

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

Conference Paper

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

Beiträge zur Jahrestagung des Vereins für Socialpolitik 2013: Wettbewerbspolitik und Regulierung in einer globalen Wirtschaftsordnung - Session: Housing, No. G10-V1

Provided in Cooperation with:

Verein für Socialpolitik / German Economic Association

Suggested Citation: Kolbe, Jens; Schulz, Rainer; Wersing, Martin; Werwatz, Axel (2013) : Location, location, location: Extracting location value from house prices, Beiträge zur Jahrestagung des Vereins für Socialpolitik 2013: Wettbewerbspolitik und Regulierung in einer globalen Wirtschaftsordnung - Session: Housing, No. G10-V1, ZBW - Deutsche Zentralbibliothek für Wirtschaftswissenschaften, Leibniz-Informationszentrum Wirtschaft, Kiel und Hamburg

This Version is available at:

<https://hdl.handle.net/10419/79732>

Standard-Nutzungsbedingungen:

Die Dokumente auf EconStor dürfen zu eigenen wissenschaftlichen Zwecken und zum Privatgebrauch gespeichert und kopiert werden.

Sie dürfen die Dokumente nicht für öffentliche oder kommerzielle Zwecke vervielfältigen, öffentlich ausstellen, öffentlich zugänglich machen, vertreiben oder anderweitig nutzen.

Sofern die Verfasser die Dokumente unter Open-Content-Lizenzen (insbesondere CC-Lizenzen) zur Verfügung gestellt haben sollten, gelten abweichend von diesen Nutzungsbedingungen die in der dort genannten Lizenz gewährten Nutzungsrechte.

Terms of use:

Documents in EconStor may be saved and copied for your personal and scholarly purposes.

You are not to copy documents for public or commercial purposes, to exhibit the documents publicly, to make them publicly available on the internet, or to distribute or otherwise use the documents in public.

If the documents have been made available under an Open Content Licence (especially Creative Commons Licences), you may exercise further usage rights as specified in the indicated licence.

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

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

February 28, 2013

*Kolbe: Deutsches Institut für Wirtschaftsforschung, Mohrenstrasse 58, 10117 Berlin Germany. Email: jkolbe@diw.de. Wersing and Schulz: University of Aberdeen Business School, Edward Wright Building, Dunbar Street, Aberdeen AB24 3QY, United Kingdom. Email: martin.wersing@abdn.ac.uk and r.schulz@abdn.ac.uk. Werwatz: 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: axel.werwatz@tu-berlin.de.

Abstract

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 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 finest 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

¹The 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, say 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 Gençay (1996) and Clapp (2004), among others, also fitted semiparametric models to house prices, but with more restrictive and less flexible location 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.

²Lack 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

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

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

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

Here, p and \mathbf{z} denote the house price and the building characteristics per square meter lot size.³ We assume that the noise term is mean independent from \mathbf{z} and \mathbf{l} , which implies $E(\varepsilon|\mathbf{z}, \mathbf{l}) = 0$.

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

³The continuous building characteristics (per sqm) may be transformed further to capture non-linearities in the hedonic price function.

of the parametric component in Eq. 2 and compute the residual $u = p - \mathbf{z}'\boldsymbol{\beta}$, which 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).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 also an expert-based land value and expert-based location rating. These assessments will serve as benchmarks for our estimated location values and are further described in Section 3 below.

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

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

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

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

damage or flooding risk and legal aspects such as rights of way or use for pipes or cables. Such easements are rather common.

[Table 1 about here]

Another important source of variation is the location of a house within the city, as indicated by the map plotted in Figure 1.

[Figure 1 about here]

The area of Berlin is 891 km², 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.

2.2 Estimation of the building component

We use the estimator proposed by Yatchew (1997) and Wang et al. (2011) for the estimation of β , the vector of coefficients of the building characteristics. The basic idea of the estimator is that the location value $a(\mathbf{l})$ 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.⁷

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

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

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 β can then be

⁷The Appendix explains the algorithm.

estimated consistently with ordinary least squares.⁸ The simulations in Wang et al. (2011) show that the estimator works well even if the unknown function $a(\bullet)$ is bumpy or has sharp boundaries.

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

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

The weights fulfill the two restrictions

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

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 $\hat{\boldsymbol{\beta}}_{\Delta_m}$ 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 2 presents the ordinary least squares estimates for the coefficients of Eq. 4, with m set to 10.⁹ The standard errors are calculated with a heteroscedasticity-robust sandwich estimator.

[Table 2 about here.]

The overall fit for Eq. 4 is remarkably good with an $R^2 = 0.830$.¹⁰ Moreover, all of the estimated coefficients have reasonable signs and most of them are statistically

⁸Wang et al. (2011) provide a technical discussion of what minimal smoothness assumptions are required for consistency.

⁹A 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).

¹⁰ R^2 is computed with $1 - s_m^2/s_p^2$, where $s_m^2 = (N - m)^{-1} \sum_{i=1}^{N-m} (\Delta_m p_i - \mathbf{z}_i' \hat{\boldsymbol{\beta}}_{\Delta_m})^2$, s_p^2 is the variance of p , and N is the number of observations (Yatchew, 1997, Proposition 1).

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 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. For instance, relative to the price of a building with a normal state of repair buildings with a poor (good) state of repair demand a price rebate (premium).

The estimator $\hat{\beta}_{\Delta_m}$ 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.¹¹ 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 slightly different orderings of the observations.

Given $\hat{\beta}_{\Delta_m}$, we can adjust the per-square meter house prices for the building component and obtain the residuals $\hat{u}_i = p_i - \mathbf{z}'_i \hat{\beta}_{\Delta_m}$. These first-stage residuals contain the location values to be extracted by the second stage of our method. Figure 2 shows box plots of these residuals for Berlin's districts.¹² For the plot, the

¹¹We assume that the building value accounts for 50% of the house price. The mean house price and floor size in Table 1 then imply an average building value per sqm of 935 Euros. The within standard deviation of the predicted building component is 27 Euro.

¹²The inner-city district of Kreuzberg-Friedrichshain has no single-family house neighborhoods and is not part of the plot.

residuals are normalized to the unit interval and the districts are ordered according to the median of the expert-based location ranking.

[Figure 2 about here.]

Within each district, the residuals show substantial variation. This suggests that location values vary among areas within any given district.

2.3 Estimation of the location value surface

The second-stage of our method has two aims: (i) to separate the location values contained in the first-stage residuals \hat{u}_i from the transaction-specific noise and (ii) 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). AWS allows to separate the underlying structure in the data 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 $a(\mathbf{l}_i)$ can be approximated well by a constant level. 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 \hat{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

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

where the weights are computed as

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

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.¹³ 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

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

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

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

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 $a(\mathbf{l}_i) = a(\mathbf{l}_j)$. The parameter λ acts as a critical value for this test statistic: if $T_{ij}^{(k)} > \lambda$ observation j does not receive weight in the estimate at i .¹⁴ Thus, the

¹³In our empirical application, we use the triangular kernel for, both, $K_{dist}(\bullet)$ and $K_{lev}(\bullet)$.

¹⁴ $K_{lev}(\bullet)$ has bounded support $[-1, 1]$.

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 increase in estimation efficiency. Both advantages will become clear when we complete our description of AWS by sketching the steps of its iterative algorithm.¹⁵

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 $ch^{(k-1)} \geq h^*$ where the parameter c controls the bandwidth growth.

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 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

¹⁵The Appendix summarizes the algorithm and gives details on the choice of smoothing parameters.

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 AWS procedure requires the choice of smoothing parameters for the distance and level penalty. Since the distance penalty is successively relaxed during the algorithm, the choice of its bandwidth, h , is much less important than for standard Kernel regression. The key smoothing parameter 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 observations. As a result, areas of homogeneous location values may not be identified. Too large values of λ , on the other hand, will result in a loss of sensitivity towards discontinuities in location values. Neighboring observations 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, we 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. We use 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 we obtain it from Monte Carlo simulations.¹⁶

In our empirical application we use binning to reduce the computational burden of AWS (Fan and Marron, 1994).¹⁷ We therefore generate a 300×300 grid with the two dimensions ‘latitude’ and ‘longitude’ and allocate the observations to bins with the grid points as centers. Each bin has an approximate size of 171×114 meters. 7,704

¹⁶The Appendix explains the simulation.

¹⁷The computational burden of obtaining estimates of $a(\mathbf{l}_i)$ at every location i is not a specific feature of AWS but common to all nonparametric smoothers.

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. We estimate location values for each nonempty bin.

Figure 3 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 shading is used to represent their magnitudes.

[Figure 3 about here.]

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 are data-driven and locally adaptive. AWS identified these areas by relaxing the distance penalty in successive iterations and implicitly testing for local homogeneity of location values. As long as the location values are sufficiently similar, relaxing the distance penalty is justified and adjacent bins are subsumed into an area.

3 Comparison with expert-based location assessments

In order to evaluate the adequacy of the estimated location values we compare them with two expert-based location assessments. Our first benchmark is an estimate of the notional value of land as if it were undeveloped. The land values are computed by GAA appraisers using the sales comparison approach based on information from transactions of undeveloped land. The estimated location value and the expert-based land value are both estimates of the true but unobserved location value. We therefore expect to find a strong positive correlation between the two of them. Our second benchmark is 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 particular location.¹⁸ For this

¹⁸The geographic unit of a location corresponds to a street block.

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. We expect that the estimated location values, as well as the expert-based land values agree in many cases with the expert-based location ratings regarding their assessment of a location.

Summary statistics of the expert-based land value and expert-based location rating are presented in the last part of Table 1. Both variables 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 to derive an assessment of a particular location.

Figure 4 shows a sunflower scatter plot of the estimated location value, $\widehat{a(\mathbf{I})}$, and the expert-based land value. The plot represents the density of observations in a region using stylized sunflowers. In a light sunflower, each petal represents one observation. In a dark sunflower, each petal represents several observations.¹⁹ All values are expressed in Euros per square meter. Since our semiparametric estimator does not identify the level of location values, we normalize $\widehat{a(\mathbf{I})}$ so that location and land values have the same median. To work at the same level of geographical detail, the plot uses the within-bin average of the expert-based land value.²⁰

[Figure 4 about here.]

The expected positive correlation between both location assessments is visible and strong, with a coefficient of correlation 0.845. The majority of paired observations lie on the 45 degree line, although a few outliers with particularly large (small) values of $\widehat{a(\mathbf{I})}$ are apparent. The outliers are attributable to few first stage residuals that are rather large or small. These residuals could be the result of mis-specifications

¹⁹A dark sunflower with p petals represents between $p96 - 96/2$ and $p96 + 96/2$ observations.

²⁰Binning removes differences in the quality of locales within a bin that is still present in the expert-based land values.

of the first stage regression or could be the result of aberrant idiosyncratic effects during the transaction.

The plot in Figure 4 shows that the expert-based land values and the estimated location values, on average, conform to each other. We also expect that the land value of a house located in a nice area is higher than if the house is located in a dilapidated area. Figure 5 shows in its right panel box plots of the estimated location value, $\widehat{a(1)}$, 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, both, the estimated location values and the expert-based land values for locations with low and medium rating, however, overlap to a large extent. In both cases, the separation for the top two levels of the expert-based rating is more pronounced.

[Figure 5 about here.]

Notably, the variation of the estimated location values is higher than the variation of the expert-based land values, except for locations with low rating. As discussed above, this is attributable to outliers in the first-stage residuals.

To formally test if the estimated location values and expert-based land values agree with the location ratings, we convert, both, the location and land values into an ordinal rating. In this conversion, the 2% largest location and land values receive the rating ‘excellent’, the next 20% values the rating ‘high’ and so forth. Constructed this way, the ordinal ratings based on the land or location values have the same marginal distribution as the expert-based location rating. Panel A and B of Table 3 give the matching frequencies of the two constructed ratings and the expert-based location rating. If one of the constructed ratings and the expert-based location ratings were identical, then the respective contingency matrix would have the marginal frequencies on the diagonal and zeros elsewhere.

[Table 3 about here.]

For both constructed ratings, this is not the case. They are also not independent of the expert-based rating, as a comparison with Panel C shows. The panel gives the frequencies that we would expect if matching were random. In case of the rating based on the estimated location values in Panel A the test statistic of the chi-square test for statistical independence is 11,025. This is a highly unlikely realization under a $\chi^2(9)$ -distribution and we reject the null of statistical independence. For the rating based on the expert-based land values in Panel B, the null hypothesis of the chi-square test for statistical independence is rejected as well.

The strength of the relationship between any two ratings can be measured with Goodman and Kruskal's γ , and Kendall's τ , respectively. Both measures are rank correlation coefficients that range from -1 (perfect inversion) to $+1$ (perfect agreement). In Panel A, we estimate $\hat{\gamma} = 0.662$ and $\hat{\tau} = 0.459$, indicating the expected positive relationship between the estimated location values and expert-based location rating. This is also true in Panel B, where the estimated rank correlation coefficients are of very similar magnitude. Locations that are rated to have a higher level of amenities have also a higher location value.

4 Conclusion

In this paper, we proposed a novel semiparametric method to extract location values out of house prices. Both stages of this method exploit the availability of the location geo-codes in the data. The first stage of the method separates the price into a building component and a land component by working with price differences of nearby observations. Such properties will have similar location values. Their price difference will thus primarily reflect differing building characteristics. 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. As illustrated by our application, these areas 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.

AWS identified these areas by relaxing the distance penalty in successive iterations and implicitly testing for local homogeneity of location values. As long as the location values are sufficiently similar, relaxing the distance penalty is justified and adjacent areas are subsumed into one.

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

Acknowledgements

We have benefited from comments on earlier versions of this paper by Jörg Polzehl and Verity Watson. 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.

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 the 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. Depending on the initial observation, the nearest neighbor algorithm can lead to slightly different ordering sequences and sometimes misses shorter routes. As a rule of thumb, if the last few stages of the sequence are comparable in length to the first stages, then the ordering is reasonable.

A.2 AWS algorithm

The AWS algorithm used to estimation the location value surface can be summarized as follows:

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

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

and presmoothed estimates

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

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

$$dist_{ij}^{(k)} = \left| \frac{\rho(\mathbf{l}_i, \mathbf{l}_j)}{h^{(k)}} \right|^2,$$

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

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

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

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.

We use the contributed package ‘aws’ of the R-Project for Statistical Computing (Polzehl, 2011) to implement AWS. The choice of smoothing parameters is explained in Appendix A.3.

A.3 Choice of smoothing parameters

Parameters for distance penalty: $h^{(0)}$, $h^{(1)}$, h^* , c ,

We set the maximal bandwidth h^* to 75 bins which allows that quite far away points \mathbf{l}_j are (potentially) used to form an estimate for point \mathbf{l}_i . With every iteration k the bandwidth 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 $ch^{(k-1)} \geq h^*$. With respect to the initial bandwidth $h^{(0)}$ and subsequent bandwidths $h^{(1)}$, we select $h^{(0)} = 9.49$ and $h^{(1)} = 1.06$, respectively. Both bandwidths are small enough so that the former contains a sufficient number of design points in the initial iteration and the latter does not increase the bandwidth too much in every iteration.

Parameter for level penalty: λ

We choose λ as the smallest value satisfying a propagation condition. This condition requires that for a model with a globally constant parameter value, $a(\mathbf{l}_i) = a$, the AWS estimator approximately behaves like its nonadaptive counterpart. Since the value λ provided by the propagation condition does not depend on a , λ can be approximately found by simulations. Formally, we search for the smallest λ which fulfills the following inequality at every location

$$\mathbf{E} \left[\left(\widehat{a(\mathbf{l}_i)} - a \right)^2 \right] \leq (1 + \alpha) \mathbf{E} \left[\left(\widetilde{a(\mathbf{l}_i)} - a \right)^2 \right], \quad \text{with } \alpha = 0.05, \quad (\text{A1})$$

where $\widehat{a(\mathbf{l}_i)}$ is the AWS estimate and $\widetilde{a(\mathbf{l}_i)}$ uses $\lambda = \infty$ which leads to a nonadaptive kernel estimate. Both estimators employ the same bandwidth h^* . The intuition of this approach is to choose the minimal λ that recovers the global constant parameter value while using the most adaptive bandwidth choice. A theoretical justification is given in Polzehl and Spokoiny (2006, Theorem 5.1). We obtain $\lambda^* = 24.18$ by the Monte Carlo simulation explained in Appendix A.4.

A.4 Monte Carlo simulation

The Monte Carlo simulation used to obtain the smoothing parameter λ can be summarized as follows:

- Generate S ($S = 500$) data sets with N ($N = 20,000$) observations from the globally constant model

$$y_i^{(s)} = a(\mathbf{l}_i)^{(s)} + \nu_i^{(s)}, \quad (\text{A2})$$

where the superscript s indicates the data set. We set $a(\mathbf{l}_i)^{(s)} = 0$ for all locations \mathbf{l}_i and $\nu_i^{(s)} \sim N(0, \sigma_\nu^2)$. The variance of the error term is $\sigma_\nu^2 = \sigma_{\hat{u}|\mathbf{l}_i}^2 = 126.57^2$, where $\sigma_{\hat{u}|\mathbf{l}_i}^2$ denotes the conditional variance of the first-stage residuals \hat{u} . We estimate $\sigma_{\hat{u}|\mathbf{l}_i}^2$ by regressing \hat{u} on a full set of location (i.e. bin) dummies. For each data set, the N observations are randomly distributed on the (x, y) -plane by drawing cartesian coordinates (x, y) from two independent uniform distributions $U_x[0, 1]$ and $U_y[0, 1]$.

- Choose some initial value $\lambda^{(j)} < \lambda^*$, where λ^* denotes the minimal λ that fulfills Eq. A1.
- Perform the following search algorithm:

1. For each data set, obtain the estimates $\widehat{a(\mathbf{l}_i)}^{(s)}$ and $\widetilde{a(\mathbf{l}_i)}^{(s)}$, where the former sets $\lambda = \lambda^{(j)}$ and the latter $\lambda = 1e10$. Both estimators use binning on a 300×300 grid and set the maximal bandwidth to $h^* = 75$ bins.

2. Compute

$$\bar{r}^{(j)} = \frac{1}{S} \sum_{s=1}^S \left\{ \frac{\sum_{i=1}^N \left(\widehat{a(\mathbf{l}_i)}^{(s)} - 0 \right)^2}{\sum_{i=1}^N \left(\widetilde{a(\mathbf{l}_i)}^{(s)} - 0 \right)^2} - 1 \right\}, \quad (\text{A3})$$

which is a global estimate of Eq. A1 given the parameter value $\lambda^{(j)}$.

3. Update $\lambda^{(j)}$ as follows:

- As long as λ has not reached its upper bound, i.e. $\bar{r}^{(j)} > \alpha$, set

$$\lambda^{(j+1)} = \lambda^{(j)} + c, \quad (\text{A4})$$

where c is some arbitrary constant and go back to step 1

- Once λ has reached its upper bound, i.e. $\bar{r}^{(j)} \leq \alpha$, set

$$\lambda^{(j+1)} = \begin{cases} \lambda^{(j)} - \left| \frac{\lambda^{(j-1)} - \lambda^{(j)}}{2} \right|, & \text{if } \bar{r}^{(j)} \leq \alpha \\ \lambda^{(j)} + \left| \frac{\lambda^{(j-1)} - \lambda^{(j)}}{2} \right|, & \text{if } \bar{r}^{(j)} > \alpha \end{cases} \quad (\text{A5})$$

and go back to step 1.

4. Terminate the algorithm if $\lambda^{(j)} - \lambda^{(j-1)} < \epsilon$, where the threshold ϵ is a sufficiently small number.

References

- Anglin, P., and Gençay, 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 Titterton, 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.

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. The value assumes that land is not contaminated or burdened with unusual legal covenants. 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
<i>Building characteristics</i>			
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		
<i>Land characteristics</i>			
Lot size	578.56	525.00	313.33
<i>Unusual features of the house</i>			
Physical	0.03		
Legal	0.18		
<i>Expert-based land values and location ratings</i>			
Land value	284.97	256.46	148.41
Location rating			
Low	0.29		
Medium	0.49		
High	0.20		
Excellent	0.02		

Table 2: 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 \times age	-0.884	0.806
Gross area \times age	-4.421	0.507***
Gross volume \times age	-1.604	0.201***
Floor size \times gross area	1.287	0.388***
Floor size \times 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

Table 3: 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 $\widehat{a(1)}$. 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: Estimated location values					
	Expert-based rating				
	Low	Medium	High	Excellent	Total
Low	0.132	0.154	0.007	0.000	0.293
Medium	0.149	0.289	0.054	0.000	0.486
High	0.019	0.043	0.130	0.011	0.202
Excellent	0.000	0.000	0.011	0.008	0.019
Total	0.293	0.486	0.202	0.019	1.000
χ^2 -stat.	1.1e+04	$\widehat{\gamma}$	0.662		
P-value	0.000	$\widehat{\tau}$	0.459		

Panel B: 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 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		

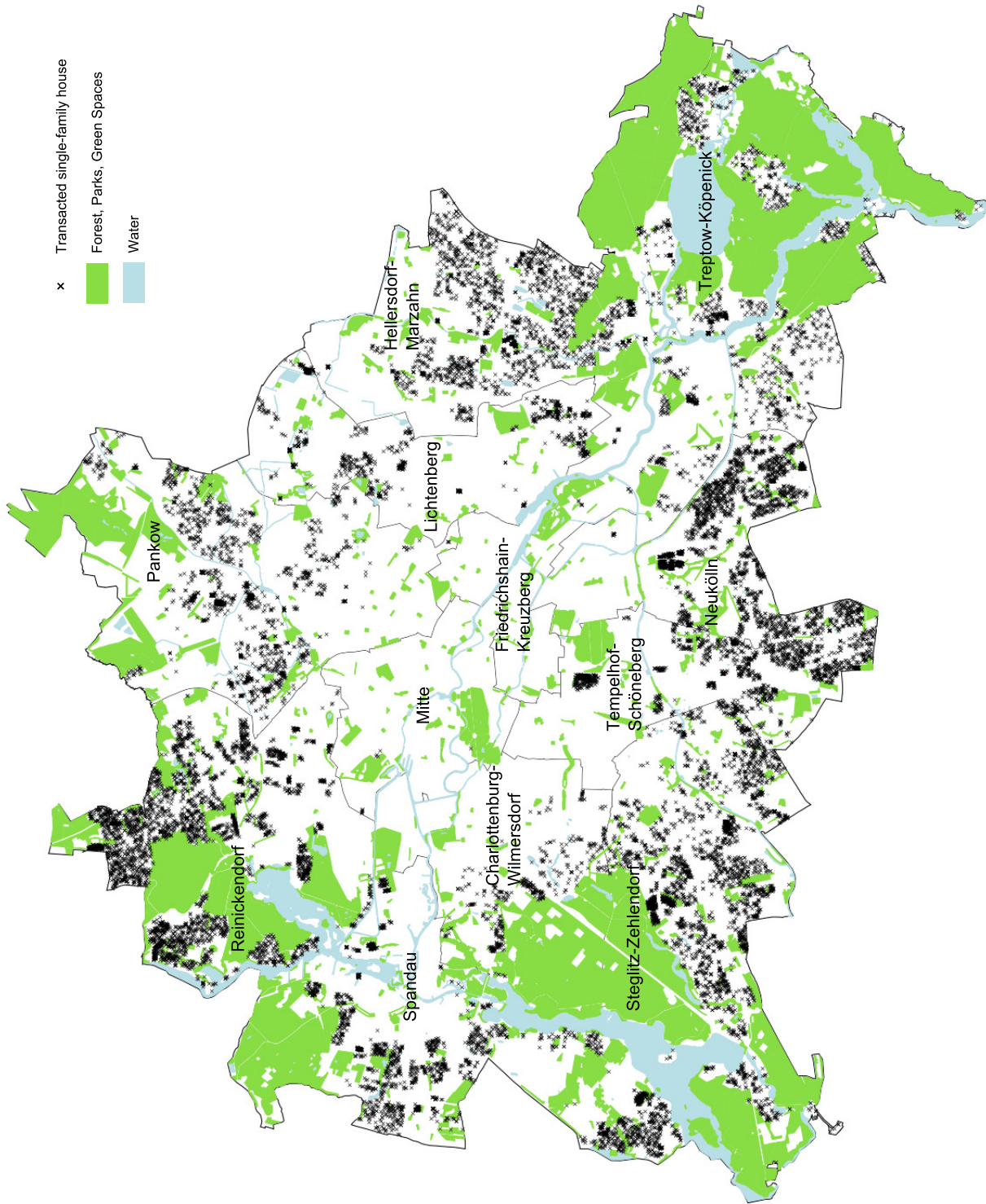


Figure 1: Transacted houses in Berlin, 1996-2010. Figure shows the location of the transacted single-family houses within in the city. Number of observations is 19,283. Solid lines represent the borders of Berlin's 12 administrative districts (as of the year 2000).

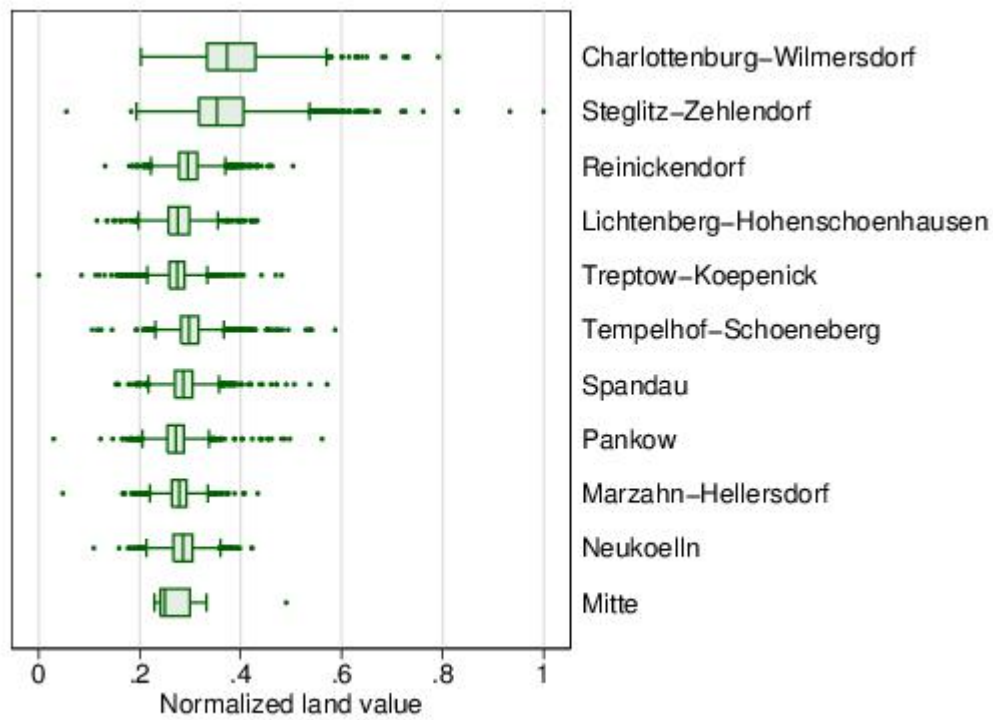


Figure 2: First-stage residuals by district. Figure shows box plots for the normalized first-stage residuals for Berlin’s administrative districts. Number of observations is 19,283. Districts are sorted in descending order with respect to the median of the expert-based location rating. Line that separates 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.

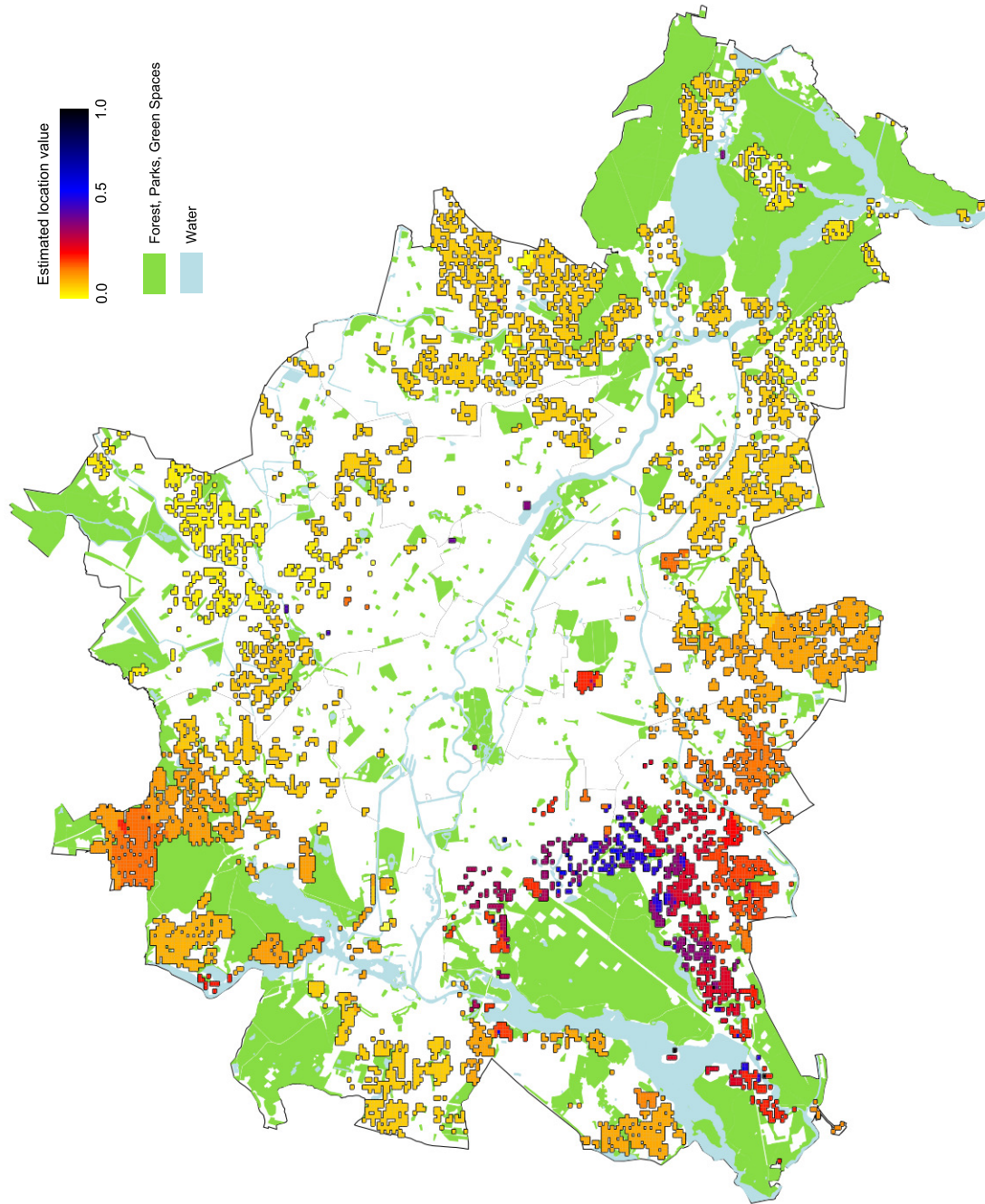


Figure 3: Estimated location value surface. Map of Berlin with the estimated location values $\widehat{a(\mathbf{l})}$.

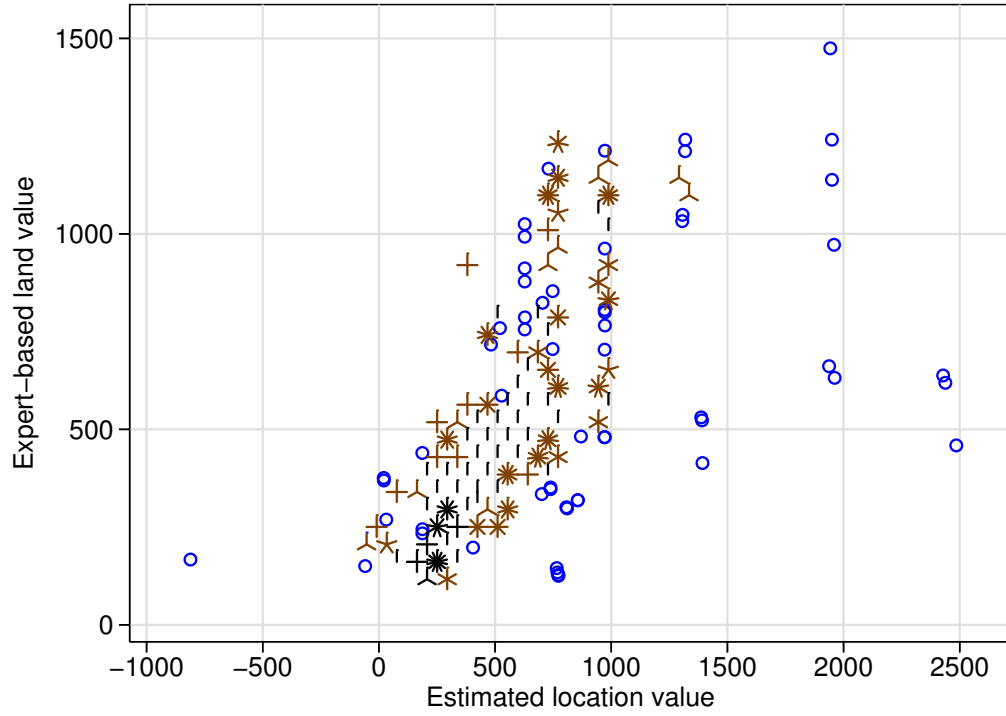


Figure 4: Sunflower plot of expert-based land values and estimated location value. Shows sunflower plot of the bin average of expert-based land value and estimated location value, $\hat{a}(l)$. Both figures are in real (year 2000) Euros. Number of observations is 7,704. Each petal of a light sunflower represents 1 observation. Each petal of a dark sunflower represents several observations. Circles represent individual observation in low density region.

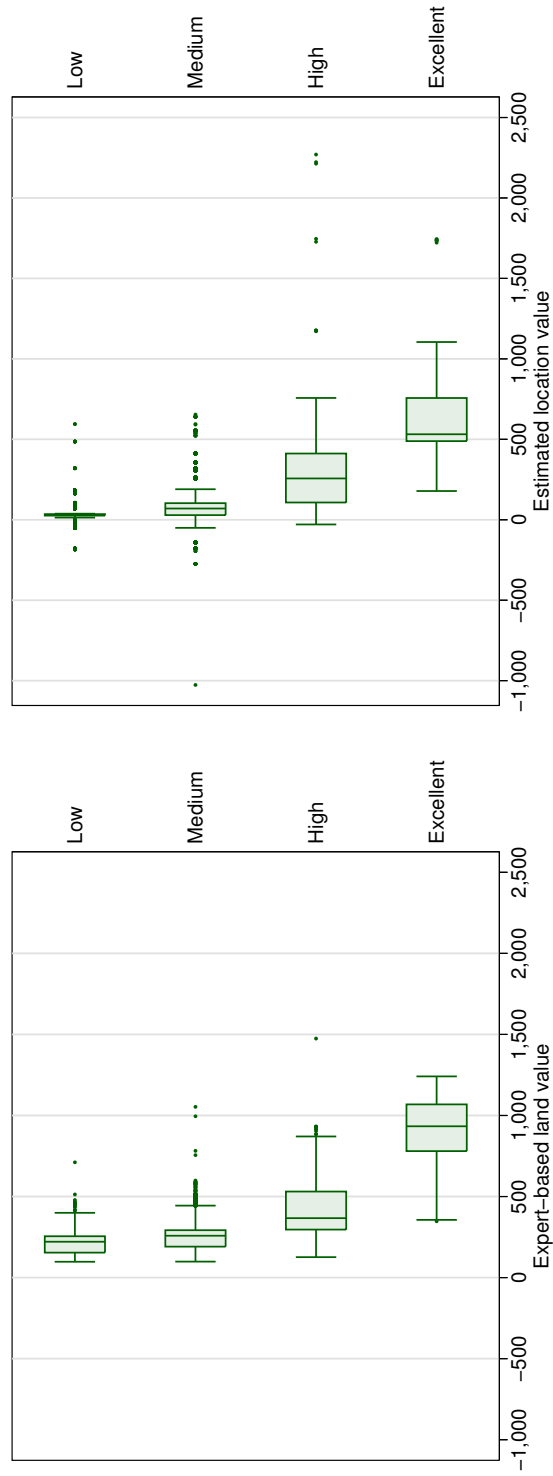


Figure 5: Box plots of expert-based land value and estimated location value. Left panel shows box plots of the expert-based land value for each of the four levels of the expert-based location rating. Right panel shows box plots of the estimated location value for each of the four levels of the expert-based 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.