DIW BERLIN

1216

Discussion Papers

Deutsches Institut für Wirtschaftsforschung 2012

Location, Location, Location: Extracting Location Value from House Prices

Jens Kolbe, Rainer Schulz, Martin Wersing and Axel Werwatz

IMPRESSUM

DIW Berlin
German Institute for Economic Research
Mohrenstr. 58
10117 Berlin

Tel. +49 (30) 897 89-0
Fax +49 (30) 897 89-200
http://www.diw.de

ISSN print edition 1433-0210
ISSN electronic edition 1619-4535

Papers can be downloaded free of charge from the DIW Berlin website:
http://www.diw.de/discussionpapers

Discussion Papers of DIW Berlin are indexed in RePEc and SSRN:
http://ideas.repec.org/s/diw/diwwpp.html
http://www.ssrn.com/link/DIW-Berlin-German-Inst-Econ-Res.html

Location, location, location: Extracting location value from house prices

Jens Kolbe*, Rainer Schulz†, Martin Wersing‡, and Axel Werwatz‡

May, 2012

*Deutsches Institut für Wirtschaftsforschung, Mohrenstrasse 58, 10117 Berlin Germany. Email: jkolbe@diw.de.

†University of Aberdeen Business School, Edward Wright Building, Dunbar Street, Aberdeen AB24 3QY, United Kingdom. Email: r.schulz@abdn.ac.uk.

‡Technische Universität Berlin, Institut für Volkswirtschaftslehre und Wirtschaftsrecht, Straße des 17. Juni 135, 10623 Berlin, Germany, and Collaborative Research Center 649 *Economic Risk*, Humboldt-Universität zu Berlin. Emails: martin.wersing@tu-berlin.de and axel.werwatz@tu-berlin.de.

Financial support from the Deutsche Forschungsgemeinschaft, CRC 649 *Economic Risk*, and the Wissenschaftsgemeinschaft Gottfried Wilhelm Leibniz, Pakt Econs, is gratefully acknowledged. The usual disclaimer applies.

Abstract

The price for a single-family house depends both on the characteristics of the building and on its location. We propose a novel semiparametric method to extract location values from house prices. After splitting house prices into building and land components, location values are estimated with adaptive weight smoothing. The adaptive estimator requires neither strong smoothness assumptions nor local symmetry. We apply the method to house transactions from Berlin, Germany. The estimated surface of location values is highly correlated with expert-based land values and location ratings. The semiparametric method can therefore be used for applications where no other location value information exists or where this information is not reliable.

Keywords: location value, adaptive weight smoothing, spatial modeling

JEL Classification: R31, C14

1 Introduction

When asking a real estate professional about the three most important characteristics of a house, the likely answer will be 'location, location, location'.[1] Naturally, the characteristics of the building itself also play a role in its desirability, but the phrase emphasizes the importance of the surrounding area. The nicest villa in an otherwise run-down neighborhood is much less desirable than the very same building in a nice suburban area with shady forests, quiet lakes, and good schools. Based on this reasoning, we expect that the house will fetch a higher price when located in a nice area than when located in a run-down area. Seen differently, the difference between the prices of the villa in the two different areas gives the location value of the nice area relative to the run-down area. Once buildings differ with respect to their characteristics, such a simple price comparison is no longer sufficient to learn about the relative value of a location. But the general notion remains: house prices contain information on the value of the location.

Location values are of interest for several reasons. They can be used for spatial analysis with respect to the influence of amenities and externalities. They can be used for studying the impact of regulation, such as zoning. They can be used to measure the effects of policy interventions, such as regeneration and revitalization. Location values can be estimated directly from transactions of undeveloped land (Colwell and Munneke, 2003). However, particularly in densely populated urban areas, few (if any) transactions of undeveloped land may occur. House sales are typically more frequent.

In this paper, we propose a flexible method to estimate location values from house prices. At the first stage, we use the semiparametric estimator of Yatchew (1997) and Wang et al. (2011) to split the house price into components related to the building and the location. At the second stage, we use adaptive weight smoothing

[1] The phrase is in use at least since the 1920ties, see William Safire's 'On language: location, location, location' in *The New York Times*, June 28, 2009.

(AWS) as pioneered by Polzehl and Spokoiny (2000, 2006) to estimate the location value surface. AWS is flexible regarding the shape of the surface and does not require smoothness assumptions. AWS identifies areas with homogenous location values by an adaptive iterative algorithm that is based on nonparametric smoothing. Unlike standard smoothers such as kernel regression, the algorithm does not require that the local areas have the same shape (e.g. rectangular or radial) at different locations.

We illustrate the methodology in an empirical application to data of geo-coded single-family house transactions from Germany's capital Berlin. Our estimated location surface provides a comprehensive characterization of the location values of Berlin's residential areas. The shape and size of areas with similar location values are completely data-driven and need not adhere to administrative boundaries. Since the true location values are not observed, we assess the adequacy of our estimates by comparing them with expert-based land values and expert-based ordinal location ratings. We find that our semiparametric method estimates location values that are highly correlated with the expert-based land values and location ratings.

Only a few previous studies have modeled location values from house price information. Cheshire and Sheppard (1995), Rosenthal (1999), and Rossi-Hansberg et al. (2010) are examples; none of these studies compares the estimated location values with benchmarks as we do.[2] Anglin and Gencay (1996) and Clapp (2004) also fitted semiparametric models to house prices, but with more restrictive and less flexible value functions.

In summary, the novel method proposed in this paper allows us to estimate location values from house prices. We find that the estimated location values are reliable in the sense that they show agreement with expert assessments based on different information. The method should prove useful for applications where location values are needed and no expert-based information is available or where such information should be complemented by data-driven flexible location value estimates.

[2]Lack of such a benchmark is the reason why location values have to be imputed in the first place.

2 Methodology and estimation

We start with the assumption that the price of a house can be split into the value B of the building and the value L of land, so that $P = B + L$. Such a zero-profit condition holds for new houses if they are produced by a competitive construction industry using a constant returns to scale technology. In the case of old houses, the condition should hold once the building value is adjusted for depreciation; the condition corresponds then to the depreciated cost approach (Bourassa et al., 2011). To make explicit that houses are heterogenous, we write

$$P = B(\mathbf{x}_B) + L(\mathbf{x}_L) \ , \tag{1}$$

where the vectors $\mathbf{x}_B$ and $\mathbf{x}_L$ collect building and land characteristics. We specify the building component as $B(\mathbf{x}_B) = \mathbf{x}_B'\boldsymbol{\beta}$. Building characteristics include continuous variables such as floor area and age and discrete variable such as cellar and building type. The land component is specified as $L(\mathbf{x}_L) = sa(\mathbf{l})$, where s measures lot size in square meters. The location value $a(\mathbf{l})$ depends on the Cartesian location coordinates $\mathbf{l} = (l_1, l_2)$, but is otherwise unspecified and flexible. The coefficient vector $\boldsymbol{\beta}$ and the location value function $a(\mathbf{l})$ are not known and have to be estimated.

Dividing both sides of Eq. 1 by the lot size s and adding the term ε for unobserved characteristics and idiosyncratic effects during the transaction, we obtain the partial-linear regression model

$$p = \mathbf{z}'\boldsymbol{\beta} + a(\mathbf{l}) + \varepsilon \ . \tag{2}$$

Here, p and $\mathbf{z}$ denote the house price and the building characteristics per square meter lot size.[3] We assume $\mathrm{E}(\varepsilon|\mathbf{z}, \mathbf{l}) = 0$.

In order to estimate the nonparametric location value function, we first remove the building value from the house price. Specifically, we obtain a consistent estimate of the parametric component in Eq. 2 and compute the residual $u = p - \mathbf{z}'\boldsymbol{\beta}$, which

[3]The continuous building characteristics may be transformed further to capture non-linearities in the hedonic price function.

equals the sum of the location value plus the transaction noise term ε. We then separate the residual into the latter two terms using AWS.

We note that our method does not allow the identification of separate constants for the building and the location value component. We can therefore estimate the relative location value surface, but additional information is required to convert the surface into levels.[4]

2.1 Data description

Our main data is provided by Berlin's Committee of Valuation Experts (GAA, Gutachterausschuss für Grundstückswerte) from their transaction database (AKS, Automatisierte Kaufpreissammlung).[5] The data covers arms-length transactions of single-family houses during the years 1996-2010. The data contains information on the transaction price, geographic location coordinates, and numerous building characteristics.

Each transaction has an expert-based land value, which is the notional value of land as if it were undeveloped. The value assumes that land is not contaminated or burdened with unusual legal covenants. The land values are computed by GAA appraisers using the sales comparison approach based on information from transactions of undeveloped land. Expert-based land values are expressed in Euros per square meter. Each transaction has also an expert-based location rating, which is provided by Berlin's Senate Department for Urban Development and the Environment. The ordinal rating uses four levels to summarize the quality of a location. For this rating, the experts consider the amount of natural amenities such as lakes and forests, the quality of existing buildings, and the access to public transport and shopping facilities.

[4]Observing the price for undeveloped land at the location where $a(\mathbf{l})$ reaches, say, its minimum would be sufficient to calibrate the surface.

[5]The GAA is entitled by law to request and collect information on all real estate transactions occurring in Berlin.

Table 1 gives summary statistics for the 19,283 observations. House prices and expert-based land values are converted into year 2000 Euros using constant-quality price and land value indices, respectively.[6] As indicated by the standard deviation, house prices show substantial variation. This is in line with the substantial variation of building characteristics, such as floor size, number of storeys, age of the building, and building type. There is also substantial variation regarding the size of the lot. Unusual features of the house in Table 1 include physical aspects such as structural damage or flooding risk and legal aspects such as rights of way or use for pipes or cables. Such easements are rather common. Another important source of variation is the location of a house within the city, as indicated by the map plotted in Figure 1. The area of Berlin is 891 km^2, where the distance from west to east is 45 km (left to right) and 38 km from south to north (bottom to top). The map shows that the amount of lakes, rivers, parks, and forests differs between suburban areas. Modern Berlin was created by incorporating many formerly independent smaller cities and towns, some of which have kept their own distinctive character, which adds to the variation of location characteristics.

The last part of Table 1 presents summary statistics for the expert-based land values and the expert-based location ratings. The expert-based values and ratings are not unrelated, because GAA appraisers will use the location ratings for their land values and the experts of the Senate department might use information on land values for their location rating exercise. But the experts will also use different information differently and we do not expect that the two expert-based assessments always conform. Panel A of Table 2 gives the matching frequencies of the two expert-based location assessments once the expert-based land values are converted into an ordinal rating. In this conversion, the 2% largest land values receive the rating 'excellent', the next 20% values the rating 'high' and so forth. Constructed this way, the land value rating has the same marginal distribution as the expert-

[6]The indices are estimated using the hedonic regression methodology described in Schulz and Werwatz (2011).

Table 1: Summary statistics for transacted single-family houses. Number of observations is 19,283. Prices and land values are in year 2000 Euros. Floor size, gross base, and lot size are in square meters. Gross volume is in cubic meters. Gross area is the sum of all base areas in all storeys, gross volume is the corresponding volume. 8,259 objects have information on the gross volume and 15,325 on the gross base. Age of the building in years at the transaction date. Attic storey means that the attic is upgraded for living. Expert-based land value per square meter is the appraised value as if land were undeveloped. Expert-based location rating is an ordinal ranking of the neighborhood of the house.

	Mean	Median	Std. Dev.
House price	273,168	231,176	177,337
Building characteristics			
Floor size	145.99	135.00	56.23
Gross area	244.24	228.00	95.69
Gross volume	666.83	612.00	262.78
Storeys	1.5	1.0	0.6
Age	42	42	29
Type			
Detached	0.55		
Semi-detached	0.22		
Row-house	0.23		
Attic storey	0.55		
Flat roof	0.12		
No cellar	0.13		
Part cellar	0.12		
State of repair			
Poor	0.08		
Normal	0.61		
Good	0.31		
Land characteristics			
Lot size	578.56	525.00	313.33
Unusual features of the house			
Physical	0.03		
Legal	0.18		
Expert-based land values and location ratings			
Land value	284.97	256.46	148.41
Location rating			
Low	0.29		
Medium	0.49		
High	0.20		
Excellent	0.02		

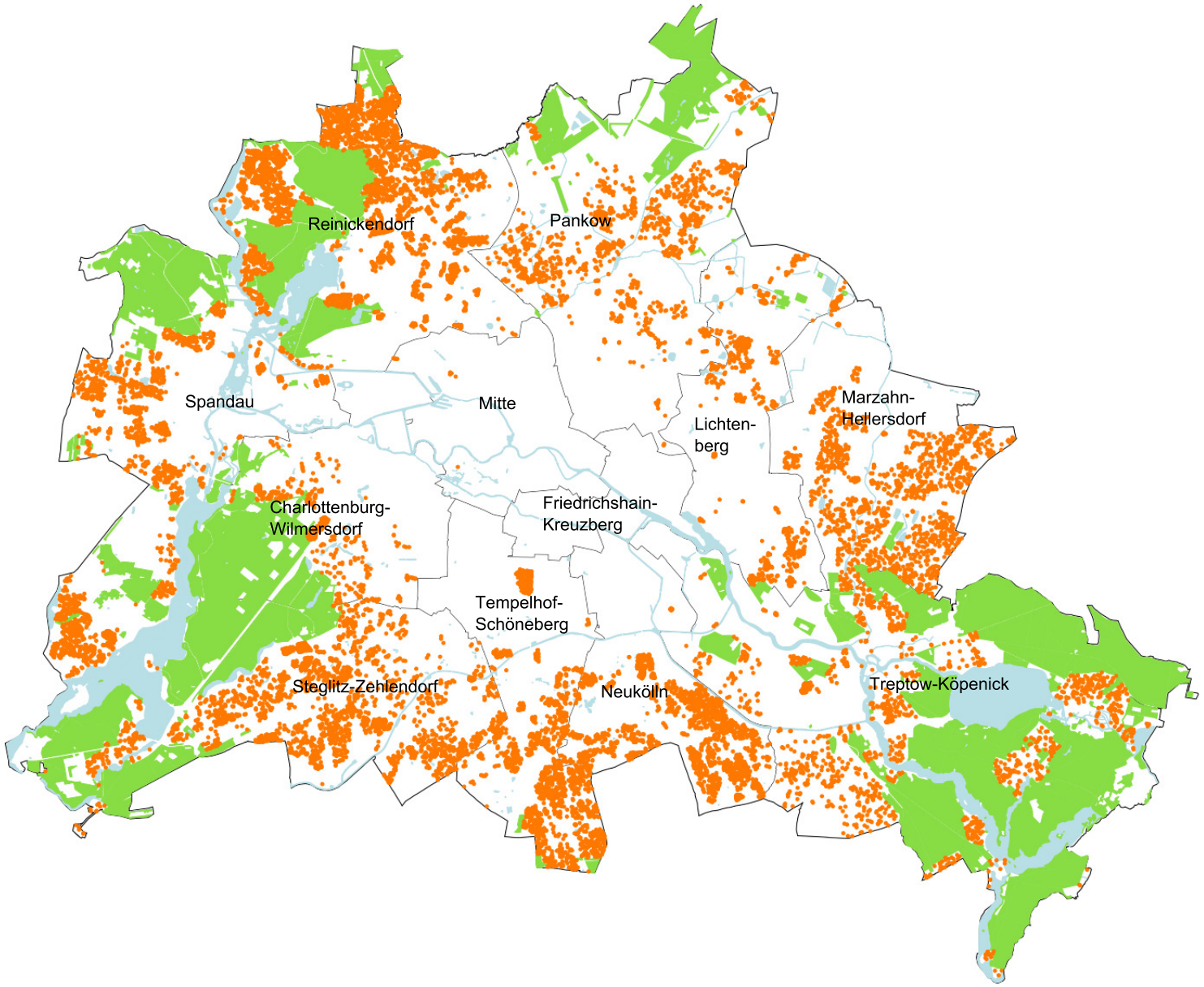

Figure 1: Transacted houses in Berlin, 1996-2010. Figure shows the location of the transacted single-family houses within in the city. Number of observations is 19,283. Solid lines represent the borders of Berlin's 12 administrative districts (as of the year 2000). Shaded blue areas are lakes and rivers. Shaded green areas are parks and forests.

based location rating. If the two ordinal location ratings were identical, then the

Table 2: Contingency tables for ordinal assessments. Panel A gives the relative frequencies of the matches of the expert-based ratings and the converted ordinal expert-based land values. Panel B gives the relative frequencies of the matches of the expert-based ratings and the converted ordinal location values $\widehat{a(\mathbf{l})}$. The Panel C gives the expected relative frequencies if expert-based ratings were randomly allocated onto itself. Pearson's χ^2-statistic is for the null that rows and columns are statistically independent. P-value is for a $\chi^2(9)$-distribution. Goodman and Kruskal's $\widehat{\gamma}$ is calculated as $\frac{N_s - N_d}{N_s + N_d}$ where N_s is the number of pairs of cases ranked in the same order and N_d is the number of pairs ranked differently. Kendall's $\widehat{\tau}$ is calculated as $\frac{N_s - N_d}{\sqrt{(N^2 - \sum N_c^2)(N^2 - \sum N_r^2)}}$ where N_c^2 and N_r^2 are the squared column and row marginals, respectively.

Panel A: Expert-based land values

	Expert-based rating				
	Low	Medium	High	Excellent	Total
Low	0.131	0.151	0.011	0.000	0.293
Medium	0.140	0.287	0.058	0.000	0.486
High	0.022	0.048	0.131	0.002	0.202
Excellent	0.000	0.000	0.003	0.016	0.019
Total	0.293	0.486	0.202	0.019	1.000
χ^2-stat.	2.1e+04	$\widehat{\gamma}$	0.634		
P-value	0.000	$\widehat{\tau}$	0.438		

Panel B: Estimated location values

	Expert-based rating				
	Low	Medium	High	Excellent	Total
Low	0.129	0.158	0.006	0.000	0.293
Medium	0.142	0.289	0.055	0.000	0.486
High	0.023	0.039	0.129	0.012	0.202
Excellent	0.000	0.000	0.012	0.007	0.019
Total	0.293	0.486	0.202	0.019	1.000
χ^2-stat.	1.1e+04	$\widehat{\gamma}$	0.644		
P-value	0.000	$\widehat{\tau}$	0.446		

Panel C: Random allocation, expected frequencies

	Expert-based rating				
	Low	Medium	High	Excellent	Total
Low	0.086	0.142	0.059	0.006	0.293
Medium	0.142	0.236	0.098	0.009	0.486
High	0.059	0.098	0.041	0.004	0.202
Excellent	0.006	0.009	0.004	0.001	0.019
Total	0.293	0.486	0.202	0.019	1.000
χ^2-stat.	0.018	$\widehat{\gamma}$	0.000		
P-value	1.000	$\widehat{\tau}$	0.000		

contingency matrix would have the marginal frequencies on the diagonal and zeros elsewhere. In Panel A, this is not the case. The two expert-based ratings are also not independent, as a comparison with Panel C shows. The panel gives the frequencies we would expect if matching were random. The null hypothesis of the chi-square test for statistical independence is rejected at the usual significance levels. Measures of strength of the relationship between the two ratings are Goodman and Kruskal's γ, and Kendall's τ, respectively. Both statistics range from -1 (perfect inversion) to $+1$ (perfect agreement). In Panel A, we estimate $\widehat{\gamma} = 0.634$ ($\widehat{\tau} = 0.438$), indicating the expected positive relationship between both expert-based ratings.

Figure 2 shows in its left panel box plots of the expert-based land values for each of the four levels of the expert-based location ratings. The median land values increase in line with the level of the expert-based rating, but the quartiles of the land values for locations with low and medium rating overlap to a large extend. The separation for the top two levels of the expert-based rating is more pronounced.

2.2 Estimation of the building component

We use the estimator proposed by Yatchew (1997) and Wang et al. (2011) for the estimation of $\boldsymbol{\beta}$. The basic idea of the estimator is that $a(\mathbf{l})$ can be neglected when working with the differences of the variables of close observations. This requires that the data are ordered to be geographically close to each other. We follow Yatchew (1997) and order the observations along a path created from the nearest-neighbor algorithm. The Appendix explains the algorithm.

Taking the differences of two nearby observations i and $i-1$ yields

$$p_i - p_{i-1} = (\mathbf{z}_i - \mathbf{z}_{i-1})'\boldsymbol{\beta} + a(\mathbf{l}_i) - a(\mathbf{l}_{i-1}) + \varepsilon_i - \varepsilon_{i-1} \; . \tag{3}$$

If the location value function is sufficiently smooth, $a(\mathbf{l}_i) - a(\mathbf{l}_{i-1})$ becomes negligible, because $\mathbf{l}_i$ and $\mathbf{l}_{i-1}$ are geographically close. The coefficient vector $\boldsymbol{\beta}$ can then be

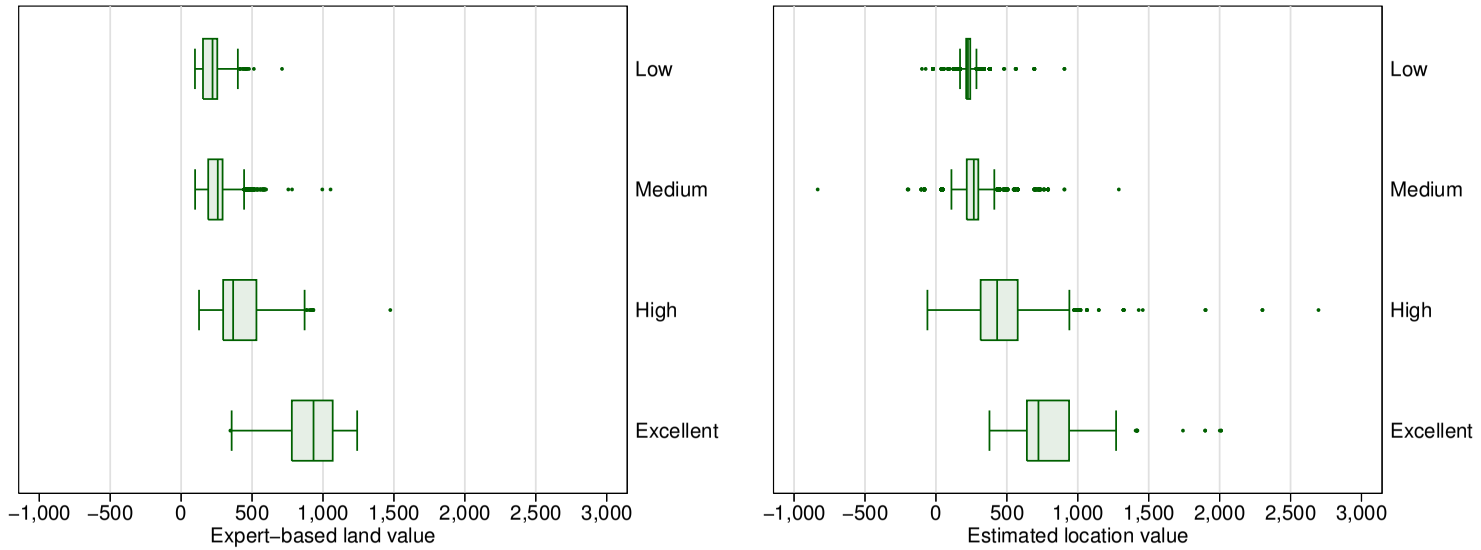

Figure 2: Box plots of expert-based land value and estimated location value. Left panel shows box plots of the expert-based land value for each of the four levels of the expert-based location rating. Right panel shows box plots of the estimated location value for each of the four levels of the expert-based location rating. Number of observations is 19,283. Line that separates the box is the median. Lower (upper) hinge of box represents 25th (75th) percentile. Length of whiskers is 1.5 times the IQR below (above) the 25th (75th) percentile.

estimated consistently with ordinary least squares.[7]

Whereas Eq. 3 is ideal for providing intuition, a version of this regression equation with weighted higher order differences will lead to a more efficient estimator. Letting $\Delta_m y_i \equiv \sum_{s=0}^{m} d_s y_{i-s}$, where y_i can be a scalar or a vector, and denoting the differencing weights with d_s, the improved estimation equation is

$$\Delta_m p_i = (\Delta_m \mathbf{z}_i)'\boldsymbol{\beta} + \Delta_m a(\mathbf{l}_i) + \Delta_m \varepsilon_i \ . \tag{4}$$

The weights fulfill the two restrictions

$$\sum_{s=0}^{m} d_s = 0 \quad \text{and} \quad \sum_{s=0}^{m} d_s^2 = 1 \ , \tag{5}$$

where the first restriction ensures that the location value function vanishes as the sample size increases and the locations become close. The second ensures that $\text{Var}[\Delta_m \varepsilon] = \sigma_\varepsilon^2$, i.e. the variance of the differenced error equals the variance of ε. The ordinary least squares estimator $\widehat{\boldsymbol{\beta}}_{\Delta_m}$ of Eq. 4 approaches asymptotic efficiency when m is chosen sufficiently large. Optimal weights for different values of m are tabulated in Hall et al. (1990, Table 1).

Table 3 presents the ordinary least squares estimates for the coefficients of Eq. 4, with m set to 10.[8] The standard errors are calculated with a heteroscedasticity-robust sandwich estimator.

The overall fit for Eq. 4 is remarkably good with an $R^2 = 0.830$.[9] Moreover, all of the estimated coefficients have reasonably signs and most of them are statistically significant at the usual levels. The price for a house increases, for instance, with both the floor size and the size and volume of all base areas in all storeys. The

[7] Wang et al. (2011) provide a technical discussion of what minimal smoothness assumptions are required for consistency. Moreover, their Monte Carlo simulations show that the estimator works well even if the unknown function $a(\cdot)$ is bumpy or has sharp boundaries.

[8] A difference order of $m = 10$ produces coefficient estimates that achieve approximately 95 percent efficiency relative to an estimator with the optimal rate of convergence (Yatchew, 1997).

[9] R^2 is computed with $1 - s_m^2/s_p^2$, where $s_m^2 = (N-m)^{-1} \sum_{i=1}^{N-m} (\Delta_m p_i - \mathbf{z}_i' \widehat{\boldsymbol{\beta}}_{\Delta_m})^2$, s_p^2 is the variance of p, and N is the number of observations (Yatchew, 1997, Proposition 1).

Table 3: Effect of building characteristics on house price. Table reports ordinary least squares estimates of Eq. 4. Continuous explanatory variables—floor size, gross area, gross volume, and age—are per sqm lot size. The gross volume of a building is used whenever the gross area was missing. Standard errors are calculated with heteroscedasticity robust sandwich estimator. *** significant at 1%-level ** significant at 5%-level * significant at 10%-level.

Dependent variable: Price per sqm lot size		
	Coef.	Std. Err.
Floor size	500.710	64.624***
Floor size squared	-0.784	0.446*
Gross area	688.681	36.731***
Gross area squared	-0.630	0.101***
Gross volume	248.891	18.367***
Gross volume squared	-0.060	0.020***
Floor size × age	-0.884	0.806
Gross area × age	-4.421	0.507***
Gross volume × age	-1.604	0.201***
Floor size × gross area	1.287	0.388***
Floor size × gross volume	0.315	0.210
Age	-198.249	106.371*
Age squared	9.235	1.144***
Semi-detached	2.497	3.492
Row house	-1.075	5.563
Good state of repair	86.035	3.672***
Poor state of repair	-73.713	3.977***
2 storeys	4.425	3.973
3 storeys	73.323	16.431***
Attic storey	10.301	3.241***
Flat roof	10.664	4.215**
No cellar	23.197	4.311***
Part cellar	-2.990	3.620
Unusual legal circumstances	-7.848	3.472**
Unusual physical circumstances	-24.015	6.204***
	Obs. 19,273	R^2 0.830

significant coefficients on the corresponding squared terms imply that these effects have diminishing rates. The age of the building, on the other hand, has a negative impact on the house price. The significant coefficient for the squared age term

implies a decreasing depreciation rate, which stays positive over the whole range of the age variable (when evaluated at the mean value of the size variables). The magnitudes of the estimated effects of the binary indicator variables are reasonable in sign and magnitude as well. Relative to the price of building with a normal state of repair buildings with a poor (good) state of repair, for instance, demand a price rebate (premium).

The estimator $\widehat{\boldsymbol{\beta}}_{\Delta_m}$ depends not only on m, but also on the ordering of observations regarding their geographical closeness. In the presented results the average distance between observations is about 95 meters with a standard deviation of 394 meters. To assess the impact of the nearest-neighbor algorithm on the estimated building values, we re-ran the regressions 50 times, each time with a different ordering. The within standard deviation of predicted building component for these runs is approximately 3% of the (average) building value.[10] Moreover, the coefficient of correlation between predicted building values from any two different runs is always well above 0.96. The results presented here are thus robust towards the specific ordering of observations.

Given $\widehat{\boldsymbol{\beta}}_{\Delta_m}$, we can adjust the per-square meter house prices for the buildung component and obtain the residuals $\widehat{u}_i = p_i - \mathbf{z}_i'\widehat{\boldsymbol{\beta}}_{\Delta_m}$. These first-stage residuals contain the location values to be extracted by the second stage of our method. Figure 3 shows box plots of these residuals for Berlin's districts.[11] For the plot, the residuals are normalized to the unit interval and the districts are ordered according to the median of the expert-based location ranking. Within each district, the residuals show substantial variation. This suggests that location values vary among areas within any given district.

[10] We assume that the building value accounts for 50% of the house price. Taking the mean house price and floor size in Table 1 then leads to an average building value per sqm of 935 Euros. The within standard deviation of the predicted building component is 27 Euro.

[11] The inner-city district of Kreuzberg-Friedrichshain has no single-family house neighborhoods and is not part of the plot.

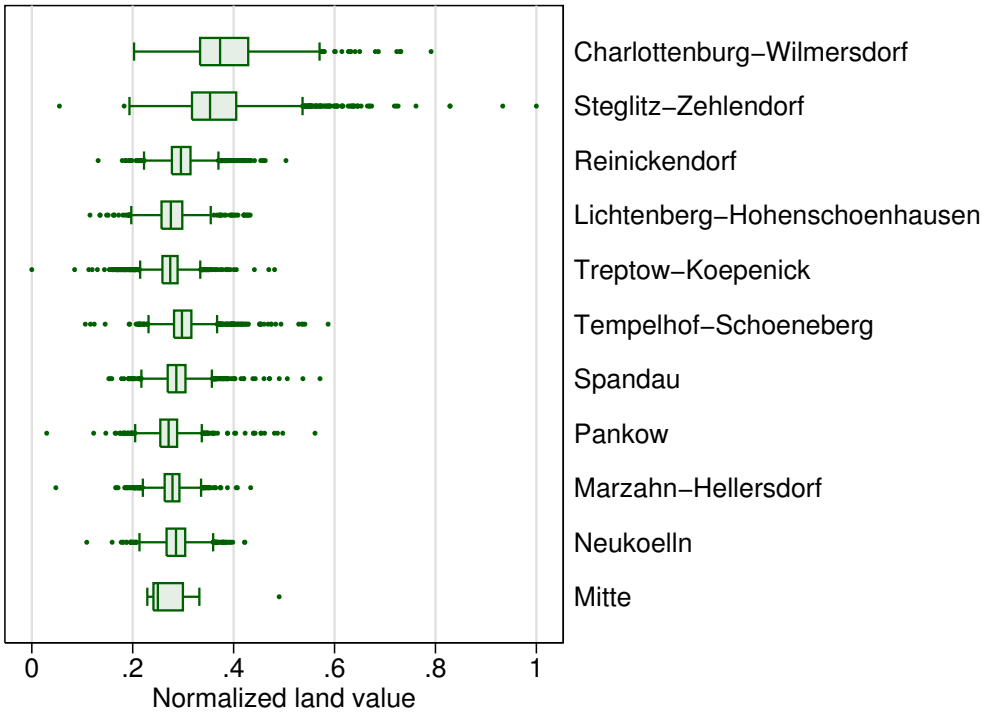

Figure 3: First-stage residuals by district. Shows box plots for the normalized first-stage residuals for Berlin's administrative districts. Number of observations is 19,283. Districts are sorted in descending order with respect to the median of the expert-based location rating. Line that separates the box is the median. Lower (upper) hinge of box represents 25th (75th) percentile. Length of whiskers is 1.5 times the IQR below (above) the 25th (75th) percentile.

2.3 Estimation of the location value surface

The second-stage of our method has two aims: (1) to separate the location values contained in the first-stage residuals $\widehat{u}_i$ from the transaction-specific noise and (2) to form areas with homogenous location values. To achieve these aims, we apply adaptive weights smoothing (AWS), a regression method developed by Polzehl and Spokoiny (2000, 2006) in the context of image denoising. AWS allows to separate the underlying structure in the data (e.g. the shape of an organ in an X-ray or ultrasonic image) from the distorting noise. AWS does not impose a priori assumptions on the form of this underlying structure (i.e. the regression function). Rather, AWS recovers the unknown regression function contained in the noisy data by an iterative, locally adaptive smoothing algorithm. In this algorithm, the local regression estimate is successively improved by searching for the largest vicinity of a nearly constant level of the regression function.

In our application, this amounts to finding the largest area around each location $\mathbf{l}_i$ in which the expected location value can be approximated well by a constant level. We denote this level at $\mathbf{l}_i$ by $a(\mathbf{l}_i)$. Similar to well-known smoothing methods such as kernel regression or nearest-neighbor estimation, AWS estimates $a(\mathbf{l}_i)$ by weighted local averaging over $\widehat{u}_j$s at $\mathbf{l}_i$. However, to determine the weight of observation j in forming the estimate of $a(\mathbf{l}_i)$, AWS does not only consider the distance between $\mathbf{l}_j$ and $\mathbf{l}_i$ (like other standard nonparametric smoothers do), but also adds a level penalty. Formally, the estimator at location $\mathbf{l}_i$ in the k-th iteration is defined as

$$\widehat{a(\mathbf{l}_i)}^{(k)} = \frac{\sum_{j=1}^{N} w_{ij}^{(k)} \widehat{u}_j}{\sum_{j=1}^{N} w_{ij}^{(k)}} \ , \tag{6}$$

where the weights are computed as

$$w_{ij}^{(k)} = K_{dist}\left(dist_{ij}^{(k)}\right) \times K_{lev}\left(lev_{ij}^{(k)}\right) \ . \tag{7}$$

The weight of observation j in the average formed at i is thus determined by a product of two kernel functions K. Both kernel functions are nonnegative and non-increasing on the positive semi-axis. That is, they give maximum weight if their

respective argument is zero and declining weights as their arguments increase. The arguments of these kernel functions are the distance penalty $dist_{ij}$ and the level penalty lev_{ij}, respectively. The distance penalty in iteration k is given by $dist_{ij}^{(k)} = |\rho(\mathbf{l}_i, \mathbf{l}_j)/h^{(k)}|^2$ where $\rho(\mathbf{l}_i, \mathbf{l}_j)$ is the Euclidean distance between the locations of observations i and j and $h^{(k)}$ is the bandwidth in iteration k. Hence, as in standard nonparametric regression, observation j will receive the more weight in the estimate at i, the closer its location to that of i. The level penalty in iteration k is computed as

$$lev_{ij}^{(k)} = \lambda^{-1} \underbrace{A_i^{(k-1)} \left\{ \widehat{a(\mathbf{l}_i)}^{(k-1)} - \widehat{a(\mathbf{l}_j)}^{(k-1)} \right\}^2}_{T_{ij}^{(k)}}, \tag{8}$$

This penalty is based on the comparison of the regression estimates at $\mathbf{l}_j$ and $\mathbf{l}_i$ in the previous iteration $(k-1)$. Hence, observation j will receive the more weight in iteration k, the closer its estimated level has been to that of observation i in the previous step. The term

$$A_i^{(k-1)} = \sum_{j=1}^{n} w_{ij}^{(k)} \tag{9}$$

equals the sum of the weights at i from the previous step and can be viewed as the local sample size that rescales the squared distance $\widehat{a(\mathbf{l}_i)}^{(k-1)} - \widehat{a(\mathbf{l}_j)}^{(k-1)}$. The product of these two terms, $T_{ij}^{(k)}$, can be viewed as a test statistic of the hypothesis $a(\mathbf{l}_i) = a(\mathbf{l}_j)$. Finally, the parameter λ chosen by the econometrician acts as a critical value for this test statistic: the larger λ, the smaller the impact of a particular deviation of $\widehat{a(\mathbf{l}_i)}$ from $\widehat{a(\mathbf{l}_j)}$ on the level penalty.

By amending the distance penalty of standard nonparametric estimation with a level penalty, AWS achieves both an extension of the scope of regression relations it can successfully tackle as well as an an increase in estimation efficiency. Both advantages will become clear when we complete our description of AWS by sketching the steps of its iterative algorithm. Further details are given in the Appendix.

In the initial step $(k = 0)$, the AWS estimator at $\mathbf{l}_i$ behaves like a standard kernel estimator by setting $w_{ij}^{(0)} = K_{dist}\left(dist_{ij}^{(0)}\right)$. That is, only the distance penalty is

considered for determining the weight of any observation j. In subsequent steps, the distance penalty is relaxed by successively increasing the location bandwidth according to the rule $h^{(k)} = ch^{(k-1)}$. The iterative algorithm terminates if $ch^{(k-1)} \geqslant h^*$ where the parameter c controls the bandwidth growth. We set the initial bandwidth $h^{(0)}$ and c according to the suggestions in Polzehl and Spokoiny (details are given in the Appendix).

Hence, successively more distant observations are considered for forming the local average at $\mathbf{l}_i$. The level penalty, which kicks in at iteration $k = 1$, ensures that this is justified. More distant observations may belong to locations where the the expected location value may be quite different from $a(\mathbf{l}_i)$, resulting in a biased estimate. This, however, is prevented by a large level penalty which effectively leads to the exclusion of such an observation from the computation of $\widehat{a(\mathbf{l}_i)}^{(k)}$. If, on the other hand, the current assessment of the expected location value at observation j, i.e. $\widehat{a(\mathbf{l}_j)}^{(k-1)}$, is close to that at observation i, then observation j does receive weight despite its potentially substantial distance in location from $\mathbf{l}_i$.

By relaxing the distance penalty and at the same time enforcing the level penalty, AWS identifies at any location the largest contiguous area of a nearly constant level of the expected location value. Unlike standard nonparametric smoothers, it thus allows more distant observations to be included in an estimate at any location as long as this is justified by homogeneity in expected location values. This not only increases the efficiency of the estimate (from the resulting increase in the local sample size), it also enables to identify shapes of regression relations that standard smoothers can not pick up. This modeling advantage is most pronounced in situations where the underlying regression function allows a piecewise constant approximation with large homogenous regions that are allowed to sharply differ at the boundaries.

The flexible AWS procedure involves several parameters that must be speci-

fied, in particular the smoothing parameters of both penalties.[12] Since the location penalty is successively relaxed during the algorithm, the choice of its bandwidth, h, is much less important for AWS than for standard Kernel regression. The key parameter of AWS is λ, the factor that scales the level penalty. Too small values of λ will result in an over-penalization of level differences between neighboring bins. As a result, areas of homogeneous location values may not be properly identified. Too large values of λ, on the other hand, will result in a loss of sensitivity towards discontinuities in location values. Neighboring bins may be joined in this case to form an area of a common level of location values when this is not warranted. To resolve this trade-off, Polzehl and Spokoiny consider the (hypothetical) situation of a constant value surface. In this case, the final estimate of AWS should coincide with high probability with the globally constant location value. They suggest using the minimal value of λ that ensures this 'propagation condition'. This value of λ does not depend on the particular globally constant location value and can be obtained from Monte Carlo simulations. This is the default value of λ in the contributed package 'aws' of the R-Project for Statistical Computing (Polzehl, 2011). We use this package to implement AWS.

AWS is designed to work on matrices. In our application, the matrix is a grid with the two dimensions 'latitude' and 'longitude' placed over the map of Berlin. A matrix where each of the 19,283 observations has its own bin has over 1.5 billion entries. This matrix is too large for the algorithm. To reduce the size of the matrix, we use binning (Fan and Marron, 1994). We generate a $300 \times 300 = 90,000$ grid and allocate the observations to bins with the grid points as centers. Each bin has an approximate size of 171×114 meters. Applied to the data, 7,704 bins contain at least one observation, 3,354 bins contain exactly one, the average count per bin is 2.5 observations, and the maximum is 49.

Figure 4 plots the estimated location values $\widehat{a(\mathbf{l})}$ for the bins within a map of Berlin. Estimates are normalized to the unit interval and coloring is used to represent

[12] Details on our choices of AWS' parameters are given in the Appendix

their magnitudes. The plot illustrates both the functioning and the advantages of AWS. Binning is visible from the somewhat angular appearance of similar colored areas but otherwise the colors, shapes and size of these areas is data-driven and locally adaptive. AWS identified these areas by relaxing the location penalty in successive iterations and implicitly testing for local homogeneity of location values. As long as the location values are sufficiently similar, relaxing the location penalty is justified and adjacent bins are subsumed into an area.

3 Comparison with expert-based location assessments

Section 2.1 showed that the expert-based land values and location ratings agree in many cases regarding their assessment of the value of a location. We also expect a strong positive relationship between the expert-based assessments and the estimated location values, $\widehat{a(\mathbf{l})}$. As the expert-based assessments do not agree in all instances, we will compare our estimated location values with each of them.

Figure 2 shows in its right panel box plots of the estimated location value, $\widehat{a(\mathbf{l})}$, for the four levels of the expert-based rating. Similar to the expert-based land values (shown in the left panel of the figure), the medians of the estimated location values increase in line with the expert-based location rating; the quartiles of the estimated location values for locations with low and medium rating overlap also. The variance of the estimated location values is higher than the variance of the expert-based land values, except for locations with low rating. This is attributable to few first stage residuals that are rather large or small. These residuals could be the result of mis-specifications of the first stage regression or could be the result of aberrant idiosyncratic effects during the transaction.

Table 2 shows in Panel B the matching frequencies for the estimated location values and the expert-based location rating. As in Section 2.1, the estimated location values are converted into an ordinal rating. Even though the matching is not perfect, the majority of pairs lie on the diagonal of the contingency table. The statistic for

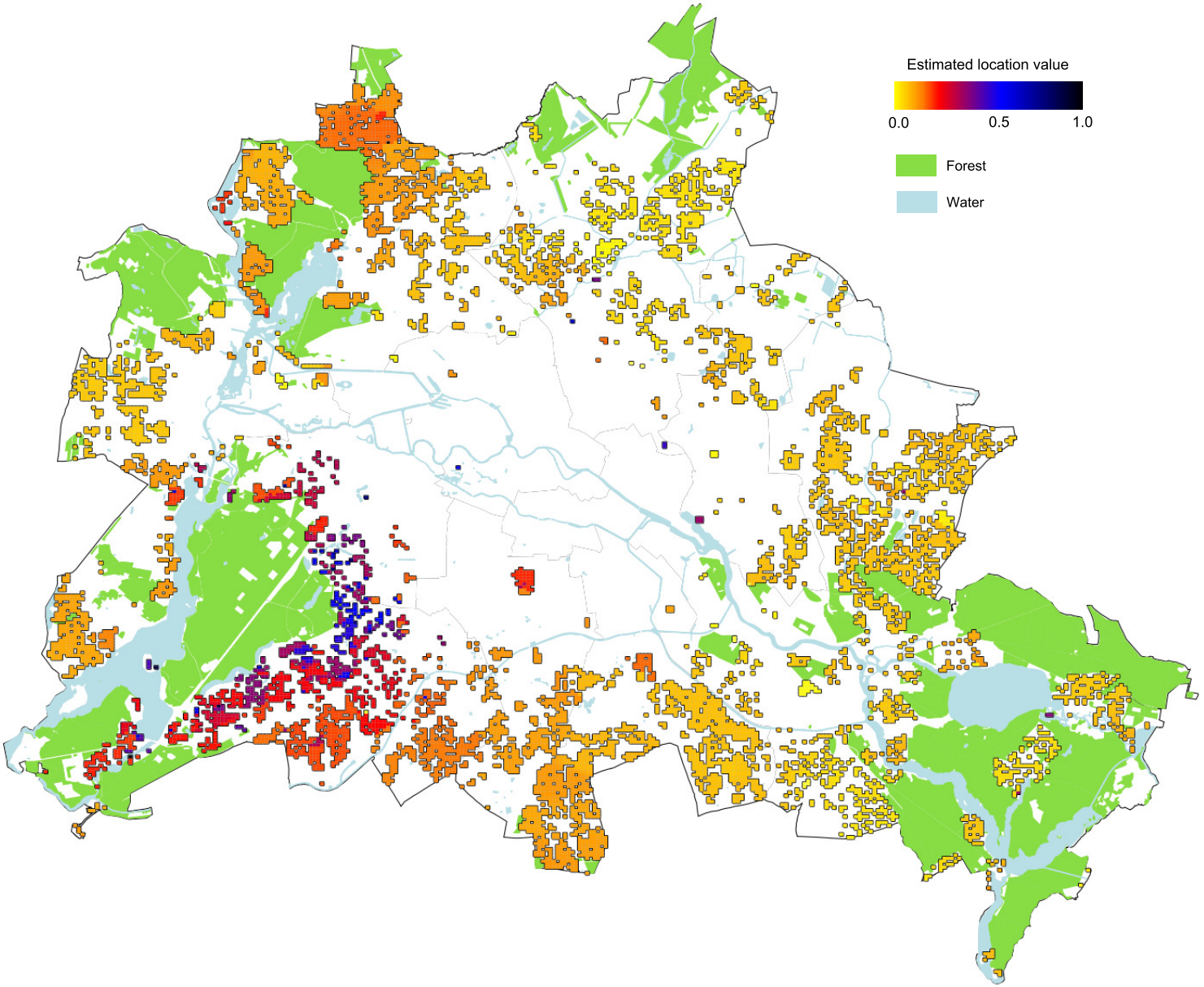

Figure 4: Estimated location value surface. Map of Berlin with the estimated location values $\widehat{a(\mathbf{l})}$.

the chi-square independence test is $10,521$, which is a highly unlikely realization under a $\chi^2(9)$-distribution. We therefore reject the null of statistical independence between the two ratings at the usual significance levels. Estimates of Goodman and Kruskal's γ ($\widehat{\gamma} = 0.644$) and Kendall's τ ($\widehat{\tau} = 0.466$) indicate the expected positive relationship between both location assessments. In order to compare the estimated

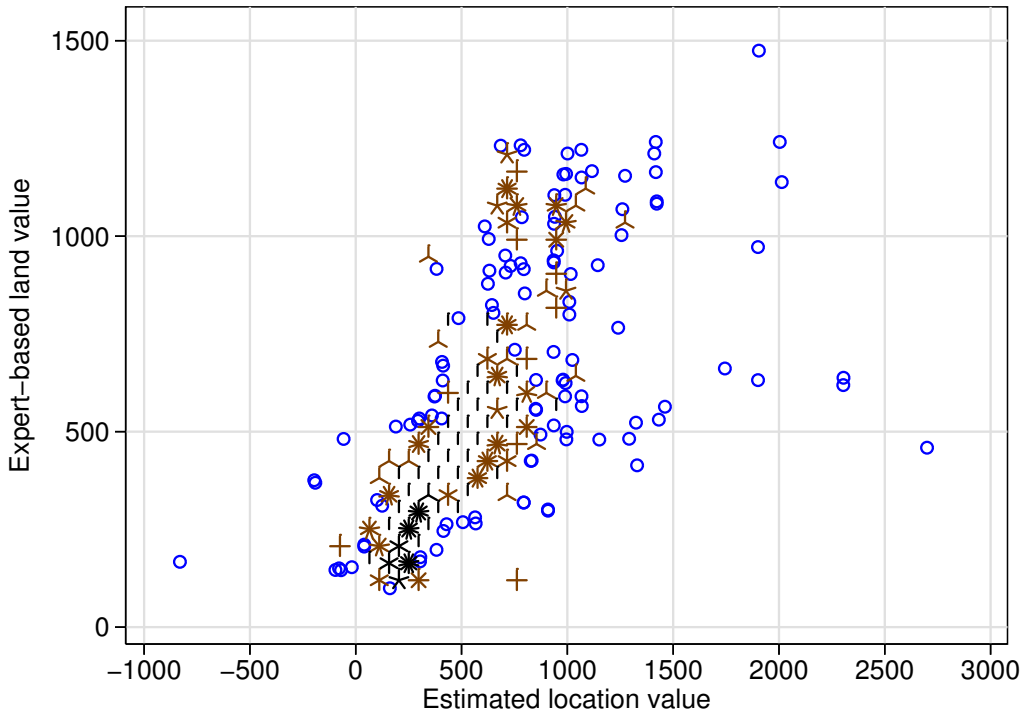

Figure 5: Sunflower plot of expert-based land values and estimated location value. Shows sunflower plot of the bin average of expert-based land value and estimated location value, $\widehat{a}(l)$. Both figures are in real (year 2000) Euros. Number of observations is 7,704. Each petal of a light sunflower represents 1 observation. Each petal of a dark sunflower represents several observations. Circles represent individual observation in low density region.

location values with the expert-based land values, we rescale $\widehat{a(\mathbf{l})}$ so that its median equals the median of the expert-based land values. The estimated location values and the expert-based land values are both estimates of the true but unobserved location value and we expect a strong positive correlation between them. Figure 5

shows a sunflower plot of the estimated location value and the expert-based land value. To work at the same level of geographical detail, the plot uses averages of the expert-based land values within bins.[13] The plot represents the density of observations using stylized sunflowers. In a light sunflower, each petal represents one observation. In a dark sunflower, each petal represents several observations.[14]

The expected positive correlation between both location assessments is visible and strong, with a coefficient of correlation 0.840.[15] The majority of paired observations lie on the 45 degree line, although a few particulary large (small) outliers are apparent again.

4 Conclusion

In this paper, we proposed a novel semiparametric method to extract location values out of house prices. The first stage of the method separates the price into a building component and a land component. The second stage employs adaptive weights smoothing (AWS), a nonparametric method to separate the residual from the first stage into the location value and a noise term. Using AWS has several advantages over standard nonparametric regression. It allows the size and shape of areas with a common location value to be completely determined by the data. They need not be symmetric or adhere to a particular shape. Moreover, unlike kernel regression, AWS does not require the location value surface to be smooth.

We apply the method to single-family house transactions from Berlin and obtain

[13] The binning procedure removes potentially variation in the estimated location values that is still present in the expert-based land values. In particular, differences in the quality of locales within a bin, as well as differences in the time trend of land values between bins are a priori removed by binning and discounting with a Berlin-wide quality-adjusted land value index. Both effects are rather small as indicated by the average standard deviation of the expert-based land value within bins, which is only 13.31 Euros.

[14] A dark sunflower with p petals represents between $p96 - 96/2$ and $p96 + 96/2$ observations.

[15] Using the 19,283 individual observations gives a coefficient of correlation of 0.836.

reliable results in the sense that they show agreement with expert-based location assessments. In particular, the estimated location values are highly correlated with, both, land values and ordinal location ratings that are provided by real estate experts. In summary, the estimated surface provides a comprehensive characterization of the relative location values of Berlin's residential areas. The methodology should thus prove useful for applications where location values are needed and no expert-based information is available or where such information should be complemented by data-driven flexible location value estimates.

A Appendix

A.1 Nearest-neighbor algorithm

The nearest-neighbor algorithm used in the estimation of the building component works as follows:

1. Initialization: Start with an arbitrary observation.
2. Iteration: Find its nearest neighbor with respect to their Euclidian distance and mark the observation as visited.
3. Stopping: Go back to Step 2 until all observations have been visited.

The resulting sequence of the visited locations provides the ordered observations. The nearest neighbor algorithm is easy to implement and computationally fast, but can lead to slightly different ordering sequences depending on the initial observation.

A.2 AWS algorithm

The AWS algorithm can be summarized as follows:

1. Initialization: The parameters λ, $h^{(0)}$, c and h^* are selected and the location weights

$$w_{ij} = K_{dist}\left(\left|\frac{\rho\left(\mathbf{l}_i, \mathbf{l}_j\right)}{h^{(0)}}\right|^2\right)$$

and presmoothed estimates

$$\widehat{a\left(\mathbf{l}_i\right)}^{(0)} = \frac{\sum_j w_{ij}^{(0)} \hat{u}_j}{\sum_j w_{ij}^{(0)}}$$

are calculated for all i, j.

2. Iteration: In each iteration k the following steps are performed for every design point, $\mathbf{l}_i$, on the grid.

   - Calculate the adaptive weights: For every point $\mathbf{l}_j$ within the bandwidth $h^{(k)}$ around point $\mathbf{l}_i$ the penalties

$$dist_{ij}^{(k)} = \left|\frac{\rho\left(\mathbf{l}_i, \mathbf{l}_j\right)}{h^{(k)}}\right|^2,$$
$$lev_{ij}^{(k)} = \lambda^{-1} A_i^{(k-1)}\left(\widehat{a\left(\mathbf{l}_i\right)}^{(k-1)} - \widehat{a\left(\mathbf{l}_i\right)}^{(k-1)}\right)^2, \quad A_i^{(k-1)} = \sum_{j=1}^{n} w_{ij}^{(k-1)}$$

     are computed and the weights are formed by $w_{ij} = K_{dist}\left(dist_{ij}^{(k)}\right) \times K_{lev}\left(lev_{ij}^{(k)}\right)$.

   - Estimation: For every design point $\mathbf{l}_i$ the updated estimate

$$\widehat{a\left(\mathbf{l}_i\right)}^{(k)} = \frac{\sum_j w_{ij}^{(k)} \hat{u}_j}{\sum_j w_{ij}^{(k)}},$$

     and the sum of weights $A_i^{(k)}$ are calculated.

3. Stopping: If $ch^{(k)} \geq h^*$, the algorithm terminates. Otherwise the bandwidth is set to $h^{(k)} = ch^{(k-1)}$ and the algorithm continues with step 2.

A.3 AWS parameters

Several parameters affect the performance of the AWS estimation procedure. These parameters control constituent features of the method and should be chosen cautiously in light of the application at hand. In the following, we describe the choice of parameters for our application.

***c*, $h^{(0)}$, $h^{(1)}$, h^*, *k*:** The number of iterations of AWS is determined by the maximal bandwidth h^*. With every iteration k the bandwith is incremented by the factor $c = (1.25)^{\frac{1}{d}}$ where $d = 2$ is the dimension of the sample space. The algorithm terminates if $ch^{(k-1)} \geq h^*$. Unlike classical nonparametric approaches, AWS does not necessarily suffer from oversmoothing. A relatively large maximal bandwidth h^* will, however, result in more iterations and thus increases the computational complexity. We set h^* to 15 which allows that quite far away points $\mathbf{l}_j$ are (potentially) used to form an estimate for point $\mathbf{l}_i$, but keeps the computional complexity reasonably low. With respect to the initial bandwidth $h^{(0)}$ and subsequent bandwiths $h^{(1)}$ we follow the suggestions in Polzehl and Spokoiny. In particular, we select $h^{(0)} = 9.6$ and $h^{(1)} = 0.74$, respectively. Both bandwidths are small enough so that the former i contains a sufficient number of design points in the initial iteration and the latter does not increase the bandwidth too much in every iteration.

***K*$_{dist}$, *K*$_{lev}$:** The kernel functions are defined on the positive semiaxis and fulfill $K_{dist}(0) = K_{lev}(0) = 1$ for the case of a perfect match (in distance or level). In general, K_{dist} scales the distance penalty and is non-decreasing, whereas K_{lev} decreases on the positive semiaxis. For both kernel functions we use the triangular kernel with

$$K_{dist}(dist_{ij}) = (1 - dist_{ij})\,\mathbf{1}_{\{+\}} \text{ and } K_{lev}(lev_{ij}) = (1 - lev_{ij})\,\mathbf{1}_{\{lev_{ij} \leq 1\}}\,, \quad \text{(A1)}$$

where lev_{ij} is always positive (see Eq. 8).

λ: The parameter λ scales the level penalty and is hence the most critical parameter of AWS. We follow Polzehl and Spokoiny and obtain $\lambda = 21.96$ by monte carlo simulations. In particular, we compare the standard deviations of two estimates for a model with a (known) globally constant parameter value $a(\mathbf{l}_i) = a$ for all i. The first estimate, $\widehat{a(\mathbf{l}_i)}$, is obtained by AWS given a certain λ; the second estimator, $\widetilde{a(\mathbf{l}_i)}$, uses a very large value of λ so that the AWS estimator converges to a nonadaptive kernel estimator. We search for the smallest λ which fulfills the following inequality

$$\mathbf{E}\left|\widehat{a(\mathbf{l}_i)}^k - a(\mathbf{l}_i)\right| \leq (1+\alpha)\,\mathbf{E}\left|\widetilde{a(\mathbf{l}_i)}^k - a(\mathbf{l}_i)\right|, \quad \text{with } \alpha = 0.05. \tag{A2}$$

The intuition of this approach is to choose the minimal λ so that AWS still recovers the global constant parameter value (with a probability α) while using the most adaptive bandwidth choice.

References

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

Bourassa, S. C., Hoesli, M., Scognamiglio, D. and Zhang, S.: 2011, Land leverage and house prices, *Regional Science and Urban Economics*, **41**, 134–144.

Cheshire, P. and Sheppard, S.: 1995, On the price of land and the value of amenities, *Economica*, **62**, 247–267.

Clapp, J. M.: 2004, A semiparametric method for estimating local house price indices, *Real Estate Economics*, **32**, 127–160.

Colwell, P. F. and Munneke, H. J.: 2003, Estimating a price surface for vacant land in an urban area, *Land Economics*, **79**, 15–28.

Fan, J. and Marron, J. S.: 1994, Fast implementations of nonparametric curve estimators, *Journal of Computational and Graphical Statistics*, **3**, 35–56.

Hall, P., Kay, J. W. and Titterington, D. M.: 1990, Asymptotically optimal difference-based estimation of variance in nonparametric regression, *Biometrika*, **77**, 521–528.

Polzehl, J.: 2011, AWS: adaptive weights smoothing, R package version 1.7-1, http://CRAN.R-project.org/package=aws.

Polzehl, J. and Spokoiny, V.: 2000, Adaptive weights smoothing with applications to image restoration, *Journal of the Royal Statistical Society Series B*, **62**, 335–354.

Polzehl, J. and Spokoiny, V.: 2006, Propagation-separation approach for local likelihood estimation, *Probability Theory and Related Fields*, **135**, 335–362.

Rosenthal, S. S.: 1999, Residential buildings and the cost of construction: New evidence on the efficiency of the housing market, *Review of Economics and Statistics* **81**, 288–302.

Rossi-Hansberg, E., Sarte, P.-D. and Owen III, R. O.: 2010, Housing externalities, *Journal of Political Economy*, **118**, 485–535.

Schulz, R. and Werwatz, A.: 2011, Is there an equilibrating relationship between house prices and replacement cost? Empirical evidence from Berlin, *Journal of Urban Economics*, **69**, 288–302.

Wang, L., Brown, L. D., and Cai, T. T.: 2011, A difference based approach to the semiparametric partial linear model, *Electronic Journal of Statistics*, **5**, 619–641.

Yatchew, A.: 1997, An elementary estimator of the partial linear model, *Economic Letters*, **57**, 135–143.