universität
innsbruck

Faculty of
Economics and Statistics

Proposing a global model to manage the bias-variance tradeoff in the context of hedonic house price models

Julian Granna, Wolfgang Brunauer, and Stefan Lang

Working Papers in Economics and Statistics

2022-12

University of Innsbruck
https://www.uibk.ac.at/EconStat/

University of Innsbruck
Working Papers in Economics and Statistics

The series is jointly edited and published by

- Department of Banking and Finance
- Department of Economics
- Department of Public Finance
- Department of Statistics

Contact address of the editor:
Faculty of Economics and Statistics
University of Innsbruck
Universitaetsstrasse 15
A-6020 Innsbruck
Austria
Tel: + 43 512 507 96136
E-mail: Dean-EconStat@uibk.ac.at

The most recent version of all working papers can be downloaded at
https://www.uibk.ac.at/fakultaeten/volkswirtschaft_und_statistik/forschung/wopec/

For a list of recent papers see the backpages of this paper.

Proposing a global model to manage the bias-variance tradeoff in the context of hedonic house price models

Julian Granna[1], Wolfgang Brunauer[2], and Stefan Lang[3]

[1] *University of Innsbruck, Universitätsstr. 15, 6020 Innsbruck, Austria. e-mail: julian.granna@uibk.ac.at*
[2] *DataScience Service GmbH, Neubaugasse 56, Vienna, Austria, e-mail: wolfgang.brunauer@datascience-service.at*
[3] *University of Innsbruck, Universitätsstr. 15, 6020 Innsbruck, Austria. e-mail: stefan.lang@uibk.ac.at*

August 31, 2022

Abstract

The most widely used approaches in hedonic price modelling of real estate data and price index construction are Time Dummy and Imputation methods. Both methods, however, reveal extreme approaches regarding regression modeling of real estate data. In the time dummy approach, the data are pooled and the dependence on time is solely modelled via a (nonlinear) time effect through dummies. Possible heterogeneity of effects across time, i.e. interactions with time, are completely ignored. Hence, the approach is prone to biased estimates due to underfitting. The other extreme poses the imputation method where separate regression models are estimated for each time period. Whereas the approach naturally includes interactions with time, the method tends to overfit and therefore increased variability of estimates.

In this paper, we therefore propose a generalized approach such that time dummy and imputation methods are special cases. This is achieved by reexpressing the separate regression models in the imputation method as an equivalent global regression model with interactions of all available regressors with time. Our approach is applied to a large dataset on offer prices for private single as well as semi-detached houses in Germany. More specifically, we a) compute a Time Dummy Method index based on a Generalized Additive Model allowing for smooth effects of the continuous covariates on the price utilizing the pooled data set, b) construct an Imputation Approach model, where we fit a regression model separately for each time period, c) finally develop a global model that captures only relevant interactions of the covariates with time. An important methodolical aspect in developing the global model is the usage of model-based recursive partitioning trees to define data driven and parsimonious time intervals.

1 Introduction

Prices of residential properties, 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 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 nonparametric 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), Brunauer, Lang, and Feilmayr (2013) and Razen and Lang (2020), 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, see e.g. Hastie, Tibshirani, and Friedman (2009). Assuming parameter stability over time could be inappropriate, but modeling each time period separately possibly poses an extreme methodology as well.

In this paper, we therefore generalize the Time Dummy and Imputation Method which allows models (and indices) between the two extreme approaches. We re-express the separate regression models in the Imputation approach as one global model, such that the selection of *statistically relevant* interactions with time through statistical model choice is possible. Moreover, 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. We finally 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.

The remainder of the paper is structured as follows. This introduction is followed by a brief presentation of the concept of hedonic price indices. Here, we review Time Dummy and Imputation Approach indices and propose our generalized index construction using global regression models. Our empirical analysis in section 3 involves, following a description of the data, a) the computation of a Time Dummy Method model based on a Generalized Additive Model allowing for smooth effects of the continuous 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 4 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.

In subsections 2.1 and 2.2 we review both approaches while subsection 2.3 provides a generalization containing the Time Dummy and Imputation Method as special cases.

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, authored by de Haan and Diewert (2011). The standard Time Dummy Variable model is formulated in a semi-logarithmic form

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

where p_{it} is the price of property i in period t as a function of K characteristics captured by z_{kit}. 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 t 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

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

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 viewing 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. 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, this involves a hedonic regression model that is run separately for each period, see Hill and Melser (2008). Within the methodology of Imputation indexes, single imputation and double imputation indices are distinguished. Single imputation indices impute solely missing observations, while double imputation indices involve imputing both missing and observed prices. Hill (2011) argues that imputing both actual and unobserved prices decreases a potential omitted variable bias. 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). The heuristic of double imputation Laspeyres indices can be summarized as follows:

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 first step of the described methodology translates into the regression model

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

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_{kit}. The subscript is added, since there is a model for each time period.
Finally, predicted property prices for period 0 and t are $ln\ \hat{p}_{i0} = \hat{\beta}_{00} + \sum_{k=1}^{K} \hat{\beta}_{k0} z_{ki0}$ and $ln\ \hat{p}_{it} = \hat{\beta}_{0t} + \sum_{k=1}^{K} \hat{\beta}_{kt} z_{kit}$, respectively.
Steps 2) and 3) of the described algorithm can be expressed with the following notation. Generally, following Hill (2011), a Laspeyres type hedonic Imputation index is written as

$$P_{0t}^{L} = \frac{\sum_{i=1}^{n_0} \hat{p}_{it}}{\sum_{i=1}^{n_0} \hat{p}_{i0}},$$

which tracks the dwellings' price evolution from the base period 0 to period t. n_0 thereby gives the total number of houses sold in the base period 0. The price of housing i sold in period t is referred to as p_{it}, while p_{i0} indicates the price of dwelling i sold in period 0. Plugging in the imputed prices obtained in step 2) in both numerator and denominator of the equation yields the hedonic double imputation (DI) Laspeyres index, which is defined as

$$P_{0t}^{HDIL} = \frac{\sum_{i=1}^{n_0}\left[exp\left(\hat{\beta}_{0t} + \sum_{k=1}^{K}\hat{\beta}_{kt}z_{ki0}\right)\right]}{\sum_{i=1}^{n_0}\left[exp\left(\hat{\beta}_{00} + \sum_{k=1}^{K}\hat{\beta}_{k0}z_{ki0}\right)\right]}. \tag{5}$$

Base period prices are imputed for properties corresponding to the period t sample, evaluated at base period 0 characteristics. As the logged price per square meter is regressed on the set of covariates, prices are converted back onto a linear scale by exponentiation. Analogously to the procedure concerning the Time Dummy Method, conversion onto a linear scale requires correction for bias. In accordance with Greene (2018), and applied by e.g. Hill (2013), Malpezzi, Chun, and Green (1998), equation (5) becomes

$$P_{0t*}^{HDIL} = \frac{\sum_{i=1}^{n_0}\left[exp\left(\hat{\beta}_{0t} + \sum_{k=1}^{K}\hat{\beta}_{kt}z_{ki0} + s_t^2/2\right)\right]}{\sum_{i=1}^{n_0}\left[exp\left(\hat{\beta}_{00} + \sum_{k=1}^{K}\hat{\beta}_{k0}z_{ki0} + s_0^2/2\right)\right]},$$

where s_t^2 are the estimated variances of the model errors ϵ_{it}.

2.3 Price indices based on a global model

From a statistical point of view, the Time Dummy Approach contains only main effects of the covariates z_{kit} thereby ignoring potential interactions of the house characteristics with time. As a consequence, the Time Dummy approach is prone to (possibly substantial) bias. On the other hand, the Imputation Method considers *all* possible interactions which may lead to highly complex models and to increased variability of estimates. We are facing here a classical bias-variance tradeoff, see in the context of regression modelling e.g. Fahrmeir et al. (2022), Chap. 3.4. We therefore generalize the Time Dummy and Imputation Method such that models between the two extreme approaches are possible. To do so, we express the separate regression models in the Imputation approach as one global model, i.e.

$$ln\ p_{it} = \beta_0 + \sum_{k=1}^{K}\beta_k z_{kit} + \sum_{t=1}^{T}\delta_t D_{it} + \sum_{t=1}^{T}\sum_{k=1}^{K}\beta_{kt}D_{it}z_{kit} + \epsilon_{it}. \tag{6}$$

Here, the Time Dummy approach is obtained as a special case by assuming $\beta_{kt} \equiv 0$. For the Imputation model β_0 is the intercept and the β_k's are the effects for time period $t = 0$, $\beta_0 + \delta_t$ is the intercept in period t and $\beta_k + \beta_{kt}$ are the effects of covariate z_k at time period t. The β_{kt}'s in (6) can also be regarded as the deviation effects for covariate z_k in time t

compared to time period $t = 0$.

The advantage of our global approach is that it allows models between the two extremes of no interactions with time (Time Dummy method) and a full interaction model (Imputation approach) by setting some of the interaction effects β_{kt}'s to zero through variable and model choice.

On the basis of the introduced model framework of the global model, we proceed by constructing a hedonic Laspeyres type price index. The steps undertaken can be summarized as follows:

1. Obtain the predicted prices of all dwellings in the base period evaluated at period t.
2. Obtain predicted prices of dwellings in the base period evaluated at the base period.
3. The sum of predicted prices evaluated at period t divided by the sum of predicted prices evaluated at period 0 gives the hedonic Laspeyres type price index at period t.

Formally, the global hedonic double imputation Laspeyres (GHDIL) index is expressed as

$$P_{0t*}^{GHDIL} = \frac{\sum_{i=1}^{n_0} \left[exp\left(\hat{\beta}_0 + \sum_{k=1}^{K} \hat{\beta}_k z_{ki0} + \hat{\delta}_t + \sum_{k=1}^{K} \hat{\beta}_{kt} z_{ki0} \right) \right]}{\sum_{i=1}^{n_0} \left[exp\left(\hat{\beta}_0 + \sum_{k=1}^{K} \hat{\beta}_k z_{ki0} + \hat{\delta}_0 + \sum_{k=1}^{K} \hat{\beta}_{k0} z_{ki0} \right) \right]}.$$

The numerator comprises the sum over the bias-corrected predicted prices of houses observed in the base period evaluated at period t. The denominator gives the sum of predicted prices of houses observed in the base period evaluated at the base period.

This kind of index design can also be referred to as a double Imputation index as both the base period and the prices in period t are imputed.

2.4 Statistical models beyond linear regression

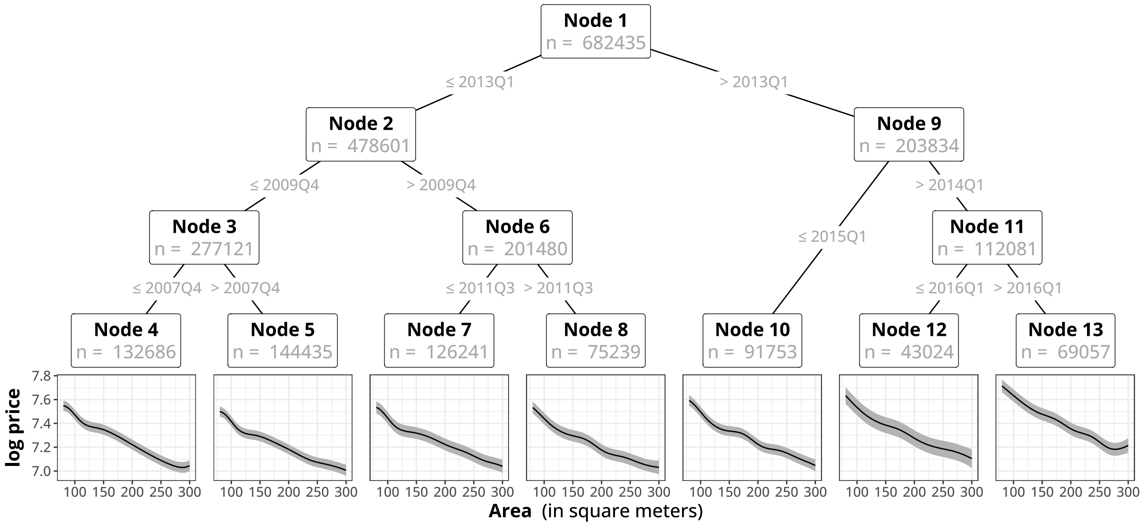
Since many effects of the z_k in (6) are possibly nonlinear, we further generalize our global model to allow for possibly nonlinear effects. We obtain

$$ln\ p_{it} = \beta_0 + \sum_{k=1}^{K} f_k(z_{kit}) + \sum_{t=1}^{T} \delta_t D_{it} + \sum_{t=1}^{T} \sum_{k=1}^{K} D_{it} f_{kt}(z_{kit}) + \epsilon_{it}, \tag{7}$$

where now the f_k's and f_{kt} are possibly nonlinear main and interaction effects with time, respectively. In the f_k's and f_{kt}'s we additionally subsume cluster specific, in particular location specific, i.i.d Gaussian random effects. In our case study in section 3 we will include 3-digit postcode locational random effects into our models in order to capture locational heterogeneity of house prices. Again, the sums $f_k + f_{kt}$ can be regarded as the (nonlinear) covariate effect of covariate z_k at time t, whereas the f_{kt}'s are the deviation effects for covariate z_k in time t compared to time period $t = 0$. Model (7) is a specific additive model, more precicely a varying coefficient model, see e.g. Fahrmeir et al. (2022), Chap. 9 for details. The model can be easily estimated using the R package mgcv by Simon Wood, see Wood (2007) and Wood (2017) for details. Here, the nonlinear functions f_k and f_{kt} are estimated based on polynomial splines, see also Appendix A for an introduction.

As elicited in the previous chapter, the choice of relevant interaction terms between time periods and the other regressors, represents a bias-variance tradeoff. The Imputation approach interacts all periods with all of the regressors, while the Time Dummy method neglects interactions completely. We tackle this tradeoff by fitting a tree-based model that allows us to identify relevant interactions with time on a model basis. Tree-based methods are especially relevant e. g. in the context of Automated Interaction Detection (AID), as introduced by Morgan and Sonquist (1963), or decision trees (Breiman et al. (2017)). Among their main advantages lies their capacity to model interactions of even high dimensions. Examples of applications of tree-based methods on real estate data include Ho, Tang, and Wong (2021), who, among other methods, utilize Random Forests and Gradient Boosting to predict property prices in Hong Kong, or Stang et al. (2022), who apply XGBoost in the context of Automated Valuation Models (AVMs) with German housing data. We employ model-based recursive partitioning as introduced by Zeileis, Hothorn, and Hornik (2008). The idea behind the approach is to fit a partitioning tree to the data, where we associate a GAM with each of the terminal nodes of the tree instead of just a simple average. Thus, we obtain an intuitively interpretable model that is able to identify interactions of the covariates with time. The tree returns time slices that are interacted with the other regressors in the context of our global model (7). In R, we use the mob()-function from the partykit-package, introduced by Hothorn and Zeileis (2015). Figure 1 illustrates the result produced by the algorithm. In this model, time, or the quarter in which an object was sold, is used as a partitioning variable and each leaf of the tree features a GAM, where the log price is regressed on a set of covariates, analogously to the models featured in Section 3.2. The algorithm, after pruning with BIC, splits the data into seven terminal nodes. As can be seen in the graph, each node is associated with a separate models, and thus, with separate effects of regressors on the dependent variable. For illustration purposes, we only plot the effect for area here. For more details on the results, see Section 3.3, and for a more thorough introduction into the concept of regression trees and model-based recursive partitioning, see Appendix A.2.2.

Figure 1: Result of model-based recursive partitioning.



Notes: The plot shows the resulting tree from the model with a GAM in each terminal node and time as the partitioning variable. The number of observations (n) is indicated in gray along with the corresponding node number and the quarter at which the data is split. The graphs on the bottom give marginal effect plots at mean of the area variable on log price in each terminal node.

3 Empirical Analysis

In this section, we introduce the dataset and the estimated models. Subsequently, we present the regression results as well as the obtained hedonic price indices.

3.1 Data

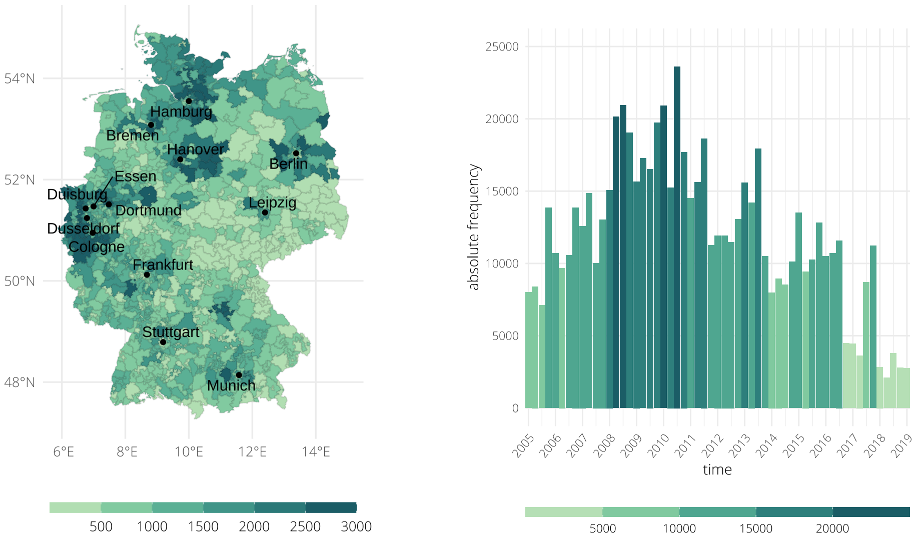
The data we utilize for our analysis is provided by the German 'F+B Forschung und Beratung für Wohnen, Immobilien und Umwelt GmbH' and comprises 682,435 observations of asking prices for private single family as well as semi-detached houses in Germany. The data set covers a time horizon from the first quarter in 2005 to the first quarter in 2019. Asking prices generally come with advantages and disadvantages. The major advantage lies in a larger sample size. This in turn implies smaller standard errors in the predicted prices and price indices next to larger variability in the explanatory variables. The major disadvantage is an 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.
A list of the variables included in our analysis along with some summary statistics is provided in Table 1. We removed extreme outliers to avoid distorted results.

variable	description	mean / rel. frequency	std. deviation	min	max
ppsm	Price per square meter	1793.600	765.660	250.000	7000.000
area	Area of flat in square meters	143.940	39.070	80.000	300.000
age	Age of flat in years	21.400	28.460	-2.000	135.000
plot.area	Plot area of object in square meters	558.210	333.040	150.000	2000.000
quarter	Quarter of last offer	25.480			
year	Year of last offer	2010.740			
PLZ3	first three digits of postcode				
alarm	Whether object has alarm system				
	0 = no	0.993			
	1 = yes	0.007			
balcony	Whether object has balcony				
	0 = no	0.767			
	1 = yes	0.233			
basement	Whether object has a basement				
	0 = no	0.583			
	1 = yes	0.417			
bright	Whether object is bright				
	0 = no	0.829			
	1 = yes	0.171			
calm	Whether flat is calm				
	0 = no	0.814			
	1 = yes	0.186			
electric.heating	Whether object has electric heating				
	0 = no	0.998			
	1 = yes	0.002			
elevator	Whether object has elevator				
	0 = no	0.999			
	1 = yes	0.001			
facilities	Degree of quality of object's facilities				
	1 = simple	0.049			
	2 = normal	0.673			
	3 = higher	0.277			
fire.place	Whether object has fire place				
	0 = no	0.871			
	1 = yes	0.129			
floor.heating	Whether object has floor heating				
	0 = no	0.800			
	1 = yes	0.200			
gallery	Whether object has a gallery				
	0 = no	0.977			
	1 = yes	0.023			
garage	Whether object has a garage				
	0 = no	0.320			
	1 = yes	0.680			
gas.heating	Whether object has gas heating				
	0 = no	0.969			
	1 = yes	0.031			
night.storage	Whether object has night storage heating				
	0 = no	0.994			
	1 = yes	0.006			
oil.heating	Whether object has oil heating				
	0 = no	0.978			
	1 = yes	0.022			
parquet	Whether object features parquet floors				
	0 = no	0.922			
	1 = yes	0.078			
quality	Degree of object's quality				
	less	0.088			
	normal	0.325			
	higher	0.511			
	luxurious	0.077			
renovation.need	Whether object is in need of renovation				
	0 = no	0.951			
	1 = yes	0.049			
villa	Whether object is a villa				
	0 = no	0.984			
	1 = yes	0.016			
wellness	Whether object features a swimming pool / whirlpool				
	0 = no	0.925			
	1 = yes	0.075			
yoc1900	Whether object is build after 1900				
	0 = no	0.017			
	1 = yes	0.983			

Table 1: List of variables included in the analysis including corresponding summary statistics.

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 2a. 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 2b displays the distribution of observations over time. Obviously, 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 for regression analysis and index creation.

Figure 2: Distribution of observations over location and time.



(a) 3-digit postcode areas.

(b) Quarters.

Notes: Panel 2a reports the density of observations in three-digit postcode areas in Germany. Dark green areas around larger cities, e. g. Cologne, Hamburg, Berlin, or Munich, refer to a higher density of houses. Light green areas indicate a lower density of observations. In panel 2b, the frequency of observations over time is given. Each bin corresponds to one quarter. Dark green bins refer to a higher frequency of houses, light green indicate low frequencies.

3.2 Models

Our analysis comprises the computation and comparison of the following models (terms in brackets refer to the corresponding shortcuts):

(TD) A model comprising data pooled over all time periods corresponding to a typical

Time Dummy Method approach. More specifically we estimate the model

$$ln\ p_{it} = \beta_0 + \sum_{t=1}^{T} \delta_t D_{it} + f_1(area_{it}) + f_2(age_{it}) + f_3(plotarea_{it}) + f_4(PLZ_{it}) + \sum_{k=1}^{K} \beta_k z_{kit} + \epsilon_{it}, \tag{8}$$

where the effects f_1-f_3 of the continuous variables *area*, *age*, and *plotarea* are modeled in a smooth, nonlinear way utilizing penalized regression splines, the $\beta_k z_{kit}$ comprise further effects of categorical dwelling characteristics, see Table 1 for a complete list of the covariates included. In analogy with Brunauer, Feilmayr, and Wagner (2012), spatial heterogeneity is captured utilizing i.i.d. Gaussian random effects $f_4(PLZ_{it})$ over 3-digit postcode dummies.

(yImp) Separate models

$$ln\ p_{it} = \beta_{0t} + f_{1t}(area_{it}) + f_{2t}(age_{it}) + f_{3t}(plotarea_{it}) + f_{4t}(PLZ_{it}) + \sum_{k=1}^{K} \beta_{kt} z_{kit} + \epsilon_{it}, \tag{9}$$

stratified for years $t = 2005, \ldots, 2019$ are built. This setting represents a typical application of a yearly Imputation Approach index.

(qImp) Separate models as in (9) now stratified for quarters rather than for years, i.e. $t = 2005/1, \ldots, 2019/4$.

(S1) For model (S1), in the first step, we fit a model-based recursive partitioning tree, where the logged price per square meter is regressed on all variables in Table 1 and the continuous variables *area*, *plotarea*, and *age* are fit using penalized splines. We choose time in quarters as the partitioning variable, such that the model returns relevant interactions between time and other covariates. In the second step, we fit a global model in the form (6) and interact each covariate with the time slices obtained from the partitioning in the first step. This allows to identify the variables, for which interaction with time is most important utilizing 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) Analogous to (S1), but leave out the interaction term between the time partitions and the postcode dummy random effects.

(S6) Analogous to (S1), but leave out the interaction terms between the time partitions and the continuous variables *area*, *plotarea*, and *age*.

The model-based recursive partitioning in models (S1)-(S6) yields the following partition of time

1. 2005 Q1 - 2007 Q4,
2. 2008 Q1 - 2009 Q4,
3. 2010 Q1 - 2011 Q3,
4. 2011 Q4 - 2013 Q1,
5. 2013 Q2 - 2015 Q1,
6. 2015 Q2 - 2016 Q1, and
7. 2016 Q2 - 2019 Q1,

resulting in only seven time periods in contrast to 15 or even 57 periods in case of (yImp) and (qImp), respectively.

3.3 Results

For model comparison, we randomly split the data into a training (comprising 682,435 observations) and a validation dataset (comprising 75,870 observations). The training data is employed to estimate the models, the validation data is used to assess the models by comparing predicted with observed prices. More specifically, we compute the root mean square error (RMSE), which, in accordance with Greene (2018), is given by

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\hat{p}_i - p_i\right)^2},$$

where $n = 75870$, $\hat{p}_i$ are the predicted prices according to the respective model and p_i are the observed prices.
Table 2 reports the out-of-sample prediction errors of the evaluated models. We draw the following conclusions:

- The Time Dummy model (TD) has the hightest prediction errors of all computed models.
- The Imputation Approach based on yearly time intervals (yImp) outperforms the approach based on quarterly intervals (qImp).

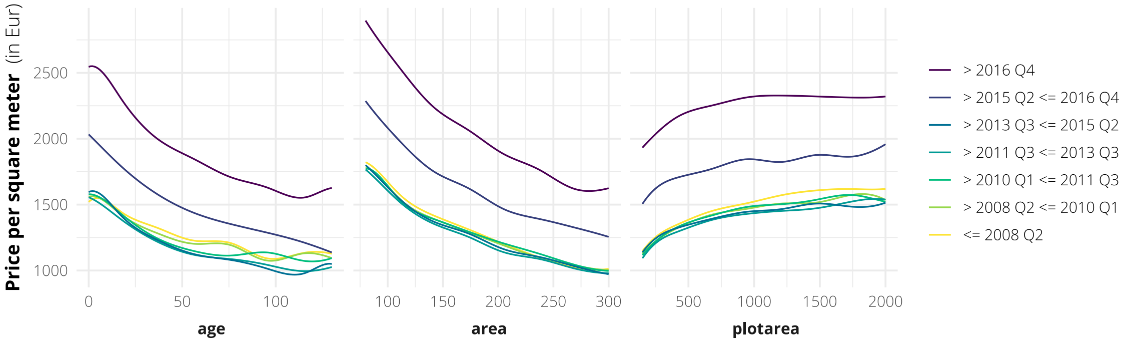
- The models (S1), (S2), (S3) and (S4) further have (slightly) lower values in RMSE. Thus, Interactions play a role, but not all interactions are equally important. Model (S2) is the globally best performing model implying that an interaction between *area* and time is not relevant. Models (S2), (S3), and (S4) further all outperform the Imputation Approach models, which indicates that the continuous covariates are not substantially interacted with time. Even if we exclude all continuous covariate interaction terms in model (S6), the predictive accuracy is not substantially reduced.
- The most important interaction 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.

	(TD)	(yImp)	(qImp)	(S1)	(S2)	(S3)	(S4)	(S5)	(S6)
RMSE	480.03	459.12	470.23	458.48	458.38	459.02	458.97	472.44	459.67

Table 2: 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) in Figure 3. The first panel indicates that 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. 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.

Figure 3: Marginal effect plots for continuous variables for model (S1).



Notes: Marginal effect plots at mean for the continuous variables area, age, and plot area. Regarded variables are varied over the given range, while other variables are fixed at their mean level. Colors of the curves refer to the time periods obtained from recursive partioning.

As seen in the second panel, 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 larger 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. Analogous to the previous graph, the lines mainly differ in their level rather than their functional form. The Figure thus further supports the conclusion that interaction with *area* is irrelevant. Finally, the marginal effect plot of *plotarea* implies that the average price of housing generally rises with increasing *plotarea* and there appears to be a saturation effect for greater values of plot area. The functional relationship seems similar over all time slices. 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 *plotarea*, there could be a relevant interaction with time, but the interaction does not appear to be large in magnitude.

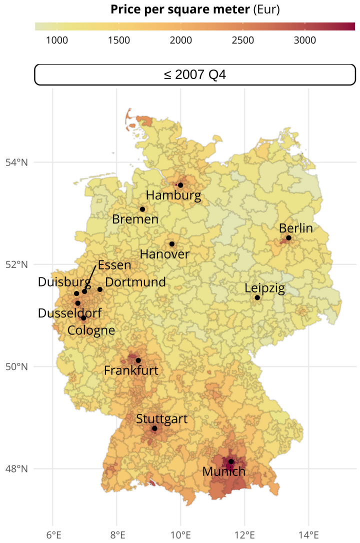
Figures 4 and 5 provide further insights into the relevance of the interaction between time and location. The chosen form of display gives insights into not only the absolute level of house prices in Germany, but also allows to identify which regions are subject to steeper price appraisals compared to other parts of the country. 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 first two panels in Figure 5 do not indicate a strong space-time interaction. Prices do not substantially change between 2008 Q2 and 2011 Q3 on average. The following graphs, however, emphasize the relevance of location-time interaction terms. Between 2011 Q4 and 2015 Q2, price changes in Germany are distributed quite heterogeneously. Areas around 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, Figures 4 and 5 emphasize 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.

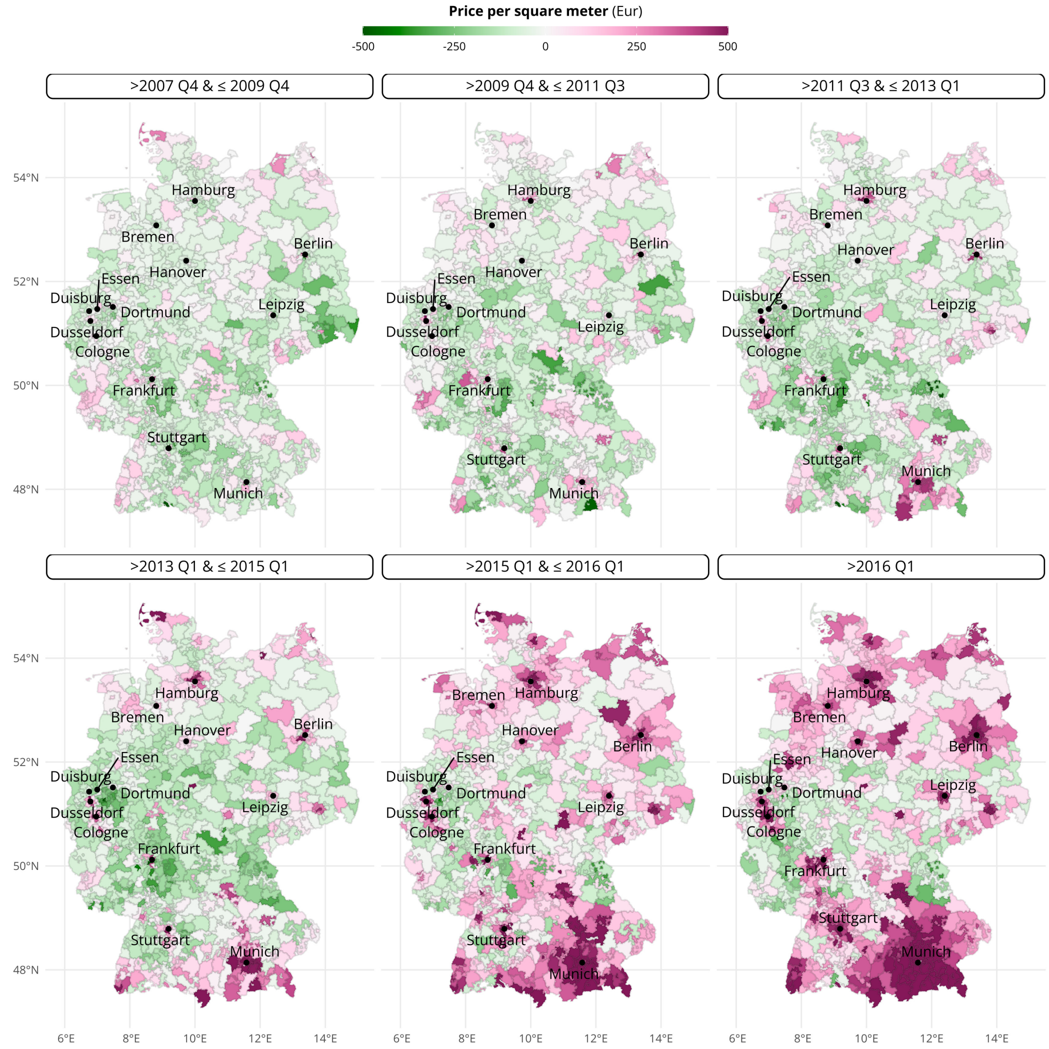
For completeness, we provide Figures 8, 9, 10, and 11 in Appendix B, where the evolution of the discrete variables' coefficients over time is depicted.

Figure 4: Marginal Effect of postcode dummies over time periods in model (S1) - first period.



Notes: This graph gives the marginal effect at mean level of the postcode dummies in the first period obtained from recursive partitioning. The graph refers to the predicted absolute level of price per square meter in Euros for houses sold between 2005 Q1 and 2008 Q2. Light areas correspond to lower average prices, dark areas refer to high average prices per square meter. Prices are again reported on a linear scale including bias correction.

Figure 5: Marginal Effect of postcode dummies over time periods in model (S1) - with reference to first period.



Notes: Graphs show the deviation of the marginal effect at mean level of the postcode dummies from the first period. Green colored polygons refer to zones subject to price drops compared with the first period, red zones refer to increases in the predicted price per square meter.

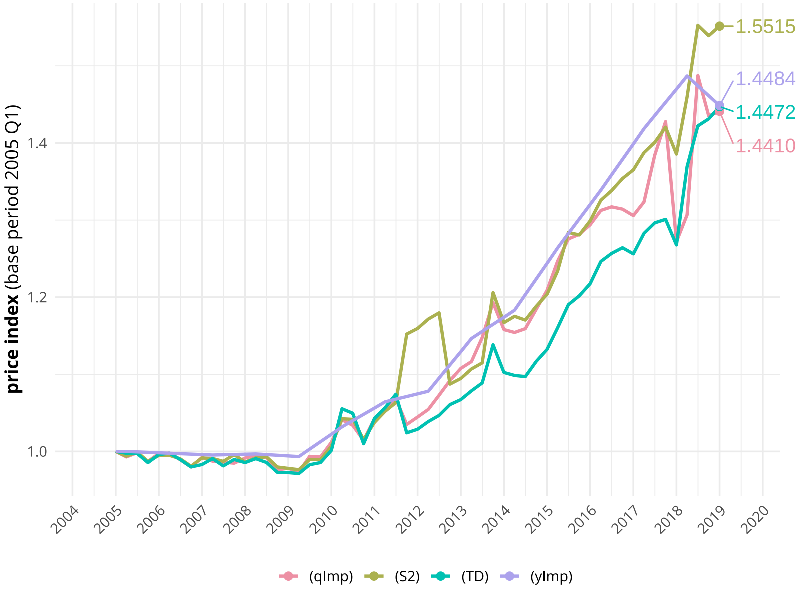
3.4 Resulting indices

Figure 6 shows the obtained hedonic indices from the models (TD), (qImp), (yImp) and the global best model (S2). 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 2010 until mid 2013, the hedonic curves are subject to steep price 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.

Figure 6: Hedonic price indices.



Notes: Resulting indices from utilized models. The red line depicts the quarterly imputation type index referring to model (qImp). The green curve indicates the evolution of the imputation index, resulting from the globally best model (S1). The index produced by the Time Dummy index from (TD) is given by the blue line. Finally, the purple line gives the yearly Imputation index from (yImp). The numbers at the ends of the curves indicate the final value of the corresponding curve.

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 roughly 44%, while the Imputation index derived from (S1) returns a price increase of close to 55% 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.

4 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 models and compare them regarding their 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 construct a global approach using 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.

References

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

Breiman, Leo, Jerome H Friedman, Richard A Olshen, and Charles J Stone (2017). *Classification and regression trees*. Routledge.

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 (2022). *Regression. Models, Methods and Applications*. Springer Verlag.

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.

Ho, Winky KO, Bo-Sin Tang, and Siu Wai Wong (2021). “Predicting property prices with machine learning algorithms”. In: *Journal of Property Research* 38.1, pp. 48–70.

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.

Hothorn, Torsten and Achim Zeileis (2015). “partykit: A Modular Toolkit for Recursive Partytioning in R”. In: *Journal of Machine Learning Research* 16, pp. 3905–3909. URL: `https://jmlr.org/papers/v16/hothorn15a.html`.

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

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

Malpezzi, Stephen, Gregory H Chun, and Richard K Green (1998). “New place-to-place housing price indexes for US Metropolitan Areas, and their determinants”. In: *Real Estate Economics* 26.2, pp. 235–274.

Morgan, James N and John A Sonquist (1963). “Problems in the analysis of survey data, and a proposal”. In: *Journal of the American statistical association* 58.302, pp. 415–434.

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

Razen, A. and S. Lang (2020). “Random scaling factors in Bayesian distributional regression models with an application to real estate data”. In: *Statistical Modelling* 20.4, pp. 347–368.

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 nonparametric regression curve fitting”. In: *Journal of the Royal Statistical Society: Series B (Methodological)* 47.1, pp. 1–21.

Stang, Moritz, Bastian Krämer, Cathrine Nagl, and Wolfgang Schäfers (2022). “From human business to machine learning—methods for automating real estate appraisals and their practical implications”. In: *Zeitschrift für Immobilienökonomie*, pp. 1–28.

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.

Wood, Simon N (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. DOI: `10.1198/106186008X319331`. eprint: `https://doi.org/10.1198/106186008X319331`. URL: `https://doi.org/10.1198/106186008X319331`.

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 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, we 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. (2022). The remainder of the section briefly outlines the concept of model-based recursive partitioning.

A.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. (2022), the GAM can be regarded as an extension of the multiple linear regression model with

$$y_i = f_1(z_{i1}) + \cdots + f_q(z_{iq}) + \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i, \tag{10}$$

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.

A.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. (2022). 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

$$f(z) = \sum_{j=1}^{d} \gamma_j B_j(z), \tag{11}$$

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

$$B_j^l(z) = \frac{z - \kappa_j}{\kappa_{j+1} - \kappa_j} B_j^{l-1}(z) + \frac{\kappa_{j+l+1} - z}{\kappa_{j+l+1} - \kappa_{j+1}} B_{j+1}^0(z), \tag{12}$$

where $\kappa_1, \ldots, \kappa_m$ are the inner m knots. As equation (12) 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. (2022) for a more detailed overview.

Finally, in order to obtain and estimate a model that is linear in its parameters, equation (11) is substituted into equation (10)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.

A.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

$$PLS(\lambda) = \sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{d} \gamma_j B_j(z_i) \right)^2 + \lambda \sum_{j=k+1}^{d} \left(\mathit{\Delta}^k \gamma_j \right)^2, \tag{13}$$

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

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

In matrix notation, equation (13) 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. (2022), 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).

A.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).

A.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.

A.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

$$\hat{\theta} = \underset{\theta \in \Theta}{\operatorname{argmin}} \sum_{i=1}^{n} \Psi(Y_i, \theta), \tag{14}$$

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 (14) formulated over all regions can then be written as the optimization problem

$$\sum_{m=1}^{M} \sum_{i \in I_m} \Psi(Y_i, \theta_m) \to \min, \tag{15}$$

over all partitions $\{R_m\}$ with the indexes $I_m, m = 1, \ldots, M$. Equation (15) 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.

B Interaction of time with discrete variables

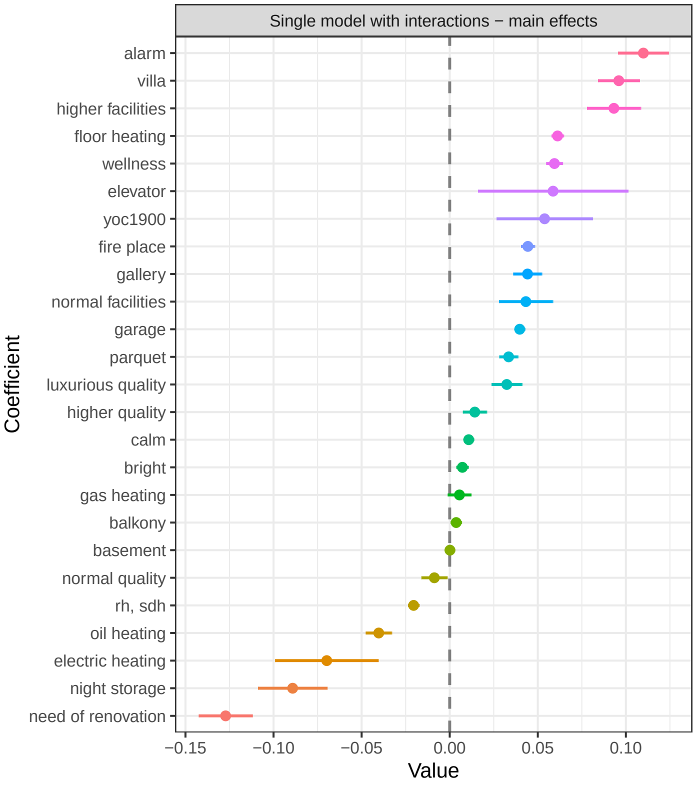


Figure 7: Main effects coefficient values and confidence intervals for (S1). The dots refer to the point estimate, while the bars give corresponding 95% confidence intervals.

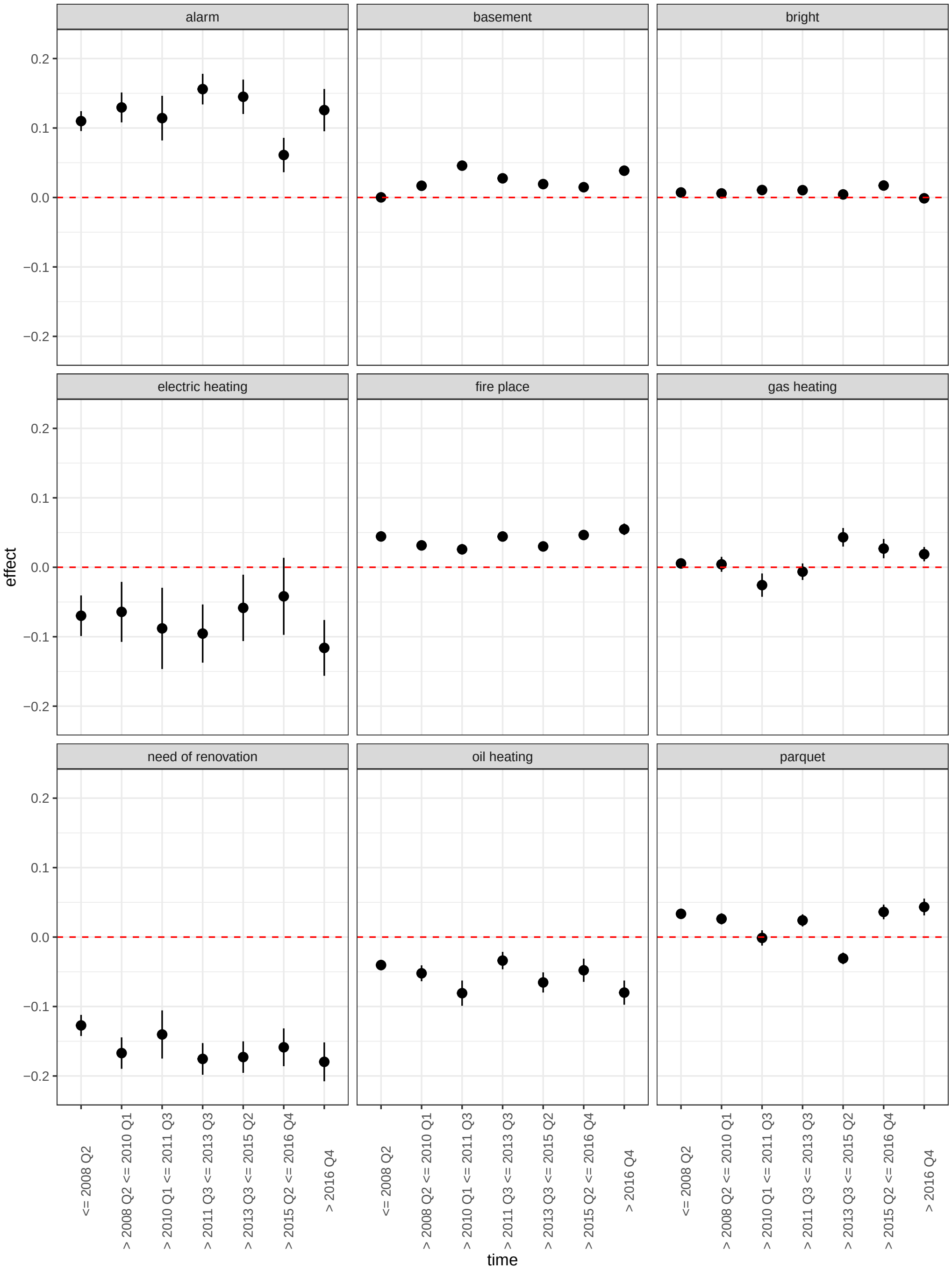


Figure 8: Interaction of dummies with time (part 1). The dot in the first period corresponds to the value of the main effect's coefficient. All subsequent periods' values refer to the sum of the main effect plus the interaction effect with the according period. Bars give 95% confidence intervals.

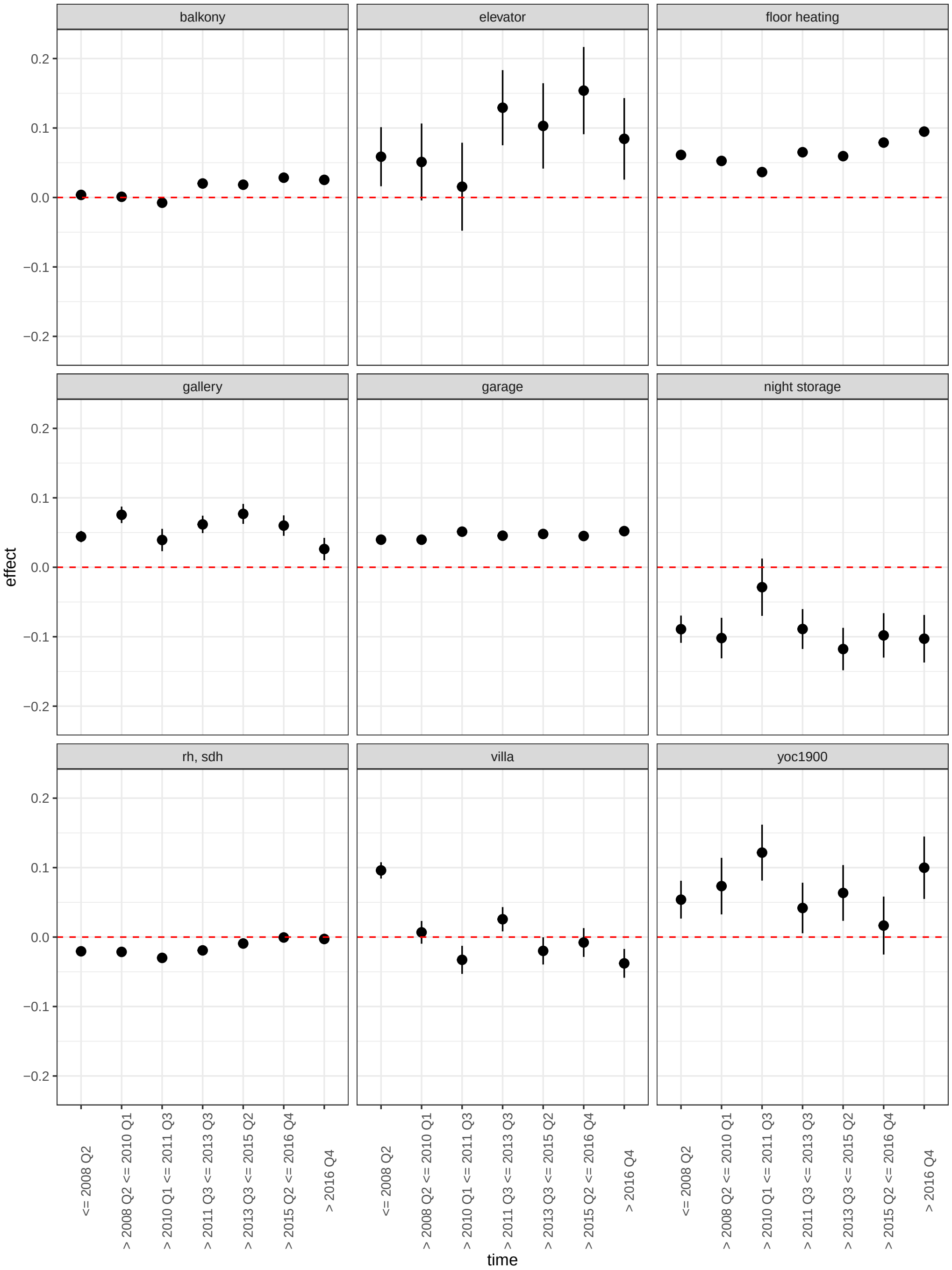


Figure 9: Interaction of dummies with time (part 2).

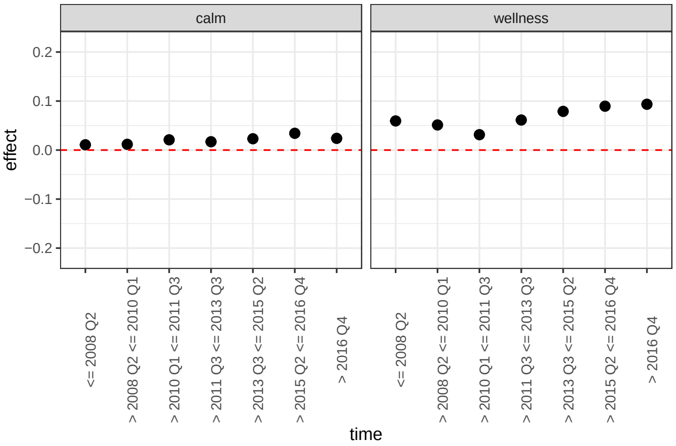
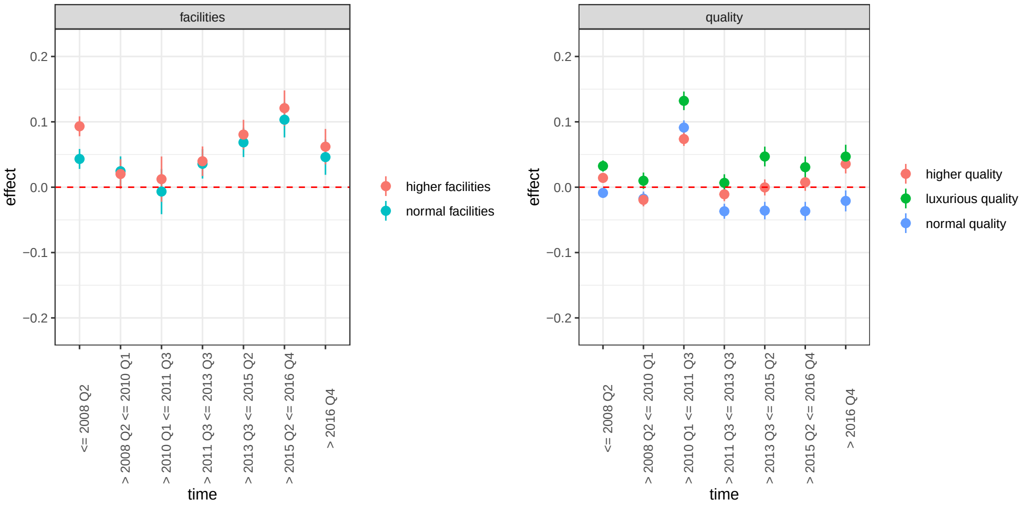


Figure 10: Interaction of dummies with time (part 3).



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

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

Figure 11: Interactions of variables *facilities* and *quality* with the time partitions.

University of Innsbruck - Working Papers in Economics and Statistics
Recent Papers can be accessed on the following webpage:

https://www.uibk.ac.at/eeecon/wopec/

2022-12 **Julian Granna, Wolfgan Brunauer, Stefan Lang:** Proposing a global model to manage the bias-variance tradeoff in the context of hedonic house price models

2022-11 **Christoph Baumgartner, Stjepan Srhoj and Janette Walde:** Harmonization of product classifications: A consistent time series of economic trade activities

2022-10 **Katharina Momsen, Markus Ohndorf:** Seller Opportunism in Credence Good Markets The Role of Market Conditions

2022-09 **Christoph Huber, Michael Kirchler:** Experiments in Finance A Survey of Historical Trends

2022-08 **Tri Vi Dang, Xiaoxi Liu, Florian Morath:** Taxation, Information Acquisition, and Trade in Decentralized Markets: Theory and Test

2022-07 **Christoph Huber, Christian König-Kersting:** Experimenting with Financial Professionals

2022-06 **Martin Gächter, Martin Geiger, Elias Hasler:** On the structural determinants of growth-at-risk

2022-05 **Katharina Momsen, Sebastian O. Schneider:** Motivated Reasoning, Information Avoidance, and Default Bias

2022-04 **Silvia Angerer, Daniela Glätzle-Rützler, Philipp Lergetporer, Thomas Rittmannsberger:** How does the vaccine approval procedure affect COVID-19 vaccination intentions?

2022-03 **Robert Böhm, Cornelia Betsch, Yana Litovsky, Philipp Sprengholz, Noel Brewer, Gretchen Chapman, Julie Leask, George Loewenstein, Martha Scherzer, Cass R. Sunstein, Michael Kirchler:** Crowdsourcing interventions to promote uptake of COVID-19 booster vaccines

2022-02 **Matthias Stefan, Martin Holmén, Felix Holzmeister, Michael Kirchler, Erik Wengström:** You can't always get what you want–An experiment on finance professionals' decisions for others

2022-01 **Toman Barsbai, Andreas Steinmayr, Christoph Winter:** Immigrating into a Recession: Evidence from Family Migrants to the U.S.

2021-32 **Fanny Dellinger:** Housing Support Policies and Refugees' Labor Market Integration in Austria

2021-31 **Albert J. Menkveld, Anna Dreber, Felix Holzmeister, Jürgen Huber, Magnus Johannesson, Michael Kirchler, Sebastian Neusüss, Michael Razen, Utz Weitzel and et al:** Non-Standard Errors

2021-30 **Toman Barsbai, Victoria Licuanan, Andreas Steinmayr, Erwin Tiongson, Dean Yang:** Information and Immigrant Settlement

2021-29 **Natalie Struwe, Esther Blanco, James M. Walker:** Competition Among Public Good Providers for Donor Rewards

2021-28 **Stjepan Srhoj, Melko Dragojević:** Public procurement and supplier job creation: Insights from auctions

2021-27 **Rudolf Kerschbamer, Regine Oexl:** The effect of random shocks on reciprocal behavior in dynamic principal-agent settings

2021-26 **Glenn E. Dutcher, Regine Oexl, Dmitry Ryvkin, Tim Salmon:** Competitive versus cooperative incentives in team production with heterogeneous agents

2021-25 **Anita Gantner, Regine Oexl:** Respecting Entitlements in Legislative Bargaining - A Matter of Preference or Necessity?

2021-24 **Silvia Angerer, E. Glenn Dutcher, Daniela Glätzle-Rützler, Philipp Lergetporer, Matthias Sutter:** The formation of risk preferences throughsmall-scale events

2021-23 **Stjepan Srhoj, Dejan Kovač, Jacob N. Shapiro, Randall K. Filer:** The Impact of Delay: Evidence from Formal Out-of-Court Restructuring

2021-22 **Octavio Fernández-Amador, Joseph F. Francois, Doris A. Oberdabernig, Patrick Tomberger:** Energy footprints and the international trade network: A new dataset. Is the European Union doing it better?

2021-21 **Felix Holzmeister, Jürgen Huber, Michael Kirchler, Rene Schwaiger:** Nudging Debtors to Pay Their Debt: Two Randomized Controlled Trials

2021-20 **Daniel Müller, Elisabeth Gsottbauer:** Why Do People Demand Rent Control?

2021-19 **Alexandra Baier, Loukas Balafoutas, Tarek Jaber-Lopez:** Ostracism and Theft in Heterogeneous Groups

2021-18 **Zvonimir Bašić, Parampreet C. Bindra, Daniela Glätzle-Rützler, Angelo Romano, Matthias Sutter, Claudia Zoller:** The roots of cooperation

2021-17 **Silvia Angerer, Jana Bolvashenkova, Daniela Glätzle-Rützler, Philipp Lergetporer, Matthias Sutter:** Children's patience and school-track choices several years later: Linking experimental and field data

2021-16 **Daniel Gründler, Eric Mayer, Johann Scharler:** Monetary Policy Announcements, Information Schocks, and Exchange Rate Dynamics

2021-15 **Sebastian Bachler, Felix Holzmeister, Michael Razen, Matthias Stefan:** The Impact of Presentation Format and Choice Architecture on Portfolio Allocations: Experimental Evidence

2021-14 **Jeppe Christoffersen, Felix Holzmeister, Thomas Plenborg:** What is Risk to Managers?

2021-13 **Silvia Angerer, Daniela Glätzle-Rützler, Christian Waibel:** Trust in health care credence goods: Experimental evidence on framing andsubject pool effects

2021-12 **Rene Schwaiger, Laura Hueber:** Do MTurkers Exhibit Myopic Loss Aversion?

2021-11 **Felix Holzmeister, Christoph Huber, Stefan Palan:** A Critical Perspective on the Conceptualization of Risk in Behavioral and Experimental Finance

2021-10 **Michael Razen, Alexander Kupfer:** Can increased tax transparency curb corporate tax avoidance?

2021-09 **Changxia Ke, Florian Morath, Anthony Newell, Lionel Page:** Too big to prevail: The paradox of power in coalition formation

2021-08 **Marco Haan, Pim Heijnen, Martin Obradovits:** Competition with List Prices

2021-07 **Martin Dufwenberg, Olof Johansson-Stenman, Michael Kirchler, Florian Lindner, Rene Schwaiger:** Mean Markets or Kind Commerce?

2021-06 **Christoph Huber, Jürgen Huber, and Michael Kirchler:** Volatility Shocks and Investment Behavior

2021-05 **Max Breitenlechner, Georgios Georgiadis, Ben Schumann:** What goes around comes around: How large are spillbacks from US monetary policy?

2021-04 **Utz Weitzel, Michael Kirchler:** The Banker's Oath And Financial Advice

2021-03 **Martin Holmen, Felix Holzmeister, Michael Kirchler, Matthias Stefan, Erik Wengström:** Economic Preferences and Personality Traits Among Finance Professionals and the General Population

2021-02 **Christian König-Kersting:** On the Robustness of Social Norm Elicitation

2021-01 **Laura Hueber, Rene Schwaiger:** Debiasing Through Experience Sampling: The Case of Myopic Loss Aversion.

University of Innsbruck

Working Papers in Economics and Statistics

2022-12

Julian Granna, Wolfgang Brunauer, and Stefan Lang

Proposing a global model to manage the bias-variance tradeoff in the context of hedonic house price models

Abstract

The most widely used approaches in hedonic price modelling of real estate data and price index construction are Time Dummy and Imputation methods. Both methods, however, reveal extreme approaches regarding regression modeling of real estate data. In the time dummy approach, the data are pooled and the dependence on time is solely modelled via a (nonlinear) time effect through dummies. Possible heterogeneity of effects across time, i.e. interactions with time, are completely ignored. Hence, the approach is prone to biased estimates due to underfitting. The other extreme poses the imputation method where separate regression models are estimated for each time period. Whereas the approach naturally includes interactions with time, the method tends to overfit and therefore increased variability of estimates.
In this paper, we therefore propose a generalized approach such that time dummy and imputation methods are special cases. This is achieved by reexpressing the separate regression models in the imputation method as an equivalent global regression model with interactions of all available regressors with time. Our approach is applied to a large dataset on offer prices for private single as well as semi-detached houses in Germany. More specifically, we a) compute a Time Dummy Method index based on a Generalized Additive Model allowing for smooth effects of the continuous covariates on the price utilizing the pooled data set, b) construct an Imputation Approach model, where we fit a regression model separately for each time period, c) finally develop a global model that captures only relevant interactions of the covariates with time. An important methodolical aspect in developing the global model is the usage of modelbased recursive partitioning trees to define data driven and parsimonious time intervals.

ISSN 1993-4378 (Print)
ISSN 1993-6885 (Online)