Kolbe, Jens; Schulz, Rainer; Wersing, Martin; Werwatz, Axel

Working Paper

How useful is listings data for research?

FORLand-Working Paper, No. 19 (2020)

Provided in Cooperation with:
DFG Research Unit 2569 FORLand "Agricultural Land Markets – Efficiency and Regulation", Humboldt-Universität Berlin

Suggested Citation: Kolbe, Jens; Schulz, Rainer; Wersing, Martin; Werwatz, Axel (2020) : How useful is listings data for research?, FORLand-Working Paper, No. 19 (2020), Humboldt-Universität zu Berlin, DFG Research Unit 2569 FORLand "Agricultural Land Markets - Efficiency and Regulation", Berlin, https://doi.org/10.18452/21038

This Version is available at:
https://hdl.handle.net/10419/213123

How useful is listings data for research?

Jens Kolbe, Rainer Schulz, Martin Wersing
and Axel Werwatz

FORLand-Working Paper 19 (2020)

Published by
DFG Research Unit 2569 FORLand, Humboldt-Universität zu Berlin
Unter den Linden 6, D-10099 Berlin
https://www.forland.hu-berlin.de

Tel +49 (30) 2093 46845, Email gabriele.wuerth@agrar.hu-berlin.de

Agricultural Land Markets – Efficiency and Regulation

How useful is listings data for research?

Jens Kolbe, Rainer Schulz, Martin Wersing,

and Axel Werwatz*

*Kolbe and Werwatz (corresponding author): Chair for Econometrics and Business Statistics, Institute for Economics and Business Law, Technische Universität Berlin, Straße des 17. Juni 135, 10623 Berlin, Germany; jkolbe@tu-berlin.de and axel.werwatz@tu-berlin.de. Schulz and Wersing: University of Aberdeen Business School, Edward Wright Building, Dunbar Street, Aberdeen AB24 3QY, United Kingdom; r.schulz@abdn.ac.uk and martin.wersing@abdn.ac.uk.

Abstract

Compared to other asset classes, information on transactions of residential real estate is scarce and available only with delay. Listing information from web-platforms is abundant and timely. Is listings data useful for research? We examine this question and find that distributions of ask and sale prices differ significantly, both because of characteristics composition and implicit pricing. Estimates of the average willingness to pay from ask data can be widely off when compared with estimates from sale data. Ask data is also not useful to predict prices of individual houses and suffer from large error variances. Quality-controlled ask and sale price indices show similar trends and we find that an ask price index can be used for nowcasting. Overall, our analysis shows that ask data has limited potential for research, and is no substitute for sale data.

Keywords: hedonic modelling, nowcasting, price prediction, stochastic dominance

JEL Classification: C14, C81, R31

1 Introduction

Online listings platforms allow home owners, real estate agents, and developers to offer properties to a wide audience with little effort and for only a small fee.[1] This appeals to those who have already decided to sell a property and to those who just want to test the market. Consequently, ask data is abundant and timely and researchers have started to use it as a substitute for sale data.

Ask data has also shortcomings, however. First, it is not clear whether a particular listing will result in a sale or not. Only successful listings are linked to market outcomes. Second, ask prices tend to overstate sale prices, as home owners are prone to overestimate the market value of their home and have an incentive to set the ask price above the market value estimate to leave leverage for haggling (Goodman and Ittner (1992), Kiel and Zabel (1999), and Horowitz (1992)).[2] Third, information on listed properties might be inaccurate. Users might not report information that makes a property unattractive and misreport other information in error. If these shortcomings affect the inference that is drawn, then ask data will be of little use to researchers.

In this paper, we examine the value of ask data for three regression applications: (1) quantification of the relationship between house prices and characteristics (*hedonic pricing*), (2) estimation of market values (*automated valuation*), and (3) measurement of house price dynamics (*price index construction*). For each application, we compare the results we obtain from ask

[1] Examples of such platforms are *Zillow* (US), *Rightmove* (UK), *Immoscout* (Austria, Germany, and Switzerland).

[2] In the data of Shimizu et al. (2016), the average ask price is about 25% to 36% higher than the average registered sale price, depending on whether a listing was successful or not. In the matched data of Haurin et al. (2010) and Carrillo (2012), the average ask price is 4% and 2% higher than the average sale price, respectively.

data with those we obtain from sale data. The three applications are used widely in practice and research. Environmental studies use *hedonic prices* to estimate the willingness to pay (WTP) for non-marketed amenities. Once estimated, benefits of a policy can be assessed, such as a ban of night flights to limit noise pollution (Taylor 2017). Practitioners use *automated valuations* to assess the tax base for property taxes or the collateral value of non-performing loans (RICS 2017). Central banks and financial regulators use *house price indices* to guide policy, for instance with respect to the timing of macroprudential measures, financial institutions use such indices for loan portfolio risk management, and individuals use them for buy and sell decisions.[3]

A paper related to ours is Shimizu et al. (2016), which examines ask and sale data distributions for condominiums in Tokyo.[4] The paper finds that price distributions differ and that this is driven mostly by the composition of characteristics in the two data sets. We conduct a similar, but methodologically improved, examination of our single-family house data from Berlin, Germany. In particular, we test explicitly whether variables in the ask data are stochastically larger than their sale data counterparts and we conduct improved inference on the estimated counterfactual price distributions (Barrett and Donald 2003, Chernozhukov et al. 2013). We also assess by how much differences in characteristics *and* implicit prices contribute to the dominance relationships. We find that both such differences have an statistically and economically significant impact on the counterfactual price distributions.

In our main contribution, we go beyond the analysis of distributions and

[3]Bauer et al. (2017) and Winke (2017) use ask data to estimate the WTP for nuclear power plant closures and, respectively, aircraft noise. Bauer et al. (2013) propose to substitute sale with ask price indices, as abundant ask data allows estimation at higher frequencies. Bauer et al. do not examine, however, the accuracy of such indices.

[4]We refer to the parts where initial ask (P_1) and final sale (P_4) data are analysed.

examine in detail what the distributional differences imply for the three regression applications. Throughout the examination of these applications, we use semiparametric additive models to impose as little structure on the data as possible.

Our main findings are as follows. Hedonic regressions fitted to ask data can lead to implicit price functions for house characteristics that are counter-intuitive. Estimates of average WTP can also be widely off when compared with the estimates from sale data. Ask prices are not very useful to predict prices of individual houses and suffer from upward bias and large error variance. Quality-controlled ask and sale price indices show similar trends and we find statistical evidence that the ask price index can be used for nowcasting. Overall, our analysis shows that ask data can be useful for research when no transaction data is available or if time is essential. However, researchers working with such data must also be aware of their shortcomings.

The rest of the paper is organised as follows. Section 2 describes the sale and the ask data sets and examines their differences. Section 3 explains the empirical methodology that we use to examine the three regression applications. Section 4 presents the results. Section 5 concludes. The web-based appendix provides further details.

2 Transaction and listings data

2.1 Preparation of the data sets

2.1.1 Data sources

The data cover the period 2007-2015. The transaction data is provided by Berlin's surveyor commission (GAA, Gutachterausschuss für Grundstückswerte in Berlin). By law, surveyor commissions are obliged to keep a detailed record of each and every real estate transaction that takes place in Berlin. To facilitate this, commissions have access to sale contracts, administrative data, and can request further clarification from parties involved in a transaction. Each observation has information on the sale price, physical and legal characteristics of the building and the plot, such as rights of way, and on legal specifics of the transaction, such as personal or business relations between the contracting parties, such a divorce and inheritance or a sale that stipulates deferred payment. We use only arm's length transactions, which leads to 17,650 observations in the GAA data.

The listings data is provided by Immoscout24 (IS24), the self-professed largest real estate platform in Germany.[5] As a multi-sided platform, IS24 puts potential sellers and buyers in contact and allows third parties, such as agents, mortgage banks, and appraisers, to advertise their services. IS24 listings are similar to classified ads, but modern technology gives much more flexibility. For instance, the content of an ad can be modified during listing's term; it is also possible to extend the term while the listing is still active.[6] Those searching

[5] In 2017, IS24 listed 470,000 properties and had about 13m visitors per month, 1.9 and 1.6 times as many as the next largest competitor (Scout24 2017).

[6] Possible terms are: two weeks, one month, three months. Listings can also be *premium*

for properties can register with IS24 and will receive afterwards personalised newsletters with updates on visited listings and links to similar properties on offer. As marketing platform, IS24 takes no responsibility that the information on listed properties is complete, correct, and that the properties are still available, i.e., have not been sold in the meantime (IS24 Terms and Conditions, Immobilien Scout 24 (2018, 5.1, 6.1, 9.1)).

For each IS24 listing, we keep only the information from the last day for which it is observed. This is the date closest to a transaction if the listing could attract a buyer. Obviously, a listing could have also ended because the seller decided to take an unsold property from the market. Such a property might then be listed again under a different identification code, perhaps with slightly varied information on the property. It is also possible that the very same property is marketed independently by several different agents at the same time. *If* the property is sold eventually, it will have produced several observations in the listing data, but only one in the transaction data. This helps to motivate why the original IS24 data has 144,274 observations, about 8.2 times as many as the GAA data.

2.1.2 Data cleaning

The IS24 data suffers from many patchy observations, the result of relying solely upon user provided information. We concentrate on observations with sale (GAA) and ask (IS24) price that have complete entries for the following *core* variables: plot area, (exterior) floor area, building age, house type, and

or *basic*. Premium listings permit a detailed presentation of the property and the ad will be placed more prominently on the web page.

administrative district in which the house is located.[7] In some parts of our examination, we use coordinates to model location values.

Despite the fairly small set of core variables, Table 1 shows that 26% of observations in the IS24 data must be removed. No observation in the GAA data must be removed, a sign of data quality.

[Table 1 about here.]

The remaining rows in Table 1 show the effects of deleting unusual observations. First, we remove observations of houses that are either still under construction or older than 100 years. Both are different from standard houses in the sense that the former do not exist yet and that the latter have existed for longer than usual. This reduces the number of observations by 14% (GAA) and 21% (IS24). Second, we apply bounds to the plot area, the floor area, and the price to floor area ratio. A researcher equipped only with listings data would use such publicly available information for data preparation.[8] We treat the GAA data equally and apply the same bounds to it. This reduces the numbers of observations by 18% (GAA) and 19% (IS24). The final data sets have 12,524 (GAA) and 68,070 (IS24) observations; we refer to the former as *sale* (index s) and the latter as *ask* data (index a).

[7]The GAA (IS24) data reports for most (all) observations *exclusively* the exterior (interior) floor area. The GAA suggests a factor of 1.25 to convert interior to exterior floor area (Gutachterausschuss für Grundstückswerte 2011, p. 44). We apply this factor to the IS24 observations, but examine alternatives in the robustness analysis.

[8]The bounds are differentiated further by location, house type, and vintage of the building. We collate the bounds from annual reports published by the GAA, see the web-based Appendix A.

2.2 The two data sets

Table 2 presents descriptive statistics. The markup of ask to sale price is 28% for the arithmetic averages ($\overline{P}_a/\overline{P}_s - 1$) and 26% for the geometric averages ($\exp\{\overline{p}_a - \overline{p}_s\} - 1$). The markups are sizable and similar to those reported in Shimizu et al. (2016).

[Table 2 about here.]

Figure 1 gives further evidence on the price distributions, where we concentrate on log prices, as it is common in the literature. The left panel shows the markups for the percentiles of the price distributions.[9] The markups are particularly high in the tails. All markups are strictly positive and statistically significant. Given the density estimates in the right panel, it seems that ask prices dominate sale prices stochastically, which would imply $F_a(p) - F_s(p) \leqslant 0$ for all $p \in [0, \max(p_a, p_s)]$.[10] The dominance is strong if the inequality is strict for some p.

[Figure 1 about here.]

We follow Barrett and Donald (2003, p.75) to test for strong dominance. Their procedure is based on the Kolmogorov-Smirnov (KS) test with statistic ($j \neq k$)

$$\widehat{d}_{j,k} = \left(\frac{N_j N_k}{N_j + N_k}\right)^{0.5} \sup_p \left\{\widehat{F}_j(p) - \widehat{F}_k(p)\right\} \tag{1}$$

[9] The mark-up at quantile τ is $\exp\{p_a(\tau) - p_s(\tau)\} - 1 \doteq p_a(\tau) - p_s(\tau)$; we estimate the right-hand side with quantile regressions of prices on a constant and an indicator that is one (zero) for the ask (sale) price.

[10] Stochastic dominance means $\text{Prob}_a\{p_a \geqslant p\} \geqslant \text{Prob}_s\{p_s \geqslant p\}$. This is equivalent to $1 - F_a(p) \geqslant 1 - F_s(p)$, which gives the inequality in the text.

Hats denote estimators and N_i the number of observations. The null hypothesis is $F_j(p) - F_k(p) \leqslant 0$ over the full support and the test statistic focusses on the most unfavourable outcome for the null. If the null is true, we expect $\widehat{d}_{j,k} \leqslant 0$. If the alternative $F_j(p) > F_k(p)$ is true for at least one p, we expect $\widehat{d}_{j,k} > 0$. The procedure works as follows. First, we test whether $\widehat{d}_{a,s} \leqslant 0$. If we cannot reject, we continue and test whether we can reject $\widehat{d}_{s,a} \leqslant 0$. If we can reject, we have established strong dominance. Table 3 presents the statistics for the price distribution in the Panel A. We conclude that ask prices dominate sale prices strictly at all of the usual significance levels (0.001, 0.01, 0.05). This implies also that $\mathrm{E}[p_a] > \mathrm{E}[p_s]$, whereas the reverse does not necessarily apply. It has been observed before that $\overline{p}_a > \overline{p}_s$, but our evidence on the whole price distributions is thus much stronger.

[Table 3 about here.]

The strong dominance of the ask price distribution could be caused *either* because the house characteristics differ between the data *or* because the characteristics are valued differently or because *both* effects play a role. Table 2 shows that houses in the ask data are on average younger and have larger floor and plot areas than those in the sale data. The estimates in Figure 2 indicate that the continuous ask variables are not only larger on average, but each along their respective whole distributions.

[Figure 2 about here.]

The dominance tests in Panel A of Table 3 confirm this at the usual significance levels.[11] Note that there are relatively many (few) detached (terraced) houses

[11]The age variable is discrete and the KS results could be too conservative. We conduct also Wilcoxon-Mann-Whitney tests, which lead to the same individual and joint test outcomes as those from Table 3.

in the ask date, which could explain the dominance of the floor and plot area variables.[12] We find it difficult to explain why the houses in the ask data are dominantly younger. There is no indication of different spatial clustering in the data, as Figure 3 shows. The distributions across Berlin's 12 administrative districts look identical and their correlation is high ($\rho = 0.97$). There seems also no differential clustering in the locations of observations with coordinates.

[Figure 3 about here.]

2.3 Decomposition of price distributions

The examination of the core variables age, floor and plot area reveals strong dominance of observations in the ask over their counterparts in the sale data. This on its own could be the cause of the strong dominance of ask over sale prices. To examine this, we use that the price distribution is

$$F_{j|k}(p) \equiv \int_{\mathcal{X}_k} F_{P_j|X_j}(p|\mathbf{x}) dF_{X_k}(\mathbf{x}) \tag{2}$$

where $F_{P_i|X_i}(p|\mathbf{x})$ is the price distribution conditional on the vector $\mathbf{x}$ of characteristics and $F_{X_i}(\mathbf{x})$ is the distribution of these vectors. We note that $F_{j|j}(p) = F_j(p)$ is the unconditional price distribution and that the counterfactual price distribution $F_{j|k}(p)$ for j ($j \neq k$) results when characteristics follow the distribution $F_{X_k}(\mathbf{x})$. We can use Eq. 2 to decompose the difference between the ask and sale price distributions as

$$F_a(p) - F_s(p) = \{F_{a|a}(p) - F_{a|s}(p)\} + \{F_{a|s}(p) - F_{s|s}(p)\} \tag{3}$$

This is similar in spirit to the decomposition at the means of Blinder (1973) and Oaxaca (1973). The first term on the right-hand side of Eq. 3 reflects

[12] t-tests (not reported) show that the proportions of the three house types are different between the data at the usual significance levels.

differences due to the composition of *characteristics* in the data and the second term reflects differences in the *implicit pricing* of these characteristics. To test whether each of the two terms on the right-hand side of Eq. 3 obeys a stochastic dominance relationship, we follow the procedure proposed by Chernozhukov et al. (2013).[13] First, we estimate the distribution functions with

$$\widehat{F}_{j|k}(p) \;=\; c + \frac{1-2c}{(G-1)N_k} \sum_{n=1}^{N_k} \sum_{g=1}^{G} \mathbf{1}\Big(\mathbf{x}'_{k,n} \widehat{\boldsymbol{\beta}}_j(\tau_g) \leqslant p\Big) \tag{4}$$

The argument of the indicator function $\mathbf{1}(\cdot)$ is the characteristics bundle of observation n from data set k evaluated at implicit prices $\widehat{\boldsymbol{\beta}}_j(\tau_g)$ estimated with a quantile regression with all observations from data set j. In particular, we regress the price on third degree polynomials of the continuous core variables, and on house type, district, and yearly time dummies.[14] Second, we use KS statistics and bootstrapped p-values to test for dominance in the terms of Eq. 3.

Panel B of Table 3 shows that ask dominate sale prices as before at the usual significance levels. The slightly different KS test statistics and p-values result from the price distributions now being estimated with Eq. 4 instead of with the raw data. As to be expected from the stochastic dominance results for the continuous house characteristics, $F_{a|a}(p) - F_{a|s}(p) \leqslant 0$ at all the usual significance level; when evaluated at the same implicit prices, the characteristics in the ask data strongly dominate those in the sale data. We also find that—once the characteristics are accounted for—the pricing of characteristics in the ask data strongly dominates those in the sale data, i.e. $F_{a|s}(p) - F_{s|s}(p) \leqslant 0$,

[13] Shimizu et al. (2016) plot point estimates for Eq. 3 and test whether differences between price and valuation distributions of ask and sale data are zero (the latter test ignores that hedonic coefficients are estimated). They do not test for stochastic dominance, although their Fig. 6 indicates that it might exist for ask over sale prices.

[14] The quantiles in Eq. 4 follow $\tau_g = c + (g-1)(1-2c)/(G-1)$. We set $c = 0.01$ and $G = 200$. Trimming at c avoids estimation of tail quantiles (Koenker 2005, p. 148).

at all the usual significance levels. Prices in the ask data dominate those in the sale data both with respect to characteristics *and* implict prices.

To gain insight into the importance of the two components, we decompose the markups in Figure 1, see web-based Appendix B for details. Overall, pricing differences are fairly small compared with characteristics differences. At the means, pricing difference contribute a tenth to the markup. This corresponds to 2.4 percentage points, which is in the range of markups reported in papers that worked with matched ask and sale data (see Fn. 2). At the medians, the contribution is of similar magnitude. At lower quantiles, the contribution can be statistically zero, whereas it can be up to a fifth at higher quantiles.

Both the significant characteristic and pricing differences have the potential to bias research results when ask instead of sale data are used. We discuss next how we implement the regression applications for which we assess the magnitude of the bias.

3 Methodology and implementation

3.1 The semiparametric hedonic model

Fully parametric linear models can impose restrictions that do not accommodate the unknown data generating process. Such models impose also restrictive assumptions on preferences (Ekeland et al. 2004). Nonparametric models provide full flexibility, but can suffer from the curse of dimensionality. Semiparametric models place *some* structure on the functional form and are a good

compromise (Bontemps et al. 2008).[15] Our full geo-additive regression model is

$$p = \mathbf{z}\boldsymbol{\gamma} + f_1(AGE) + f_2(FA) + f_3(PA) + f_4(LAT, LON) + f_5(NOI) + \varepsilon \quad (5)$$

see Kammann and Wand (2003). For a given data set and observation, p is the price reported, the row vector $\mathbf{z}$ contains dummy variables for the constant, quarters, discrete house characteristics, and—depending on the specification—for the districts. The column vector $\boldsymbol{\gamma}$ contains the coefficients for these discrete variables. The continuous variables are building age (AGE), floor (FA) and plot (PA) area, longitude and latitude coordinates (LAT, LON), and the local noise level (NOI). Below, we will collect subsets of these variables in the vector $\mathbf{x}$, a deviation from the notation used above. The impact of the continuous variables on the price are considered by smooth, but unspecified, functions f_j. The error term ε represents the part of the price left unexplained by the model.

We model the nonparametric functions in Eq. 5 with regression splines

$$f_j(x) = \sum_{k=1}^{K_j} b_{jk}(x)\beta_{jk} = \mathbf{b}_j(x)\boldsymbol{\beta}_j \quad (6)$$

where $\mathbf{b}_j(x)$ is the row vector of K_j basis functions evaluated at x and $\boldsymbol{\beta}_j$ is the column vector of coefficients. The vector of coefficients determines the shape of f_j and has to be estimated. We use cubic splines as basis for the univariate functions in Eq. 5 and a thin plate spline for the function of the geo-coordinates (in which case x is a vector). Given the basis dimensions K_j, the

[15] Haupt et al. (2010) find that a log-log specification performs better out-of-sample than semi- and nonparametric specifications (Anglin and Gencay 1996, Parmeter et al. 2007). The house transcations used in these studies contain only one continuous variable. As we work with up to six continuous variables, we expect that parametric restrictions will have a detrimental effect on performance. Our robustness analysis points in this direction.

vector of all basis functions $\mathbf{b}(\mathbf{x})$ with $\mathbf{x}$ the vector of continuous characteristics of an observation, the stacked coefficient vectors $\boldsymbol{\beta}$ and $\boldsymbol{\gamma}$ are then estimated separately for each of the two data sets as

$$\left(\hat{\boldsymbol{\gamma}}, \hat{\boldsymbol{\beta}}\right) = \arg\min_{\boldsymbol{\gamma},\boldsymbol{\beta}} \left[\sum_{n=1}^{N} \{p_n - \mathbf{z}_n\boldsymbol{\gamma} - \mathbf{b}(\mathbf{x}_n)\boldsymbol{\beta}\}^2 + \sum_{j=1}^{J} \lambda_j \boldsymbol{\beta}_j' \mathbf{D}_j \boldsymbol{\beta}_j\right] \tag{7}$$

The term $\boldsymbol{\beta}_j' \mathbf{D}_j \boldsymbol{\beta}_j$ evaluates $\int \left[f_j''(x)\right]^2 dx$ and becomes large if f_j is very wiggly and small if the function is fairly straight.[16] The smoothing parameter λ_j determines the degree at which wiggliness of the estimate of f_j is penalised. To prevent excess smoothing, we select the parameters with double cross-validation criterion (DCV), see Wood (2017, pp. 260).

3.2 Willingness to pay

Once the hedonic regression is estimated, we compute for each characteristic the average marginal willingness to pay (WTP) in monetary terms and compare by how much the estimates differ between the two data sets. For a continuous characteristic, we use

$$\text{WTP}_j = \frac{1}{N} \sum_{n=1}^{N} \frac{\hat{f}_j(x_{j,n})}{\partial x_j} \exp\{\hat{p}(\mathbf{z}_n, \mathbf{x}_n)\} \tag{8}$$

where we compute the derivative numerically with finite differences and $\hat{p}(\cdot)$ is the prediction from Eq. 5. For a discrete characteristic, we use

$$\text{WTP}_j = \frac{1}{N} \sum_{n=1}^{N} \left(\exp\{\hat{\gamma}_j\} - 1\right) \exp\left\{\hat{p}(\mathbf{z}_{n,j}^0, \mathbf{x}_n)\right\} \tag{9}$$

[16] The elements of $\mathbf{D}_j$ are discussed in Wood (2017, Sec. 5.3 and 5.5). The basis dimension K_j sets an upper limit on the flexibility of f_j. We choose K_j with the informal diagnostic tests described in Wood (2017, p. 343). Since $\mathbf{z}$ contains the constant, we impose the identification restriction $\mathbf{b}_j(x)\boldsymbol{\beta}_j = 0$ for each j during estimation. The web-based Appendix C provides details.

where $\mathbf{z}^0_{n,j}$ is the discrete variable vector for observation n with the entry for variable j set to zero. To compute heteroscedasticity-robust standard errors for the WTP estimates, we use the pairs bootstrap (Freedman 1981).

3.3 Automated valuation

We use a rolling window design to split the data into estimation and validation samples. The first validation sample contains all sale observations from 2009Q1. To predict prices with the ask data, we use the observations from 2007Q2 to 2009Q1, estimate the pricing function $\hat{p}_a(\cdot)$ from Eq. 5, and assess this function at the characteristics $(\mathbf{z}, \mathbf{x})$ of the observations in the first validation sample. The choice of estimation sample considers that ask data are available instantly. For the sale data, we proceed similarly, but use observations from 2007Q1 to 2008Q4 to estimate $\hat{p}_s(\cdot)$. The lag of one quarter considers that sale data is not instantly available. The validation and estimation windows are then rolled out quarterly and predictions are computed until the last validation sample in 2015Q4 is reached. The price predictions for the final sample are computed for the ask (sale) data set based on the estimated price function 2014Q1 to 2015Q4 (2013Q4 to 2015Q3). We compute the prediction errors $e_{j,n} \equiv p_{s,n} - \hat{p}_j(\mathbf{z}_n, \mathbf{x}_n)$ from ask ($j = a$) and sale ($j = s$) data for further analysis.[17]

3.4 Price index construction and nowcasts

We fit Eq. 5 separately for the ask and sale data and use the estimated quarterly time dummy coefficients to compute quality-controlled price indices (Diewert

[17] Due to the estimation lag, $\hat{p}_s(\cdot)$ ignores the time dummy for the current quarter in $\mathbf{z}$.

et al. 2007). As the sale price index become available only with the delay, whereas the ask price index becomes available in real time, we examine the potential for nowcasting with the regression

$$\Delta I_t^s = \phi_0 + \phi_1 \Delta I_t^a + \phi_2 \Delta I_{t-1}^s + \epsilon_t \tag{10}$$

where I_t^a (I_t^s) is the ask (sale) price index for period t and the operator Δ produces either the quarter-on-quarter or the year-on-year growth rate. We estimate Eq. 10 with OLS and use robust standard errors to control for further structure in the short series. As the resulting index series have only 36 observations each, the examination will be limited.

4 Results

4.1 Willingness to pay

We examine whether WTP estimates from the ask and the sale data differ statistically and economically. It is known that estimates from hedonic regressions can suffer from omitted variable bias, but Kuminoff et al. (2010) have shown in a simulation study that spatial modelling can reduce such bias.[18] We consider two spatial models in our regressions. First, as listings may provide only coarse location information, we run regressions that model the spatial structure with district dummies as spatial fixed effects. Second, we run regressions that model location finely with the geospatial function $f_4(LAT, LON)$.[19]

[18] While omitted variable bias might pose problems for ask and sale data, it is not the source of the *comparative* differences in the application results.

[19] Hill and Scholz (2018) find that finely graded postcode spatial fixed effects can work as well as a nonparametric function of coordinates, at least in a price index application.

As some observations report no coordinates, these regressions are fitted with smaller samples.[20]

Figure 4 shows the estimates of the functions f_j in Eq. 5 for the three continuous house characteristics age, floor and plot area. The upper (lower) panel shows the estimates that result when location is modelled with spatial fixed effects (geospatial function). The noise variable is not included in these regressions.

[Figure 4 about here.]

Evidently, as the ask data has more observations, the functions are estimated more precisely. It also seems that all functions become smoother once the geospatial function is used. The functions for the areas, while not identical, seem similar whether estimated with ask or sale data. However, we expect these functions to increase monotonically, but the function for plot area estimated with ask data shows several ups and downs that counter intuition. The functions for age differ substantially. When estimated with sale data, the function falls monotonically up to an age of 60 years, where it increases and falls again. This non-monotonic shift can be explained by a premium for houses that survived WWII. When estimated with ask data, the function increases over the first ten years, which is counterintuitive, and exhibits for higher ages several ups and downs. This erratic behaviour is hard to explain other than being artefact of the ask data. Figure 5 shows contour plots of the estimated geospatial functions. Both look similar and pick up the high quality of amenities in the south-westerly neighborhoods of Berlin. Assessed at the locations of the sold houses, the correlation between the two estimated functions is high ($\rho = 0.94$).

[20] 13% (2%) of the ask (sale) observations have no coordinates, see bottom row of Table 4.

[Figure 5 about here.]

Table 4 presents in columns (1)-(4) the estimated WTPs for the core house characteristics.

[Table 4 about here.]

We note that the standard errors for the WTPs are smaller when the ask data is used (result of the larger sample sizes) and that the errors become smaller when geospatial functions are used in all instances but one (age in (3) and (4)). In case of the ask data, only the use of the geospatial function leads to an intuitive negative WTP for age, although it remains insignificant at the usual levels (2). The counterintutive age function from Figure 4 shows up here. When estimated with the sale data, the estimated WTP for age is both times negative at the usual significance levels, irrespective of the spatial modelling approach. Regarding the house types, terraced is the reference type and the WTPs for the other two types aligns with intuition only when the geospatial function is included in the regressions. As to be expected from Kuminoff et al. (2010), it seems that the geospatial function deals with omitted variable bias, as it leads to more intuitive WTP estimates.

In the examination so far, we have used only those variables that are in the ask as well as in the sale data. The sale data, however, is of higher quality and contains additional variables that have not been used yet, see the second part of Panel A in Table 2. Table 4 (5) gives the WTP estimates when we no longer omit these variables. The WTP estimates for the formerly omitted variables seem sensible. The regression continues to use the geospatial function to control for other omitted variables. As (5) is our most complete model, we use its estimates as a benchmark. Comparison of (4) with the benchmark

shows that the formally omitted variables have only a fairly small effect on the estimated WTPs for the core variables. In all but one case, the point estimates are about 1.1 times the benchmark. The exception is the WTP for a detached house, which is 0.7 times the benchmark. Things look different when we compare (2) with the benchmark. In all but one case, the estimates from the ask data are about 1.6 times the benchmark. Inflated WTP estimates can be expected given the ask data's dominant characteristic and implicit price distributions. The only exception is the WTP for age, which is only 0.1 times the benchmark. This reflects the counterintuitive age function that results for the ask data.

Finally, we examine what such deviations imply for benefit assessment. Figure 6 plots nightly noise levels in Berlin for 2012, the darker the shading, the higher the noise. For example, the dark strip from left to right in the upper part corresponds to the noise emitted by Otto Lilienthal airport in Tegel; the noise emitted by inner-city motorways is also visible.

[Figure 6 about here.]

Estimated with sale data, the function f_5 in Figure 6 stays reasonably flat at zero up to a level of 50db—the level of noise in a quiet suburban neighbourhood—and becomes increasingly negative at higher noise levels. Estimated with ask data, however, the function puts a doubtful premium on silence—30db corresponds to rustling leaves—and exhibits non-monotonic behaviour. The estimated WTPs that result from these two functions are reported in Table 5.

[Table 5 about here.]

The estimates are negative—noise is a disamenity—and significantly different at the 0.05 level (p-value is 0.03).[21] The difference between the point estimates seems economically small, which ignores that noise usually affects many households. The difference becomes EUR533,800 per km^2 after we factor in that in Berlin the average density is 1,700 households per km^2. Obviously, a policy maker who uses the cost-benefit criterion to decide on a night flight ban may come to the wrong decision when the benefit is estimated with ask data.

4.2 Automated valuation

Table 6 presents performance measures for the out-of-sample predictions for regressions fitted separately to ask and sale data. The specification for the ask data is (2) and the sale data is (5) from Table 4.

[Table 6 about here.]

The prediction errors $e_{a,n}$ do not perform as well as the errors $e_{s,n}$. The threshold proportions are less than 0.9 times of those for the latter and the MSE is 1.5 times as large. The negative bias of the errors $e_{a,n}$ is not surprising given that the distribution of implicit prices in the ask dominates those in the sale data.[22] However, the errors $e_{s,n}$ show also bias, which reflects the quarterly lag of the data used for estimation. The absolute magnitude of the bias is close to the quarter-on-quarter growth rate of quality-controlled sale prices, see Figure 7. One could suspect that the differential performance of

[21] The point estimates correspond to a reduction of the average ask (sale) price by 0.4% (0.3%). Winke (2017, p. 1284) finds a reduction of 1.7% and reports that previous studies found reductions between 0.1% to 3.6%.

[22] $\hat{p}_a(\mathbf{z}, \mathbf{x})$ are effectively imputed ask prices and the bias of 3.5% falls well within the range of markups observed in studies that use matched data, see Fn. 2.

the errors comes mainly from the tendency of ask prices to be larger than sale prices. However, the bias is fairly unimportant for the MSE of the two sets of errors. The inferior performance of the $e_{a,n}$ errors comes mainly from their high variance, the result of fewer variables that can be used and their tilted pricing differences.

4.3 Price indices and nowcasts

Figure 7 shows the quality-controlled ask and sale price indices, the former (latter) based on specification (2) ((5)) from Table 4. The two indices have overall the same upward trend, but the trend masks some differences that are visible in the quarter-on-quarter growth rates. As both indices control for observed characteristics, these difference are due to differential valuations, wider coverage of characteristics and a random element.

[Figure 7 about here.]

Table 7 assesses the strength of the relation between the two indices and gives results for the price index growth rate regression from Eq. 10. As (1) and (4) show, the contemporaneous rates of the two indices are positively correlated, but the relation is stronger for the year-on-year than the quarter-on-quarter growth rates ($\hat{\rho} = 0.77$ versus $\hat{\rho} = 0.41$). The former are usually less volatile, which makes it more likely to detect a relationship—if it exists—in small samples like ours. As (2) and (3) show for quarter-on-quarter growth rates, ΔI^s_{t-1} provides no information for the current growth rate. For a nowcast, it is best to use $\Delta \hat{I}^s_t = 0.006 + 0.532 \Delta I^a_t$. For the year-on-year growth rates, the lag has on its own already high predictive power, see (5). But even in this instance, the inclusion of the current growth rate of the ask price index improves

explanatory power, see $\bar{R}^2$ in (6).[23]

[Table 7 about here.]

Taken together, the examination provides some evidence that an ask price index can lead to better nowcasts. Longer time series are needed to obtain clearer in-sample results and to extend the examination to out-of-sample nowcasts.

4.4 Robustness checks

We conducted several robustness checks to assess the sensitivity of our results to methodological choices, see the web-based Appendix D for details. First, we examined the sensitivity of the decomposition of the markups with respect to the specification of the conditional price distributions. The examination led to similar results to the ones reported here. Second, we used a function instead of a simple ratio to convert interior into exterior floor area. The function is flexible, considers also building age, and is estimated from sale data that have information on both area variables. The resulting exterior floor area is highly correlated with the conversion used here. Third, we implemented the regression applications with parametric models. The estimates in Figure 4 might suggest that commonly applied parametric specifications such as polynomials could produce similar results as those reported here. This is indeed the case, but the predictive accuracy of the parametric models is throughout lower than those for the semiparametric models used here. We see this as justification to present here the results from the semiparametric models.

[23] The estimated coefficient ϕ_1, however, has a p-value of 0.08 and is not significant at our usual significance levels.

5 Conclusion

Unlike transaction data, information from listings web-platforms is timely and abundant. It could thus become a novel data source for research in real estate and urban economics. In this paper, we have examined the value of listings information for three common research applications: (1) quantification of the relationship between house prices and characteristics, (2) prediction of market values, and (3) measurement of house price dynamics.

The results of our study are as follows. First, we find that the substantial differences in unconditional ask and sale price distributions are driven by differential characteristics compositions. We also establish that valuation difference play a significant role as well. While the valuation differences are not relevant for lower quantiles of the distributions, they become relevant from at least for the median onwards. While the composition result points to the possibility that ask data can used as substitute for transaction data as long as one controls appropriately for observed characteristics, the pricing result points in the direction that the substitutional potential of ask data is doubtful. This motivated the further steps of our examination. Second, we find that the estimated WTPs for house characteristics can, at times, differ quite substantially when estimated using ask instead of sale data. Third, we find that these differences lead also to inferior sale price predictions when the regression model is fitted to ask prices. The lower accuracy of these predictions is not only the result of the bias, that asking prices exhibit even after controlling for house characteristics, but also a larger variance of the predictions. Fourth, we find that constant quality ask and sale price indices paint a roughly similar picture of the general price trend. Moreover, time series regressions indicate that an ask price index might be useful for nowcasting a sale price index. Though, the

short time-series dimension of our data restricts our ability to draw conclusive inferences. In sum, our results show that listings information must be used with caution in empirical research.

Acknowledgements

We thank Robert Hill, Helmut Lütkepohl, Bryan MacGregor, Aleksandar Petreski, Verity Watson, and audiences at University of Aberdeen Business School, KTH Royal Institute of Technology Stockholm, and AREUEA 2019 International Conference for helpful comments. The usual disclaimer applies. Kolbe and Werwatz thank the DFG Research Unit 2569: Agricultural Land Markets - Efficiency and Regulation for financial support.

References

Anglin, P. M. and Gencay, R.: 1996, Semiparametric estimation of a hedonic price function, *Journal of Applied Econometrics* **11**, 633–648.

Barrett, G. F. and Donald, S. G.: 2003, Consistent tests for stochastic dominance, *Econometrica* **71**, 71–104.

Bauer, T. K., Braun, S. T. and Kvasnicka, M.: 2017, Nuclear power plant closures and local housing values: Evidence from Fukushima and the German housing market, *Journal of Urban Economics* **99**, 94–106.

Bauer, T. K., Feuerschütte, S., Kiefer, M., an de Meulen, P., Micheli, M., Schmidt, T. and Wilke, L.-H.: 2013, Ein hedonischer Immobilienindex auf Basis von Internetdaten: 2007-2011, *AStA Wirtschafts- und Sozialstatistisches Archiv* **7**, 5–30.

Blinder, A. S.: 1973, Age discrimination: Reduced form and structural estimates, *Journal of Human Resources* **8**, 436–455.

Bontemps, C., Simioni, M. and Surry, Y.: 2008, Semiparametric hedonic price models: Assessing the effects of agricultural nonpoint source pollution, *Journal of Applied Econometrics* **23**, 825–842.

Carrillo, P. E.: 2012, An empirical stationary equilibrium search model of the housing market, *International Economic Review* **53**, 203–234.

Chernozhukov, V., Fernandez-Val, I. and Mellie, B.: 2013, Inference on counterfactual distributions, *Econometrica* **81**, 2205–2268.

Diewert, W. E., Heravi, S. and Silver, M.: 2007, Hedonic imputation versus time dummy hedonic indexes, *in* W. E. Diewert, J. S. Greenless and C. R.

Hulten (eds), *Price Index Concepts and Measurement*, Studies in Income Wealth, University of Chicago Press, Chicago, chapter 4, pp. 161–202.

Ekeland, I., Heckman, J. J. and Nesheim, L.: 2004, Identification and estimation of hedonic models, *Journal of Political Economy* **112**, S60–S109.

Freedman, D. A.: 1981, Bootstrapping regression models, *Annals of Statistics* **9**, 1218–1228.

Goodman, J. L. and Ittner, J. B.: 1992, The accuracy of home owner's estimates of house value, *Journal of Housing Economics* **2**, 339–357.

Gutachterausschuss für Grundstückswerte: 2011, *Bericht über den Berliner Grundstücksmarkt 2010/11*, Senatsverwaltung für Stadtentwicklung, Berlin. Kulturbuch-Verlag Berlin.

Haupt, H., Schnurbus, J. and Tschernig, R.: 2010, On nonparametric estimation of a hedonic price function, *Journal of Applied Econometrics* **25**, 894–901.

Haurin, D. R., Haurin, J. L., Nadauld, T. and Sanders, A.: 2010, List prices, sale prices and marketing time: An application to U.S. housing markets, *Real Estate Economics* **38**, 659–685.

Hill, R. J. and Scholz, M.: 2018, Can geospatial data improve house price indexes? A hedonic imputation approach with splines, *Review of Income and Wealth* **64**, 737–756.

Horowitz, J.: 1992, The role of the list price in housing markets: Theory and an econometric model, *Journal of Applied Econometrics* **7**, 115–129.

Immobilien Scout 24: 2018, AGB für die Nutzung der über die Website www.immobilienscout24.de zugänglichen Services, *Web document*, Immobilien Scout 24 GmbH, Berlin. Accessed on 20.09.2018.

Kammann, E. E. and Wand, M. P.: 2003, Geoadditive models, *Journal of the Royal Statistical Society: Series C (Applied Statistics)* **52**, 1–18.

Kiel, K. A. and Zabel, J. E.: 1999, The accuracy of owner-provided house values: The 1978-1991 American Housing Survey, *Real Estate Economics* **27**, 263–298.

Koenker, R.: 2005, *Quantile Regression*, Econometric Society Monographs, Cambridge University Press, Cambridge.

Kuminoff, N. V., Parmeter, C. F. and Pope, J. C.: 2010, Which hedonic models can we trust to recover the marginal willingness to pay for environmental amenities?, *Journal of Environmental Economics and Management* **60**, 145–160.

Oaxaca, R.: 1973, Male-female wage differentials in urban labor markets, *International Economic Review* **14**, 693–709.

Parmeter, C. F., Henderson, D. J. and Kumbhakar, S. C.: 2007, Nonparametric estimation of a hedonic price function, *Journal of Applied Econometrics* **22**, 695–699.

RICS: 2017, The future of valuations, *Insight paper*, Royal Institution of Chartered Surveyors, London.

Scout24: 2017, Capital markets day Berlin November 2017, *Presentation slides*, Scout24 AG, München.

Shimizu, C., Nishimura, K. G. and Watanabe, T.: 2016, House prices at different stages of the buying/selling process, *Regional Science and Urban Economics* **59**, 37–53.

Silverman, B. W.: 1986, *Density Estimation for Statistics and Data Analysis*, Monographs on Statistics and Applied Probability, Chapman and Hall, London.

Taylor, L. O.: 2017, Hedonics, *in* P. A. Champ, K. J. Boyle and T. C. Brown (eds), *A primer on nonmarket valuation*, second edn, Vol. 13 of *The Economics of Non-Market Goods and Resources*, Springer, Dordrecht, chapter 7, pp. 235–292.

Winke, T.: 2017, The impact of aircraft noise on apartment prices: A differences-in-differences hedonic approach for Frankfurt, Germany, *Journal of Economic Geography* **17**, 1283–1300.

Wood, S. N.: 2017, *Generalized additive models. An introduction with R*, Texts in Statistical Sience, 2 edn, CRC Press, Boca Raton.

Table 1: Effects of data cleaning. Gives the number of observations in the original data and after each step of the data cleaning procedure. Missing values refers to observations that lack entries for some of the core variables. Old refers to a house that has a building which is older than 100 years at the date of transaction or the last listing day. Bounds for plot area, floor area, and transaction price per floor area come from annual reports of the GAA.

	GAA	IS24
Original data	17,650	144,274
After removing		
missing values	17,650	106,193
old or under construction	15,242	83,952
outwith bounds	12,524	68,070

Table 2: Summary statistics for sale and ask data sets. Panel A gives also information for variables other than the core variables. Prices are in 000' Euros. Age of building at the date of sale or end of listing, respectively. Floor and plot area are in sqm.

	Mean	Std. Dev.	Min	Max
Panel A. Sale data ($N = 12,524$)				
Price (P_s)	250.50	129.07	40.00	1450.00
ln Price (p_s)	12.33	0.45	10.60	14.19
Age	44.94	30.37	0.00	100.00
Floor area	144.78	52.79	41.00	642.00
Plot area	544.96	267.04	112.00	1500.00
Detached	0.56			
Semi-detached	0.28			
Terraced house	0.17			
Listed building	0.05			
Prefabricated	0.10			
Converted attic	0.53			
Swimming pool	0.01			
Flat roof	0.15			
No basement	0.18			
Backland development	0.17			
Lake/River access	0.01			
Condition of building				
poor	0.04			
average	0.62			
good	0.34			
Neighborhood amenity rating				
poor	0.30			
average	0.51			
good	0.18			
excellent	0.01			
Panel B. Ask data ($N = 68,070$)				
Price (P_a)	320.96	189.96	45.00	2020.00
ln Price (p_a)	12.56	0.46	10.71	14.52
Age	32.02	27.92	0.00	100.00
Floor area	187.21	74.93	50.00	650.00
Plot area	583.49	267.37	100.00	1500.00
Detached	0.68			
Semi-detached	0.23			
Terraced house	0.09			

Table 3: Stochastic dominance tests. The statistic $\widehat{d}_{j,k}$ $(j \neq k)$ tests the null hypothesis that distribution j dominates distribution k weakly. In Panel A, the variable tested for is $-AGE$, $\widehat{d}_{j,k}$ is the signed two sample KS test statistic, defined in Eq. 1. The p-values for the null are calculated as $\exp\{-2(\widehat{d}_{j,k})^2\}$, see Barrett and Donald (2003, p.78). In Panel B, $\widehat{d}_{j,k}$ is the KS maximal t-statistic as defined in Chernozhukov et al. (2013, p. 2222). The standard error of $\widehat{d}_{j,k}$ is calculated using the bootstrap interquartile range of the KS test statistic. The p-values for the null are calculated as $R^{-1}\sum_r 1(\widehat{d}^*_{j,k,r} > \widehat{d}_{j,k})$, where $\widehat{d}^*_{j,k,r}$ is the $r'th$ bootstrap test statistic, see Barrett and Donald (2003, p. 82). The number of bootstrap replications is 200.

	$\widehat{d}_{a,s}$	P-value	$\widehat{d}_{s,a}$	P-value
Panel A. Marginal distributions				
Price	0.000	1.000	20.577	0.000
Age	0.286	0.849	24.689	0.000
Floor area	0.000	1.000	34.102	0.000
Plot area	0.027	0.999	9.007	0.000
Panel B. Price decomposition				
Price	0.000	0.915	48.777	0.000
Characteristics	0.000	0.860	53.132	0.000
Implicit prices	1.076	0.705	10.978	0.000

Table 4: Willingness to pay for house characteristics. Reports WTP estimates and regression diagnostics for penalized least squares estimates of Eq. 5. WTPs for continuous (categorical) house characteristics are computed with Eq. 8 (Eq. 9). Standard errors are computed using the pairs bootstrap with 200 replications. $\bar{R}^2$ is the adjusted coefficient of determination. DCV is the double cross-validation score. Significant at ***0.001 level, **0.01 level, *0.05 level.

	Ask data				Sale data					
	(1)		(2)		(3)		(4)		(5)	
	WTP	Std. Err.	WTP	Std. Err.	WTP	Std. Err.	WTP	Std. Err.	WTP	Std. Err.
Age	243.99***	47.25	-78.96	43.88	-1173.73***	103.73	-1461.39***	106.05	-1262.78***	99.84
Floor area	1419.07***	10.39	1293.92***	7.41	1047.99***	17.46	906.73***	14.82	828.25***	14.96
Plot area	1465.18***	7.95	1380.25***	7.46	1152.82***	18.14	1011.70***	14.60	938.46***	14.74
Detached	6881.76***	1424.74	17771.60***	1263.12	-10291.94**	3284.71	5910.15*	2899.58	7905.92**	2957.12
Semi-detached	-1042.73	1205.12	5074.97***	1011.21	-7918.00**	2468.32	4130.49	2198.43	3882.04	2187.97
Listed									9267.35*	3808.55
Prefabricated									-6873.87***	1776.82
Converted attic									3734.25**	1281.93
Swimming pool									15163.56**	5268.21
Flat roof									-4874.36**	1579.35
No basement									-21796.71***	1568.11
Backland develop.									238.60	1463.06
Waterfront									59740.22***	6510.13
Poor condition									-54231.26***	2462.78
Good condition									28998.44***	1659.65
Poor amenities									-6233.90***	1702.58
Good amenities									16535.98***	2856.80
Excel. amenities									37159.69***	9963.76
$f_4(LAT, LON)$		No		Yes		No		Yes		Yes
District dummies		Yes		No		Yes		No		No
Time dummies		Yes		Yes		Yes		Yes		Yes
DCV		0.048		0.039		0 068		0.056		0.050
$\bar{R}^2$		0.775		0.816		0.665		0.731		0.760
N		68,070		59,502		12,524		12,218		12,218

Table 5: Willingness to pay for noise levels. Reports WTP estimates and regression diagnostics for penalized least squares estimates of Eq. 5. Specification for ask (sale) data identical to (2) ((5)) from Table 4 plus noise function $f_5(NOI)$. WTPs are computed with Eq. 8. Standard errors are computed using the pairs bootstrap. Number of bootstrap replications is 200. $\bar{R}^2$ is the adjusted coefficient of determination. DCV is the double cross-validation score. Significant at ***0.001 level, **0.01 level, *0.05 level.

	Ask data		Sale data	
	WTP	Std. Err.	WTP	Std. Err.
Noise level	-1141.27***	79.73	-827.79***	122.18
DCV		0.039		0.050
$\bar{R}^2$		0.819		0.762
N		59,502		12,218

Table 6: Assessment of prediction errors. Shows performance statistics for 9,152 out-of-sample prediction errors. ±10% (±25%) reports the proportion of errors which are in absolute terms no larger than 10% (25%).

Data	MSE	Bias	Var.	Med.	MAE	±10%	±25%
Ask	0.077	-0.035	0.076	-0.021	0.214	0.304	0.671
Sale	0.051	0.014	0.051	0.027	0.176	0.364	0.751

Table 7: Nowcast regressions. Reports estimates of Eq. 10. Asymptotic p-value is for the two-sided null hypothesis that the respective coefficient is zero. t-statistics are computed using Newey-West standard errors, at most four lags. $\bar{R}^2$ is the adjusted coefficient of determination.

	Quarter-on-quarter								Year-on year							
	(1)		(2)		(3)		(4)		(5)		(6)		(7)		(8)	
	Coeff.	P-val.	Coeff.	P-val.	Coeff.	P-val.	Coeff.	P-val.	Coeff.	P-val.	Coeff.	P-val.	Coeff.	P-val.	Coeff.	P-val.
ΔI_t^a	0.532	0.001	0.498	0.003	0.520	0.007	0.485	0.017	0.843	0.000	0.691	0.000	0.327	0.077	0.397	0.040
ΔI_{t-1}^a			-0.161	0.277			-0.181	0.342			0.122	0.362			-0.163	0.373
ΔI_{t-1}^s					-0.049	0.791	0.035	0.881					0.526	0.009	0.592	0.014
Constant	0.006	0.013	0.008	0.006	0.007	0.006	0.008	0.005	0.015	0.015	0.018	0.010	0.013	0.010	0.013	0.017
$\bar{R}^2$		0.172		0.132		0.115		0.104		0.597		0.549		0.678		0.675
N		35		34		34		34		32		31		31		31

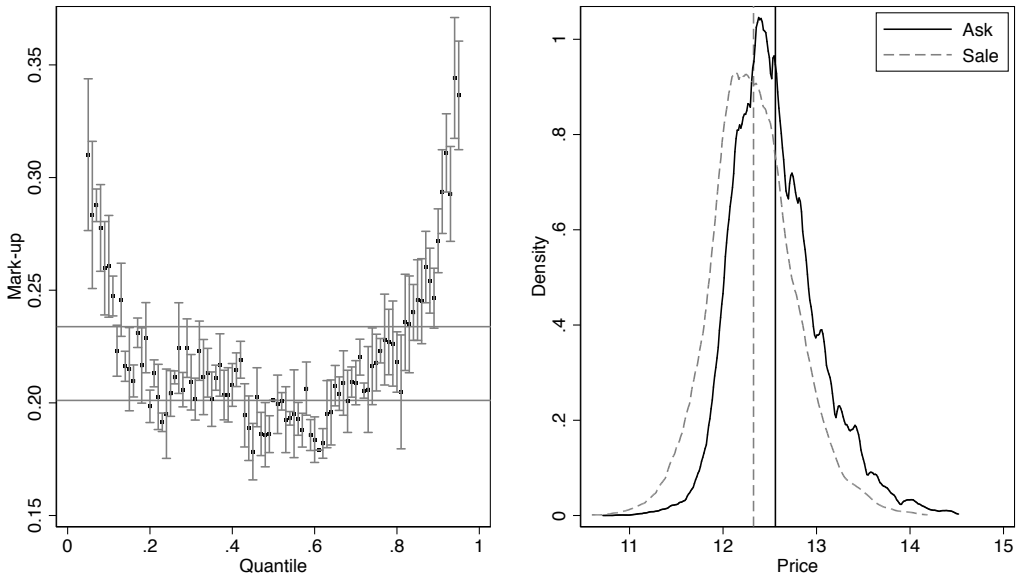

Figure 1: Distributions of ask and sale prices. Left panel shows markups of ask over sales prices at different quantiles. Horizontal lines give markups at the medians (20.1%) and means (23.4%). Whiskers give pointwise confidence intervals at the 0.95 level. Right panel shows kernel density estimates of the distributions of prices from ask data set (solid black) and from sale data set (dashed gray). Bandwidths are chosen with Silverman's (1986) rule-of-thumb. Vertical lines are the respective means of the ask and sale prices.

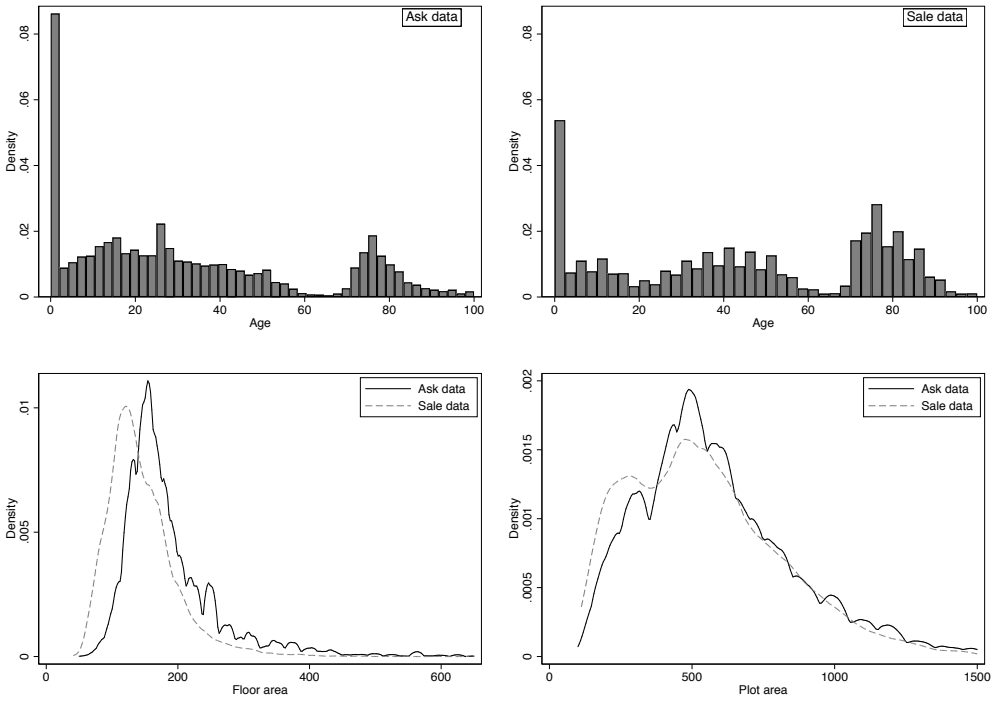

Figure 2: Distributions of house characteristics. Top panel shows histograms of building age for the observations in the ask and the sale data, respectively. Lower panel shows kernel density estimates of floor area (left) and plot area (right) for the observations in the data. Bandwidths are chosen with Silverman's (1986) rule-of-thumb.

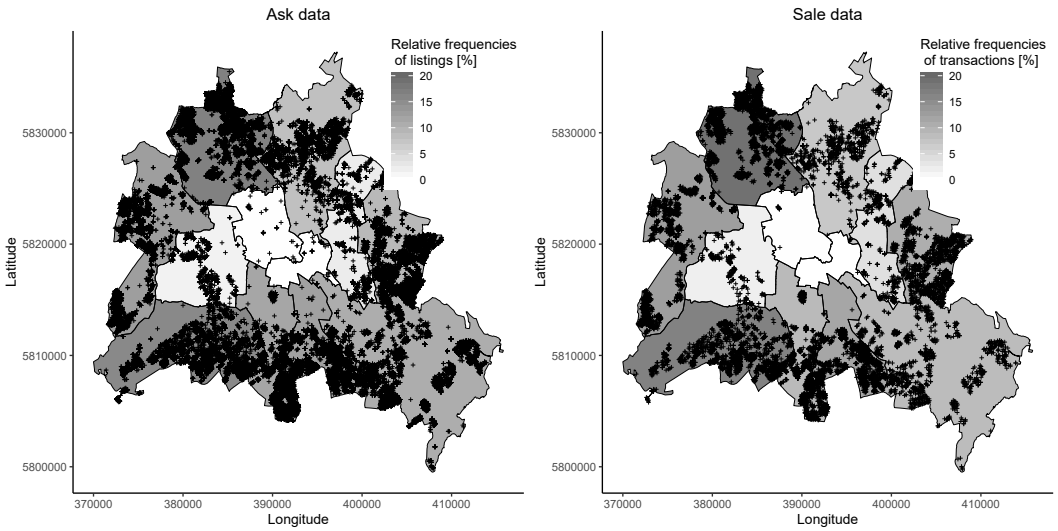

Figure 3: Spatial distribution of ask and sale observations. Shows the relative frequency of observations in the ask and sales data across Berlin's 12 administrative districts. Crosses give locations of the 59,502 (12,218) individual observations in the ask (sale) data for which we have coordinates.

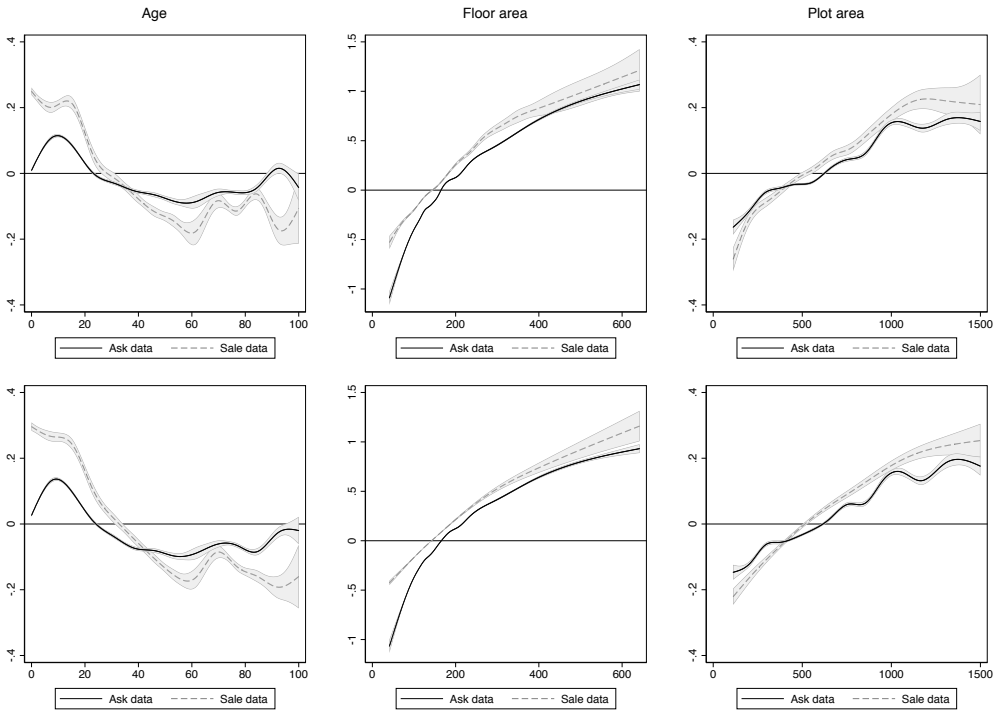

Figure 4: Estimates of components of semiparametric regression model. Upper panel shows estimates of $f_1(AGE)$, $f_2(FA)$, and $f_3(PA)$ for regression in Eq. 5 that uses spatial fixed effects. Lower panel shows estimates for the same functions, but controls with the geospatial function $f_4(LAT, LON)$. The corresponding estimated functions are shown in Figure 5. Noise variable NOI is not included in the regressions. Functions are normalized to have a mean of zero. Shaded areas are pointwise confidence intervals at the 0.95 level, computed using heteroscedasticity robust standard errors.

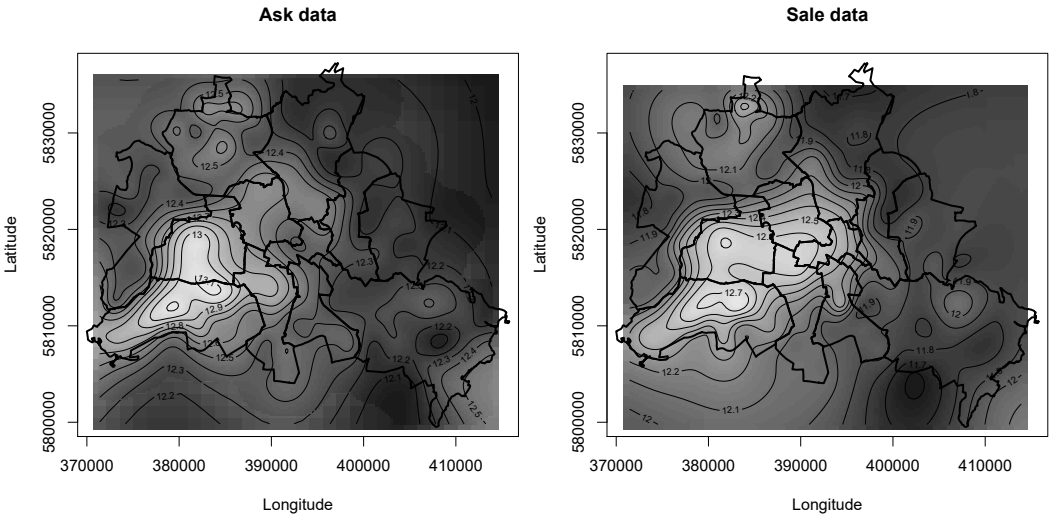

Figure 5: Location value surface. Shows contour plot of estimated geospatial functions $\widehat{f_4}(LAT, LON)$ from ask (right panel) and sale data (right panel). Location value surface is evaluated at median values of continuous house characteristics and modal values of discrete house characteristics.

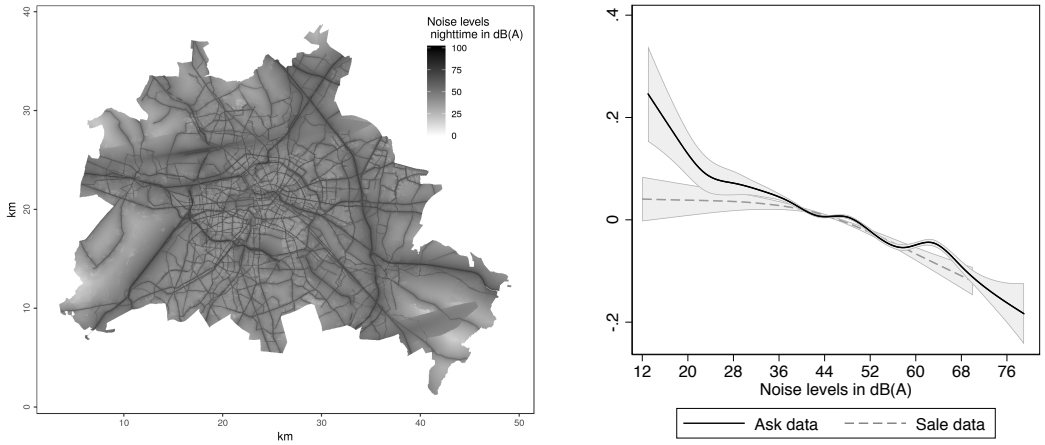

Figure 6: Noise levels and WTP. Left panel shows noise levels in decibel (dB(A)) in Berlin at night for 2012. The data comes from Berlin's Senate Department for Urban Development and Housing. Right panel shows estimates of the function f_5 from specifications (2) and (5) in Table 4, when including f_5. Shaded areas are 0.95 pointwise confidence intervals, computed using heteroscedasticity robust standard errors.

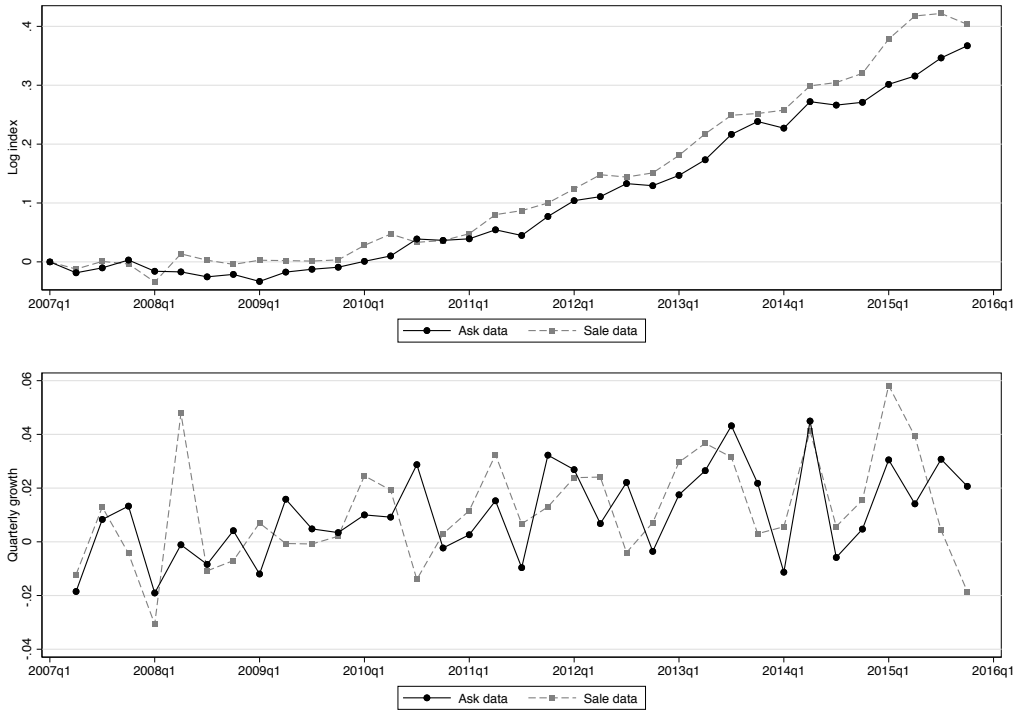

Figure 7: Quarterly quality-adjusted house price indices. Upper panel shows ask and sale price indices for Berlin 2007Q1-2015Q4, lower panel shows quarter-on-quarter growth rates. The growth rate of the ask (sale) price index is 1.0% (1.2%) with volatility of 1.7% (2.0%).

Web-based appendix for "How useful is listings data for research?"

Jens Kolbe, Rainer Schulz, Martin Wersing,

and Axel Werwatz*

August 23, 2019

*Kolbe and Werwatz: Institut für Volkswirtschaftslehre und Wirtschaftsrecht, Technische Universität Berlin, Straße des 17. Juni 135, 10623 Berlin, Germany; jkolbe@tu-berlin.de and axel.werwatz@tu-berlin.de. Schulz and Wersing: University of Aberdeen Business School, Edward Wright Building, Dunbar Street, Aberdeen AB24 3QY, United Kingdom; r.schulz@abdn.ac.uk and martin.wersing@abdn.ac.uk.

A Preparation of data sets

A.1 Variable bounds

The GAA annual market reports provide tables for houses transacted in a given year with minimum, average, and maximum of: lot area, floor area, ratio of price to floor area (Gutachterausschuss für Grundstückswerte 2011, 2012, 2013, 2014, 2015). The tables are provided separately for detached, semi-detached, and terraced houses. Within the tables, the information detailed further by the part of the city the house is located in (East, West) and the period during which the house was constructed (before 1949, since 1949). We use the information for the years 2010-2014 and select for each house type, location, and vintage, the minimum of the minima and the maximum of the maxima for each of the three variables. Table A1 summarizes the variable bounds.

[Table A1 about here.]

B Decomposition of markups

B.1 Decomposition at quantiles

To examine the contribution of house characteristics and implicit prices to the ask price markups (see Fig. 1), we assess the quantile decomposition

$$Q_a(\tau) - Q_s(\tau) = \left\{Q_{a|a}(\tau) - Q_{a|s}(\tau)\right\} + \left\{Q_{a|s}(\tau) - Q_{s|s}(\tau)\right\} \tag{B1}$$

where $Q_{j|k}(\tau)$ is the τth quantile of the distribution of the price in data set j, given the characteristics in data set k. We obtain the quantiles by inverting the

estimated distribution functions $\hat{F}_{j|k}(p)$ (see Eq. 4). In order to conduct inference, we compute pointwise and uniform confidence bands using the bootstrap procedure described in Chernozhukov et al. (2013, Algorithms 2 & 3). The intervals are constructed at the 0.95 level and take the estimation uncertainty about $\hat{F}_{j|k}(p)$ into account.

B.2 Decomposition at means

The Blinder-Oaxaca decomposition is

$$\mathrm{E}[p_a] - \mathrm{E}[p_s] = \{\bar{\mathbf{x}}_a - \bar{\mathbf{x}}_s\}\,\boldsymbol{\beta}_a + \{\boldsymbol{\beta}_a - \boldsymbol{\beta}_s\}\,\bar{\mathbf{x}}_s \tag{B2}$$

where the vector $\boldsymbol{\beta}_k$ collects the implicit prices in data set j and the vector $\bar{\mathbf{x}}_j$ collects the mean values of house characteristics in data set k. We estimate the implicit prices by running a linear regression of the log price on the continuous core variables, and house type, district and yearly time dummies. We include the continuous variables as third-degree polynomials, which is analogous to the quantile regressions used to estimate Eq. 4. In order to conduct inference, we compute standard errors according to the suggestions in Jann (2008). To allow for heteroscedastic error terms, we estimate the covariance matrix of the implicit prices using the Huber/White estimator.

B.3 Results

Table B1 presents the decomposition of the ask price markups.

[Table B1 about here.]

In Panel A, the estimated markup at the median is slightly lower than the markup reported in Fig. 1. This is because the markups are estimated from

Eq. 4, rather than not the empirical distribution functions (EDFs) of ask and sale prices. The upper-left (right) panel of Figure B1 shows a Q-Q plot for the ask (sale) price distribution estimated from Eq. 4 and the EDF. For, both, ask and sale prices the distribution $\hat{F}_j(p)$ resembles closely the corresponding EDF. This is reflected in the estimated markups, which exhibit a similar U-shaped pattern as the markups in Fig. 1.

[Figure B1 about here.]

Panel B (C) of Table B1 shows the estimated contributions of characteristics (implict prices). As indicated by the Q-Q plots in the lower panel of Figure B1, the contribution of characteristics differences in the two data sets accounts for the greater part of the markups. Nonetheless, according to the pointwise *and* uniform confidence bands, implicit price contribute also to the markups for $\tau > 0.3$, at least, at the 0.05 level.

C Semiparametric regression model

C.1 Choice of smoothing parameters

We select $\boldsymbol{\lambda} = \{\lambda_1, \ldots, \lambda_5\}$ as by minimizing the DCV score

$$\hat{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda}} \frac{N \sum_{i=1}^{N} \left(p_i - \hat{p}_i(\boldsymbol{\lambda})\right)^2}{\left\{N - 1.5\mathrm{tr}\left(\mathbf{H}(\boldsymbol{\lambda})\right)\right\}^2} \tag{C3}$$

where $\hat{p}_i(\boldsymbol{\lambda})$ is the predicted price for a given set of $\boldsymbol{\lambda}$ values and $\boldsymbol{H} = \left(\mathbf{X}^\mathsf{T}\mathbf{X} + \mathbf{S}_{\boldsymbol{\lambda}}\right)^{-1}\mathbf{X}^\mathsf{T}$ is the hat matrix of the penalized least squares estimator of Eq. 7. Here, $\mathbf{X}$ is the design matrix collecting the basis functions for the continuous and all dummy variables. The matrix $\mathbf{S}_{\boldsymbol{\lambda}}$ collects the penalty

terms, see Wood (2017, pp. 249-50). DCV is a consistent estimator of the mean squared error of the regression model and DCV prevents excess smoothing (Wood 2017, pp. 260-61).

C.2 Basis dimensions for splines

The exact size of the basis dimensions $\mathbf{K} = \{K_1, \ldots, K_5\}$ is not as critical as the smoothing parameters $\boldsymbol{\lambda}$, as they only set an upper bound on the flexibility of the functions $f_j(\cdot)$. We use the informal diagnostic tests of Wood (2017, pp. 343), as well as visual inspection of the estimated functions, and set $k_1 = k_2 = k_3 = k_5 = 14$ for the univariate functions and $k_4 = 150$ for the geospatial function. Residual diagnostics provide evidence that these values are sufficiently large to provide adequate flexibility, see Section C.2.1 below.

C.2.1 Residual diagnostics

The upper left panel of Figure C1 (C2) shows a scatter plot of the fitted values against the residuals from the ask price regression (1) ((2)) in Tab. 4. The residuals are well behaved; no obvious pattern remains after fitting the data.

[Figure C1 about here.]

[Figure C2 about here.]

The three remaining panels of Figure C1 (C2) show nonparametric functions of the building age, floor area, and, respectively, plot area fitted to the regression residuals. We model each function using a cubic spline with basis dimension $\tilde{K}_j = 2 \cdot K_j$ to check if a higher basis dimensions reveals additional structure in the data. We fit the regressions using penalized least squares.

[Figure C3 about here.]

[Figure C4 about here.]

For specification (1) ((2)), each of fitted functions shows an erratic behavior. This could indicate that the basis dimensions K_j are not sufficiently large. However, even after quadrupling K_j the erratic behavior persist, see Figures C3 (C4). Moreover, the fitted functions have very similar shapes (not reported) to those shown in Fig. 4. We attribute the erratic behavior of the smoothed residuals to an artefact of the ask data, rather than a misspecification of the splines.

Figure C5 (C6 , C7) show the same diagnostic plots for the sale price regressions from Tab. 4. There is no structure left in the fitted sale data that could be captured by more flexible functions. In fact, the smoothed residuals are completely flat as one would expect in the absence of model misspecifications.

[Figure C5 about here.]

[Figure C6 about here.]

[Figure C7 about here.]

D Robustness checks

D.1 Decomposition of price distributions

To examine how sensitive the markup decomposition is with respect to the model specification of Eq. 4, we re-estimated the price distributions using log-

linear hedonic (quantile) regressions. Log-linear functional forms are frequently employed in empirical research,[1] since they are often less prone to omitted variables bias than more flexible specifications. This is particularly true when one can only control for unobserved location effects crudely, as we do by using district dummies, see Cropper et al. (1988) and Kuminoff et al. (2010).

Figure D1 shows a Q-Q plot for the estimated and empirical ask (sale) price distribution in its upper-left (right) panel. The estimated distribution $\hat{F}_j(p)$ resemble closely the corresponding EDF. Moreover, the Q-Q plots are comparable to those in Figure B1. Both, the log-linear and polynomial quantile regressions produce distribution estimates, $\hat{F}_j(p)$, that well approximate the corresponding EDFs.

[Figure D1 about here.]

Table D1 reports the decomposition based on the log-linear quantile regressions.

[Table D1 about here.]

The estimated markups and contributions of characteristics and implicit are qualitatively similar to those reported in Table B1; the markups are sizable and, both, characteristics and implicit prices contribute to them at *all* quantiles. Relative to Table B1, the log-linear quantile regressions produce larger implicit price estimates for the ask than the sale data, particularly at lower quantiles; see also the lower-right panel of Figure D1. Given the nonlinearities in the hedonic price function at the mean (see Fig. 4), we prefer to allow for

[1] Shimizu et al. (2016), for example, use a log-linear specification in their analysis of ask and sale data distributions.

some flexibility in the quantile regressions, as well, and prefer to present the results from Table B1 in the main paper.

D.2 Nonparametric imputation of exterior floor area

The relationship between the interior and the exterior floor area will be affected by design of the building. Applying a fixed conversion factor, regardless of building type and design, may therefore introduce additional measurement error to the ask data. The ask data, however, does not provide (reliable) information about many characteristics, such as an attic converted into living space, that would presumably produce a more refined conversion of the interior to exterior floor area. To examine how a feasible method might affect our analysis, we thus proxy the building design by the age of the building.[2] Specifically, we fit the varying coefficient model

$$FA = \alpha + \beta(CY) \cdot IFA + \varepsilon \tag{D4}$$

where FA is the exterior and IFA is the interior floor area. α is a constant and $\beta(AGE)$ is a smooth coefficient that is allowed to vary with the year that the house was constructed (CY). We model $\beta(\cdot)$ using cubic regression splines and fit the model via penalised least squares. We choose the smoothing parameter by double cross-validation.

Table D2 provides summary statistics for 1,513 observations in the sale data that report the exterior *and* interior floor area. Relative to the full sample (see Tab. 2), houses are of significantly lower age but otherwise comparable.

[Table D2 about here.]

[2]While we observe the building type itself, splitting the data accordingly, would result in too small sample sizes to fit Eq. D4.

Figure D2 shows the estimated conversion factor in its upper left panel. The point estimates varies between about 1.00 and 1.26 and is thus of similar magnitude as our fixed conversion factor. The pointwise confidence bands indicate that the estimation uncertainty can be sizable.

[Figure D2 about here.]

The upper right panel correlates the exterior floor area computed from the fitted model and the fixed conversion factor of 1.25. Both measures are highly correlated ($\hat{\rho} = 0.956$). Still, the fitted values explain more of the variation in the actual exterior floor area (the R^2 of the fitted model is 0.561 vs 0.354 for the fixed conversions factor); see also the lower panels of Figure D2. Taken together, we do not expect that the exterior floor area estimated from Eq. D4 would significantly change the results presented in the main paper. Moreover, given the high estimation uncertainty, we prefer the fixed conversion for our analysis.

D.3 Automated valuation

To assess the predictive accuracy of a parametric hedonic model, we re-ran the prediction experiment and fitted

$$p = \mathbf{z}\boldsymbol{\gamma} + g_1(AGE; \boldsymbol{\beta}_1) + g_2(FA; \boldsymbol{\beta}_2) + g_3(PA; \boldsymbol{\beta}_3) + g_4(LAT, LON; \boldsymbol{\beta}_4) + \varepsilon \quad \text{(D5)}$$

where $g_j(\cdot)$ is a p_j'th degree polynomial in continuous variable j ($p_j \in \{1, 2 \ldots, 7\}$). All variables are defined as in Section 3.3. For each estimation window, we select $\mathbf{p} = \{p_1, p_2, p_3, p_4\}$ as

$$\hat{\mathbf{p}} = \arg\max_{\mathbf{p}} \left[1 - \frac{\sum_{i=1}^{N} (p_i - \hat{p}_{-i})}{\sum_{i=1}^{N} (p_i - \bar{p})} \right] \quad \text{(D6)}$$

where $\hat{p}_{-i}$ is the leave-one-out estimator for observation i. We calculate the predictive residuals, $\epsilon_i = p_i - \hat{p}_{-i}$, from the ordinary residuals and diagonal elements of the hat matrix (Myers 1990, pp. 172-73).

Table D3 presents performance measures for the out-of-sample predictions from the parametric regressions fitted separately to ask and sale data.

[Table D3 about here.]

Comparing the performance of the prediction errors $e_{a,n}$ and $e_{s,n}$ leads to qualitatively similar conclusions as in Section 4.3. Using ask data leads to prediction errors that are severely biased *and* significantly more dispersed than prediction errors from sale data. Furthermore, comparing the results in Table D3 to those in Tab. 6 reveals that the parametric model produces prediction errors with inferior performance. This is true for, both, quadratic and absolute loss functions.

E Software

E.1 Decomposition of price distributions

To implement the stochastic dominance tests and markup decomposition we employ the user-written `Stata` commands `cdeco` and `oaxaca`. The former can be installed from `https://sites.google.com/site/blaisemelly/home/computer-programs/inference-on-counterfactual-distributions`. The latter can be installed from the Boston College Statistical Software Components (SSC) archive using the command `ssc install oaxaca`.

E.2 Non- and semiprametric regression models

To estimate the various non- and semiparamteric regression models we employ the `gam()` function from the `R` package `mgcv`, see `https://cran.r-project.org/web/packages/mgcv/index.html`. Wood (2017) provides an excellent introduction to generalized additive models and the `mgcv` package.

References

Chernozhukov, V., Fernandez-Val, I. and Mellie, B.: 2013, Inference on counterfactual distributions, *Econometrica* **81**, 2205–2268.

Cropper, M. L., Deck, L. B. and McConnel, K. E.: 1988, On the choice of functional form for hedonic price functions, *Review of Economic and Statistics* **70**, 668–675.

Gutachterausschuss für Grundstückswerte: 2011, *Bericht über den Berliner Grundstücksmarkt 2010/11*, Senatsverwaltung für Stadtentwicklung, Berlin. Kulturbuch-Verlag Berlin.

Gutachterausschuss für Grundstückswerte: 2012, *Bericht über den Berliner Grundstücksmarkt 2011/12*, Senatsverwaltung für Stadtentwicklung und Umwelt, Berlin. Kulturbuch-Verlag Berlin.

Gutachterausschuss für Grundstückswerte: 2013, *Bericht über den Berliner Grundstücksmarkt 2012/13*, Senatsverwaltung für Stadtentwicklung und Umwelt, Berlin. Kulturbuch-Verlag Berlin.

Gutachterausschuss für Grundstückswerte: 2014, *Bericht über den Berliner Grundstücksmarkt 2013/14*, Senatsverwaltung für Stadtentwicklung und Umwelt, Berlin.

Gutachterausschuss für Grundstückswerte: 2015, *Bericht über den Berliner Grundstücksmarkt 2014/15*, Senatsverwaltung für Stadtentwicklung und Umwelt, Berlin.

Jann, B.: 2008, The Blinder-Oaxaca decomposition for linear regression models, *Stata Journal* **8**, 453–479.

Kuminoff, N. V., Parmeter, C. F. and Pope, J. C.: 2010, Which hedonic models can we trust to recover the marginal willingness to pay for environmental amenities?, *Journal of Environmental Economics and Management* **60**, 145–160.

Myers, R. H.: 1990, *Classical and modern regression with applications*, Duxbury Press, Belmont, California.

Shimizu, C., Nishimura, K. G. and Watanabe, T.: 2016, House prices at different stages of the buying/selling process, *Regional Science and Urban Economics* **59**, 37–53.

Wood, S. N.: 2017, *Generalized additive models. An introduction with R*, Texts in Statistical Sience, 2 edn, CRC Press, Boca Raton.

Table A1: Bounds for data cleaning. Reports lower and upper bounds used for data cleaning procedure. Floor and plot area are in sqm. Source: Gutachterausschuss (2011, 2012, 2013, 2014,2015)

Detached	Plot area		Floor area		Price per sqm	
West, vintage						
< 1949	400	1500	50	650	650	3530
≥ 1949	400	1500	50	625	780	2990
East, vintage						
< 1949	400	1500	50	510	410	2630
≥ 1949	400	1495	50	440	910	3185
Semi-detached						
West, vintage						
< 1949	215	700	80	455	665	3655
≥ 1949	175	700	65	360	1005	3055
East, vintage						
< 1949	230	700	40	330	430	2790
≥ 1949	190	700	60	210	1005	3350
Terraced houses						
West, vintage						
< 1949	130	695	65	470	720	3512
≥ 1949	115	700	75	335	895	3160
East, vintage						
< 1949	115	695	60	285	495	2085
≥ 1949	100	665	65	285	1095	2695

Table B1: Decomposition of markups. Shows decomposition of the ask and sale price distributions. Standard errors for mean decomposition are computed using heteroscedasticity robust covariance estimator. 0.95 pointwise (uniform) confidence intervals for quantile decomposition are computed using bootstrap standard errors (the inverse of the bootstrapped KS t-statistic). Number of bootstrap replications is 200.

	Estimated Effect	Pointwise Std. Err.	Pointwise Conf. Interv.		Uniform Conf. Bands	
Panel A. Markup						
Mean	0.234	0.004	0.225	0.242		
Quantile						
0.1	0.241	0.006	0.230	0.253	0.223	0.260
0.2	0.210	0.005	0.201	0.219	0.196	0.224
0.3	0.197	0.004	0.189	0.205	0.184	0.210
0.4	0.193	0.004	0.185	0.201	0.180	0.205
0.5	0.194	0.004	0.186	0.203	0.181	0.207
0.6	0.202	0.005	0.193	0.211	0.188	0.216
0.7	0.218	0.005	0.208	0.228	0.202	0.234
0.8	0.241	0.006	0.229	0.252	0.222	0.259
0.9	0.285	0.008	0.270	0.300	0.261	0.309
Panel B. Characteristics						
Mean	0.210	0.005	0.201	0.219		
Quantile						
0.1	0.246	0.005	0.236	0.256	0.234	0.258
0.2	0.212	0.004	0.204	0.221	0.201	0.223
0.3	0.192	0.004	0.184	0.201	0.182	0.203
0.4	0.181	0.004	0.173	0.189	0.171	0.191
0.5	0.177	0.004	0.169	0.185	0.166	0.187
0.6	0.177	0.004	0.169	0.185	0.166	0.188
0.7	0.185	0.004	0.176	0.193	0.173	0.196
0.8	0.199	0.005	0.190	0.209	0.187	0.211
0.9	0.237	0.006	0.225	0.250	0.221	0.253
Panel C. Implicit prices						
Mean	0.024	0.004	0.017	0.032		
Quantile						
0.1	-0.005	0.006	-0.017	0.007	-0.025	0.015
0.2	-0.003	0.004	-0.011	0.005	-0.016	0.011
0.3	0.004	0.003	-0.002	0.011	-0.007	0.016
0.4	0.011	0.003	0.005	0.017	0.001	0.022
0.5	0.018	0.003	0.012	0.023	0.008	0.027
0.6	0.025	0.003	0.019	0.031	0.015	0.035
0.7	0.033	0.003	0.027	0.040	0.023	0.044
0.8	0.041	0.004	0.034	0.049	0.029	0.054
0.9	0.048	0.005	0.039	0.057	0.032	0.064

Table D1: Decomposition of markups, log-linear model. Shows decomposition of the ask and sale price distributions. Standard errors for mean decomposition are computed using heteroscedasticity robust covariance estimator. 0.95 pointwise (uniform) confidence intervals for quantile decomposition are computed using bootstrap standard errors (the inverse of the bootstrapped KS t-statistic). Number of bootstrap replications is 200.

	Estimated Effect	Pointwise Std. Err.	Pointwise Conf. Interv.		Uniform Conf. Bands	
Panel A. Markup						
Mean	0.234	0.004	0.225	0.242		
Quantile						
0.1	0.262	0.005	0.252	0.273	0.248	0.277
0.2	0.219	0.004	0.210	0.227	0.208	0.230
0.3	0.196	0.004	0.189	0.204	0.186	0.207
0.4	0.184	0.004	0.176	0.192	0.174	0.194
0.5	0.178	0.004	0.170	0.185	0.167	0.188
0.6	0.178	0.004	0.169	0.186	0.167	0.189
0.7	0.187	0.005	0.177	0.196	0.174	0.199
0.8	0.211	0.006	0.200	0.222	0.196	0.226
0.9	0.276	0.008	0.260	0.291	0.255	0.296
Panel B. Characteristics						
Mean	0.219	0.005	0.210	0.228		
Quantile						
0.1	0.169	0.003	0.162	0.176	0.160	0.177
0.2	0.150	0.003	0.144	0.156	0.142	0.157
0.3	0.139	0.003	0.134	0.145	0.132	0.147
0.4	0.135	0.003	0.129	0.140	0.128	0.142
0.5	0.133	0.003	0.127	0.139	0.126	0.140
0.6	0.138	0.003	0.132	0.144	0.130	0.146
0.7	0.152	0.004	0.145	0.159	0.143	0.161
0.8	0.180	0.004	0.172	0.189	0.169	0.191
0.9	0.250	0.006	0.238	0.261	0.235	0.265
Panel C. Implicit prices						
Mean	0.015	0.004	0.007	0.0228		
Quantile						
0.1	0.093	0.005	0.083	0.104	0.079	0.108
0.2	0.069	0.004	0.061	0.077	0.059	0.080
0.3	0.057	0.003	0.050	0.063	0.048	0.066
0.4	0.049	0.003	0.043	0.055	0.041	0.057
0.5	0.044	0.003	0.039	0.050	0.037	0.052
0.6	0.040	0.003	0.034	0.046	0.032	0.048
0.7	0.035	0.003	0.029	0.041	0.026	0.044
0.8	0.031	0.004	0.023	0.038	0.020	0.041
0.9	0.026	0.005	0.016	0.036	0.012	0.039

Table D2: Summary statistics for observations in sale data that report the interior and exterior floor area. Number of observations is 1,513. Age of building at the date of sale. Exterior and interior floor area are in sqm.

	Mean	Std. Dev.	Min	Max
Age	22.53	29.43	0.00	100.00
Construction year	1988	29.24	1908	2015
Floor area				
exterior	141.76	46.29	45.00	451.00
interior	125.15	35.13	42.00	552.00

Table D3: Assessment of prediction errors from parametric hedonic model. Shows performance statistics for 9,152 out-of-sample prediction errors. ±10% (±25%) reports the proportion of errors which are in absolute terms no larger than 10% (25%).

Data	MSE	Bias	Var.	Med.	MAE	±10%	±25%
Ask	0.089	-0.024	0.088	-0.012	0.231	0.282	0.610
Sale	0.059	0.013	0.059	0.027	0.189	0.334	0.696

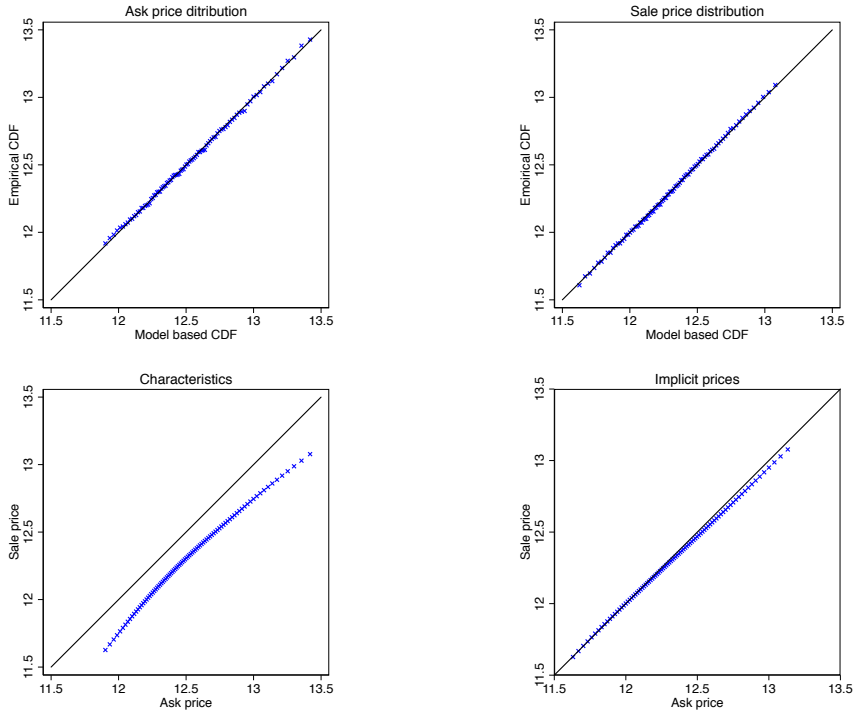

Figure B1: Q-Q plots for price distributions. Upper-left (right) panel compares $\hat{F}_{a|a}$ ($\hat{F}_{s|s}$) estimated from Eq. 4 with their EDFs. Lower-left panel compares $\hat{F}_{a|a}$ and $\hat{F}_{a|s}$. Lower-right panel compares $\hat{F}_{a|s}$ and $\hat{F}_{s|s}$. Solid black lines are the 45 degree line.

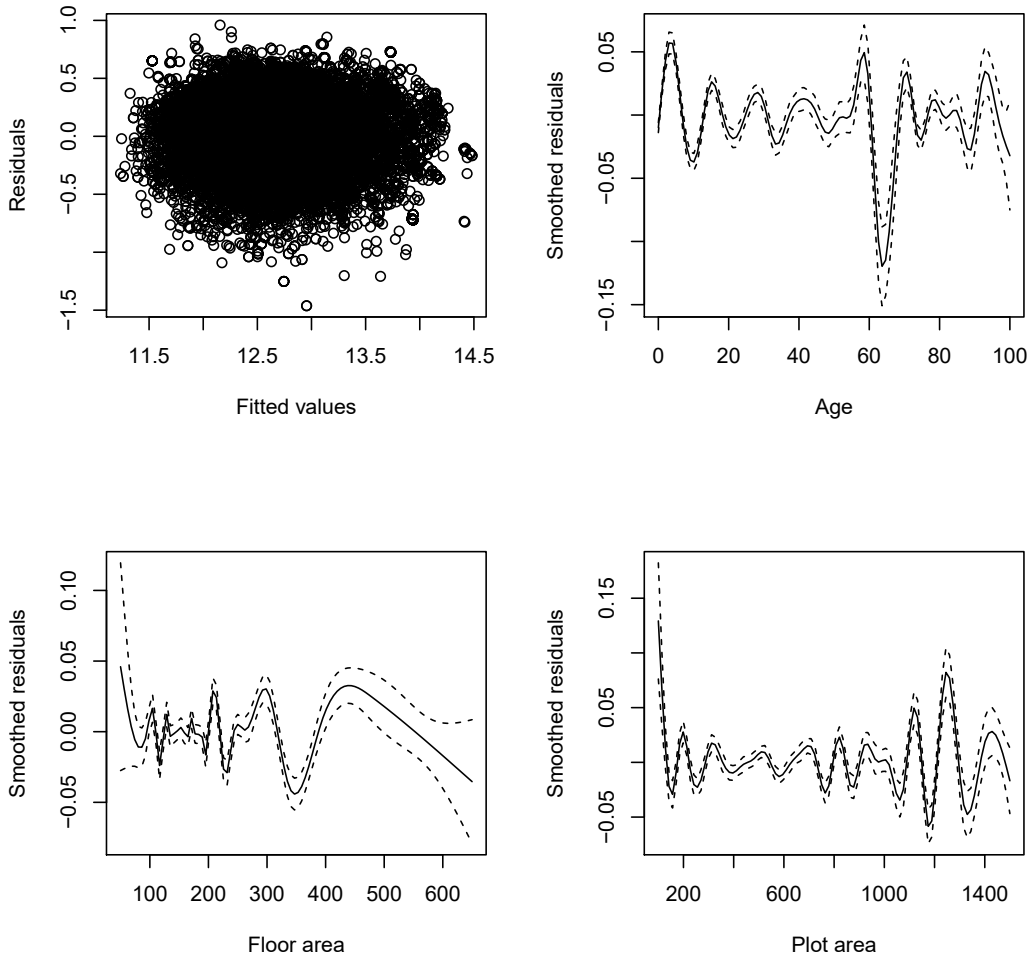

Figure C1: Residual diagnostics (1). Model specification fits ask price to core variables and spatial fixed effects (column (1) in Table 4).

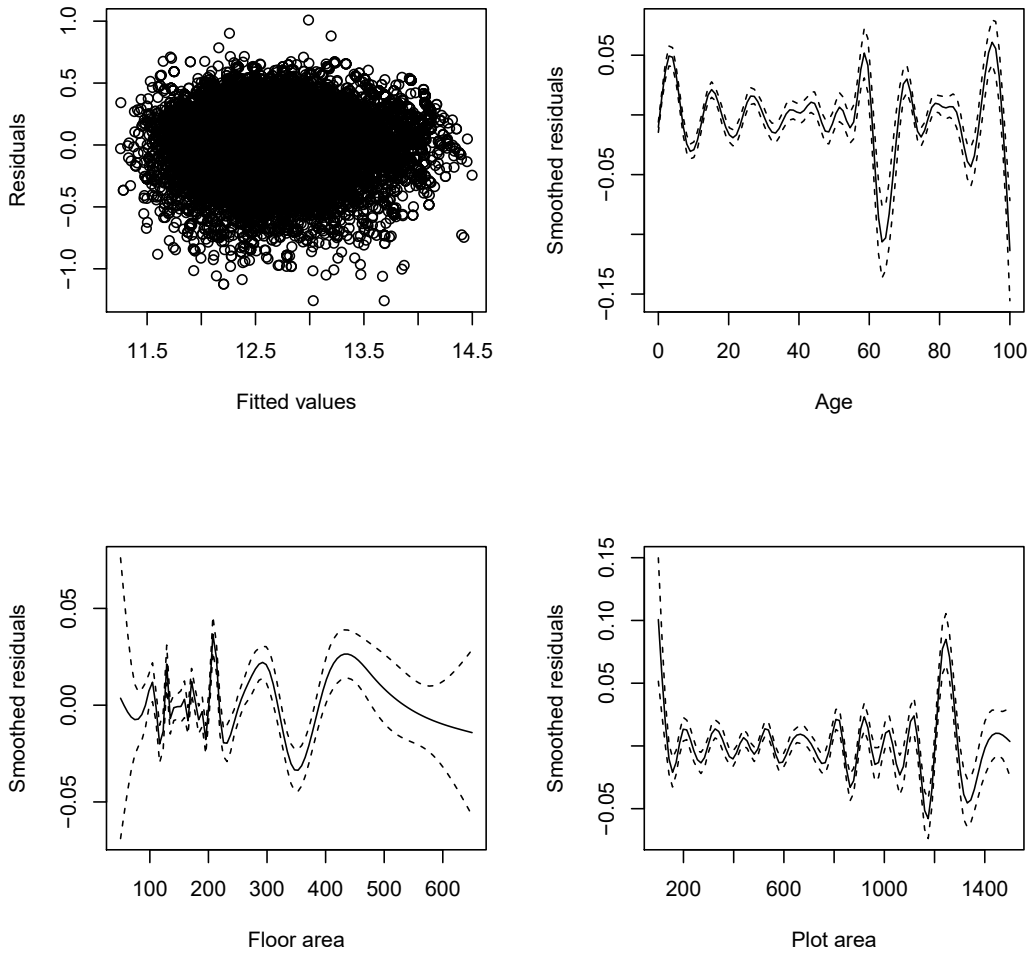

Figure C2: Residual diagnostics (2). Model specification fits ask price to core variables and geospatial smooth (column (2) in Table 4).

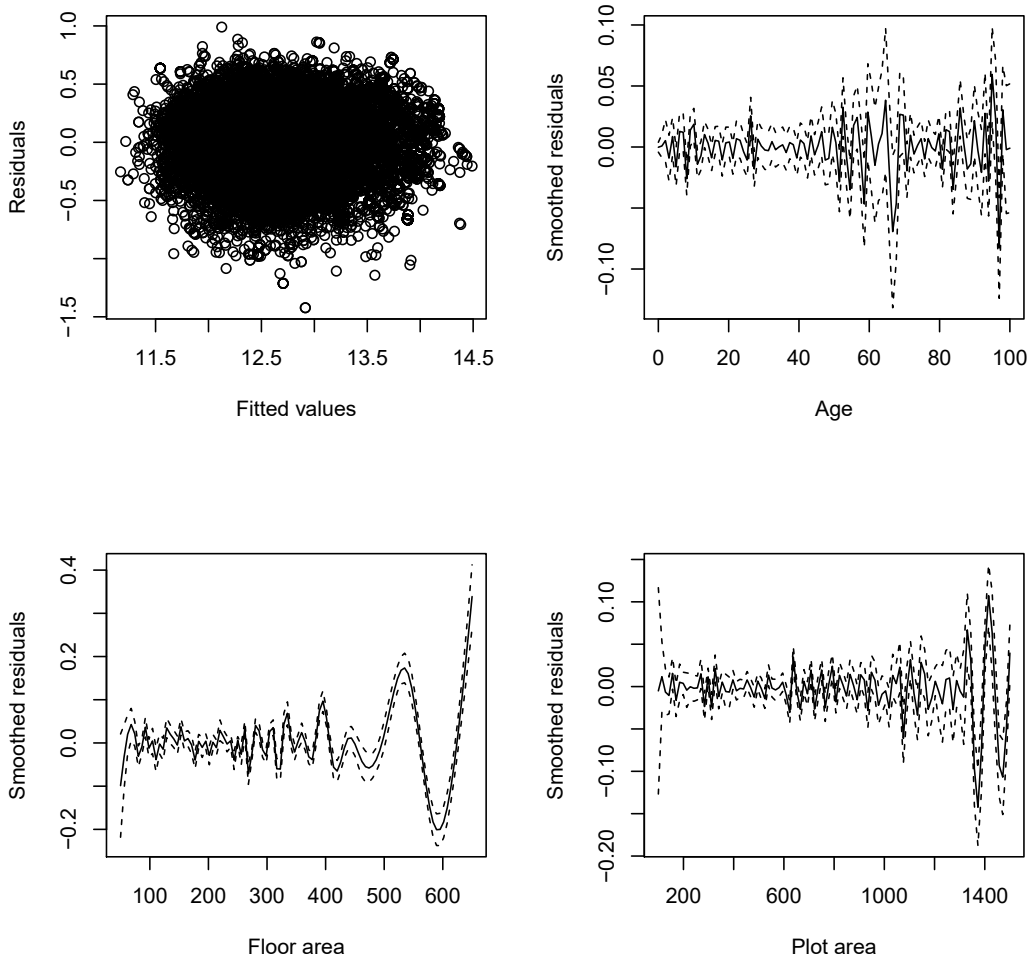

Figure C3: Residual diagnostics (1b). Model specification fits ask price to core variables and spatial fixed effects (column (1) in Table 4).

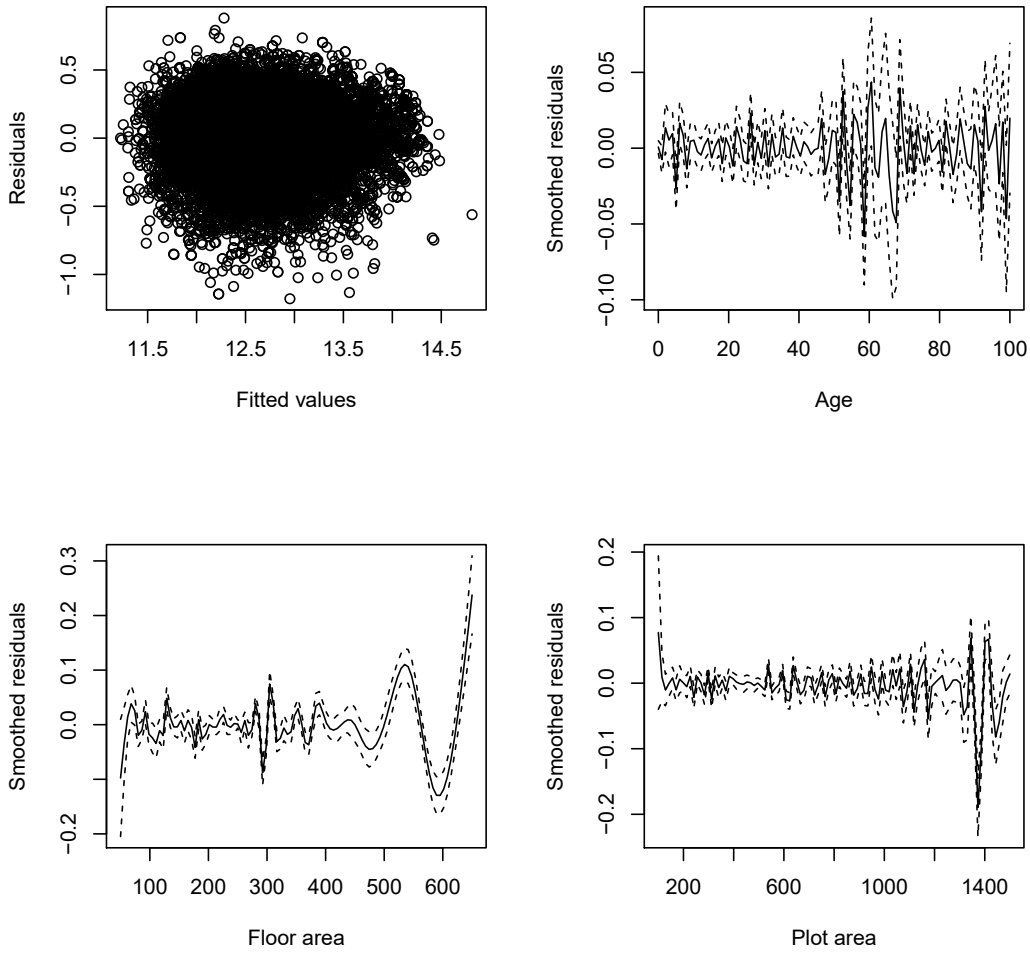

Figure C4: Residual diagnostics (2b). Model specification fits ask price to core variables and geospatial smooth (column (2) in Table 4).

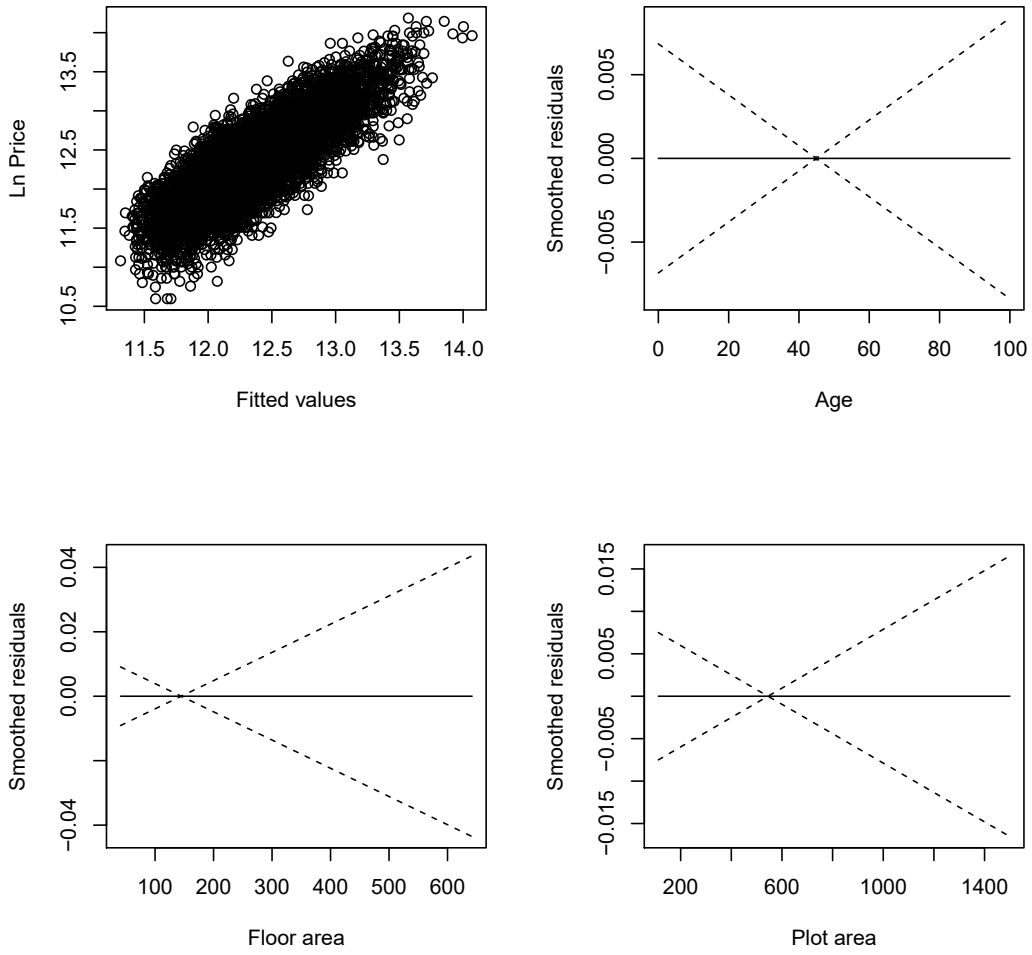

Figure C5: Residual diagnostics (3). Model specification fits sale price to core variables and spatial fixed effects (column (3) in Table 4).

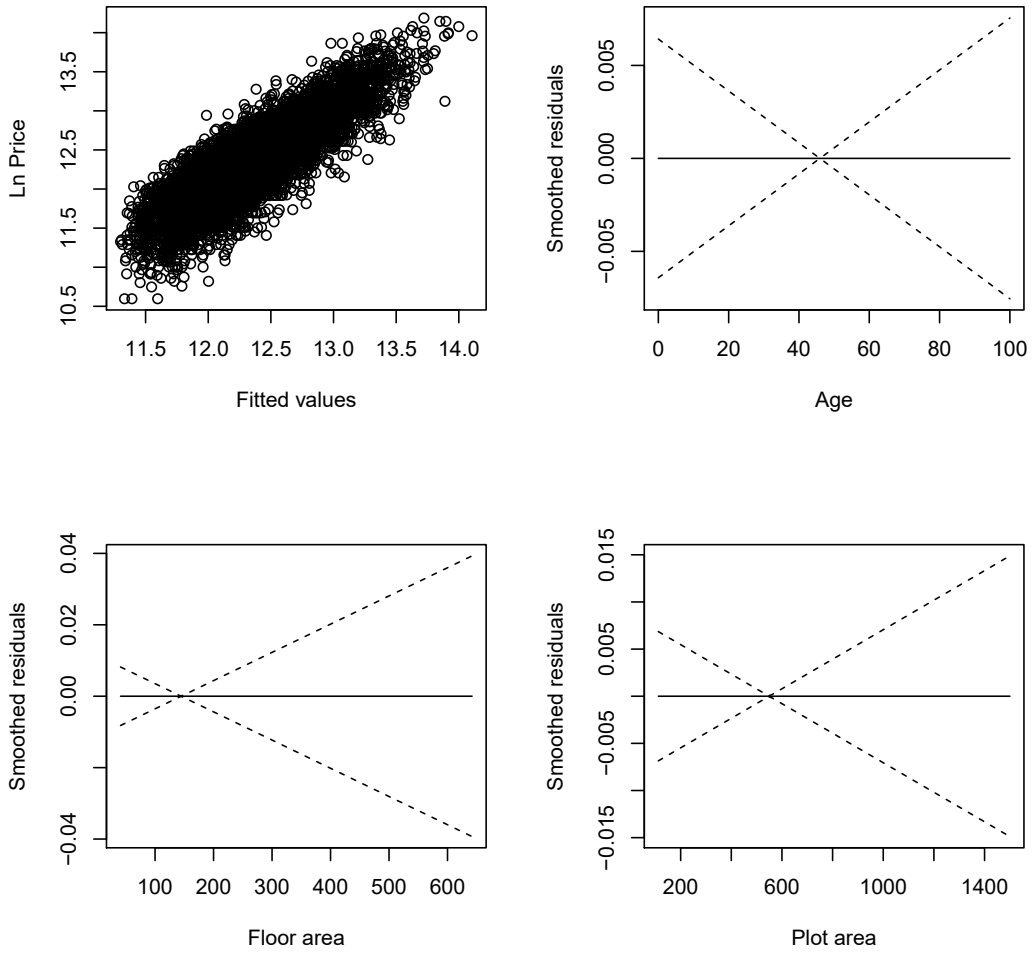

Figure C6: Residual diagnostics (4). Model specification fits sale price to core variables and geospatial smooth (column (4) in Table 4).

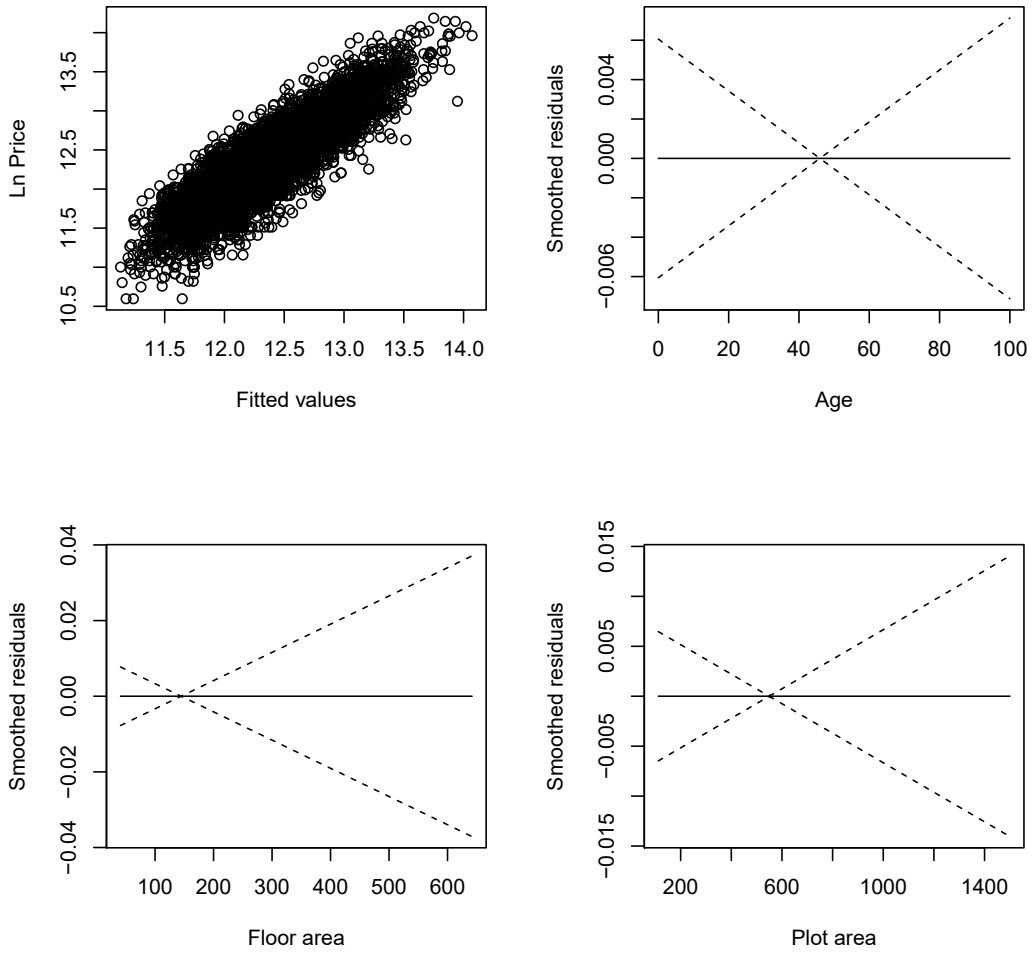

Figure C7: Residual diagnostics (5). Model specification fits sale price to all variables and spatial fixed effects (column (5) in Table 4).

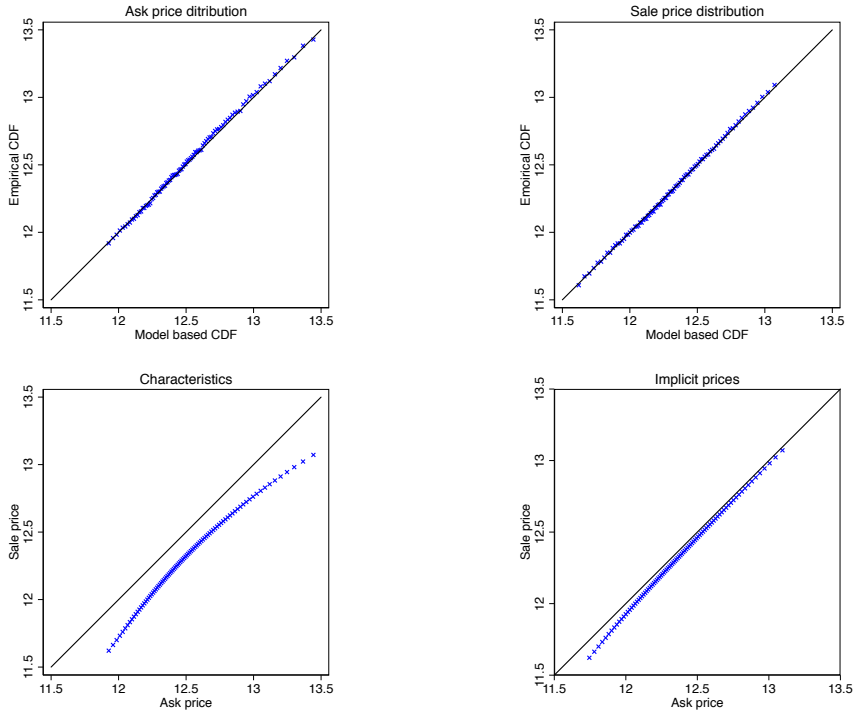

Figure D1: Q-Q plots for price distributions, log-linear specification. Upper-left (right) panel compares $\hat{F}_{a|a}$ ($\hat{F}_{s|s}$) estimated from Eq. 4 with their EDFs. Lower-left panel compares $\hat{F}_{a|a}$ and $\hat{F}_{a|s}$. Lower-right panel compares $\hat{F}_{a|s}$ and $\hat{F}_{s|s}$. Solid black lines are the 45 degree line.

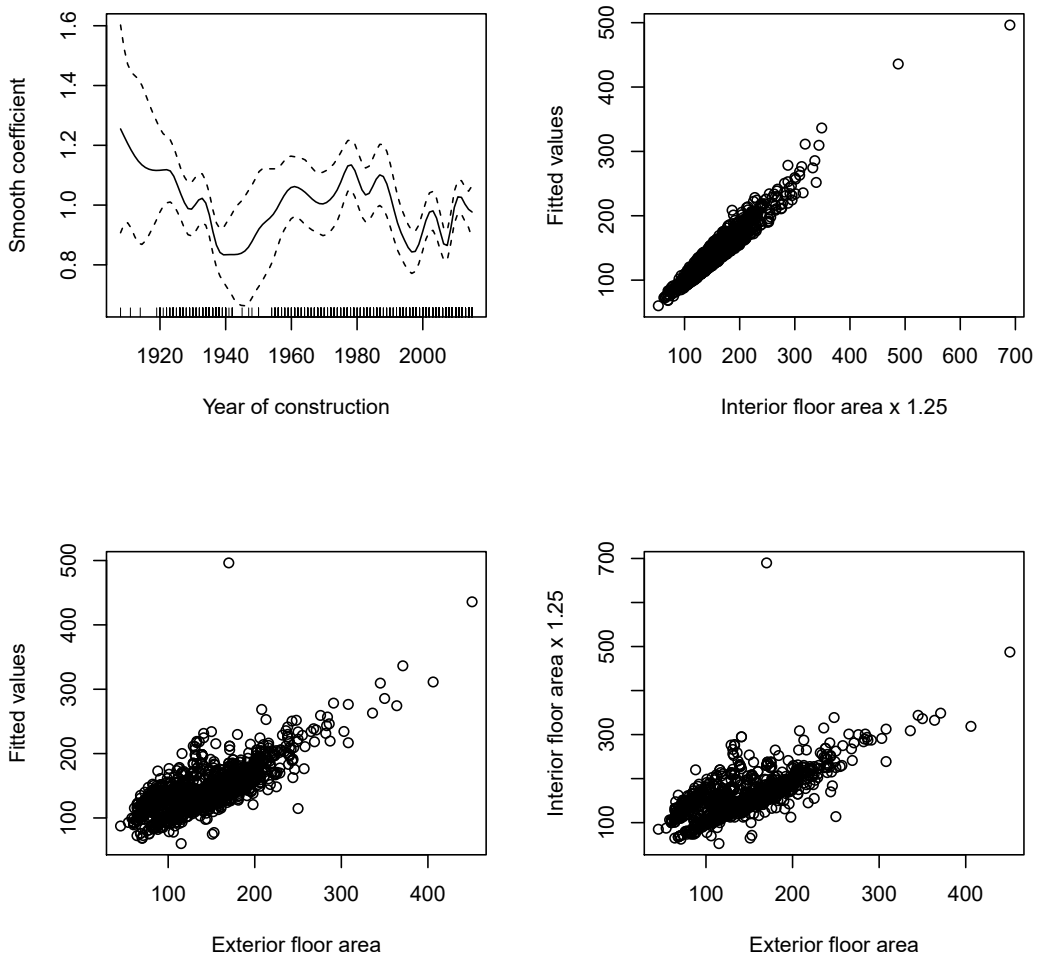

Figure D2: Varying coefficient model and floor area conversion. Shows estimates from Eq. D4. Number of observations is 1,513.