponds to the full combined parameter. Equation (12) formulated over all regions can then be written as the optimization problem

$$\sum_{m=1}^{M} \sum_{i \in I_m} \Psi(Y_i, \theta_m) \rightarrow \min, \tag{13}$$

over all partitions $\{R_m\}$ with the indexes $I_m, m = 1, \ldots, M$. Equation (13) corresponds to a single model corresponding to each terminal node in a tree. To decide whether a possible split is necessary, a fluctuation test is utilized. The fitting of a model-based recursive partitioning model can then be summarized in the following algorithm:

1. In a possible node, fit the model with $\hat{\theta}$ to all corresponding observations by minimizing the objective function Ψ, in our case, least squares.

2. Utilizing a fluctuation test, evaluate whether the parameter estimates are stable with respect to every ordering in the partitioning variables j. If there is significant parameter instability, choose the variable j which corresponds to the highest degree of instability. If there is no significant instability in the parameters, stop.

3. Calculate the split point s that locally minimizes Ψ.

4. Split the current node into a set of daughter nodes and repeat the previous steps.

For a more detailed description of the steps, see Zeileis, Hothorn, and Hornik (2008). The algorithm as outlaid above relies on pre-pruning based on significant parameter instability in each node. To increase power and prediction accuracy, the authors propose some form of post-pruning. We employ the `party`-package by Hothorn, Hornik, and Zeileis (2006), which includes the `lmtree()` function. The function includes a `prune` option that we use for post-pruning using the BIC model selection criterion.

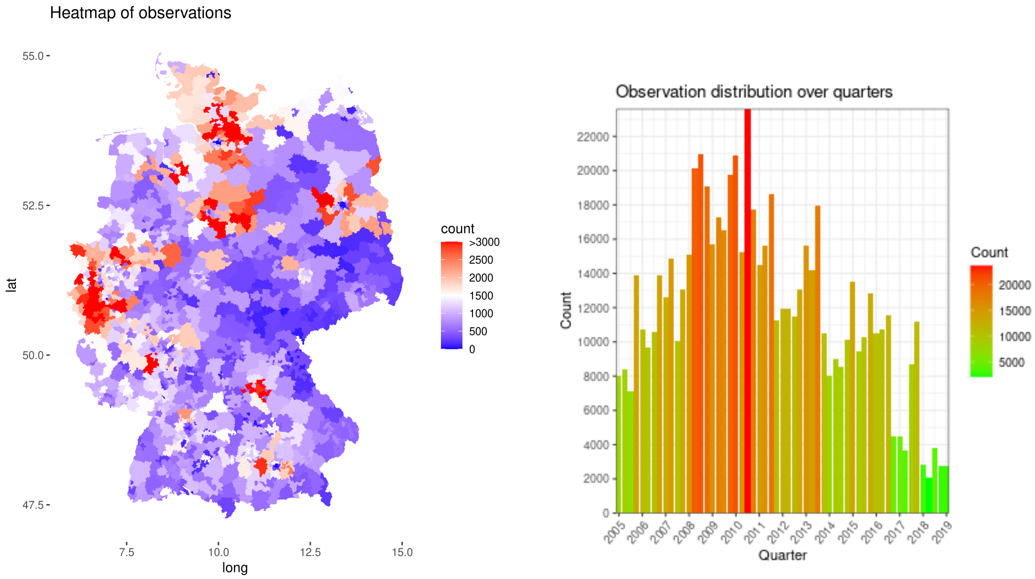
4 Empirical Analysis

In this section, we shortly introduce the utilized dataset and the models which we fit. Subsequently, we present the results as well as the obtained hedonic price indices.

4.1 Data

The data we utilize for our analysis is provided by the German 'F+B Forschung und Beratung für Wohnen, Immobilien und Umwelt GmbH' and comprises 682,435 observations of offer prices for private single family as well as semi-detached houses in Germany. The data set covers a time horizon from the first quarter in 2005 (2005 Q1) to the first quarter in 2019 (2019 Q1). Offer prices generally come with some advantages as well as disadvantages. The major advantage of offer prices lies in a higher number of observations. This, of course, comes along with smaller standard errors in the predicted prices and greater variation in the explaining variables, which in turn yields less variant house price indices. The major disadvantage is an upward bias of the offer prices. As the last offer price is usually greater than, but never beneath the actual selling price, offer prices are on average higher than selling prices. In our work, we disregard this upward bias of the prices, as a bias correction can easily be done even after the computation of indices. In our application it is hence irrelevant. In Table 2, a list of the variables included in our analysis is provided. In order to assess the locational distribution of the investigated houses, a heatmap of the observations in three-digit postcode areas is presented in Figure 1a. Red areas indicate higher counts of houses, while blue areas refer to little observations in the regarded area. More data is accumulated in the central north, far west, and in the Berlin area. Especially in the rural east of Germany and in rural Bavarian areas, the density of observations is lower.

Figure 1b gives a histogram of the observation density over time. Most observations are accumulated between 2008 and 2013. Especially in 2018 and 2019 there are less observations, but given the high absolute count, the data is still sufficient to construct local models.



(a) Heatmap of observations in three-digit postcode areas.

(b) Histogram of observation density over time.

Some summary statistics are provided in Table 3. We removed extreme outliers to avoid distorted results. The utilized data does not contain unavailable data points. This is vital to ensure comparability between the employed models.

4.2 Models

Our analysis comprises the computation and comparison of the following models. The terms in brackets refer to the corresponding abbreviations.

(TD) A model comprising data pooled over all time periods corresponding to a typical Time Dummy Method approach. The model equation results from equation (1), supplemented with a term to account for the semiparametric part of the model:

$$ln\ p_i = \sum_{q=1}^{Q} f_q(z_{iq}) + \beta_0 + \sum_{k=1}^{K} \beta_k x_{ik} + \sum_{\tau=1}^{T} \delta_\tau D_{i\tau} + \epsilon_i. \tag{14}$$

Thus, the log-price is regressed on a set of characteristics. The continuous variables *area*, *age*, and *plot area* enter the formula in the first term on the right-hand side of the regression. The respective effects are modeled in a smooth, nonlinear way utilizing penalized regression splines as introduced in Section 3.1. β_0 gives the intercept, x_{ik} is the matrix of dwelling i's characteristic k, while β_k is the shadow price of the corresponding characteristic. A complete list of the covariates included in the regarded models is provided in Table 2. $D_{i\tau}$ is the time dummy variable, whose value is 1, if dwelling i comes from period τ (and 0 otherwise). δ_τ is the vector of the respective shadow prices. The error term is finally given by ϵ_i. In analogy with Brunauer, Feilmayr, and Wagner (2012), spatial heterogeneity is captured utilizing random effects over 3-digit postcode dummies. Hence, among other advantages, it is possible to obtain predictions even for observations, whose postcode is not included in the training data.

(yImp) Secondly, a model is built stratified into years is built. This setting represents a typical application of yearly Imputation Approach index.

$$ln\ p_{it} = \sum_{q=1}^{Q} f_{qt}(z_{iqt}) + \beta_{0t} + \sum_{k=1}^{K} \beta_{kt} z_{ikt} + \epsilon_{it}, \tag{15}$$

which is fit separately for each year. Hence, inclusion of a Time Dummy is obsolete. The model formula thus corresponds to the equation (4) plus a term added to again account for the smooth modeling of the metric covariates.

(qImp) The data is stratified and modeled separately for each quarter, instead of each year as in (yImp). The model equation is analogous to equation (15).

(S1) For model (S1), in the first step, we fit a model-based recursive partitioning tree. We partition a linear model, where the logged price per square meter is modeled through all variables given in Table 2. The metric variables *area*, *plot area*, and *age* enter the model as a cubic polynomial. We choose *quarter* as the partitioning variable. In the second step, we next fit a global model again and interact each covariate with the time periods obtained from the partitioning in the first step. This makes it possible to evaluate the necessity of included interaction terms and to utilize standard model selection criteria.

(S2) Analogous to (S1), but we leave out the interaction term between the time partitions and *area*.

(S3) Analogous to (S1), but we leave out the interaction term between the time partitions and *plotarea*.

(S4) Analogous to (S1), but we leave out the interaction term between the time partitions and *age*.

(S5) To assess the importance of time-location interaction, we again fit a global model analogously to (S1), but leave out the interaction between the postcode dummies and the time regions.

(S6) Finally, a global model is fit accordingly to (S1), but interaction terms with the included metric variables *area*, *plot area*, and *age* are left out.

4.3 Results

In order to evaluate the tradeoff between variance and bias, we first randomly assign the data into a training data set (comprising 682,435 observations) and a validation data set (comprising 75,870 observations). The training data set is employed to compute the models, the training data set is used to compute predictions, which in turn are compared with the observed prices. Predictions and actual prices are compared on a linear scale rather than on a log scale to assure better comparability. As applied in the prior sections, we correct the predictions for bias when back-transforming them into linear scale.

In our application, we use the root mean squared error (RMSE) for evaluating the models. Following Greene (2018), the RMSE can be written as

$$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}, \tag{16}$$

where y_i are the observed prices and $\hat{y}_i$ are the corresponding predicted prices. In our case, since we aim at comparing the price p_i with its estimated counterpart $\hat{p}_i$, and under consideration of correction for bias, equation (16) becomes

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{N}\left(exp\left(\hat{p}_i + s^2/2\right) - exp(p_i)\right)^2}.$$

Table 1 reports the out-of-sample prediction error of the evaluated models. The worst model with regard to the out-of-sample prediction accuracy is the Time Dummy (TD) model. Model (TD) is outperformed by both yearly (yImp) and quarterly (qImp) Imputation Approach models with respect to RMSE. Thereby, (yImp) clearly outperforms (qImp). This indicates a potential overfit of the (qImp) model with regard to model complexity. The out-of-sample prediction comparison indicates that the assumption of parameter stability in the context of the Time Dummy approach is too restrictive. However, the yearly Imputation Approach model (yImp) is not the globally best model.

The model-based recursive partitioning yields separate time partitions over the regarded data, which we then interact with the other covariates in (S1), (S2), (S3), (S4), (S5), and (S6). These are

1. 2005 Q1 - 2008 Q2,
2. 2008 Q3 - 2010 Q1,
3. 2010 Q2 - 2011 Q3,
4. 2011 Q4 - 2013 Q3,
5. 2013 Q4 - 2015 Q2,

6. 2015 Q3 - 2016 Q4, and
7. 2017 Q1 - 2019 Q1.

Thus, we obtain only seven time periods, for which models are fit in contrast to 15 separate models for (yImp), and 57 separate models for (qImp), respectively. Further, our seven time periods are chosen on a model basis, as opposed to classic Imputation approach, where periods are naively chosen. Models (S1), (S2), (S3) next to (S4) have lower values in RMSE. This implicates that tight stratification also results in decreasing out-of-sample prediction precision and thus poses an overfitting to the data. Interactions play a role, but not all interactions are equally important. Through our approach, we are able to reduce bias and variance while minimizing model complexity. One further advantage of our model design is that it allows to switch interactions with time on or off at will and we are thus able to draw conclusions about which interactions play a role and hence, which covariate are (in)stable over time. (S2) is the globally best performing model implicating that an interaction between *area* and the time regions does not play a role. (S2), (S3), and (S4) further all outperform the Imputation indices, which indicates that the metric covariates are not substantially interacted with time. Even if we exclude all metric covariate interaction terms with the partitions as for (S6), our predictive accuracy is not substantially higher.
The most important interaction, with regard to variation in the obtained RMSEs, is that with location. Model (S5) has a substantially higher RMSE than both (yImp) and (S2). The RMSE is rather much closer to that of the Time Dummy (TD) fit, which represents the case of no included interactions at all.
We further carried out a stepwise BIC algorithm to choose appropriate interactions between the discrete covariates with the time regions. However, this procedure does not yield any further improvements regarding RMSE. Thus, we omit the corresponding results.

	(TD)	(yImp)	(qImp)	(S1)	(S2)	(S3)	(S4)	(S5)	(S6)
RMSE	480.03	459.12	470.23	454.44	454.40	458.88	458.84	472.44	459.52

Table 1: Out-of-sample prediction accuracy of evaluated models in terms of root mean squared error (RMSE).

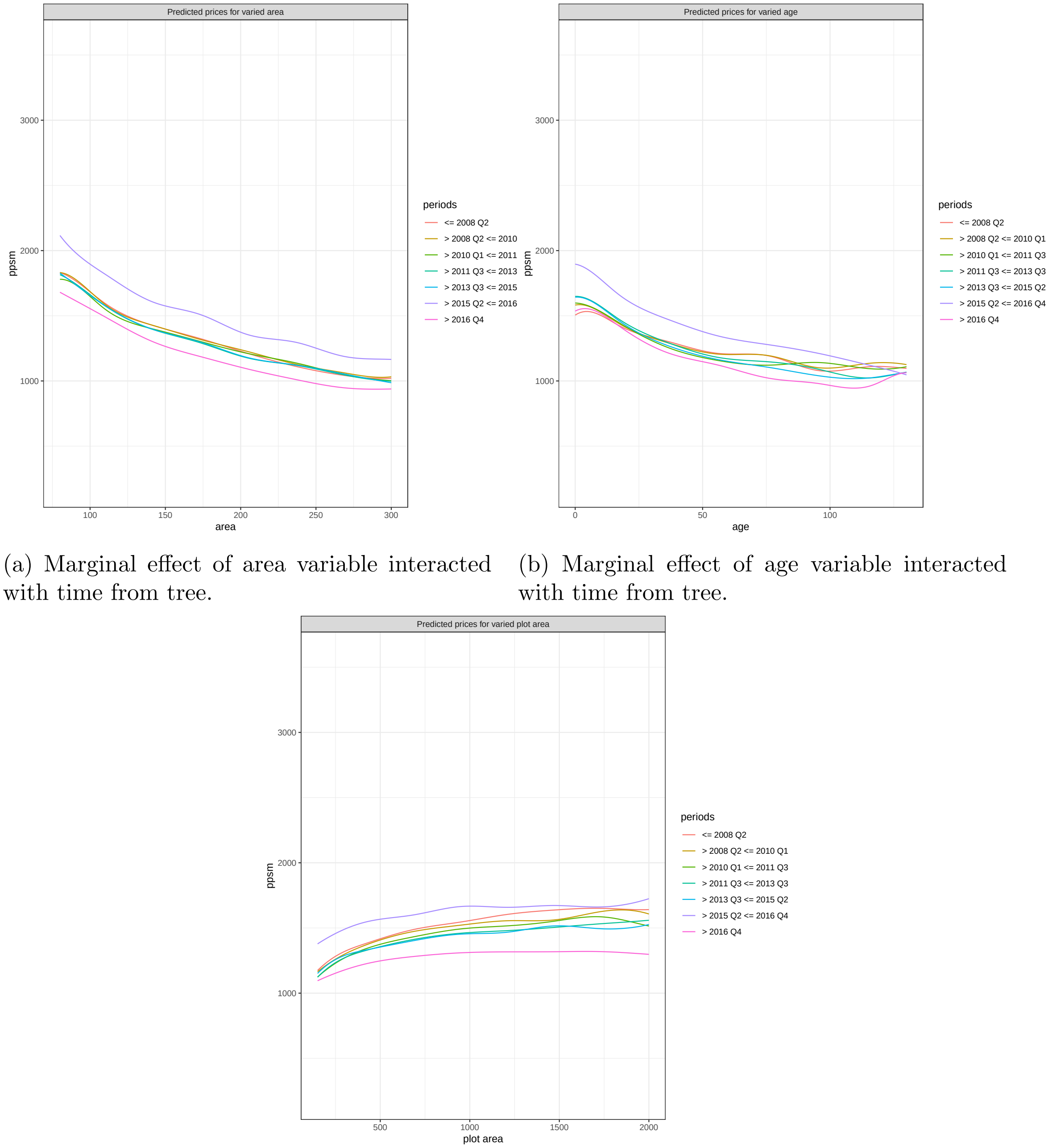
To discuss the relevance of each covariate's interaction with time in depth, we provide marginal effect plots for model (S1), where we include all interaction terms, in Figures 2 and 3. In the corresponding graphs, we predict prices per square meter by varying the regarded variable over the given range while holding the other variables fixed at their mean. Again, we compute the smearing estimator as shown in equation 7 to obtain unbiased results. Through this approach, we are able to provide prices on a linear scale, which facilitates interpretation. Since we interact each of the covariates with the seven time regions obtained by the model-based recursive partitioning algorithm, we obtain one graph for each partition. Thus, we get seven curves referring to the corresponding time interval.

The marginal effect of area is depicted in Figure 2a. Generally, the average price square meter monotonously descends with rising area of the underlying dwelling. The slope of descent appears to be greater (in absolute value) for smaller values in area than for greater ones. This can be interpreted as a form of bulk discount, where the price of an additional price per square meter declines for larger objects. All curves are aligned close to parallel to each other and differ only in their level. However, a shift in level is irrelevant regarding a possible interaction between the depicted variables. The Figure supports the conclusion that interaction with area is irrelevant. Thus, it seems inappropriate that parameter instability is implicitly assumed within (yImp) and (qImp).

Figure 2b provides the marginal effect curves of the age variable. Again, the average price per square meter declines with increasing values in age, i. e. older buildings are associated with lower average prices per square meter. Analogous to the previous graph, the lines mainly differ in their level rather than their functional form.

Finally, the marginal impact of plot area on the price per square meter for each partition is given in Figure 2c. The average price of housing generally rises with increasing area and there appears to be a saturation effect for greater values of plot area. The functional relationship appears to be quite similar over all time regions. However, the curves appear slightly flatter for later than for earlier years.

The graphs emphasize, why out-out-sample prediction accuracy is only improved slightly, if at all, when introducing the corresponding interaction terms. For age and plot area, there could be a relevant interaction with time, but the interaction does not appear to be large in magnitude.



(a) Marginal effect of area variable interacted with time from tree.

(b) Marginal effect of age variable interacted with time from tree.

(c) Marginal effect of plot area variable interacted with time from tree.

Figure 2: Marginal effects for metric covariates area, age, and plot area interacted with the time periods received from model-based recursive partitioning. The y-axis reports the price per square meter in Euros.

Figure 3 provides further insights into the relevance of an interaction between time and location in the data. It gives the predicted price per square meter for varied 3-digit postcode dummies over the obtained partitions. The first graph thereby provides the absolute price level in the corresponding regions, while all subsequent graphs merely give the deviation in predicted price per square meter in the postcode areas. This simplifies the identification of possible interactions. Further, the chosen form of display gives insights into which regions in Germany are subject to steeper price appraisals compared to other parts of the country. Prices are again reported on a linear scale including bias correction. Generally, regions in former West Germany, especially metropolitan areas like Munich, Stuttgart, Berlin, or Hamburg, are associated with higher price levels compared to former East Germany in general, especially rural regions. Some regions in and around Munich have average prices per square meter much higher than 3000 Euros, while some rural regions in former East Germany feature prices of below 1000 Euros per square meter.

The two subsequent graphs do not indicate a strong location-time interaction. Prices do not structurally rise or fall between 2008 Q2 and 2011 Q3 on average. For the fourth graph and the following, however, the relevance of location-time interaction becomes visible. Between 2011 Q4 and 2015 Q2, price changes in Germany are distributed quite heterogeneously. Areas aroung large cities like (especially) Munich, Hamburg, Nuremberg, but also Dresden, face steep price increases of close to and beyond 500 Euros per square meter, while other urban regions are even associated with price drops. For the period 2017 Q1 and following, there is an overall upward shift in the price level. Blue zones almost completely disappear and the map is dominated by (deep) red areas. However, price increases in urban areas are on average higher than for rural regions.

Overall, Figure 3 emphasizes the importance of the time-location interaction in the data. An interaction is visible underlining the results of the out-of-sample prediction accuracy comparison.

For completeness, we provide Figures 6, 7, 8, and 9, where the evolution of the coefficients over time is depicted. We track the respective development by plotting the main effect's coefficient in the base perios, plus the respective interaction term coefficient in the subsequent periods together with the corresponding confidence intervals.

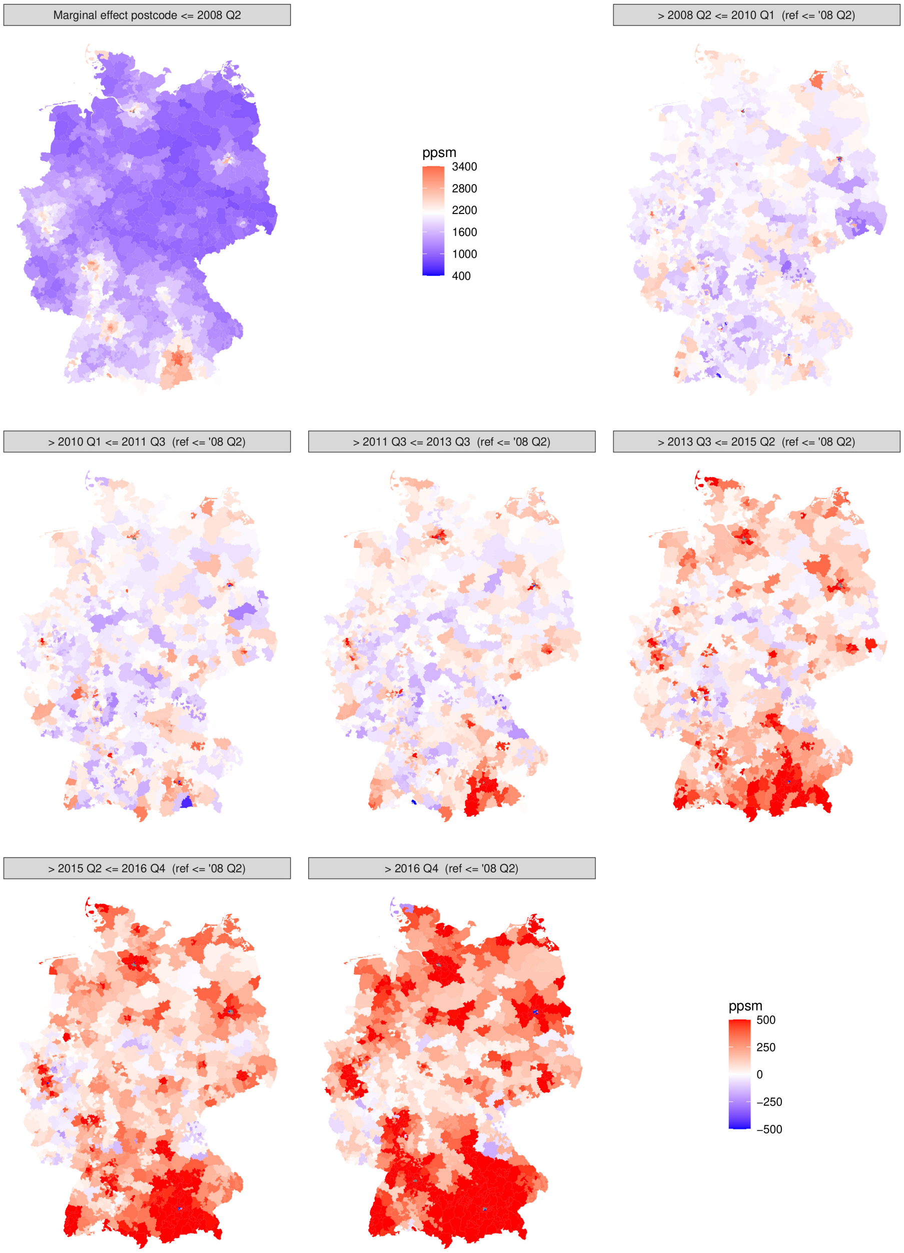

Marginal effect postcode <= 2008 Q2
ppsm
3400
2800
2200
1600
1000
400
> 2008 Q2 <= 2010 Q1 (ref <= '08 Q2)
> 2010 Q1 <= 2011 Q3 (ref <= '08 Q2)
> 2011 Q3 <= 2013 Q3 (ref <= '08 Q2)
> 2013 Q3 <= 2015 Q2 (ref <= '08 Q2)
> 2015 Q2 <= 2016 Q4 (ref <= '08 Q2)
> 2016 Q4 (ref <= '08 Q2)
ppsm
500
250
0
−250
−500

Figure 3: Marginal Effect of postcode dummies over time periods. The first graph refers to the absolute level of price per square meter in Euros. Subsequent graphs show the deviation from the first figure. Blue colored polygons refer to zones subject to price drops, red zones refer to increases in the predicted price per square meter.

4.4 Resulting indices

Figure 4 shows the obtained hedonic indices from the utilized models (TD), (qImp), (yImp) and the global best model (S2). The latter is computed by an adapted Imputation Approach: We take the base periods observations and vary their values in the *quarter* variable over the regarded time span. Thus, we are able to construct a Laspeyres type index. The curves correspond to the price evolution on a linear scale including bias correction.

Until the second quarter in 2009, the underlying houses are not subject to price increases. The Time Dummy index (TD) reaches a local minimum of close to 1 at that time. From circa 2010 until circa mid 2013, the hedonic curves are subject to steep prices raises. From 2010 to 2019, all hedonic indices begin to gradually rise. This overall trend does not appear to end within the investigated time horizon. The general form and level of all hedonic indices does not vary from each other until roughly mid 2011.

Although the indexes' RMSEs are not substantially different in value, we find that these small variations translate into relatively large differences in the corresponding price indices: Over the complete time span, the quarterly Imputation Approach (qImp) index indicates a price increase of just under 50%, while the Imputation index derived from (S1) returns a price increase of close to 60% over the 15 regarded years. All hedonic indexes are relatively close to each other until roughly mid 2011. In the subsequent periods, (TD) model's index runs underneath the other investigated indices. The functional form is similar in shape, but the general level is shifted downwards. Taking into consideration the higher RMSE of the model compared with e. g. (S1), this implicates that the index resulting from (TD) is downward biased.

Regarding variation in the investigated hedonic indices, the index obtained from (qImp) is the most volatile. This finding implicates that regarding the bias variance tradeoff, the model is too complex, which yields less biased, but highly volatile estimates. The inflated RMSE of the corresponding model supports this finding. The index of the yearly Imputation Approach model (yImp) is less volatile and proceeds parallel to the (S2) index. However, it provides only a yearly index and hence no information about the underlying quarters.

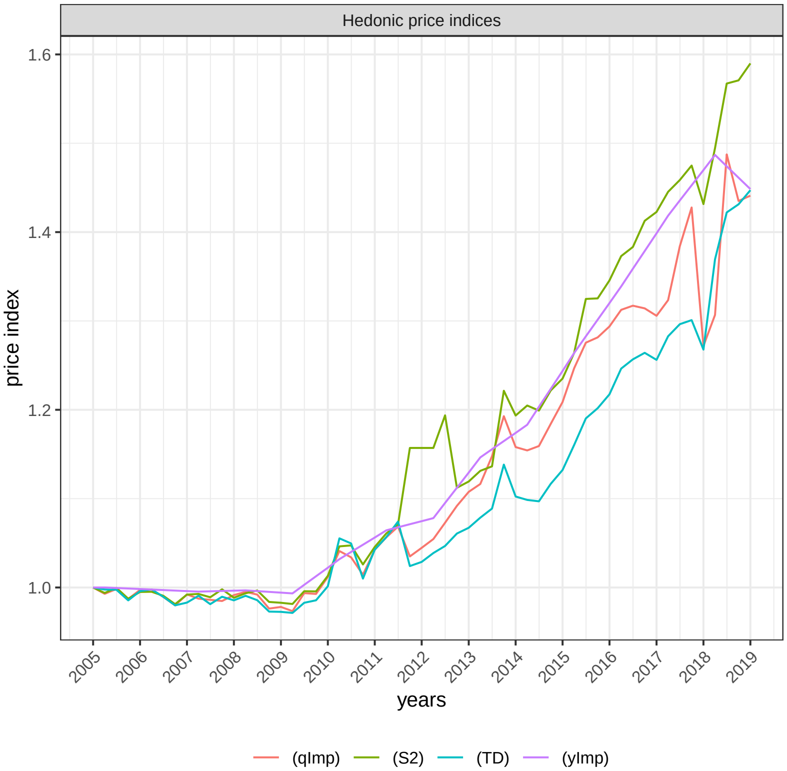


Figure 4: Resulting indices from utilized models. The blue line shows the evolution of the mean house price. The index produced by the Time Dummy index from (TD) is given by the purple line. The light green line depicts the quarterly imputation type index referring to (qImp). The dark green line gives the yearly Imputation index from (yImp). Finally, the red graph indicates the evolution of the imputation index, resulting from (S1).

5 Discussion

Hedonic Price Indices play a major role in assessing quality-adjusted price changes of housing over time. Within its class, the Imputation Approach next to the Time Dummy Method play a prominent role. A great problem is to capture interactions of the covariates with time. We construct hedonic indices for a range of various model and compare them regarding the underlying assumptions, predictive accuracy, and resulting indices.

Based on our analysis, the following main findings emerge. First, pooling the data over both space and time appears too restrictive and the implicit constant parameter assumption seems to be violated, which is indicated by the lower out-of-sample prediction accuracy with regard to RMSE. Imputation Approach indices outperform those based on the Time Dummy Method. We find the hedonic house price index resulting from the pooled model to be downward biased.

However, typical Imputation approach indices, that naively stratify data into periods, pose extreme methodologies, too. We show that stratification into too many periods leads to inflated RMSEs, even utilizing a large data set, like in our case. The resulting indices are often very volatile (qImp). The quality of the respective index further highly depends on the underlying data. Regarding more regional data, stratification into even years could lead to inflated variation in the estimated prices, which in turn translates into volatile price indexes. Naive stratification further rules out the possibility to exclude possibly irrelevant interaction terms with time. Evaluating the relevance of the regarded interactions is not possible either.

We apply model-based recursive partitioning to identify relevant interactions of the covariates with time and fit a global model, enabling us to employ standard model selection criteria to select relevant interaction terms. This approach further enables us to make statements about the variables for which parameter stability plays a role. We find that the most important interaction is that between time and location. Excluding the corresponding interaction from our model leads to a relatively big inflation in RMSE, i. e. a decrease in prediction accuracy. The exclusion of other interaction terms only leads to small losses in predictive accuracy. For the exclusion of the interaction term with *area*, we even find an improvement in RMSE. Since we stratify the data on a model basis, our approach is more flexible and we expect it to be more suitable compared with the classic approaches for other data sets. For smaller samples, e. g. regional data, the algorithm would likely select less time periods and we expect the advantage of model-based recursive partitioning to be even larger with respect to out-of-sample prediction accuracy.

The investigation of such regional data remains subject of future work. The same holds for partitioning the model over variables other than time in order to investigate relevant interactions. Finally, recursive partitioning that directly incorporates smooth effects of the covariates would be of great use.

References

Balk, Bert M (1995). “Axiomatic price index theory: A survey”. In: *International Statistical Review* 63.1, pp. 69–93.

Brunauer, Wolfgang A, Stefan Lang, and Wolfgang Feilmayr (2013). “Hybrid multilevel STAR models for hedonic house prices”. In: *Jahrbuch für Regionalwissenschaft* 33.2, pp. 151–172.

Brunauer, Wolfgang, Wolfgang Feilmayr, and Karin Wagner (2012). “A new residential property price index for Austria”. In: *Statistiken–Daten und Analysen Q* 3, pp. 90–102.

Dierckx, Paul (1995). *Curve and surface fitting with splines*. Oxford University Press.

Diewert, W Erwin (2007). *Index numbers*. Department of Economics, University of British Columbia.

— (2009). “The Paris OECD-IMF workshop on real estate price indexes: conclusions and future directions”. In: *Price and Productivity Measurement* 1, pp. 87–116.

Eilers, Paul HC and Brian D Marx (1996). “Flexible smoothing with B-splines and penalties”. In: *Statistical science*, pp. 89–102.

Fahrmeir, Ludwig, Thomas Kneib, Stefan Lang, and Brian Marx (2013). “Regression Models”. In: *Regression. Springer, Berlin, Heidelberg.*

Greene, William H (2018). *Econometric analysis 8th edition*. Pearson Education.

Gu, Chong (2013). *Smoothing spline ANOVA models*. Vol. 297. Springer Science & Business Media.

Hastie, Trevor and Robert Tibshirani (1987). “Generalized additive models: some applications”. In: *Journal of the American Statistical Association* 82.398, pp. 371–386.

Hastie, Trevor, Robert Tibshirani, and Jerome Friedman (2009). “The elements of statistical learning - Data Mining, Inference, and Prediction, Second Edition”. In: *Springer-Verlag New York.*

Hill, Robert (2011). “Hedonic price indexes for housing”. In: *OECDStatistics Working Papers.*

— (2013). “Hedonic price indexes for residential housing: A survey, evaluation and taxonomy”. In: *Journal of economic surveys* 27.5, pp. 879–914.

Hill, Robert and Daniel Melser (2008). “Hedonic imputation and the price index problem: an application to housing”. In: *Economic Inquiry* 46.4, pp. 593–609.

Hill, Robert and Michael Scholz (2018). “Can geospatial data improve house price indexes? A hedonic imputation approach with splines”. In: *Review of Income and Wealth* 64.4, pp. 737–756.

Hothorn, Torsten, Kurt Hornik, and Achim Zeileis (2006). “Unbiased Recursive Partitioning: A Conditional Inference Framework”. In: *Journal of Computational and Graphical Statistics* 15.3, pp. 651–674.

ILO et al. (2004). “Consumer price index manual: Theory and practice”. In: *ILO Publications, Geneva.*

Jensen, Johan Ludwig William Valdemar et al. (1906). “Sur les fonctions convexes et les inégalités entre les valeurs moyennes”. In: *Acta mathematica* 30, pp. 175–193.

Kennedy, Peter E (1981). “Estimation with correctly interpreted dummy variables in semilogarithmic equations”. In: *American Economic Review* 71.4, p. 801.

O’Sullivan, Finbarr (1986). “A statistical perspective on ill-posed inverse problems”. In: *Statistical science*, pp. 502–518.

Sakamoto, Yosiyuki, Makio Ishiguro, and Genshiro Kitagawa (1986). “Akaike information criterion statistics”. In: *Dordrecht, The Netherlands: D. Reidel* 81.

Silverman, Bernhard W (1985). “Some aspects of the spline smoothing approach to non-parametric regression curve fitting”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 47.1, pp. 1–21.

Triplett, Jack (2004). “Handbook on hedonic indexes and quality adjustments in price indexes”. In:

Wahba, Grace (1990). *Spline models for observational data.* SIAM.

Waltl, Sofie R (2016). “A hedonic house price index in continuous time”. In: *International Journal of Housing Markets and Analysis.*

Wood, Simon N (2001). “mgcv: GAMs and generalized ridge regression for R”. In: *R news* 1.2, pp. 20–25.

— (2017). *Generalized additive models: an introduction with R.* CRC press.

Wood, Simon (2007). “The mgcv package”. In: *www. r-project. org.*

Yu, Kam and Marc Prud’homme (2010). “Econometric issues in hedonic price indices: the case of internet service providers”. In: *Applied Economics* 42.15, pp. 1973–1994.

Zeileis, Achim, Torsten Hothorn, and Kurt Hornik (2008). “Model-based recursive partitioning”. In: *Journal of Computational and Graphical Statistics* 17.2, pp. 492–514.

de Boor, Carl (1978). *A practical guide to splines.* Vol. 27. Springer-Verlag New York.

de Haan, Jan (2010). “Hedonic Price Indexes: A Comparison of Imputation, Time Dummy and ’Re-Pricing’ Methods.” In: *Jahrbucher fur Nationalökonomie & Statistik* 230.6.

de Haan, Jan and Erwin W Diewert (2011). “Handbook on residential property price indexes”. In: *Luxembourg: Eurostat.*

A Description of covariates

Variable name	**description**
plot area	plot area of the object in m^2
age	age of object when sold
plot area	plot area of object in m^2
PLZ_3st	first three digits of postcode
alarm	object has alarm system (*yes* / *no*)
bright	object is bright (*yes* / *no*)
fire place	object has fireplace (*yes* / *no*)
basement	object has basement (*yes* / *no*)
need of renovation	object is in need of renovation (*yes* / *no*)
wellness	object has a swimming pool, sauna, or whirlpool (*yes* / *no*)
gas heating	object has gas heating (*yes* / *no*)
electric heating	object has electric heating (*yes* / *no*)
oil heating	object has central oil heating (*yes* / *no*)
night storage	has night storage heating system (*yes* / *no*)
parquet	object has parquet flooring (*yes* / *no*)
calm	object is located in calm area (*yes* / *no*)
garage	object has a garage (*yes* / *no*)
gallery	object has a gallery (*yes* / *no*)
floor heating	object has floor heating (*yes* / *no*)
balcony	object has balcony (*yes* / *no*)
elevator	object has elevator (*yes* / *no*)
yoc1900	object was built before 1900 (*yes* / *no*)
villa	object is a villa (*yes* / *no*)
facilities	degree of quality of object's facilities (*simple* / *normal* / *higher*)
quality	degree of object's quality (*less* / *normal* / *higher* / *luxurious*)
type	whether object is *single family home* or *semi-detached home* (versus single family home)
quarter	quarter of last offer for object
year	quarter of last offer for object

Table 2: List of variables included the analysis.

Statistic	Mean	St. Dev.	Min	Pctl(25)	Pctl(75)	Max
log price	7.351	0.508	4.605	7.114	7.663	8.854
plot area	564.835	339.358	150	306	719	2,000
area	143.341	39.364	80	118	160	300
age	23.869	30.814	−2	0	41	135

Table 3: Summary statistics of continuous variables.

B Interaction of time with discrete variables

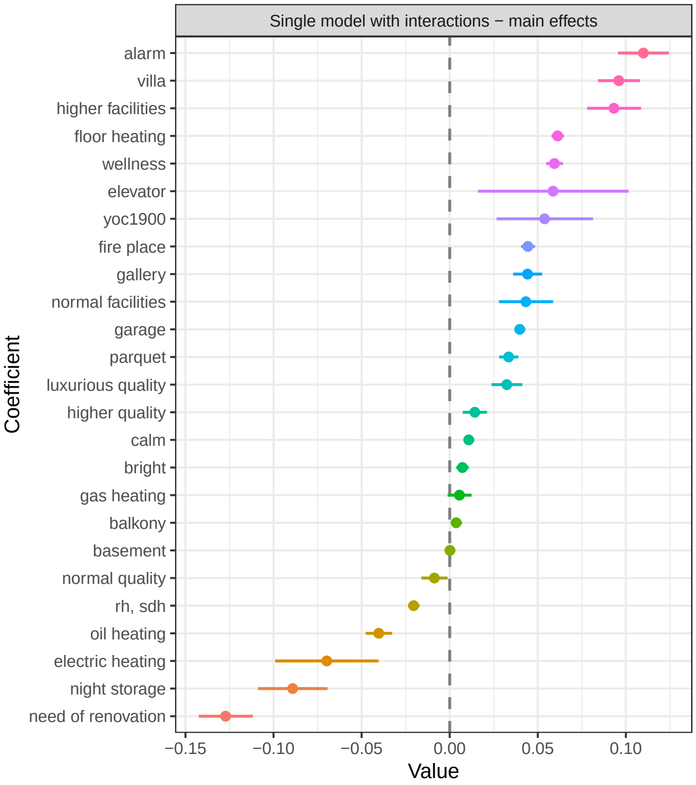


Figure 5: Main effects coefficient values and confidence intervals for (S1). The dots refer to the point estimate, while the bars give corresponding 95% confidence intervals.

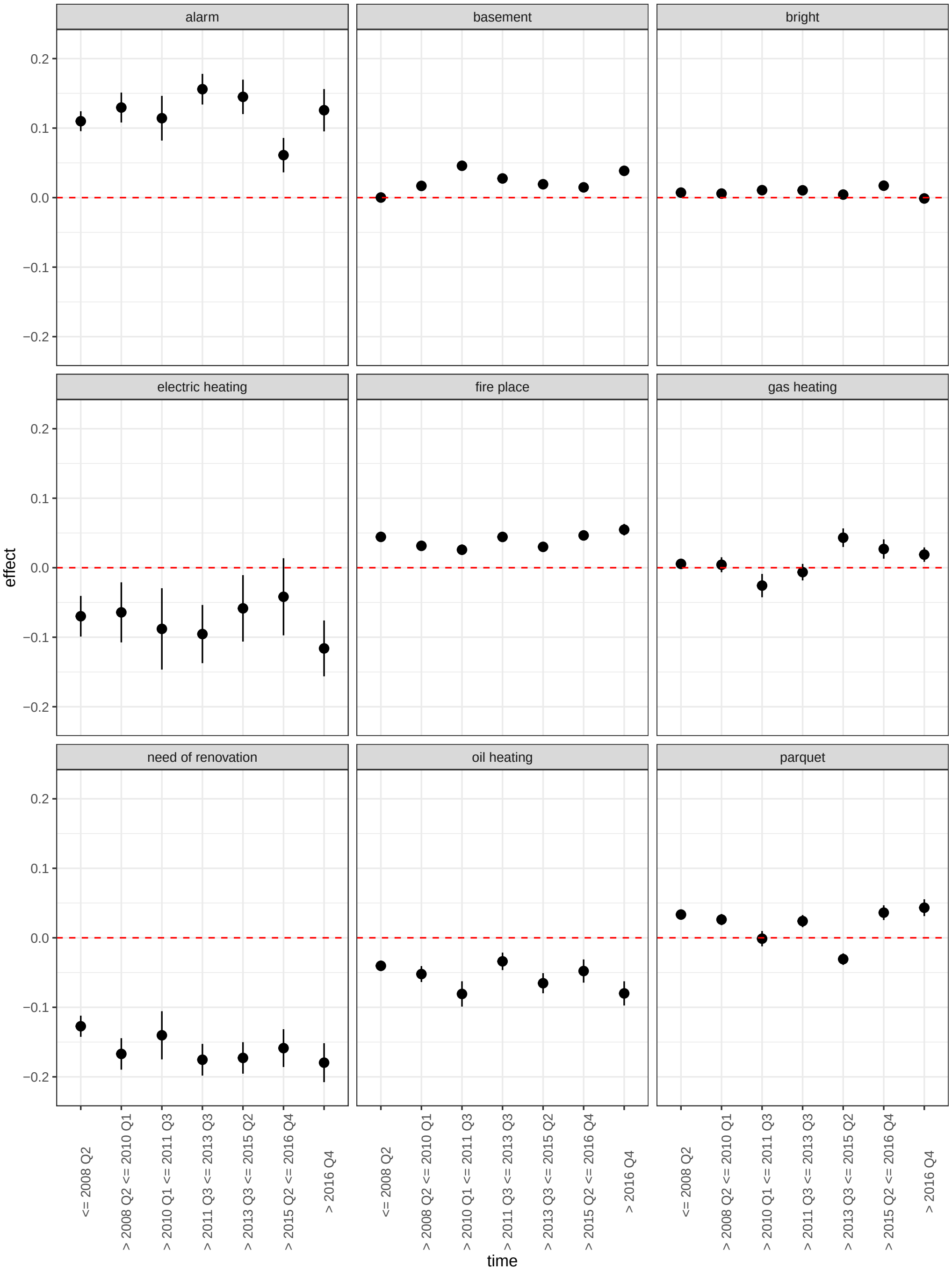


Figure 6: Interaction of dummies with time (part 1). The dot in the first period corresponds to the value of the main effect's coefficient. All subsequent periods' values refer to the sum of the main effect plus the interaction effect with the according period. Bars give 95% confidence intervals.

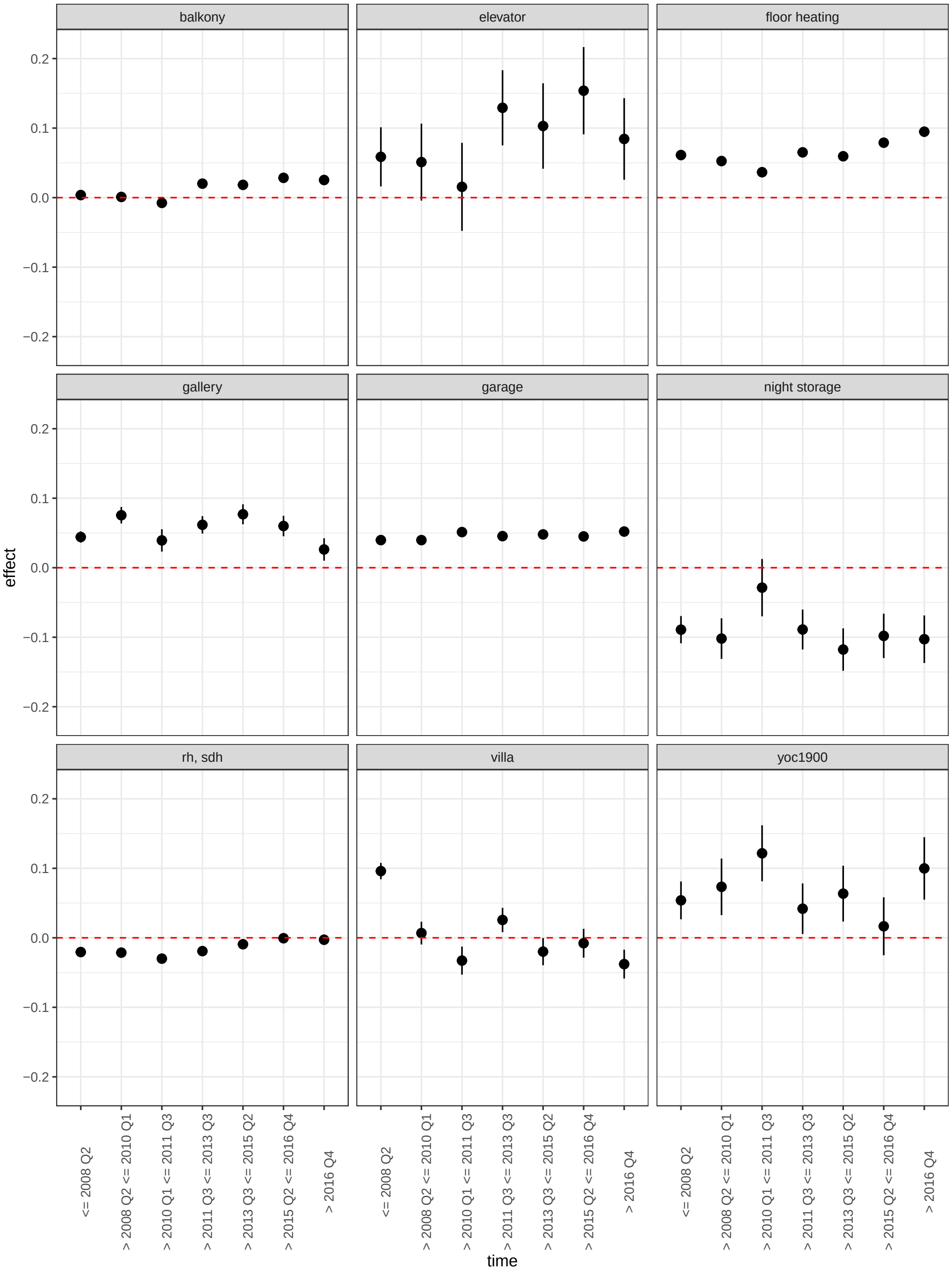


Figure 7: Interaction of dummies with time (part 2).

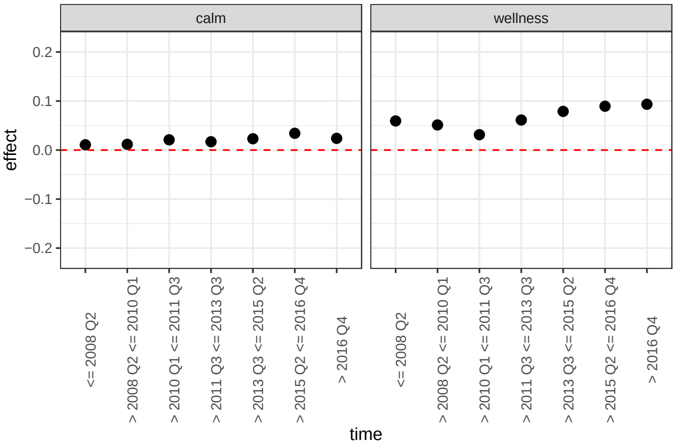
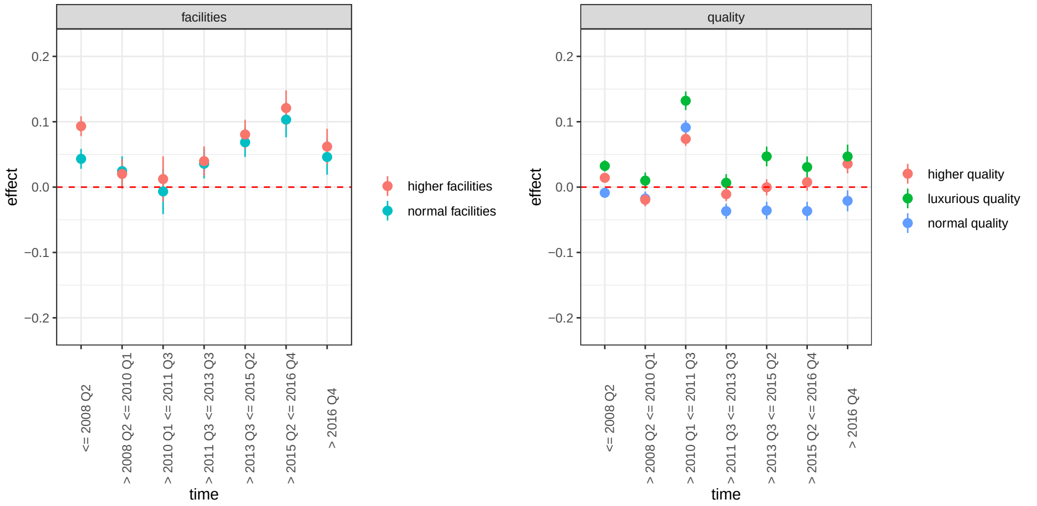


Figure 8: Interaction of dummies with time (part 3).



(a) Interaction of dummies with time (part 4).

(b) Interaction of dummies with time (part 5).

Figure 9: Interactions of variables *facilities* and *quality* with the time partitions.