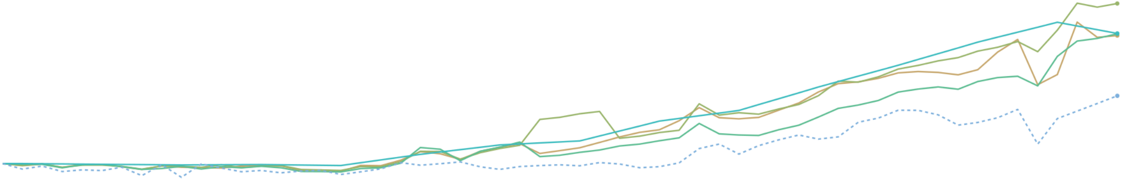
universität
innsbruck



ERES ’21: Managing the Bias-Variance Tradeoff in the Context of House Price Prediction and Hedonic Indices

An Application for German Housing Data

Julian Granna, Wolfgang Brunauer, Stefan Lang

www.uibk.ac.at/statistics

Contents

1. **Hedonic House Price Indices**
2. **Time Dummy Method and Imputation Approach**
3. **Model-Based Recursive Partitioning**
4. **Data**
5. **Model Results**
6. **Resulting Indices**
7. **Discussion and Outlook**

Motivating Hedonic House Price Indices

Hedonic indices have become the Gold Standard for index construction.[5] They can be characterized as

- expressing house prices as a function of their characteristics,
- within the framework of a regression model.

On the basis of hedonic regression, quality-adjusted price indices can be constructed. Among them, the most relevant[1] are

- Time-Dummy Method, and the
- Imputation Approach.

Time Dummy vs Imputation Approach

Time Dummy Method: A regression model is fit to the pooled data set.

- Does not capture interactions (between time and other covariates).
- Implicitly, constant parameters are assumed over all time periods.
- Very simple; little variant, but biased.

Imputation Approach: A model is fit separately for each time period .

- Implicitly includes interaction effect between time periods and all covariates.
- Relaxes constant parameter assumption over time.
- High model complexity; increased variance, less biased.
- Stratification into time periods naively chosen.

→ Both methods are extremes!

Strategy

Strategy: Find relevant interactions with time using model-based recursive partitioning.

- Split up the time horizon into regions using model-based recursive partitioning with a linear model in the terminal nodes.
- Fit a global Generalized Additive Model and interact all covariates with the obtained time partitions (equivalent to separately fitting model in each partition).
- Check which interactions with time are most important by graphical inspection and out-of-sample prediction accuracy.

Generalized Additive Models

The hedonic indices are computed employing a Generalized Additive Model (GAM) framework. In this context, the effects are estimated using **Penalized Splines**. These offer several advantages (towards e.g. high polynomials):

- splines are very flexible,
- do not imply a specific relationship between the dependent variable and its regressors,
- typically yield a better fit, especially for nonlinear relationships, and
- they are local.

Model-Based Recursive Partitioning

Like GAMs, model-based recursive partitioning is considered a technique of supervised statistical learning.

Tree-based methods:

- Regression Trees
- Model-based recursive partitioning
- Random Forests
- Boosting
- . . .

Model-Based Recursive Partitioning

- introduced by Zeileis et al. (2008)[7]

Idea: Model-based recursive partitioning, is a tree-based method, where each leaf in a regression tree is not associated with a simple average, but instead with a fitted model, e. g. a linear model.

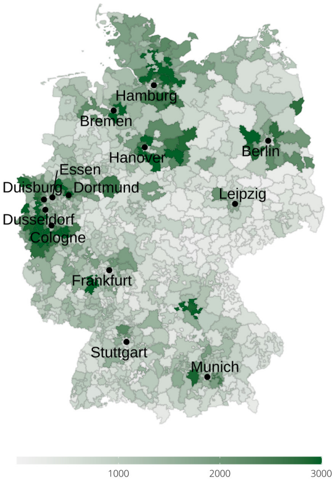
Data

- provided by the F+B Forschung und Beratung für Wohnen, Immobilien und Umwelt GmbH
- 682,435 observations of offer prices for private single family as well as semi-detached houses in Germany
- time horizon from 2005 Q1 to 2019 Q1

Data

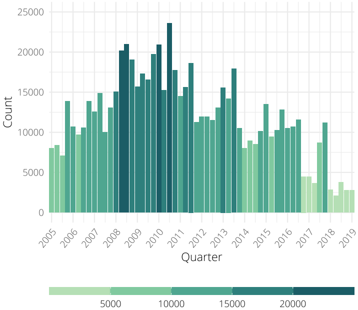
Distribution of Observations within Germany

Number of observations in each 3-digit postcode area



Distribution of observations over time

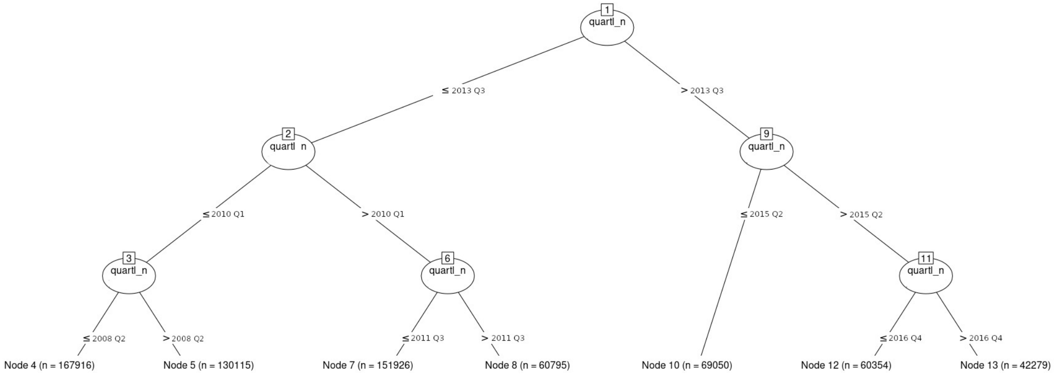
Number of observations in each quarter



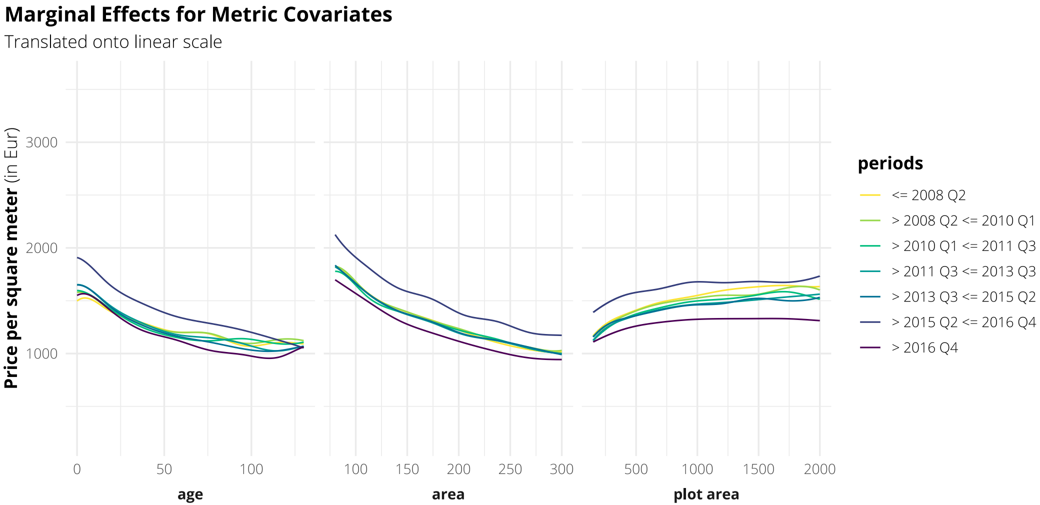
Variables

- the analysis includes the **metric variables:**
  plot area, (living) area, age
- next to the **categorical variables:**
  garage, renovation, quality, basement, alarm, calm, elevator, floor heating, gallery, gas heating, electric heating, oil heating, night storage, bright, fire place, balcony, parquet, wellness, floor heating, year of construction < 1900, type, facilitie, quality
- plus location: 3-digit postcode dummies
- plus time (quarter/year)

Model Results: Model-Based Recursive Partitioning



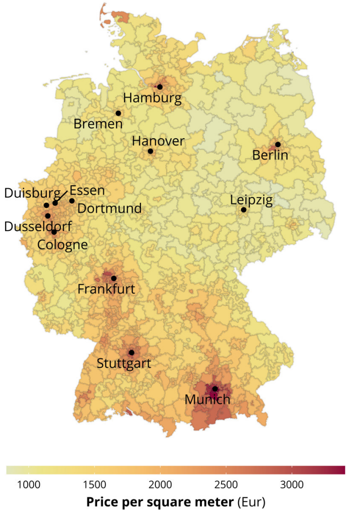
Model Results: Effects Interacted with Time

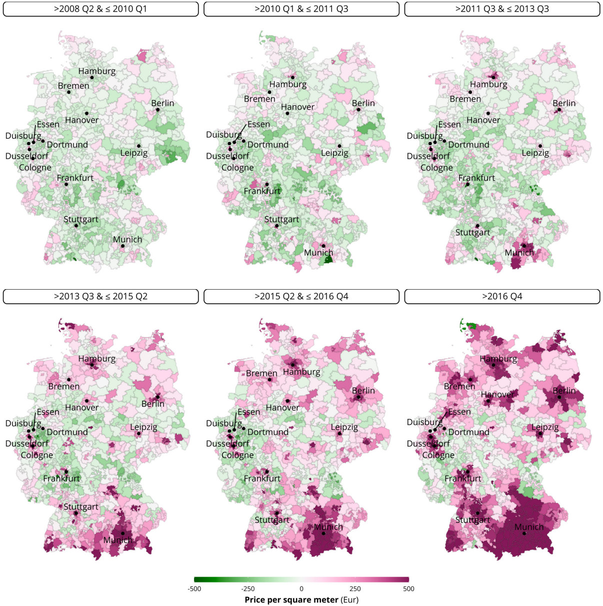


Marginal Effects for 3-Digit Postcode

Translated onto linear scale

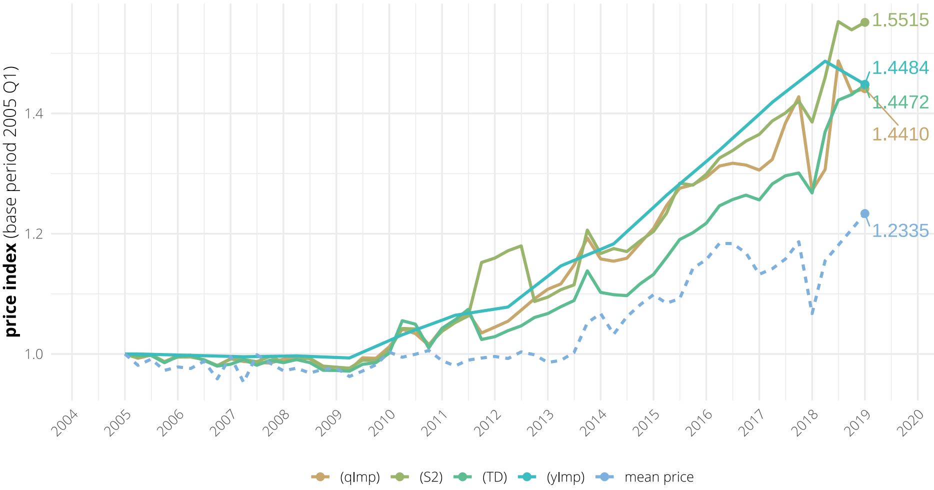
≤ 2008 Q2




>2008 Q2 & ≤ 2010 Q1
>2010 Q1 & ≤ 2011 Q3
>2011 Q3 & ≤ 2013 Q3
>2013 Q3 & ≤ 2015 Q2
>2015 Q2 & ≤ 2016 Q4
>2016 Q4
Hamburg
Bremen
Berlin
Hanover
Essen
Duisburg
Dortmund
Dusseldorf
Cologne
Leipzig
Frankfurt
Stuttgart
Munich
-500
-250
0
250
500
Price per square meter (Eur)

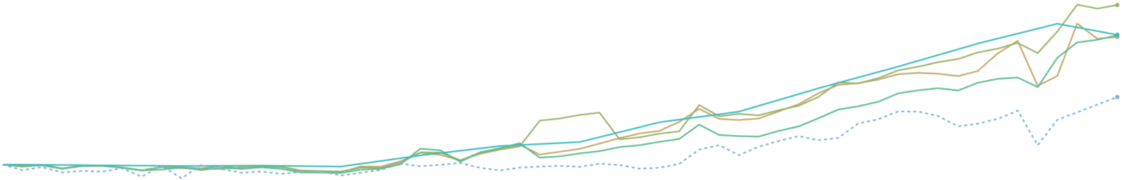
Hedonic Price Indices

Comparison of indices resulting from utilized models



Summary

- Interaction with time plays an important role
- Neglecting interaction leads to bias
- Naive stratification leads to less biased results, but is paid for with high complexity (and variance).
- Application of model-based recursive partitioning more appropriate to identify and capture interaction with time.
- Outlook: We only include interactions with time. Other interactions, like with location are expected to be crucial, too. But: Model-based recursive partitioning not well suited for that task.
- To Do: Clustering mechanism to identify local (or alike) clusters. Apply model-based recursive partitioning with GAMs in terminal nodes directly.



Thank you for your attention!

Julian Granna, Wolfgang Brunauer, Stefan Lang

www.uibk.ac.at/statistics

References I

W. Brunauer, W. Feilmayr, and K. Wagner.
A New Residential Property Price Index for Austria.
Statistiken - Daten und Analysen Q3, 90-102, 2012.

P. H. C. Eilers and B. D. Marx.
Flexible smoothing with B-splines and penalties.
Statistical Science, 11(2):89–121, 1996.

L. Fahrmeir, T. Kneib, and S. Lang.
Regression.
Springer-Verlag Berlin Heidelberg, 2007.

T. Hothorn and A. Zeileis.
partykit: A modular toolkit for recursive partytioning in r.
The Journal of Machine Learning Research, 16(1):3905–3909, 2015.

References II

ILO, IMF, OECD, UNECE, Eurostat, and The World Bank.
Consumer Price Index Manual: Theory and Practice.
ILO Publications, Geneva., 2004.

J. L. W. V. Jensen.
Sur les fonctions convexes et les inégalités entre les valeurs moyennes.
Acta mathematica, 30:175–193, 1906.

A. Zeileis, T. Hothorn, and K. Hornik.
Model-based recursive partitioning.
Journal of Computational and Graphical Statistics, 17(2):492–514, 2008.

Appendix: Hedonic Indices - Time Dummy Method

Standard Time Dummy Model in semi-logarithmic form

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

- in simplest form, model (1) is fit over the whole data
- price index P_{0t}^{TD} is then simply derived by $P_{0t}^{TD} = exp(\hat{\delta}_t)$, or, to avoid bias,[6] by

$$P_{0t}^{TD} = exp(\hat{\delta}_t + \frac{1}{2} Var(\hat{\delta}_t))$$

go back

Appendix: Hedonic Indices - Imputation Approach

Adjusted semi-logarithmic equation that is fit for each time period separately:

$$ln\ p_{it} = \beta_{0t} + \sum_{k=1}^{K} \beta_{kt} z_{ikt} + \epsilon_{it} \tag{2}$$

Index Construction

1. Fit a regression model in each period.
2. Apply standard index formulae to obtain price index. In our application: Laspeyres.

go back

Appendix: GAMs Notation

Generalized Additive Models

Can be regarded as an extension of the multiple linear regression model:[3]

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

with dependent variable y_i, intercept β_0, a set of regressors $x_{i1}, \ldots, x_{ik}$, and smooth effects of the regressors, $f(z_{i1}), \ldots, f(z_{iq})$.

go back

Appendix: GAMs Basis Splines

For the construction of Basis Splines

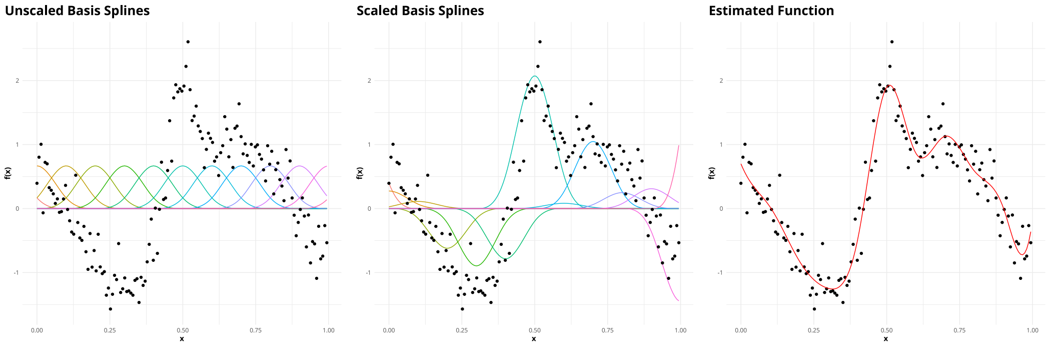
- Divide range of regressors into m equidistant segments, or **knots**.
- Construct $(l+1)$ polynomial fractures, where l is the degree of the spline.
- The fractures are put together so that they are $(l-1)$-times continuously differentiable.
- Finally, the linear combination of $d = m + l - 1$ basis functions gives the representation of f(z):

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

where B_j are the basis functions and γ_j the corresponding coefficients.

go back

Basis Splines



Unscaled Basis Splines

Scaled Basis Splines

Estimated Function

go back

Appendix: GAMs Penalized Splines

Main challenge lies in choice in number of knots as a choice between a good fit and model complexity. Penalized Splines are a possible solution to this. In terms of a penalized least squares optimization:[2]

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

where Δ^k are the differences of k-th order. Thus, the second term poses a penalty on the coefficients. λ is often chosen via minimization of AIC or GCV.

go back

Appendix: Regression Trees

Idea: Partition the characteristic space into rectangles and fit an average in each space.

Consider dependent variable Y with p explanatory variables and n observations. The goal is to identify splitting variables and split points. Crete M regions $R_1, R_2, \ldots, R_m$ and model the response as a constant c_m in each region, so that

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

where $I(x \in R_m)$ is an indicator variable which is 1, if observation x is in region R_m and 0 else.

go back

Appendix: Regression Trees

With minimization of sum of squares, one obtains 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. However, choice of optimal partitions over all variables is computationally unfeasible. Solution:

Greedy algorithm

1. In the current node choose splitting point for all covariates.
2. Select splitting variable with greatest reduction in residual sum of squares.
3. Repeat previous steps in all daughter nodes.
4. Stop, when minimum threshold of observation is reached in all terminal nodes.

go back

Appendix: Model-Based Recursive Partitioning

Idea: Model-based recursive partitioning, is a tree-based method, where each leaf in a regression tree is not associated with a simple average, but instead with a fitted model, e. g. a linear model.

Suppose a global parametric model $\mathcal{M}(Y, \theta)$ with observations Y and parameter vector θ. Obtain model by minimization objective function $\Psi(Y, \theta)$:

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

Then, divide the characterstic space into regions, so that:

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

becomes the optimization problem.

Appendix: Model-Based Recursive Partitioning Algorithm

The following process describes the model-based recursive partitioning algorithm:

Model-Based Recursive Partitioning Algorithm

1. Fit model with $\hat{\theta}$ to all corresponding observations by minimizing objective function Ψ, in our case, least squares.
2. (Do a fluctuation test to test for parameter instability.)
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.
5. Prune the tree.

→ implemented in the `partykit`-package.[4]

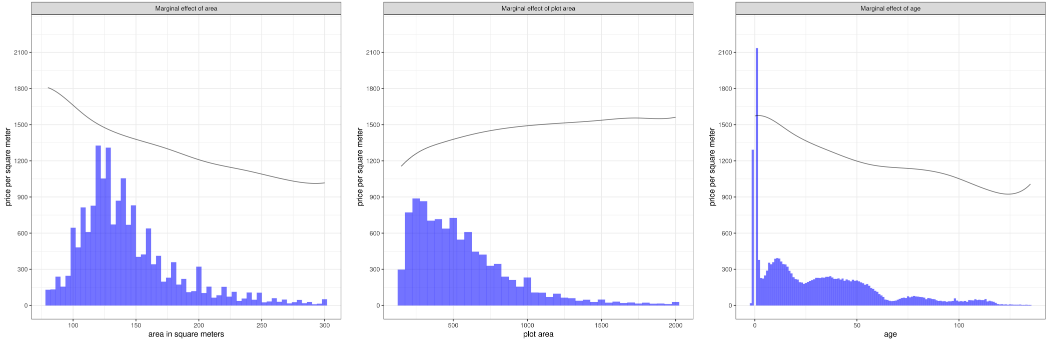
go back

Appendix: Results - Prediction Accuracy

Model	RMSE	EDF
Time Dummy	480.03	771.89
Yearly Imputation	459.12	9,343.94
Quarterly Imputation	470.23	35,257.08
tree-based: all interactions	458.33	4,935.49
tree-based: no interaction w/ area	458.27	4,891.51
tree-based: no interaction w/ plot area	458.89	4,891.77
tree-based: no interaction w/ age	458.85	4,887.35
tree-based: no interaction w/ postcode	472.44	1,058.33
tree-based: no interaction w/ metric	459.53	4,800.77

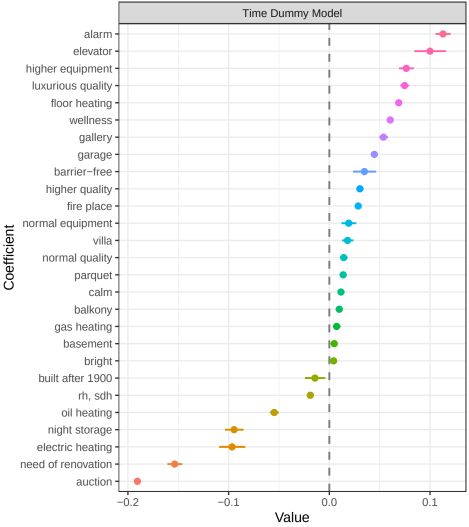
go back

Appendix: Time Dummy Method



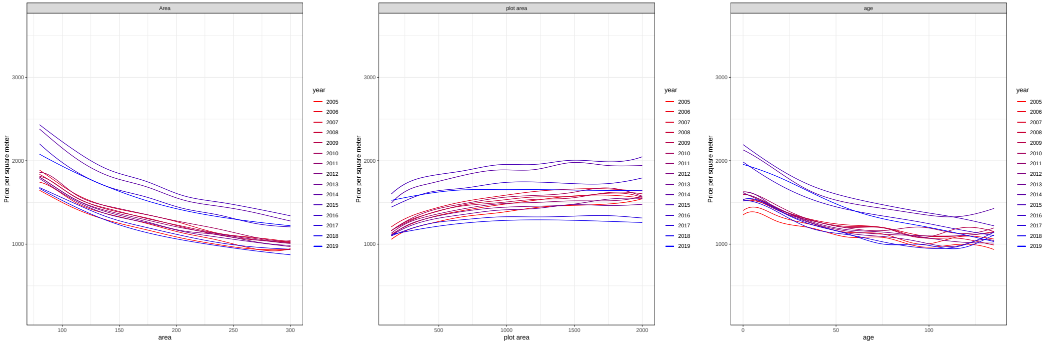
go back

Appendix: Time Dummy Method



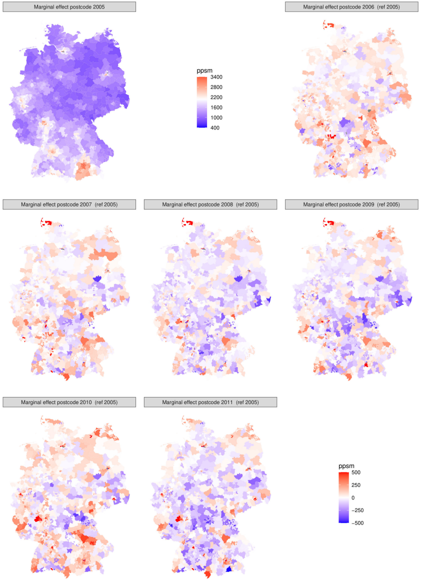
go back

Appendix: Imputation Approach



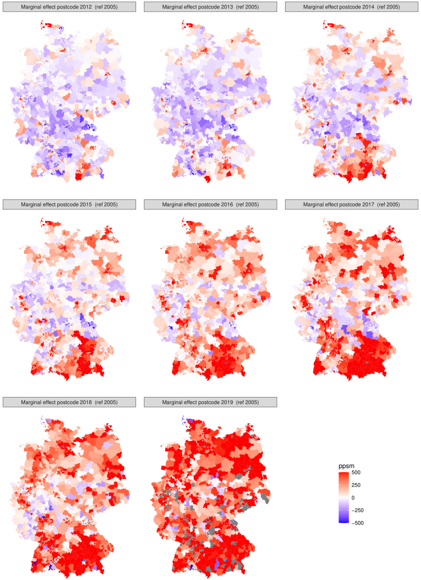
go back

Appendix


Marginal effect postcode 2005
ppsm
3400
2800
2200
1600
1000
400
Marginal effect postcode 2006 (ref 2005)
Marginal effect postcode 2007 (ref 2005)
Marginal effect postcode 2008 (ref 2005)
Marginal effect postcode 2009 (ref 2005)
Marginal effect postcode 2010 (ref 2005)
Marginal effect postcode 2011 (ref 2005)
ppsm
500
250
0
−250
−500

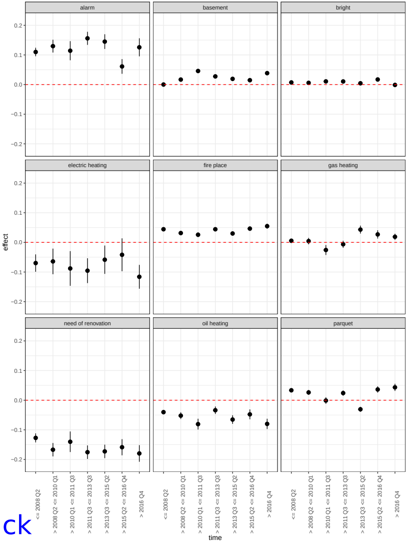
go back

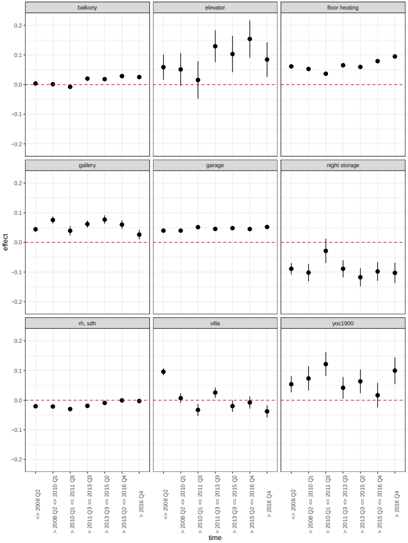
Appendix


Marginal effect postcode 2012 (ref 2005)
Marginal effect postcode 2013 (ref 2005)
Marginal effect postcode 2014 (ref 2005)
Marginal effect postcode 2015 (ref 2005)
Marginal effect postcode 2016 (ref 2005)
Marginal effect postcode 2017 (ref 2005)
Marginal effect postcode 2018 (ref 2005)
Marginal effect postcode 2019 (ref 2005)
ppsm
500
250
0
−250
−500

go back

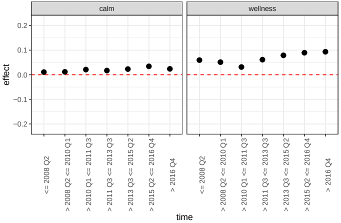
Appendix

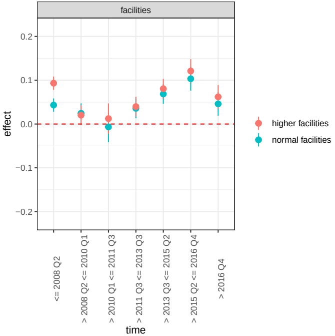


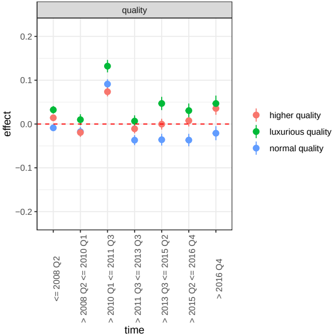


go back

Appendix: Imputation Approach







go back

Appendix: Summary Statistics

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: Summary statistics of continuous variables.

go back