Data-Driven Regionalization of Housing Markets

Marco Helbich,* Wolfgang Brunauer,† Julian Hagenauer,* and Michael Leitner‡

*Institute of Geography, University of Heidelberg
†Credit Risk Methods Development, Strategic Risk Management & Control, Bank Austria—Member of UniCredit Group
‡Department of Geography and Anthropology, Louisiana State University

This article presents a data-driven framework for housing market segmentation. Local marginal house price surfaces are investigated by means of mixed geographically weighted regression and are reduced to a set of principal component maps, which in turn serve as input for spatial regionalization. The out-of-sample prediction error of a hedonic pricing model is applied to determine a "near-optimal" number of spatially coherent and homogeneous submarkets. The usefulness of this method is demonstrated with a detailed data set for the Austrian housing market. The results provide evidence that submarkets must always be considered, however they are defined, and that the proposed submarket taxonomy on a regional level significantly improves predictive quality compared to (1) a traditional pooled model, (2) a model that uses an ad hoc submarket definition based on administrative units, and (3) a model incorporating an alternative submarket definition on the basis of aspatial *k*-means clustering. Moreover, it is concluded that the Austrian housing market is characterized by regional determinants and that geography is the most important component determining the house prices. *Key Words: Austria, hedonic modeling, mixed geographically weighted regression, prediction accuracy, real estate, spatial regionalization.*

本文提出一个住宅市场区隔的数据导向架构。地方的边际住宅价格水平, 将透过混合的地理加权回归方法检视之, 并将简化成一组主要成份地图, 进而提供做为空间区域化的输入。本文运用特征价格模型的样本外预测误差, 决定空间上一致且均质之次级市场的 "近似最优解" 数量, 并将透过奥地利住宅市场的详细数据集, 证明该方法的有效之处。研究结果显示, 次级市场永远必须被考虑, 不论它们如何被定义。研究结果亦证实, 区域层级的次级市场分类学, 和下列模型相较之下, 更能够显著地提升预测质量: (1) 传统的混合模型; (2) 利用根据行政区划分特别定义的次级市场之模型; (3) 包含根据分群算法定义的替代性次级市场之模型。本文更进一步结论:奥地利住宅市场的特征在于区域决定因素, 而地理是决定住宅价格的最重要元素。 *关键词: 奥地利, 特征价格模型, 混合地理加权回归, 预测的正确性, 房地产, 空间区域化。*

Este artículo presenta una estructura controlada por datos sobre la segmentación del mercado de vivienda. Las superficies del precio de la vivienda marginal a nivel local se investigan por medio de regresión ponderada geográficamente mixta, superficies de las cuales se deriva un conjunto de mapas de componentes principales, que a la vez contribuyen como insumo a la regionalización espacial. La predicción de error por fuera de muestra de un modelo hedónico de precios se aplica para determinar el número "cercano a lo óptimo" de sub-mercados espacialmente coherentes y homogéneos. La utilidad de este método se demuestra a través de un conjunto pormenorizado de datos del mercado de vivienda austriaco. Los resultados ponen en evidencia que los sub-mercados siempre deben ser tomados en cuenta, sin importar cómo se les defina, y que la taxonomía del sub-mercado propuesta a nivel regional mejora significativamente la cualidad predictiva en comparación con (1) un modelo mancomunado tradicional, (2) un modelo que utilice una definición ad hoc del sub-mercado basada en unidades administrativas, y (3) un modelo que incorpore una definición alternativa de sub-mercado con base en agrupamiento aespacial de *k*-medios. Además, se concluye que el mercado de vivienda austriaco está caracterizado por determinadores regionales y que la geografía es el componente más importante para la determinación de los precios de las casas. *Palabras clave: Austria, modelización hedónica, regresión ponderada geográficamente mixta, exactitud de la predicción, finca raíz, regionalización espacial.*

In real estate research, house prices are usually modeled by hedonic regression models, where structural attributes of the house, its neighborhood, and its location serve as explanatory variables (Bourassa, Cantoni, and Hoesli 2007). Because housing is treated as a heterogeneous good, hedonic price theory is applicable, where an object is valued for its utility-bearing characteristics and its price is decomposed into its

Annals of the Association of American Geographers, 10X(X) XXXX, pp. 1–19 © XXXX by Association of American Geographers
Initial submission, October 2011; revised submission, March 2012; final acceptance, May 2012
Published by Taylor & Francis, LLC.

individual value-adding components (Rosen 1974). The hedonic price function results from spatial equilibrium conditions of supply and demand for housing characteristics, ensuring stationarity of the effects on prices over space within a market.

Extended research has established that housing markets are segmented in submarkets; that is, that hedonic price functions can vary across space (e.g., Straszheim 1975; Schnare and Struyk 1976; Goodman 1978; Palm 1978; Maclennan and Tu 1996; Goodman and Thibodeau 1998, 2007; Watkins 2001; Bourassa, Hoesli, and Peng 2003; Hwang and Thill 2009). The lack of clear guidance from economic theory has resulted in a lack of a coherent terminology, however (Watkins 2001). According to Palm (1978, 218), "a housing submarket may be defined as a collectivity of buyers and sellers with a distinct pattern of price-attribute valuations," which results in geographical areas with constant marginal prices (Goodman and Thibodeau 2007). Reasons for functional disequilibria are immobility and durability of real estate, information constraints, search costs, and spatially varying differences in socioeconomic and demographic housing characteristics (Palm 1978; Goodman and Thibodeau 1998). Therefore, submarkets can be defined (1) on the supply side, where structural and neighborhood housing characteristics serve as discriminating factors; (2) on the demand side, based on household income or other sociodemographic characteristics; or (3) on a combination of both (Goodman and Thibodeau 1998, 2007), as is proposed in this research.

In most applications, submarkets are operationalized by disaggregating the entire market into discrete and disjoint regions. Their consideration in hedonic modeling improves explanatory power, mitigates model violations resulting from neighborhood effects, and leads to more precise model predictions (e.g., Straszheim 1975; Bourassa, Hoesli, and Peng 2003; Bourassa, Cantoni, and Hoesli 2007; Hwang and Thill 2009). A submarket definition, however, requires that the delimitation coincides with the true data generating process (Fotheringham, Charlton, and Brunsdon 2002) as well as the economic process, which are two assumptions that are difficult to fulfill. Not meeting these requirements could result in an estimation bias and residual spatial correlation (LeSage and Pace 2009). Furthermore, the results can be affected by the modifiable areal unit problem (Openshaw 1984), meaning that different numbers of submarkets and their spatial arrangement could affect the regression output (Fotheringham and Wong 1991). To avoid these problems, Fotheringham, Charlton, and Brunsdon (2002) and Páez, Fei, and Farber (2008), among others, proposed spatially bounded soft-market segmentations, where the hedonic price function shows instability over space, meaning that marginal prices continuously change their influence on the house price. The authors argue that continuous representations describe reality more closely and more appropriately than rigid, clear-cut submarket boundaries.

Nevertheless, these appealing properties come with serious methodological drawbacks, extensively discussed in Wheeler and Tiefelsdorf (2005), Griffith (2008), Wheeler and Páez (2009), and Páez, Farber, and Wheeler (2011). Such limitations and the comprehensive literature review in the next section underpin the need for a generic data-driven framework, building on the strengths of both continuous and discrete approaches. Due to the fact that housing submarkets are not observable theoretical constructs, data-driven approaches, not relying on the intuition and expert knowledge of assessors, seem to be a rational option that should be considered. Therefore, this article combines a semi-local regression technique and a regionalization algorithm to derive discrete submarket definitions. This explicitly spatial approach has the advantage that little prior knowledge is needed on the exact housing submarket boundaries as the data-driven modeling process guides the specification of housing market segmentations. Subsequently, these results can be used within the traditional regression framework. This supports the recommendations by Can (1992) as well as Goodman and Thibodeau (1998), who demand more empirical justification and less arbitrariness in submarket segmentation. Introducing a detailed database for the Austrian housing market, the advantages of this novel approach are demonstrated with a model competition complementing this research. The following research questions will be answered in this article:

- Is the proposed data-driven technique suitable to derive housing submarkets?
- How well do the derived submarkets perform as additional predictors?
- How well do the empirically derived submarkets perform in a hedonic regression compared to a pooled model,[1] a completely unpooled model, an ad hoc predefined submarket definition using federal states, and an alternative submarket definition using *k*-means–based submarkets?

The following section briefly discusses different approaches and previous empirical findings on housing

submarkets. The subsequent section presents the modeling framework, followed by a definition of the study area and a brief introduction of the data. The results of the empirical analysis are discussed next. Finally, the article concludes with major implications of the research findings and makes suggestions for directions of future work.

Housing Market Segmentation: A Literature Review

Even though the modeling properties of housing submarkets are well known, their empirical delineation raises many methodological questions (Goodman and Thibodeau 2007). Basically, there are three main approaches to model housing submarkets. First, submarkets are defined exogenously by means of predefined spatial units, such as administrative units (Bourassa et al. 1999), school districts (Goodman and Thibodeau 2003), stratifications of metropolitan areas (Adair, Berry, and McGreal 1996), or regions (Bischoff and Maennig 2011). Second, submarkets are considered to have spatially continuous boundaries (Páez, Fei, and Farber 2008). Third, multivariate statistical methods, such as clustering algorithms (Hwang and Thill 2009) or neural networks (Kauko 2004), are used to define submarkets. Each of these three approaches is discussed in more detail next.

Ad Hoc Housing Market Segmentations

In most applications, the *fixed effects model* (e.g., in the framework of an ordinary least squares [OLS] or autoregressive model) is used to model spatial heterogeneity, with submarkets being modeled using dummy variables, which results in the intercept varying over space. Slope heterogeneity can be controlled for through spatial interaction effects of the submarket dummy variables with explanatory covariates (e.g., Kestens, Theriault, and Des Rosiers 2004).

Early examples for the application of ad hoc predetermined submarket definitions using fixed effects models are Schnare and Struyk (1976), who analyzed family housing prices in suburban Boston (United States). They concluded that marginal prices vary over space, even though they have small effects on the overall house price. Comparing the estimated standard errors of regressions between the market-wide and the stratified model, they found that both model types have the same efficiency in terms of prediction power. This supports the argument that submarkets are of marginal importance for predictions. Additionally, due to data loss through market segmentation, the stratified model could result in a lower reliability of model estimates. Goodman (1978) rejected these findings, comparing hedonic coefficients for New Haven, Connecticut (United States). He found significant price differences between submarkets, which are not established by a single equation model, and thus promoted the use of disaggregated submarkets. Likewise, Adair, Berry, and McGreal (1996) divided the metropolitan area of Belfast, Ireland, into an inner, middle, and outer submarket and emphasized their heterogeneous structure, which is expressed in different combinations of significant covariates. On a national scale, Bischoff and Maennig (2011) found heterogeneity in rental housing markets by distinguishing between East and West Germany. Similar, Páez, Fei, and Farber (2008) stressed the importance of market segmentation for Toronto, Canada, reporting superior accuracy of local regressions (i.e., consideration of submarkets) compared to geostatistical models. Recently, Bourassa, Cantoni, and Hoesli (2007) compared several fixed effects models like aspatial OLS and autoregressive models in terms of their predictive performance for Auckland, New Zealand. They concluded that, because of its simplicity, OLS in combination with submarket dummies is more appropriate than more sophisticated approaches (e.g., conditional autoregressive or geostatistical models). If the number of submarkets or interactions becomes large, however, the fixed effects model approach can result in the *incidental parameter problem* (Neyman and Scott 1948), meaning that the number of degrees of freedom is not sufficiently large for parameter estimation. Therefore, a trade-off between model bias resulting from ignoring submarkets and parameter variability due to the smaller sample size has to be taken into consideration (Bourassa, Hoesli, and Peng 2003).

Random or *mixed effects models* partially solve the incidental parameter problem in addition to model heterogeneity as well as correlation structures within housing submarkets explicitly (Jones and Bullen 1994). The simplest case of a random effects model is the one-way intercept model, where intercepts vary across submarkets. Random effects can be approximated as a weighted average of the mean of the observations in the submarkets (with a dummy specification) and the overall mean of the entire study area (Gelman and Hill 2007). The weights are determined by the amount of information within each submarket. For small subsample sizes (little information), estimates tend to be

close to the global mean, whereas for large subsamples, estimates tend to be similar to the unpooled dummy estimate. One major advantage of this model strategy is that the closer the estimates are to the pooled (global) model, the less effective degrees of freedom are used. If observations belong to several (nested) levels of spatial units, the model turns into a *multilevel* or *hierarchical regression* problem (Goldstein 2011).

Empirical examples of multilevel or hierarchical models can be found, among others, in Jones and Bullen (1994), Orford (2000), and Brunauer et al. (2010) using a Bayesian framework. Also, Goodman and Thibodeau (1998) proposed hierarchical models to study Dallas, Texas, housing submarkets on the basis of the quality of public education. In a more recent study, the same authors (Goodman and Thibodeau 2003) compared two alternative submarket definitions: first, a combination of adjacent census tracts and, second, aggregated zone improvement plan code districts. They conclude that results strongly depend on how performance is measured. If random effects are correlated with the predictors, the regression yields biased and inconsistent results. In contrast, fixed effects models produce unbiased and consistent results. In the case of a rather small number of submarkets (similar to our study presented here), where the loss of degrees of freedom is not substantial, the application of the fixed effects model seems to be preferred.

Spatially Continuous Housing Market Segmentations

In contrast to discrete submarket boundaries applied to fixed or random and mixed effects models, Fotheringham, Charlton, and Brunsdon (2002) as well as Páez, Fei, and Farber (2008) proposed local regression methods, namely, geographically weighted regression (GWR), to model spatially continuous segmentations. GWR possesses the following limitations (e.g., Wheeler and Tiefelsdorf 2005; Griffith 2008; Wheeler and Páez 2009), however: First, a number of data points are repeatedly used in parameter estimations, resulting in multiple comparisons. Second, GWR per se results in artificial parameter smoothness (Páez, Farber, and Wheeler 2011). Third, local multicollinearity can falsely induce parameter variability and inflates parameter variance. Fourth, a strong correlation between the GWR parameters might be present. Finally, the resulting standard errors are just approximations, and the classical statistical test procedures are pseudo-counterparts of the traditional test procedures. As a consequence, GWR results in highly volatile parameter estimations, possibly as a result of a globally optimized bandwidth selection through cross-validation (Farber and Páez 2007). Recently, Páez, Farber, and Wheeler (2011) concluded that mitigating these negative side effects requires sample sizes of more than 1,000 observations. Nevertheless, GWR should be applicable to exploratory data analysis only (Wheeler and Páez 2009), and house price predictions should continue to be based on traditional econometric or geostatistical techniques (see, e.g., Bourassa, Cantoni, and Hoesli 2007; Diggle and Ribeiro 2007; LeSage and Pace 2009).

Housing Market Segmentations Using Multivariate Statistics and Neural Computation

A different research line compared to ad hoc and continuous segmentations relies on multivariate statistics, particularly clustering algorithms, and neural networks to examine housing submarkets (e.g., Abraham, Goetzmann, and Wachter 1994; Bourassa et al. 1999; Kauko, Hooimeijer, and Hakfoort 2002; Case et al. 2004; Hwang and Thill 2009). Early attempts at this alternative approach can be found in work by Abraham, Goetzmann, and Wachter (1994), who applied *k*-means clustering in combination with bootstrapping to investigate cluster significance. For the years 1977 to 1992 they discovered diversity and fluctuation within the U.S. housing market. Goetzmann and Wachter (1995), focusing on rent and vacancy data in U.S. metropolitan areas, detected clusters of similar cities using the method of Abraham, Goetzmann, and Wachter (1994). On a local scale, Bourassa et al. (1999) analyzed dwelling markets of the Australian cities of Sydney and Melbourne by expanding the research design originally introduced by Maclennan and Tu (1996). Both studies combined principal component analysis (PCA) with *k*-means cluster analysis. Bourassa et al.'s (1999) results identified five distinct submarkets. Tests concerning prediction accuracy for the city of Melbourne showed that submarkets derived on the basis of individual data have a higher accuracy compared to alternative submarket constructions. Single-market models have the lowest accuracy. Pricing models with geographically concentrated submarkets yield the lowest prediction errors. To impose contiguity constrains, Bourassa, Cantoni, and Hoesli (2010) included spatial coordinates as additional variables in a hierarchical clustering, arguing that the Ward algorithm is less sensitive to initial seed variations compared to *k*-means. Applying the fuzzy *c*-means

clustering algorithm on the Buffalo–Niagara Falls metropolitan statistical area, Hwang and Thill (2009) stressed the advantage of noncrisp cluster boundaries, which contradicts the findings of Bourassa et al. (1999). Nevertheless, further analysis within the traditional statistical framework needs a defuzzification,[2] which again results in crisp submarket boundaries.

Kauko, Hooimeijer, and Hakfoort (2002) and Kauko (2004) demonstrated the usefulness of unsupervised neural networks (e.g., self-organizing maps [SOMs; Kohonen 2001]) to analyze nonlinear relationships in housing markets. A comparison between the cities of Helsinki, Finland, and Amsterdam, The Netherlands, showed that the latter has a more fragmented housing market. SOMs result in soft and data-driven market segmentations. To account for spatial dependencies Bação, Lobo, and Painho (2005) stressed the need to adapt Kohonen's (2001) original SOM. Nevertheless, SOMs and their variants strongly depend on the subjective choice of several input parameters (e.g., number of neurons, map topology, learning rate), making them unsuitable for this study.

With the exception of Bação, Lobo, and Painho (2005), all of the previously mentioned clustering algorithms neglect spatial dependency, spatial proximity, and topological relationships between adjacent objects immanent in housing data (Dubin 1992). To overcome these serious limitations, a "regionalization" approach has proven to be promising (Miller 2010). This approach is a form of clustering that groups data into spatially homogenous and contiguous submarkets. Several techniques (e.g., Openshaw and Rao 1995; Assunção et al. 2006; Guo 2008) are available to handle spatial effects during clustering, although none of them has ever been applied in real estate studies. Recently, Assunção et al. (2006) proposed the Spatial 'K'luster Analysis by Tree Edge Removal (SKATER) algorithm. Compared to former techniques (e.g., Openshaw and Rao 1995), it produces more homogenous regions and reduces computational burden. Guo (2008) criticized SKATER, however, because, first, the contiguity constraint imposed by the minimum spanning tree (MST) is static, notwithstanding that the relations of adjacent regions might change during the clustering. Second, the MST cannot guarantee that all objects within a cluster are similar to each other. Thus, Guo (2008) promoted a family of algorithms called regionalization with dynamically constrained agglomerative clustering and partitioning. Nevertheless, SKATER has a lower theoretical computational complexity compared to Guo's (2008) algorithms and is thus more suitable to cluster medium to large data sets similar to the one presented in this article.

To summarize, although the relevance of submarkets stated by Maclennan (1982) is widely accepted in practice, there is no consensus about the methods to delimit or conceptualize hedonic submarkets empirically and which method is more appropriate. It seems, however, that each of these three main approaches has its individual advantages and disadvantages. Therefore, a combination of the strengths of each approach seems to be suitable for delimiting submarkets more efficiently, while explicitly considering spatial effects during the regionalization process. For a model comparison, OLS linked with submarket dummies is a suitable method to evaluate housing segmentation.

Methodology

The study design is summarized in Figure 1. The framework consists of four main steps:

1. A semi-local regression is applied to investigate nonstationary covariates of housing prices. To get area-wide estimation surfaces of marginal prices, these local pointwise regression estimates are interpolated.
2. A PCA reduces the information variability of these resulting parameter surface maps.
3. These principal component (PC) maps are clustered with the SKATER regionalization algorithm and the resulting submarkets are evaluated by means of hedonic regressions.
4. A model competition tests the predictive accuracy of the data-driven submarkets against no segmentation and commonly used segmentations (i.e., federal states and *k*-means–based submarkets). Finally, covariate variability is investigated by an unpooled hedonic model.

The entire analysis is performed within the R environment (R Development Core Team 2012). It must be noted that the proposed methodology (Figure 1) is not at all limited to the study area of this research but can, in principle, be generalized and transferred to any other study area's housing markets.

Geographically Weighted Regression

The GWR is a locally weighted regression analysis, introduced by Brunsdon, Fotheringham, and Charlton (1996), to model spatial autocorrelation as well as spatial heterogeneity. For each of the parameter estimation

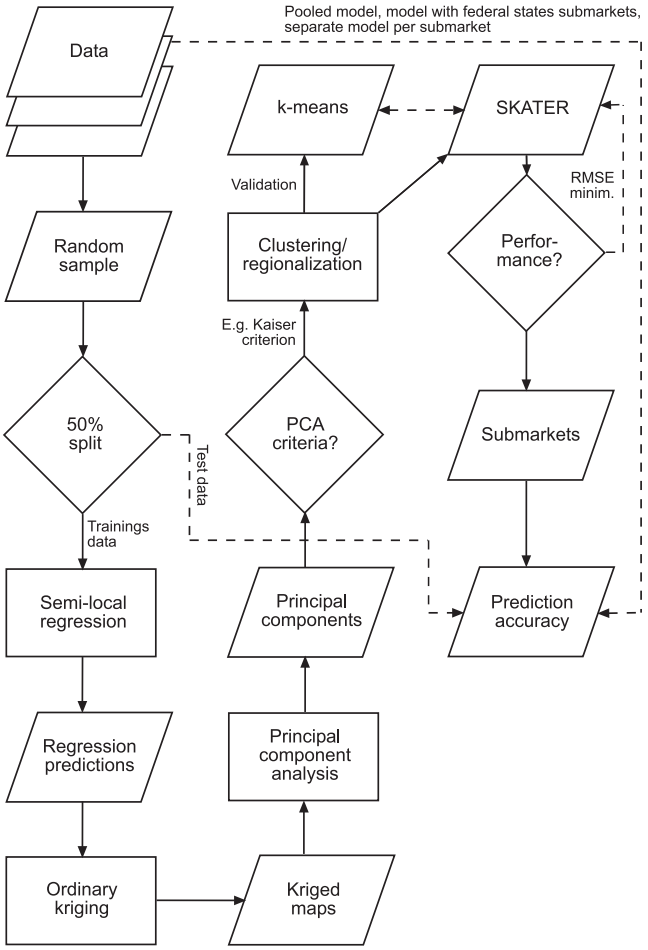

Figure 1. Simplified research design to model dwelling submarkets. SKATER = Spatial 'K'luster Analysis by Tree Edge Removal; RMSE = root mean square error; PCA = principal component analysis.

steps, only a subset of data is taken into account, where data near the regression point have a higher influence than data further away. The weighting itself is based on a kernel function (e.g., Gaussian), although the choice of the kernel function has little impact on estimation results. The crucial point in the GWR is the choice of the bandwidth. Commonly, the bandwidth is allowed to vary across space, depending on the density of the data points. This improves the goodness of fit, if the data points are irregularly distributed across space. Thus, in densely populated areas, the kernel has a shorter bandwidth compared to regions with longer interpoint distances (Fotheringham, Charlton, and Brunsdon 2002).

The basic GWR assumes that all predictors have a nonstationary behavior and vary across space. If this is not the case, a mixed GWR (MGWR) is more appropriate, reducing the degrees of freedom and thus resulting in a more parsimonious model. MGWR keeps coefficients with nonsignificant variation constant, meaning that, for example, their effect on housing prices is stationary, whereas others are allowed to vary across space. Stationarity (spatially constant effects) can be investigated by a test statistic put forward by Leung, Mei, and Zhang (2000). A multistep algorithm proposed by Fotheringham, Charlton, and Brunsdon (2002) is used to estimate the following equation:

$$y_i = \sum_{j=1}^{k} a_j x_{ij} + \sum_{l=1}^{m} b_l (u_i, v_i) x_{il} + e_i$$

where y_i is the logarithmically transformed sales price of observation i, a_j are the global coefficients of covariates x_{ij}, and $b_l(u_i, v_i)$ are the local coefficients of covariates x_{il} at the coordinates (u_i, v_i) of observation i. As the (M)GWR only provides point estimates, spatial interpolation (e.g., ordinary kriging; Pebesma 2004) is employed to get area-wide estimates of marginal housing prices required for further analysis.

As discussed in the literature review, significant spatial variation of the parameters points to the existence of submarkets. Thus, MGWR can be regarded as a data-driven way to explore parameter variability of the hedonic model across space and soft-market housing segmentation, respectively. The rather high volatility of nonstationary parameters seems artificial and not robust, however, and is thus hardly applicable for prediction purposes. This requires a way to reduce the information and to operationalize the results in a more parsimonious submarket definition.

Principal Component Analysis

Noise and multicollinearity of kriged coefficient surfaces have serious effects (e.g., instability) on the discriminating power of regionalization algorithms. The PCA provides a solution to this problem (e.g., Maclennan and Tu 1996). Its objective is a linear dimensional reduction of data resulting in a small set of orthogonal PCs, reflecting the most inherent variability of the multivariate attribute space. Technically, the PCA is an orthogonal transformation decomposing the covariance or correlation matrix of the input data into its eigenvectors and eigenvalues. The latter represent the amount of information gathered by each principal component (Jolliffe 2002). Several criteria (e.g., Kaiser criterion, scree plot) exist to select an appropriate number of components (Reimann et al. 2008). In addition, the final subset of PCs serves as input data for regionalization by the SKATER algorithm to derive spatially

compact and homogeneous regions, serving as submarkets in a hedonic regression.

Spatial Regionalization

Although being aware of Guo's (2008) criticism, the SKATER algorithm has the ability to derive spatially coherent and homogeneous submarkets. The SKATER algorithm transforms the regionalization problem into a graph partitioning problem (Assunção et al. 2006). A spatially contiguous graph is built by connecting regions that are geographically adjacent. Additionally, each edge is assigned the similarity between the two contiguous regions. By pruning the resulting graph into subgraphs, homogenous and spatially contiguous regions are obtained. To reduce the complexity of the partitioning of the graph, the SKATER algorithm starts with a MST. A MST is a graph with no circuits, minimum costs, and minimum number of edges connecting all vertices. The costs are defined as the sum of the dissimilarities over all edges. The dissimilarity measure is dependent on the attribute space and is measured by the squared Euclidean distance. For partitioning, edges are sequentially removed from the MST until the desired number of regions is achieved. Because of the complexity of finding optimal edges for removal, the SKATER algorithm uses a heuristic approach based on two objective functions f_1 and f_2. f_1 measures the change in homogeneity of the clustering, when partitioning the tree by removing an edge, and the second function f_2 measures the change in homogeneity of the least homogenous tree that results from cutting out an edge. Starting from the central vertex of the MST, incident edges are evaluated. If the removal of one of these edges results in costs according to f_1 superior to the best solution found so far, the edge is stored as a candidate solution. Then, a new vertex from the set of evaluated edges is chosen based on the cost function f_2, which penalizes edges that result in imbalanced trees when cutting them out. Then, the incident edges of the selected vertex are evaluated again. This procedure is iteratively repeated, until a stopping criterion (e.g., the candidate solution has not changed for a certain number of iterations) is reached. Restrictions such as population-balanced regions are not considered in this regionalization procedure. A significant task in regionalization is the determination of the number of submarkets. Because no a priori knowledge about the actual number of submarkets is present and generally applied indexes (e.g., Davies and Bouldin 1979) do not consider spatial contiguity, a model-driven approach applying the predictive performance of the hedonic regression is required.

To determine a near-optimal number of submarkets, two opposing strategies are pursued: First, the prediction error of a pooled stepwise hedonic model with additional submarket variables or submarket interaction effects between the covariates (integrating intercept and slope heterogeneity), based on the SKATER-derived submarket partition, is estimated. This is referred to as the regionalized model. This procedure is repeated with different numbers of SKATER clusters. As additional decision guidance, Akaike's information criterion (AIC; Akaike 1974) is used. The AIC describes a trade-off between bias and model variance, penalizing overfitting (Burnham and Anderson 2002). Second, for each partition, separate stepwise hedonic models are estimated for each submarket and their average prediction error is determined. This is referred to as the single equation model and represents the notion of extreme heterogeneity, or distinct markets, where covariates follow completely different patterns over space.

To evaluate out-of-sample prediction accuracy of the competing models, the housing data set is split into a training and a test set, which have not been used for model building (Jain, Duin, and Mao 2000). To reveal differences between the estimated values and the true values for each model the root mean square error (RMSE) is calculated. Concerning Goodman and Thibodeau's (2007) result that model suitability depends on the performance measure, the mean absolute error (MAE) is also calculated for the model competition. Smaller RMSE and MAE values denote a smaller prediction error and thus a more accurate hedonic model.

Finally, a model competition is conducted. The SKATER-based model competes against (1) a pooled market model without considering submarkets, which serves as the reference model (the so-called global model); and (2) two alternative models, one using federal states[3] (the so-called ad hoc model) and the other one using *k*-means clusters as submarket delimitation (henceforth the *k*-means–based model).

Study Site and Data

The empirical study is based on 3,887 geocoded single-family homes located in Austria. The housing segment of single-family homes is typical for suburban and rural areas, compared to urban areas being primarily characterized by apartments. Attached to each home is the transaction price for the years 1998 to 2009, nine

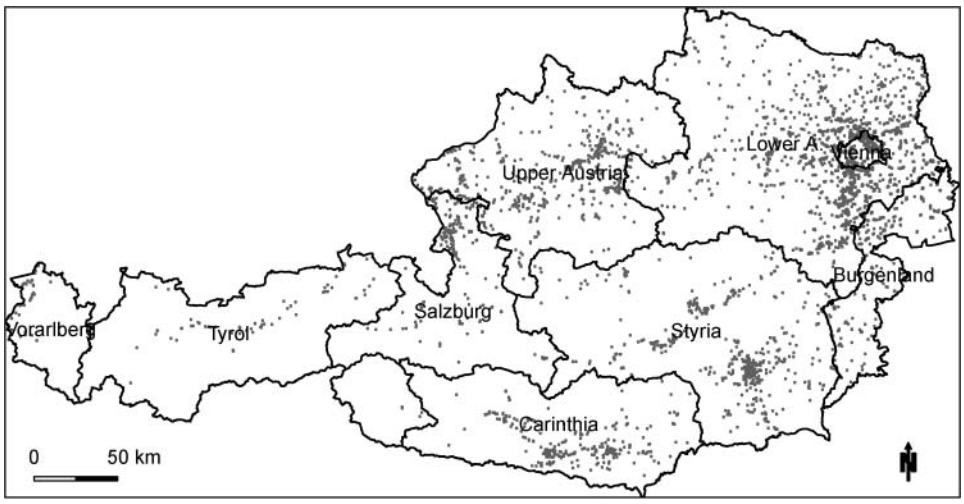

Figure 2. Study site: Dwelling locations (gray points) and Austrian federal states (black lines).

house-specific structural variables (e.g., total floor area), and two temporal covariates (e.g., year of purchase). These data have been collected by UniCredit Bank Austria AG and represent transactions that are associated with Bank Austria AG. Figure 2 shows the study site and the individual location of each house. Clearly, some east–west divide is evident, with more houses in the eastern federal states.

The neighborhood of each house is described by five attributes (e.g., proportion of academics) for the year 2001 at two different levels of aggregation: the Austrian administrative unit of municipality (equivalent to the U.S. census tract) and the enumeration district level (equivalent to the U.S. census block group), both published by Statistics Austria. Additionally, more up-to-date data sets for 2009 are used for two of the five attributes describing each house's neighborhood. The appendix lists all covariates, their anticipated effects on house price, and descriptive statistics.

Results

Exploring Nonstationarity

Initially, the exploratory data analysis indicates that house prices substantially vary across Austria (Moran's $I = 0.288$, $p < 0.001$), with the highest prices paid in and around Vienna, in most provincial capitals (e.g., Salzburg, Innsbruck), and in some smaller cities dominated by the tourism industry. This is a first indication of possible spatial instability in the hedonic price function, which is now further explored by the MGWR.

For this purpose the entire data set is partitioned into a regionalization data set as well as a training and a test data set for validation of the modeled submarkets. Due to excessively time-consuming computations on a standard desktop computer, a 50 percent random sample of the entire data set for the actual regionalization was selected for running the MGWR. Subsequently, the remaining 50 percent of the entire data set was randomly divided, using a common 85–15 percent split. The larger of the two samples was used for hedonic regression analysis, and the smaller sample was applied to calculate an unbiased measure of the prediction accuracy.

The hedonic function was expressed with the natural logarithm of the transaction price as response variable. Similarly, some of the continuous covariates were also logarithmically transformed (see Table A1).[4] A MGWR was estimated to explore global and local effects on house prices. The model performance shows a pseudo-coefficient of determination (R^2) ranging from 0.24 to 0.49 and regression assumptions like homoscedasticity, residual independence, and normality not being violated. The $F(3)$-statistic (Leung, Mei, and Zhang 2000) was consulted to discriminate between global and local effects. This resulted in six stationary and ten nonstationary covariates, all having expected signs. The stationary covariates (e.g., condition of the house) are independent of location and show identical behavior across Austria; thus, they are not relevant for the regionalization performed later. All nonstationary effects show significant spatial variation (at least $p < 0.05$) and include the following (each covariate's effect on house price is shown in parentheses):

- Structural covariates: Log total floor area (+ eff.), log plot space (– eff.), low quality of the heating system (– eff.).

- Temporal effects: Age of building at time of sale (– eff.), year of purchase (+ eff.).
- Neighborhood covariates: Unemployment rate (– eff.), purchase power index (+, – eff.), proportion of academics (+ eff.), age index (– eff.), log population density (+ eff.).

As an example, Figure 3 shows the parameter estimates of four selected nonstationary covariates, interpolated with ordinary kriging. As expected, both structural covariates quantifying the size of the dwellings, namely, plot space and total floor area, are positively related to house prices. Plot space has the highest positive effect in the north and northeast of Vienna. For this region, a 10 percent increase of plot space induces an increase in dwelling prices of up to 2 percent. This relationship can also be interpreted as a plot space elasticity of 0.2. This effect is reduced, for example, in the province of Salzburg, where the plot space elasticity is only 0.1. The total floor area elasticity is lower in and around Vienna (0.3) compared to the western part of Austria, where the elasticity is nearly twice as high (0.5), particularly in the province of Salzburg. The neighborhood variables show a similar spatially heterogeneous behavior. The proportion of academics has a positive influence on house prices. The higher the proportion of academics, the higher the dwelling prices are. The maximum marginal value is reached in the north of the city of Salzburg, where a 1 percent increase of academics results in a more than 3 percent increase in house prices, when holding all other effects constant. The age index measures the average age of inhabitants. A high population age index, reflecting a rather old population, serves as a proxy for structural weakness and is expected to have a negative effect on house prices (see Brunauer, Lang, and Umlauf 2010). The effect of this covariate is almost negligible in Vienna and the city of Linz. In other parts of Austria, the effect is marginal but negative and ranges from –0.01 (for most parts of the province of Lower Austria) to –0.05 (province of Tyrol).

Reduction of Multicollinearity Using PCA

Possible methodological limitations of the (M)GWR found by Wheeler and Tiefelsdorf (2005), like parameter collinearity, are explored here as well. Spearman's ρ correlation coefficients confirm significant intercorrelations of the nonstationary parameter surfaces (e.g., the coefficient surface age of the building is correlated with the unemployment rate; $\rho = 0.495$, $p < 0.001$). Elimination of these multicollinearity effects and thus enhancement of the performance and quality of regionalization requires the calculation of a PCA on the scaled parameter surfaces. The Kaiser criterion suggests that a PC needs to possess eigenvalues greater than one, which in our study results in three PCs. This result is also supported by the scree plot (Figure 4A) with the declining eigenvalues flattening out after the first three PCs. Moreover, these three PCs explain approximately 84 percent of the overall variance (see Jolliffe 2002). PC1 explains 56 percent of the total variance, PC2 17 percent approximately one sixth, and PC3 12 percent. These results are similar to those of Bourassa et al. (1999), whose first three PCs explain 82 percent of the variance.

The biplot (Figure 4B) shows the loadings (eigenvectors) as well as the first and second scores of the data points (Reimann et al. 2008). The loadings represent the direction of each PC and express the relationship with the original variables, whereas the scores are the new data points projected on the new coordinates for each PC. Both grouping effects of variables and similar loadings are visible. Most of the MGWR surfaces are moderately correlated (indicated by the arrow's angles) and show a balanced variability of each MGWR surface for PC1 and PC2 (shown by the length of arrows). Additionally, PC1 seems relatively balanced concerning positive and negative loadings. For example, the logged plot space and logged population density load positive on PC1, whereas logged total floor area has a high negative loading. PC1 shows a tendency to summarize structural characteristics of dwellings. In contrast, the second PC emphasizes the neighborhood characteristics slightly more. For instance, the proportion of academics has a relatively high positive loading, whereas age of the building has a negative loading on PC2. PC3 is difficult to interpret, representing structural, temporal, and neighborhood characteristics, dominated by the age index.

The use of census data in hedonic modeling is prevalent to describe neighborhood characteristics located in specific submarkets. To ensure such a nested structure as well as to facilitate operability and integrity with these data sets, the three PCs are aggregated by averaging the cells with a resolution of 1,000 m over all municipalities, needed for subsequent regionalization.

Regionalization of the Housing Market

The regionalization algorithm is initiated with two to fifteen submarkets and each time a stepwise hedonic

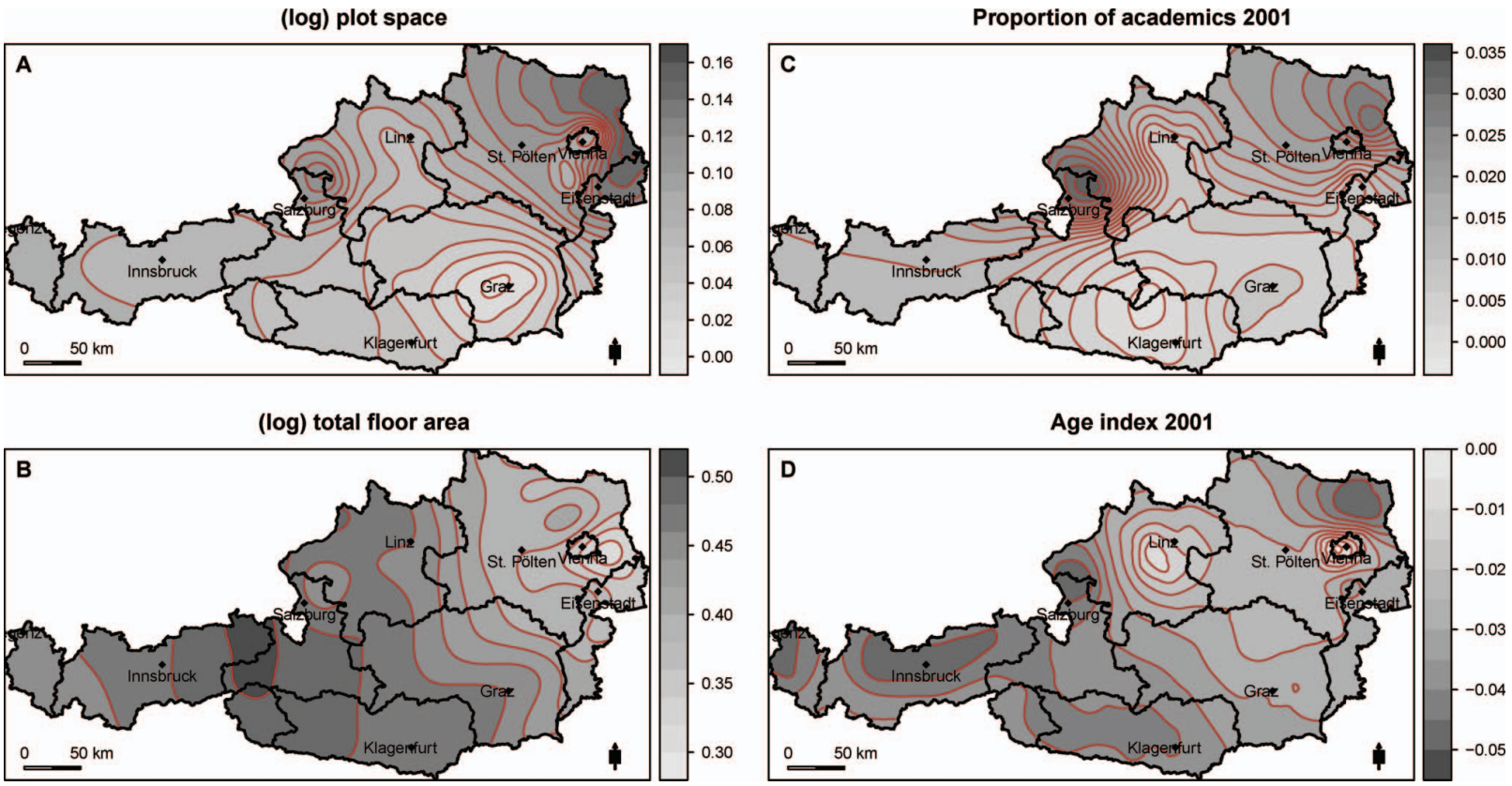

Figure 3. Nonstationary parameter estimations by mixed geographically weighted regression (MGWR). Panels A and B represent structural characteristics and panels C and D show neighborhood characteristics. Points show the capitals of the federal states. The black lines are federal state boundaries. (Color figure available online.)

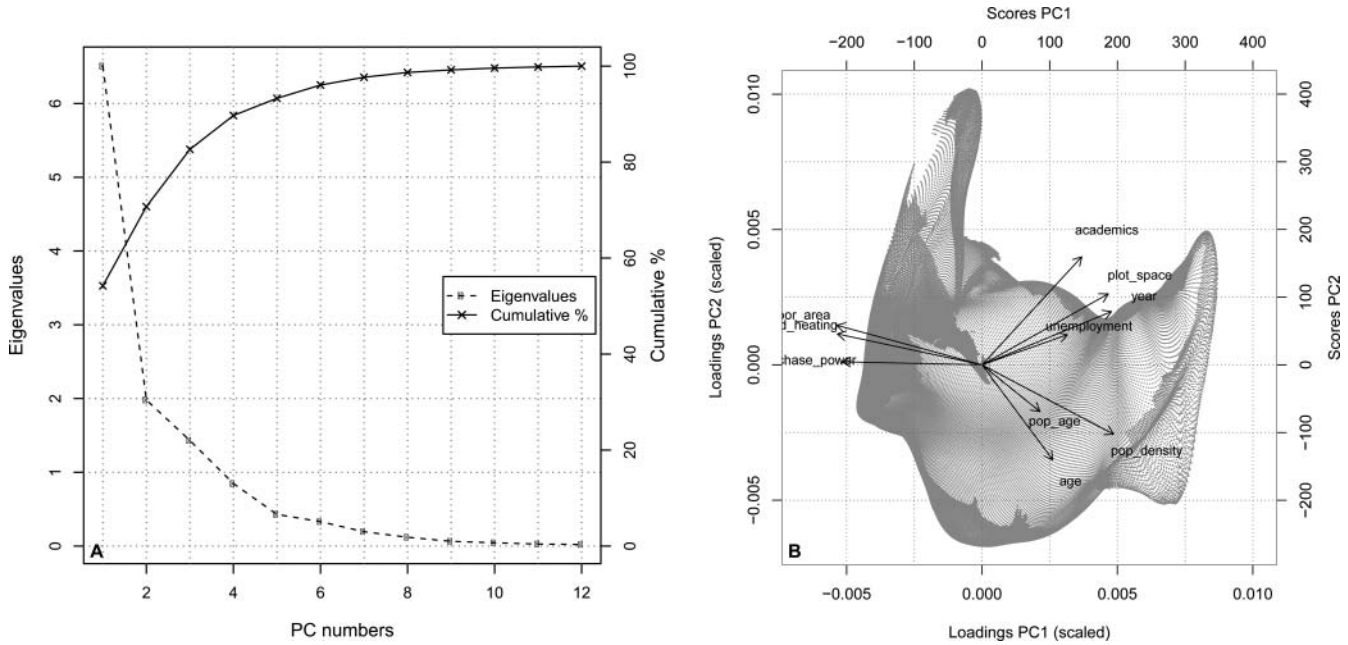

Figure 4. Principal component analysis results: Panel A shows the scree plot and the cumulative percentage of the explained variance. Panel B shows the biplot of the first two principal components. PC = principal component.

regression is estimated, using the covariates in Table A1. The left panel of Figure 5 shows the regression performance of the regionalized model. Dividing Austria into eleven submarkets results in the lowest AIC score and, when compared to all other partitions, a relatively small RMSE of 0.324. A larger number of submarkets improves the RMSE only slightly (e.g., –0.0008) but, at the same time, increases the AIC.

In comparison, the average RMSEs of the best single equation model, recommending only two submarkets, lead to a rather poor prediction error of more than 0.339, with an increase in the number of submarkets resulting in a higher RMSE. The single equation model with eleven submarkets exceeds an RMSE of 0.388. The single equation approach furthermore illustrates the incidental parameter problem. In other words, the number of possible submarkets is constrained, as otherwise the number of observations is not sufficiently high for estimation purposes. These results mirror Bourassa, Hoesli, and Peng's (2003, 15) conclusion that "too much homogeneity may not be a good thing in practice." Similarly, Adair, Berry, and McGreal (1996) provided evidence that just a few submarkets on a macrolevel are preferable.

The final submarket partition of Austria is shown in the right panel of Figure 5. With the exception of SKATER submarket 5 (SK5), all submarkets show significant differences compared to the reference cluster SK2 (Table 1). Local expert knowledge confirms these results. For instance, Vienna and its proper surroundings (cluster 2) represent one submarket. Here particularly, southern municipalities in immediate proximity to Vienna achieve high housing prices. Helbich and Leitner (2009) and Helbich (2012) concluded that these are effects of suburbanization and postsuburbanization processes boosting land and housing prices. The extent of this region reflects a commuting distance of approximately as high as thirty-five minutes to the city boundary of Vienna by motorized individual transport. For the year 2001, Statistics Austria (2007) reported for Lower Austria that more than 25 percent of all daily commuters needed sixteen to thirty-five minutes to get to work. Compared to all other clusters, the submarket represented by cluster 1 has relatively high housing prices because the available land is scarce in alpine areas and mostly limited to the valley floors. Additionally, the tourism industry has pushed housing prices higher. In contrast, the submarket in the north of Vienna (cluster 6) reflects an economically weak region suffering from aging of the population and outmigration (Fassmann, Görgl, and Helbich 2009).

Validation of Modeled Submarkets

Table 1 presents detailed results of the model competition using out-of-sample RMSE (see earlier). On the basis of paired Wilcoxon signed rank tests, the null hypothesis that the models including submarkets perform statistically equal to the global model without submarkets can be rejected at $p < 0.05$. In addition, prediction errors are noticeably lower for the submarket models. The ad hoc model reduces the RMSE by about 0.005 and the regionalized model by about 0.008 compared to the global model. Even the adjusted R^2 of the regionalized model is lower than the ad hoc model (44 percent vs. 45 percent explained variance). Additionally, as advised in Jain, Duin, and Mao (2000), the regionalization is cross-checked against the widely used aspatial *k*-means algorithm in housing segmentation (e.g., Maclennan and Tu 1996). To define a suitable number of clusters, this study follows Venables and Ripley (2002), who iteratively minimized the within-group sum of squares. As in Bourassa, Cantoni, and Hoesli (2010), the results suggest eight to ten clusters, which are then used in the hedonic regression. All *k*-means RMSEs are higher than the SKATER results, marginally lower or equal than the ad hoc results, and lower than the global model. It should be noted that simply using the MGWR model results in the highest RMSE of 0.338. What is more, the regionalized model results in a significantly superior out-of-sample prediction performance than the alternative counterparts (Wilcoxon test $p < 0.05$). No significant difference in the prediction performance is found between the ad hoc model and the *k*-means–based model. In contrast to the findings of Goodman and Thibodeau (2007), who claimed that the model preference depends on the measured performance criterion, the results between different performance measures are consistent in our research, as the MAE also supports the conclusions from the RMSE.

Independent of the prediction accuracy of these competing models, all estimated coefficients for the structural, temporal, and neighborhood effects of the models have expected signs and are highly significant. Similar to Bourassa et al. (1999), the results suggest that internal homogeneity and external heterogeneity are better represented in the modeled submarkets than in exogenously defined alternative submarkets.

Finally, the results of the regionalized model with eleven submarkets are compared to the single equation model, where a separate model, one for each of the eleven SKATER submarkets, is estimated. As earlier, for each regression a stepwise variable selection is

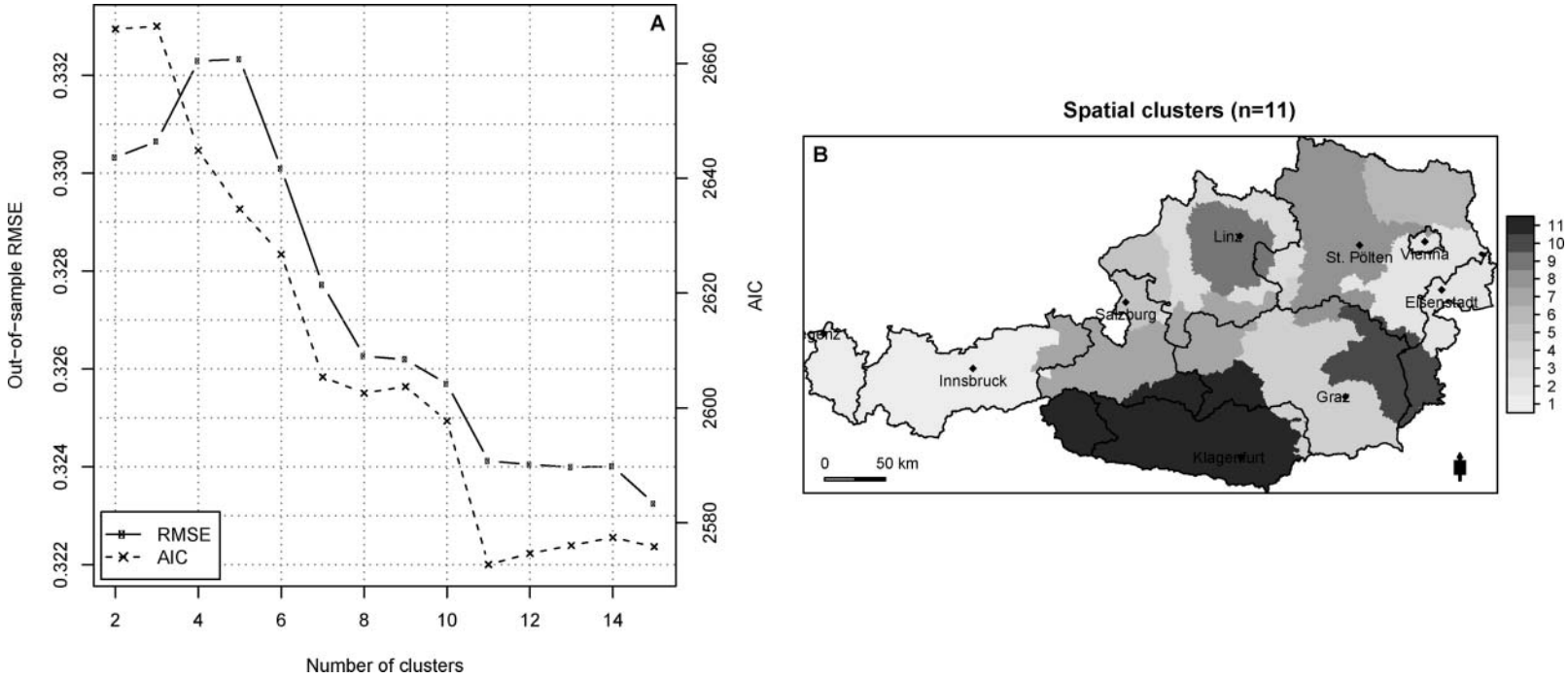

Figure 5. Model-driven determination of the number of submarkets (left) and the resulting regionalization across Austria (right). Areas in gray represent the SKATER submarkets and black lines are the federal state boundaries. RMSE = root mean square error; AIC = Akaike information criterion; SKATER = Spatial 'K'luster Analysis by Tree Edge Removal.

Table 1. Estimated ordinary least squares models

	Regionalized model 11 SKATER submarkets			Global model: Pooled data			Ad hoc model 8 submarkets			*k*-means–based model 8 submarkets		
	Coefficient	*SE*	*t*-value	Coefficient	*SE*	*t*-value	Coefficient	*SE*	*t*-value	Coefficient	*SE*	*t*-value
Intercept	10.011	0.213	46.999***	10.346	0.203	50.903***	9.203	0.221	41.578***	10.113	0.214	47.285***
Structural covariates												
lnarea_total	0.425	0.020	20.977***	0.452	0.020	22.480***	0.424	0.020	21.374***	0.438	0.020	21.698***
lnarea_plot	0.094	0.014	6.482***	0.058	0.014	4.085***	0.104	0.014	7.277***	0.084	0.014	5.879***
cond_house1	–0.027	0.016	–1.715†	–0.026	0.016	–1.672†	–0.040	0.015	–2.582**	–0.026	0.016	–1.684
heat1	–0.119	0.026	–4.491***	–0.109	0.027	–4.064***	–0.111	0.026	–4.256***	–0.127	0.026	–4.781***
bath1	–0.069	0.025	–2.767**	–0.059	0.025	–2.347*	–0.060	0.024	–2.431*	–0.061	0.025	–2.446*
attic1	–0.026	0.013	–1.987*	–0.038	0.013	–2.896**	–0.027	0.013	–2.083*	–0.023	0.013	–1.743
cellar1	0.124	0.015	8.315***	0.126	0.015	8.377***	0.135	0.015	9.208***	0.122	0.015	8.154***
garage1	–0.079	0.013	–5.991***	–0.078	0.013	–5.811***	–0.085	0.013	–6.468***	–0.080	0.013	–6.035***
terr1	0.068	0.013	5.137***	0.065	0.013	4.893***	0.066	0.013	5.083***	0.068	0.013	5.106***
Temporal covariates												
age	–0.006	0.000	–15.189***	–0.006	0.000	–14.528***	–0.006	0.000	–15.978***	–0.006	0.000	–14.778***
time	0.009	0.003	3.254***	0.006	0.003	2.301*	0.011	0.003	3.947***	0.008	0.003	2.798**
Neighborhood covariates												
enumd_unempl	–1.129	0.499	–2.263*	–1.799	0.465	–3.870***				–1.463	0.496	–2.949**
muni_pp_ind	0.004	0.001	4.190***	0.005	0.001	6.437***	0.007	0.001	7.275***	0.004	0.001	3.767***
muni_acad	0.013	0.002	5.639***	0.011	0.002	5.444***	0.009	0.002	4.426***	0.012	0.002	5.422***
muni_age_ind	–0.031	0.004	–8.030***	–0.039	0.004	–10.983***	–0.020	0.004	–5.234***	–0.032	0.004	–8.411***
ln_muni_popd	0.060	0.006	9.529***	0.063	0.006	10.679***	0.020	0.007	2.901**	0.057	0.006	9.013***
Submarkets												
SK1/ Vbg+Tyr/ k1	0.156	0.040	3.919***				0.249	0.038	6.627***	–0.182	0.034	–5.293***
SK3/Car/ k2	–0.091	0.035	–2.640**				–0.090	0.029	–3.058**	–0.168	0.022	–7.602***
SK4/Sty/k3	–0.129	0.023	–5.612***				–0.050	0.022	–2.279*	0.136	0.040	3.418***
SK5/Upp/k4	0.038	0.026	1.442				0.040	0.020	1.961*	0.001	0.024	0.024
SK6/Sal/k5	–0.083	0.025	–3.401***				0.255	0.029	8.788***	–0.068	0.022	–3.055**
SK7/Vie/k6	0.104	0.042	2.463*				0.304	0.033	9.141***	–0.078	0.024	–3.241**
SK8/Bgld/k7	–0.053	0.023	–2.247*				–0.130	0.029	–4.439***	–0.088	0.025	–3.594***
SK9	–0.055	0.025	–2.177*									
SK10	–0.237	0.038	–6.277***									
SK11	–0.137	0.029	–4.767***									
F-test (*p*-value)	0.001			0.001			0.001			0.001		
Adj. R^2 (%)	44			42			45			43		
RMSE	0.324			0.332			0.327			0.327		

Note: SE = standard error; SK = SKATER; Vbg = Vorarlberg; Tyr = Tyrol; Car = Carinthia; Sty = Styria; Upp = Upper Austria; Sal = Salzburg; Vie = Vienna; Bgld = Burgenland; RMSE = root mean square error.

†$p < 0.1$.

*$p < 0.05$.

**$p < 0.01$.

***$p < 0.001$.

Table 2. Single equation model for each SKATER submarket

	SK1	SK2	SK3	SK4	SK5	SK6	SK7	SK8	SK9	SK10	SK11
Intercept	++++	++++	++++	++++	++++	++++	++++	++++	++++	++++	++++
Structural covariates											
lnarea_total	+++	++++	++++	++++	++++	++++	+++	++++	++++	++++	++++
lnarea_plot	+	+++			+++	++++	+++	++			
cond_house1			−−			−	+			−−	−
heat1	−−	−−−−		−		−−−					
bath1					−−						
attic1							−				
cellar1	+	+++		++++	++	++++		+++			
garage1		−−−−	−	−−				−	−−−−		−−
terr1		++++			++	+++			+++		
Temporal covariates											
age		−−−−	−−−−	−−−−	−−−−	−−−−	−	−−−−	−−−−		−−−−
time	+++	++++		+					++++		
Neighborhood covariates											
enumd_unempl	+	−−−			−		−	++		−−	
muni_pp_ind		+++			++		++				++++
muni_acad	+	++	++	++++	++	++	−	++++	+++	++	
muni_age_ind		−−		−−−−	−−−−	−−−		−−−			−−
ln_muni_popd	+	++++						+++	+++		
Adj. R^2 (%)	24	46	26	35	49	57	39	42	41	41	28
RMSE	0.511	0.345	0.362	0.357	0.377	0.335	0.511	0.348	0.369	0.399	0.348
n	100	1,003	134	402	293	328	85	309	304	107	239

Note: For all models, F-test $p < 0.001$. SK = SKATER; + = positive effect; − = negative effect; RMSE = root mean square error.
Significance levels: ++++ = 0.001; +++ = 0.01; ++ = 0.05; + = 0.10; −−−− = 0.001; −−− = 0.01; −− = 0.05; − = 0.10.

applied. The results are summarized in Table 2 and indicate local differences in the hedonic price function. Within each submarket, different covariates are significant to explain house prices. Total floor area has a highly significant positive effect and age has a negative effect across all submarkets. Furthermore, a tendency for submarkets with a smaller number of houses (n) to lead to a higher RMSE and thus less reliable estimates is noticeable. These findings are consistent with Adair, Berry, and McGreal (1996), where the spatial heterogeneity is also expressed in different combinations of covariates to explain house prices. Such single equation models, in particular the ones with a small sample size, can result in estimates having an unexpected sign. This is demonstrated in the model SK7 with only eighty-five houses. For example, the covariate poor condition of the house shows a positive effect, whereas the proportion of academics shows a negative effect, both of which are unexpected. Remaining submarkets having a larger number of observations show effects on house prices that are expected. To summarize, separate models for each submarket result in serious shortcomings, including unstable estimates due to a reduced sample size, being less prone to misspecifications, and being affected by statistical artifacts.

Conclusions

This article promotes a data-driven spatial regionalization framework for housing market segmentation. Well-established approaches, such as the ad hoc submarket definition, specify housing submarkets exogenously and have their administrative units' mimic the spatial extent (e.g., Adair, Berry, and McGreal 1996). Such an approach suffers from arbitrarily chosen boundaries not corresponding with spatial economic processes. It might induce the modifiable areal unit problem, which can further result in biased estimates of the hedonic price function. Therefore, local spatial analysis techniques are proposed to estimate local varying submarkets (e.g., Páez, Fei, and Farber 2008). Nevertheless, it is shown that due to serious methodological drawbacks this line of research is more capable for exploratory analysis and less suitable for predictions. To avoid the arbitrariness of ad hoc

submarket definitions, clustering of housing as well as socioeconomic attributes is applied, which results in homogeneous submarkets (e.g., Bourassa et al. 1999). Even though it is well known in real estate that space matters, submarket definitions by means of clustering algorithms have disregarded this important maxim so far. This might have led to erroneous conclusions.

In contrast, this research addresses and resolves these drawbacks. For the delimitation of housing market segmentation, a data-driven framework is proposed. First, stationary and nonstationary structural, temporal, and neighborhood effects on house prices are explored using the MGWR. The former represent economically connected real estate markets (e.g., through similar federal policy, like governmental subsidies), and the latter result in a continuous and local definition of a housing segmentation. These estimated MGWR coefficients tend to be highly correlated and show distinctive volatility. Hence, in a second step, the MGWR coefficients are reduced to a small number of orthogonal PCs, serving as input data for the regionalization. Finally, the SKATER algorithm is used to investigate homogenous and spatially contiguous housing market segmentation, explicitly accounting for spatial effects, which have been neglected so far. To determine an appropriate number of submarkets, a model-driven approach in the form of a hedonic regression is utilized. The out-of-sample prediction performance is applied to determine the "near-optimal" number of submarkets. Compared to previous approaches, this framework needs less theory and prior knowledge on the exact housing market segmentation and the underlying spatial economic process.

Finally, using a data set of 3,887 geocoded single-family homes throughout Austria from 1998 to 2009, the proposed methodological framework is empirically tested. A trade-off between prediction accuracy and model parsimony shows that eleven submarkets are most suitable for Austria. The modeled submarkets are further validated within a hedonic regression framework. It is demonstrated that a hedonic model considering the modeled SKATER submarkets significantly improves prediction quality compared to a pooled market-wide model. These results are in line with Goodman and Thibodeau (2003), as well as Hwang and Thill (2009). More important, the ad hoc submarket definition using federal states is significantly outperformed with this novel approach as well as the commonly used *k*-means–based segmentation, which results in comparatively higher prediction errors. In contrast to Bourassa, Cantoni, and Hoesli (2010), the results show that ad hoc submarkets are competitive to *k*-means–based submarkets. Due to a straightforward application of the ad hoc approach, the empirical application recommends considering administrative spatial indicators instead of the more sophisticated *k*-means clustering. Additionally, a different single hedonic model is estimated for each of the eleven submarkets to analyze the instability of structural, temporal, and neighborhood covariates across each submarket. The results show that the type of covariates and their significance levels differ across the submarkets and that the reliability of the estimates strongly depends on the sample size within each submarket. In accordance with the results in Abraham, Goetzmann, and Wachter (1994) for the U.S. housing market, it can be concluded that the Austrian housing market is based on strong regional determinants. In other words, geography is the essential component determining the housing market's characteristic.

To conclude, this article proposes a powerful methodological framework for housing segmentation, being superior in terms of prediction accuracy compared to former approaches. The empirical model competition clearly demonstrates that submarkets, however defined, must always be considered in hedonic modeling. Nevertheless, there are some limitations. First, the methodology has limited applicability for large data sets because each modeling component is computationally demanding. Second, this data-driven submarket definition must be further analyzed by a more extensive clustering and regionalization algorithm comparisons (e.g., Hagenauer and Helbich 2012) and alternative methodologies to reach its full potential. Such alternative methodologies include multilevel modeling or additive (mixed) modeling. For this reason, future additional investigations into the performance of submarkets will be necessary.

Acknowledgments

Marco Helbich was supported by the Alexander von Humboldt Foundation. We thank the four anonymous reviewers for their constructive comments that greatly improved the article.

Notes

1. A pooled model estimates a single regression for the whole study area, not accounting for submarket specific effects. In contrast, a completely unpooled model estimates a separate equation for each submarket. The SKATER, ad hoc, and *k*-means models lie in between these two extremes and can model spatial heterogeneity through intercept or slope variation.

2. Defuzzification translates the fuzzy clustering to a crisp representation by assigning an object to the cluster having the highest membership degree.
3. Due to a sparse sample size in Vorarlberg and Tyrol, both regions are considered as one spatial unit.
4. If both sides of the equation appear as a log–log specification, the coefficient β can be interpreted as elasticity. A 1 percent change of x results in a change of y by $\beta \times 100$ percent. In a log-linear specification, a change of one unit of x results in a change of y by $100 \times (\exp(\beta) - 1)$ percent. For small values, however, β can be interpreted as semielasticity, meaning that a one-unit change in x results in a $\beta \times 100$ percent change in y. See Greene (2008) for a more detailed discussion.

References

Abraham, J. M., W. N. Goetzmann, and S. M. Wachter. 1994. Homogeneous groupings in metropolitan housing market. *Journal of Housing Economics* 3 (3): 186–206.

Adair, A. S., J. N. Berry, and W. S. McGreal. 1996. Hedonic modelling, housing submarkets and residential valuation. *Journal of Property Research* 13 (2): 67–83.

Akaike, H. 1974. A new look at the statistical model identification. *IEEE Transactions on Automatic Control* 19 (6): 716–23.

Assunção, R. M., M. C. Neves, G. Câmara, and C. Da Costa Freitas. 2006. Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees. *International Journal of Geographical Information Science* 20 (7): 797–811.

Bação, F., V. Lobo, and M. Painho. 2005. The self-organizing map, the Geo-SOM, and relevant variants for geosciences. *Computers & Geosciences* 31 (2): 155–63.

Bischoff, O., and W. Maennig. 2011. Rental housing market segmentation in Germany according to ownership. *Journal of Property Research* 28 (2): 133–49.

Bourassa, S. C., E. Cantoni, and M. Hoesli. 2007. Spatial dependence, housing submarkets, and house price prediction. *The Journal of Real Estate Finance and Economics* 35 (2): 143–60.

———. 2010. Predicting house prices with spatial dependence: A comparison of alternative methods. *Journal of Real Estate Research* 32 (2): 139–60.

Bourassa, S. C., F. Hamelink, M. Hoesli, and B. MacGregor. 1999. Defining housing submarkets. *Journal of Housing Economics* 8 (2): 160–83.

Bourassa, S. C., M. Hoesli, and V. S. Peng. 2003. Do housing submarkets really matter? *Journal of Housing Economics* 12 (1): 12–28.

Brunauer, W., S. Lang, and N. Umlauf. 2010. Modeling house prices using multilevel structured additive regression. Working paper, Faculty of Economics and Statistics, University of Innsbruck, Austria. http://econpapers.repec.org/paper/innwpaper/2010–19.htm (last accessed 11 May 2012).

Brunauer, W., S. Lang, P. Wechselberger, and S. Bienert. 2010. Additive hedonic regression models with spatial scaling factors: An application for rents in Vienna. *The Journal of Real Estate Finance and Economics* 41 (4): 390–411.

Brunsdon, C., S. A. Fotheringham, and M. Charlton. 1996. Geographically weighted regression: A method for exploring spatial nonstationarity. *Geographical Analysis* 28 (4): 281–98.

Burnham, K., and D. Anderson. 2002. *Model selection and multimodel inference: A practical information-theoretic approach*. New York: Springer.

Can, A. 1992. Specification and estimation of hedonic house price models. *Regional Sciences and Urban Economics* 22 (3): 453–74.

Case, B., J. Clapp, R. Dubin, and M. Rodriguez. 2004. Modeling spatial and temporal house price patterns: A comparison of four models. *Journal of Real Estate Finance and Economics* 29 (2): 167–91.

Davies, D. L., and D. W. Bouldin. 1979. A cluster separation measure. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2:224–27.

Diggle, P., and J. Ribeiro. 2007. *Model-based geostatistics*. New York: Springer.

Dubin, R. A. 1992. Spatial autocorrelation and neighborhood quality. *Regional Science and Urban Economics* 22 (3): 433–52.

Farber, S., and A. Páez. 2007. A systematic investigation of cross-validation in GWR model estimation: Empirical analysis and Monte Carlo simulations. *Journal of Geographical Systems* 9 (4): 371–96.

Fassmann, H., P. Görgl, and M. Helbich. 2009. *Atlas der wachsenden stadtregion* [Atlas of the growing urban region]. Vienna: PGO.

Fotheringham, S. A., M. Charlton, and C. Brunsdon. 2002. *Geographically weighted regression: The analysis of spatially varying relationships*. Chichester, UK: Wiley.

Fotheringham, S. A., and D. W. S. Wong. 1991. The modifiable areal unit problem in multivariate statistical analysis. *Environment and Planning A* 23 (7): 1025–44.

Gelman, A., and J. Hill. 2007. *Data analysis using regression and multilevel/hierarchical models*. Cambridge, UK: Cambridge University Press.

Goetzmann, W. N., and S. M. Wachter. 1995. Clustering methods for real estate portfolios. *Real Estate Economics* 23 (3): 271–310.

Goldstein, H. 2011. *Multilevel statistical models*. Chichester, UK: Wiley.

Goodman, A. C. 1978. Hedonic prices, price indices and housing markets. *Journal of Urban Economics* 5 (4): 471–84.

Goodman, A. C., and T. G. Thibodeau. 1998. Housing market segmentation. *Journal of Housing Economics* 7 (2): 121–43.

———. 2003. Housing market segmentation and hedonic prediction accuracy. *Journal of Housing Economics* 12 (3): 181–201.

———. 2007. The spatial proximity of metropolitan area housing submarkets. *Real Estate Economics* 35 (2): 209–32.

Greene, W. H. 2008. *Econometric analysis*. 6th ed. Upper Saddle River, NJ: Pearson.

Griffith, D. A. 2008. Spatial-filtering-based contributions to a critique of geographically weighted regression (GWR). *Environment and Planning A* 40 (11): 2751–69.

Guo, D. 2008. Regionalization with dynamically constrained agglomerative clustering and partitioning (REDCAP).

International Journal of Geographical Information Science 22 (7): 801–23.
Hagenauer, J., and M. Helbich. 2012. Contextual neural gas for spatial clustering and analysis. *International Journal of Geographical Information Science*. doi: 10.1080/13658816.2012.667106.
Helbich, M. 2012. Beyond postsuburbia? Multifunctional service agglomeration in Vienna's urban fringe. *Journal of Economic & Social Geography* 103 (1): 39–52.
Helbich, M., and M. Leitner. 2009. Spatial analysis of the urban-to-rural migration determinants in the Viennese metropolitan area: A transition from sub- to postsuburbia? *Applied Spatial Analysis and Policy* 2 (3): 237–60.
Hwang, S., and J.-C. Thill. 2009. Delineating urban housing submarkets with fuzzy clustering. *Environment and Planning B: Planning and Design* 36 (5): 865–82.
Jain, A. K., R. P. W. Duin, and J. Mao. 2000. Statistical pattern recognition: A review. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22 (1): 4–37.
Jolliffe, I. T. 2002. *Principal component analysis*. New York: Springer.
Jones, K., and N. Bullen. 1994. Contextual models of urban house prices: A comparison of fixed- and random-coefficient models developed by expansion. *Economic Geography* 70 (3): 252–72.
Kauko, T. 2004. A comparative perspective on urban spatial housing market structure: Some more evidence of local sub-markets based on a neural network classification of Amsterdam. *Urban Studies* 41 (13): 2555–79.
Kauko, T., P. Hooimeijer, and J. Hakfoort. 2002. Capturing housing market segmentation: An alternative approach based on neural network modelling. *Housing Studies* 17 (6): 875–94.
Kestens, Y., M. Theriault, and F. Des Rosiers. 2004. The impact of surrounding land use and vegetation on single-family house prices. *Environment and Planning B: Planning and Design* 31 (4): 539–67.
Kohonen, T. 2001. *Self-organizing maps*. New York: Springer.
LeSage, J., and K. R. Pace. 2009. *Introduction to spatial econometrics*. Boca Raton, FL: CRC Press.
Leung, Y., C.-L. Mei, and W.-X. Zhang. 2000. Statistical tests for spatial nonstationarity based on the geographically weighted regression model. *Environment and Planning A* 32 (1): 9–32.
Maclennan, D. 1982. *Housing economics: An applied approach*. London: Longman.
Maclennan, D., and Y. Tu. 1996. Economic perspectives on the structure of local housing systems. *Housing Studies* 11 (3): 387–406.
Miller, H. J. 2010. The data avalanche is here. Shouldn't we be digging? *Journal of Regional Science* 50 (1): 181–201.
Neyman, J., and E. L. Scott. 1948. Consistent estimates based on partially consistent observations. *Econometrica* 16 (1): 1–32.
Openshaw, S. 1984. *The modifiable areal unit problem*. Norwich, UK: Geo Books.
Openshaw, S., and L. Rao. 1995. Algorithms for reengineering 1991 census geography. *Environment and Planning A* 27 (3): 425–46.
Orford, S. 2000. Modelling spatial structures in local housing market dynamics: A multilevel perspective. *Urban Studies* 37 (9): 1643–71.
Páez, A., S. Farber, and D. Wheeler. 2011. A simulation-based study of geographically weighted regression as a method for investigating spatially varying relationships. *Environment and Planning A* 43 (12): 2992–3010.
Páez, A., L. Fei, and S. Farber. 2008. Moving window approaches for hedonic price estimation: An empirical comparison of modelling techniques. *Urban Studies* 45 (8): 1565–81.
Palm, R. 1978. Spatial segmentation of the urban housing market. *Economic Geography* 54 (3): 210–21.
Pebesma, E. 2004. Multivariable geostatistics in S: The gstat package. *Computers & Geosciences* 30 (7): 683–91.
R Development Core Team. 2012. *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.
Reimann, C., P. Filzmoser, R. G. Garrett, and R. Dutter. 2008. *Statistical data analysis explained: Applied environmental statistics with R*. Chichester, UK: Wiley.
Rosen, S. 1974. Hedonic prices and implicit markets: Product differentiation in pure competition. *Journal of Political Economy* 82 (1): 34–55.
Schnare, A., and R. Struyk. 1976. Segmentation in urban housing markets. *Journal of Urban Economics* 3 (2): 146–66.
Statistics Austria. 2007. Commuters (based on census data 2001). http://www.statistik.at/web_de/statistiken/bevoelkerung/volkszaehlungen_registerzaehlungen/pendler/index.html (last accessed 2 February 2012).
Straszheim, M. 1975. *An econometric analysis of the urban housing market*. Cambridge, MA: National Bureau of Economic Research.
Venables, W. N., and B. D. Ripley. 2002. *Modern applied statistics with S*. New York: Springer.
Watkins, C. A. 2001. The definition and identification of housing submarkets. *Environment and Planning A* 33 (12): 2235–53.
Wheeler, D., and A. Páez. 2009. Geographically weighted regression. In *Handbook of spatial analysis*, ed. M. M. Fischer and A. Getis, 461–86. Oxford, UK: Elsevier.
Wheeler, D., and M. Tiefelsdorf. 2005. Multicollinearity and correlation among local regression coefficients in geographically weighted regression. *Journal of Geographical Systems* 7 (2): 161–87.

Correspondence: Institute of Geography, University of Heidelberg, Berliner Straße 48 D-69120, Heidelberg, Germany, e-mail: helbich@uni-heidelberg.de (Helbich); hagenauer@uni-heidelberg.de (Hagenauer); Credit Risk Methods Development, Strategic Risk Management & Control, Bank Austria—Member of UniCredit Group A-1090, Vienna, Austria, e-mail: wolfgang.brunauer@unicreditgroup.at (Brunauer); Department of Geography and Anthropology, Louisiana State University, Baton Rouge, LA 70803, e-mail: mleitne@lsu.edu (Leitner).

Appendix

Table A1. Description and descriptive statistics of the variables

Abbreviation	Description	Source	Effect	Minimum	Mean	Maximum
lnp	Log purchase price of the dwelling	BA		10.309	11.917	13.218
Structural covariates						
lnarea_total	Log of total floor area (except cellar)	BA	+	3.778	4.846	6.204
lnarea_plot	Log of plot space	BA	+	4.382	6.498	7.824
cond_house	Condition of the house (0 = good, 1 = poor)	BA	–	0.000		1.000
heat	Quality of the heating system (0 = high, 1 = poor)	BA	–	0.000		1.000
bath	Quality of the bathroom/toilet (0 = high, 1 = poor)	BA	–	0.000		1.000
attic	Attic (0 = no, 1 = yes)	BA	–	0.000		1.000
cellar	Cellar (0 = no, 1 = yes)	BA	+	0.000		1.000
garage	Quality of the garage (0 = high, 1 = poor)	BA	–	0.000		1.000
terr	Terrace (0 = no, 1 = yes)	BA	+	0.000		1.000
Temporal covariates						
age	Age of building at time of sale	BA	–	1.000	24.360	81.000
time	Year of purchase (1998-2009)	BA	+	0.000	7.224	11.000
Neighborhood covariates						
enumd_unempl	Unemployment rate (2009) enumeration district	MB	–	0.000	0.034	0.148
muni_pp_ind	Purchase power index (2009) municipality level	SA	+	65.000	102.796	148.500
muni_acad	Proportion of academics (2001) municipality level	SA	+	3.716	15.099	40.795
muni_age_ind	Age index (2001) municipality level	SA	–	33.040	39.321	46.136
ln_muni_popd	Log population density (2001) municipality level	SA	+	–3.519	0.727	4.857

Note: BA = Bank Austria AG; MB = Michael Bauer Research; SA = Statistics Austria.