Forecasting Housing Prices under Different Submarket Assumptions

Zhuo Chen[1], Seong-Hoon Cho[2], Neelam Poudyal[3], Roland K. Roberts[2]

Selected Paper prepared for presentation at the American Agricultural Economics Association

Annual Meeting, Portland, OR, July 29-August 1, 2007

[1] Corresponding author: Visiting Scholar, Chicago Center of Excellence in Health Promotion Economics, University of Chicago. Correspond to: 1423E Druid Valley Dr. NE, Atlanta, GA 30329, USA (e-mail:zchen1@cdc.gov)
[2] Department of Agricultural Economics, University of Tennessee, Knoxville, TN 37996, USA (e-mail:{scho9, rrobert3}@utk.edu)
[3] Department of Forestry, Wildlife and Fisheries, University of Tennessee, Knoxville, TN 37996, USA (e-mail: npoudyal@utk.edu

Acknowledgments: The authors thank Tim Kuhn and Gretchen Beal of the Knoxville-Knox County Metropolitan Planning Commission and Keith G. Stump of the Knoxville-Knox County-Knoxville Utilities Board Geographic Information System for providing urban growth boundary, school district, and land value data. However, the authors are responsible for any remaining errors. The views expressed are those of the authors. No official endorsement by their employers is intended, nor should be inferred.

Forecasting Housing Prices under Different Submarket Assumptions

Abstract

This research evaluated forecasting accuracy of hedonic price models based on a number of different submarket assumptions. Using home sale data for the City of Knoxville and vicinities merged with geographic information, we found that forecasting housing prices with submarkets defined using expert knowledge and by school district and combining information conveyed in different modeling strategies are more accurate and efficient than models that are spatially aggregated, or with submarkets defined by statistical clustering techniques. This finding provided useful implications for housing price prediction in an urban setting and surrounding areas in that forecasting models based on expert knowledge of market structure or public school quality and simple model combining techniques may outperform the models using more sophisticated statistical techniques.

JEL classification: C53; R21

Keywords: Clustering, Forecasting, Hedonic price, Housing Submarket

Forecasting Housing Prices under Different Submarket Assumptions

1. Introduction

Fueled by low mortgage rates, housing market values had increased rapidly in the United States (US) since the late 1990s. The average housing price rose by more than 45% between 1996 and 2003 after adjusting for inflation (Baker, 2005). Economists are engaged in an ongoing debate about whether the US housing market is over-inflated. Despite disagreement about housing market inflation, an agreement exists on the fact that the housing market has been an important force in the overall economy during the late 1990s and early this century. Rapid appreciation in house values had boosted homeowners' confidence, thereby powering consumer spending and driving the overall economy.

More recently, the US housing market appeared to be cooling. Total existing-home sales declined by 6% in April, 2006 compared with the same month in 2005 (NAR, 2006). Given the size of the housing market in the US economy, its future direction is of concern not only to home buyers and sellers but also to policy makers. Accurate housing price forecasts can provide valuable information to these parties for making decisions about housing market. Hedonic price models have been commonly used in modeling the relationship between housing price and the physical attributes and geographic characteristics of homes (Song, 1995). With a few notable exceptions (e.g., Goodman and Thibodeau, 1998, 2003; Bourassa *et al.* (1999, 2003)), previous research on housing prices has largely focused on the parametric specification of the model, i.e., using ordinary least square or Box-Cox transformed regressions to examine the interaction between housing prices and regressors. Goodman and Thibodeau (1998), among others, have argued that market segmentation is an important feature that should be modeled in housing price research. Market segmentation is the process of defining the suitability of a submarket for a specific housing property. Grigby *et al.* (1987) defined a submarket as a set of dwellings that are

reasonably close substitutes for one another, but relatively poor substitutes for dwellings in other submarkets.

The extant literature on hedonic price forecasting and submarket analysis (e.g., Michael and Smith,1990; Goodman and Thibodeau,1998; Dale-Johnson,1982; and Bourassa *et al.*, 1999,,2003) has shown that submarket identification is an important factor for the successful modeling of housing prices. Submarket identification is important because property prices in different submarkets are determined by different functional relationships. The task of delineating a large geographic area into several relatively homogenous housing submarkets raises numerous theoretical and methodological questions (Palm, 1978). Many studies used predefined or otherwise convenient geographical boundaries to identify submarkets (see discussions in Bourassa *et al.* (1999). Other literature has adopted more systematic methods, e.g., principal component analysis and clustering, to delineate submarkets.

Dale-Johnson (1982) and Bourassa *et al.* (1999, 2003) used factor analysis and clustering technique to assign individual observations to different submarkets. Goodman and Thibodeau (1998) proposed identifying housing submarket boundaries with estimated parameters from a hierarchical clustering analysis. Their model delineated submarkets where variation in public school quality explained variation in the hedonic coefficient for geographic sizes of school zones. Goodman and Thibodeau (2003) extended their earlier work by comparing the hedonic prediction accuracy for different methods of delineating housing submarkets, i.e., no spatial disaggregation, by zip code districts, by census tracts, and using the Goodman-Thibodeau technique to delineate submarkets. The concept behind the Goodman-Thibodeau technique is that all homes within a spatially concentrated area share amenities associated with the property's location. Consequently, the housing characteristics that determine a property's market value are nested in a hierarchical structure.

Comparisons among delineation methods revealed that spatial disaggregation yields significant gains in prediction accuracy. One difficulty with delineating submarkets using predefined geographical regions, such as zip codes, census tracts, and school zones, is that in many cases no obvious basis for internal homogeneity exists within these regions (Bourassa *et al.*, 1999). Some researchers used individual dwelling data to avoid complications from the assumption of internal homogeneity within predefined regions. For example, Day (2003) used hierarchical clustering techniques to identify property submarkets defined by a combination of property types, locations, and socioeconomic characteristics of the inhabitants of each housing unit.

Bourassa *et al,* (1999) concluded that the classification derived from the clustering procedure was significantly better than other submarket classification methods. Finally, they suggested using the same clustering procedure with a larger data set to investigate the difficult issue of determining the optimal number of submarkets. On the other hand, Bourassa *et al.* (2003) found that prediction accuracy of housing price using real estate appraisers' defined submarket is better than statistically defined submarket. The authors used the *k*-means clustering to define the submarkets and did not examine those results using alternative statistical techniques. Thus, comparing models based on submarkets defined by real estate appraisers with those with a more comprehensive set of submarket structures that are derived from a variety of delineation criteria and forecasting algorithm are needed to confirm the robustness of their conclusions. For example, *k*-means clustering has been criticized for relying on an *a priori* number of submarkets while two-step clustering uses no clear *a priori* information about the number of segments (SPSS, 2006). Furthermore, the forecasting accuracy based on expert defined submarket need to be compared with those based on submarket by school district zones because school quality has been found as one of the most important predictors of housing prices (Bogart and Cromwell,

1997; Goodman and Thibodeau, 1998, 2003, 2006; Hayes and Taylor, 1996; Brasington, 1999; Haurin and Brasington, 1996; Song, 1998).

The objective of this research was to investigate the accuracy of alternative housing market segmentation criteria in hedonic housing price forecasting. Our hypotheses are that allowing for market segmentation in housing-price models increases their forecasting accuracy and that the criteria used to segment housing markets greatly affect their forecasting accuracy. A significant contribution of this work was the use of a more comprehensive set of housing submarket delineation criteria than previous works (Goodman and Thibodeau, 2003; Bourassa *et al.*, 2003) for estimation and comparison of hedonic housing price models. Segmentation criteria used in the paper include no segmentation, three statistical clustering methods, predefined delineations of geographic areas, high school district zones and areas defined through expert knowledge. The forecasting accuracy was compared for models developed from these six segmentation criteria and for simple and weighted averages of the latter five models. To our knowledge, this list of segmentation criteria is more comprehensive than others found in the literature.

2. Data Description and Sources

Geographically digitalized data from three sources were used in this study: (a) property parcel records from the Knoxville - Knox County - Knoxville Utilities Board (KUB) Geographic Information System (KGIS, 2006) (b) data extracted from the 2000 US census (GeoLytics, 2006) and (c) geographical information from the 2004 Environmental Systems Research Institute (ESRI) Maps and Data (ESRI, 2006)..Knox County property parcel records contain detailed information about structural attributes of the properties. Census-block group data describe neighborhood characteristics. Distance characteristics were computed using the ESRI data.

Knox County includes 12 high school districts, 83 census tracts, and 234 census block groups. Of the 23,002 transactions that took place in Knox County between January 1998 and December 2002, 22,979 have complete information for use in the analysis (KGIS, 2006). To mitigate the impact of outliers, the numbers of observations were reduced to 18,380 by excluding transactions with sale prices less than the tenth percentile or higher than the ninetieth percentile.

Descriptive statistics for the variables used in the analysis are presented in Table 1. A typical sample home had 1,796 square feet of finished area, had 3 bedrooms, and rested on a 25,483 square-foot (0.59-acre) lot. About 71% of sample homes had a fireplace, 21% had all brick exterior walls, 4% had a pool, and 63% had a garage. In 2000, average travel time to work was 23 minutes, average per capita income was $24,300, and the average unemployment rate was 3%. The distance variables, i.e., distance to downtown, distance to railway, distance to golf course, distance to park, distance to water body, and distance to sidewalk were calculated using the ArcGIS 9.1 software (ESRI, 2006).

To assess how accurately sample regression coefficients capture the corresponding population regression coefficients, estimated regression coefficients were used to evaluate forecasting accuracy with an independent set of validation data. This practice is known as "cross-validation." For cross validation (Fetcher *et al.*, 2004; Goodman and Thibodeau, 2003) the pooled sample of 18,425 transactions was divided into an estimation sub-sample and a validation sub-sample randomly using Stata 9.0 (Stata, 2006). The estimation sub-sample contained 16,583 (approximately 90%) observations and the validation sub-sample included 1,842 (10%) observations.

Table 1 compares the descriptive statistics of the pooled sample and the estimation and validation sub-samples. The average transaction price for the pooled sample was $114,598; whereas for the estimation and validation sub-samples they were $114,684 and $113,835

respectively. The difference between the sub-samples is trivial and statistically insignificant, implying that cross validation results are reliable.

3. Submarket Identification

No universally accepted method exists in the housing literature to identify the optimal number of submarkets (e.g., Bourassa *et al.*,1999; 2003; Johnson, 1982; Michaels and Smith, 1990; Schnare and Struyk, 1976; Goodman and Thibodeau, 1998; 2003). We used six methods for delineating housing submarkets. These submarket models were: 1) no segmentation in the housing market; 2) the k-means clustering method; 3) a clustering method using a two-step procedure without housing prices; 4) a clustering method using a two-step procedure with housing prices; 5) using *a priori* information of high school districts; and 6) districts using *a priori* information from experts.

3.1 No Market Segmentation (Baseline model)

The no-segmentation model was estimated as a baseline model for comparisons with the forecasting accuracy of the models estimated with variables representing housing submarkets.

3.2 K-Means Clustering

The k-means clustering technique was used to identify submarkets based on census tract-level data (Bourassa *et al.*, 1999; 2003; Day, 2003). This method requires *a priori* specification of the number of submarkets. We started by initially specifying 15 as the number of submarkets. This starting number was chosen based on a logical reduction in the number of submarkets defined by local experts (realtors) described below. The number of clusters in the model was gradually lowered until the number of sales transaction was at least 500 transactions within the smallest submarket cluster following Goodman and Thibodeau (2003); Bourassa *et al.* (2003). This procedure yielded five clusters (see Figure 1). However, the k-means clustering approach has

been criticized for requiring an *a priori* assumption of the number of groups and its inability to handle categorical variables.

3.3 Two-step Clustering with Price

The first step of the two-step cluster method begins with pre-clustering observations for individual sales transactions by constructing a likelihood function and selecting the optimal number of clusters using either the Bayesian Information Criterion or the Akaike Information Criterion (AIC). A matrix containing Euclidean distances between all pairs of pre-clustered observations is then created. In the second step, these pre-clustered groups of original observations are treated as individual observations and re-grouped. Because a large number of original observations are grouped into a much smaller number of pre-clusters, traditional methods such as agglomerative hierarchical clustering are typically used to re-group the pre-clusters.

In contrast to the aforementioned k-means clustering method, the two-step clustering method determines the number of clusters without an *a prior* assumption about the initial number of groups. While the two-step clustering method can use continuous and categorical variables for clustering, it is preferred over k-means clustering when categorical variables are used (SPSS, 2006). The optimal number of clusters was determined by using AIC and the procedure yielded four clusters (see Figure 2).

3.4 Two-step Clustering without Price

The sale price was used as a clustering variable in the above method and in previous studies. Nevertheless, *a prior* inclusion of sale price in clustering may introduce the problem of overfitting. To determine whether the inclusion of sale price has an effect on submarket delineation and subsequent out-of-sample forecasting accuracy, sale price was excluded from the two-step clustering method. This procedure yielded three clusters (see Figure 3).

3.5 High School Districts

Existing literature indicated that school quality is a strong predictor of housing price (e.g., Bogart and Cromwell,1997; Hayes and Taylor, 1996; Brasington, 1999; Haurin and Brasington, 1996; Song 1998). Many housing submarket studies relied on this *a priori* intuition about school quality by defining school districts as submarkets and found strong support of doing so (e.g., Goodman and Thibodeau, 1998; 2003; 2006). In our study, Knox County's 12 high school districts were used to define the submarket structure (see Figure 4).

3.6 Expert-Defined Submarkets

The final market segmentation method used boundaries drawn by local realtors who have first-hand expert knowledge and understanding of the local housing market. We drew the boundaries in GIS form based on interviews with local realtors and a sub-area map found on the web site of the Knoxville Area Association of Realtors® Internet Data Exchange Program[1] (2006) . The map was created to help realtors and home buyers search for residential homes of specific types in particular areas of town. The sub-area map was developed by a group of professional realtors based on settlement patterns, housing and neighborhood characters, and housing choices of new buyers. The map has 23 sub-areas, including 9 smaller sub-areas in downtown Knoxville and 14 larger sub-areas. We used the 14 larger sub-areas as submarkets and aggregated the downtown Knoxville sub-areas into one submarket for a total of 15 expert-defined submarkets (see Figure 5). The aggregation of the 9 downtown sub-areas was justified given that these areas share many common submarket characters for the downtown neighborhood.

Frequency distributions for the no-market segmentation, k-means clustering, two-step clustering with price, two-step clustering without price, high school districts, and expert-defined submarkets are reported in Table 2.

4. Empirical Hedonic Model

The general hedonic model used in this study was:

$$\ln y_i^j = \beta_0^j + \sum_k \beta_k^j x_{ik}^j + \varepsilon_i^j \qquad i = 1, 2, ..., N^j; \; j = 1, 2, ..., J \tag{1}$$

where $\ln y_i^j$ is the natural logarithm of the sale price of house *i* in submarket *j*; x_{ik}^j is *k*th variable of structure, neighborhood, distance, and time characteristics; J is the total number of submarkets; N^j is the number of observations in submarket *j*; and ε_i^j is a residual capturing the random disturbance for submarket *j*.

Nine key structural characteristics were available and included in this study: finished area, age, lot size, number of stories, number of bedrooms, and if the structure included a fireplace, garage, all brick walls, and/or pool. Condition and quality variables were also included. The latter two variables were defined on a scale of 1 to 6 for the following categories as rated by the Knox County tax assessors' office: excellent, very good, good, average, fair, and poor. These structural, quality, and condition variables served as control variables and were typically found to be important in explaining housing price variation in the literature (e.g., Song, 1995; Bin and Polasky, 2004; Bourassa *et al,*. 2003).

Some of the variables representing neighborhood characteristics were extracted from the 2000 U.S. Census for census-block groups (GeoLytics, 2006),,including population density, per capita income, travel time to work, vacancy rate, unemployment rate, rural, and rural-urban interface areas. Population density was included to capture the effect of population pressure on land and natural resources (Katz and Rosen, 1987; Song, 1998). Per capita income and the unemployment rate were included as measures of economic conditions within a neighborhood (Down, 2002; Song, 1998; Phillips and Goodstein, 2000). Since residents in urban areas value access to employment (Small and Song, 1992; Song, 1995), average travel time to work was

included as a spatial measure of distance to the employment hub. Vacancy rate was a proxy for prevailing housing market conditions (Dowall and Landis, 1982). Differences often occur between rural and urban areas with regard to public services such as roads and law enforcement. These differences were captured using dummy variables for houses within rural areas and within the rural-urban interface, using houses within urban areas as the reference dummy variable.

Another set of neighborhood variables included boundary dummy variables and high school dummy variables. The study area (City of Knoxville and Knox County, Tennessee) adopted an urban growth boundary (UGB) in 2001. Based on Public Chapter 1101 of the Growth Policy Act in 1998 (TACIR, 2006), the county, the city, and towns within the county identified three classifications of land: rural areas, area enclosed by UGBs, and planned growth areas (PGAs).[2] Dummy variables for the UGB that excludes city area (Henceforth it is noted as UGB for simplicity) and PGA were included to capture the potential effects of growth boundary statutes.

Distance variables included distances to downtown Knoxville, nearest water body, nearest greenway, nearest railroad, and nearest sidewalk. These distance variables were intended to capture the effects on housing prices of proximities to various amenities and disamenities. Size of the nearest park was included to measure the premium for being close to more park amenities, which was a significant factor in explaining property value (Lutzenhiser and Netusil, 2001).

Previous studies found that a logarithmic transformation of distance variables generally performs better than a simple linear functional form because the transformation captures the declining effects of these distance variables (Bin and Polasky, 2004; Iwata, Murao, and Wang, 2000; Mahan, Polasky, and Adams, 2000). Logarithmic transformations of some quadratically

specified distance variables were tried yielding no improvement in model fit. Thus, a simple logarithmic transformation for distance-related variables was used in this study.

To account for potential annual and seasonal variation, year and month dummy variables were included with 1998 as the reference year and January as the reference month. Housing prices may vary seasonally—that is, prices are typically higher in spring and summer months irrespective of the overall trend. More buyers tend to be in the market during these months, increasing the demand for housing.

5. Forecast-Combining Algorithms

Several forecast-combining algorithms have been proposed in the literature. The common conclusion from the literature was that forecasting accuracy can be substantially improved by combining multiple individual forecasts (Clemen, 1989). For this reason, two forecast-combining algorithms were evaluated and compared with forecasts from the six individual housing submarket models.

In some cases, a simple average of individual forecasts was found to produce superior forecasting accuracy than more complicated algorithms (e.g., Armstrong *et al.*, 1983; Makridakis and Hibon, 2000). Hence, the simple average of forecasts generated by the five methods of delineating housing submarkets (excluding no market segmentation) was used as a simple forecast-combining algorithm. Predicted housing prices were generated from the individual models as the expectation of the log-normal distribution using the estimated mean and standard errors in the corresponding hedonic price function.

For comparison with the simple average, an adapted version of a more complex forecasting algorithm, a variation of the “encompass combining algorithm” proposed by Granger and Ramanathan (1984) was used. Their algorithm assigns higher weights to more accurate

forecasting methods. In the first stage of their algorithm, the actual house price (y) was regressed on the set of price forecasts from the five submarket models $(\hat{y}^m s)$ using ordinary least square,

$$y_i = \alpha_0 + \sum_{m=1}^{5} \alpha_m \hat{y}_i^m + \xi_i, \qquad i = 1,...,N, \tag{2}$$

where m represents the five submarket models. In the second stage, parameter estimates $(\hat{\alpha}_m, m = 0,1,...,5))$ were used to construct the encompass combined forecast y^{ec},

$$y_i^{ec} = \alpha_0 + \sum_{m=1}^{M} \hat{\alpha}_m \hat{y}_i^m, \qquad i = 1,...,N.. \tag{3}$$

6. Measures of Forecasting Accuracy

Forecasts were obtained from the aforementioned models using the estimation sample and the validation sample was used to compare and test the forecasting accuracy of those models. Recent literature has discussed the comparison of alternative forecasting models (e.g., Chen and Lian, 2005; Chen and Yang, 2004; Fletcher *et al.*, 2004; Fetcher *et al.*, 2000; Makridakis, 1993; Campbell, 2002). The forecasting literature has provided a long list of accuracy measures, e.g., absolute percentage error (APE), proportional prediction error (PPE), and root mean square error (RMSE), just to name a few. In the third of a series of forecasting competitions (M3 competition) between various forecasting procedures using data series provided by Makridakis and colleagues, Makridakis and Hibon (2000) observed that different accuracy measures produce different rankings of forecasting models. Chen and Yang (2004) and Kunst and Jumah (2004), among others, echoed their conclusion, pointing out an empirical difficulty associated with selecting a particular accuracy measure. Generally, APE (also referred as Mean APE in forecasting literature) is commonly used. Ahlburg (1992) found that of 17 research papers, ten used APE. Nonetheless, APE is criticized for asymmetry and instability when the original value of y is close to zero (Koehler, 2001). However, in our study, housing prices were generally nontrivial, mitigating issues of instability.

In this study, we used mean error, APE, and PPE in evaluating forecasting accuracy following Goodman and Thibodeau (2003). APE was calculated as (Makridakis, 1993):

$$\text{(mean) } APE^r = \frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i - \hat{y}_i^r}{y_i}\right| \times 100\%, \; r = 1, 2, \ldots, 8, \tag{4}$$

where y_i is the true price of the house and $\hat{y}^r$ represents the predicted price that was generated by the *r*th forecasting procedure. PPE is the error divided by the true price of the house and it is expressed as (Goodman and Thibodeau,2003).

$$\text{(mean) } PPE^r = \frac{1}{N}\sum_{i=1}^{N}\left(\frac{y_i - \hat{y}_i^r}{y_i}\right) \times 100\%. \tag{5}$$

The PPE differs from mean error by penalizing deviations from lower housing prices more than deviations from higher housing prices. Suppose two houses with prices of $50,000 and $500,000 have a same prediction error of $5,000, their contributions to mean error are the same but the PPEs are different, 10 percent and 1 percent respectively.

7. Empirical Results

Table 3 shows the mean error, APE, and PPE calculated with the validation samples for the eight forecasting models. Means in Table 3 relate to forecasting accuracy while standard deviations indicate forecasting efficiency. Mean errors and mean PPE have little implication for forecasting accuracy because the negative values of errors and PPEs could cancel the positive values of errors and PPEs respectively. Negative values of mean errors and PPEs revealed that all models consistently underestimated housing prices. In contrast to the mean values, standard deviations of the error and the PPE have implications for efficiency of the models. The error standard deviation for the expert-defined submarket model is $23,456, or $2,141 (9%) less than the no-market-segmentation model. While the expert-defined submarket model has the smallest error

standard deviation among submarket model, the high school districts, simple average and the encompass combining algorithm generated about the same error standard deviation. Based on the error standard deviation, there is not much difference among these four models in terms efficiency in forecasting. The PPE standard deviation for the expert-defined submarket model is 3.4% lower than the PPE standard deviation for the no-market-segmentation model. Although, the evaluation of PPEs indicates that the expert-defined submarket model is the most efficient forecasting model, school district model and other two statistical algorithms produce the very similar result.

Because the mean APE does not cancel positive and negative values, it can be used to evaluate efficiency as well as accuracy. The mean APE for the expert-defined submarket is 14.38%, or 1.13% less than the no-market-segmentation model. Interestingly, the high school district model and two forecast-combining algorithms produced similar mean APEs compared with the expert-defined submarket model, while the encompass combining algorithm projected a slightly lower mean APE than the simple average. The standard deviations for the two-step clustering with price and high school district models were estimated to be about 4% lower than the no-market-segmentation model. Based on the mean APE, the expert-defined submarket model, high school district and the two forecast-combining algorithms are the most accurate forecasting models, while the two-step clustering with price and high-school-district models are the most efficient forecasting models. Hence, the above mentioned results indicate the consistently similar accuracy of high school district with that of expert defined model, the high school district model is found to be even more efficient than that defined with expert knowledge.

The three forecasting accuracy measures have revealed that the expert-defined submarket model, high school district and the encompass combining algorithm performed relatively well. The mean APE result indicates that these three models are the most accurate forecasting models

and the standard deviations of mean errors, PPEs, and APEs suggest that these models have the most efficient forecasting ability. Overall, the five submarket models and the two forecast-combining algorithms performed better than the no-market-segmentation model. The mean APE and standard errors of mean errors, PPEs, APEs consistently showed that the no-market-segmentation model has relatively low accuracy and efficiency.

Of the three measures of prediction error (error, PPE, and APE), the value of APE is the only one that addresses prediction accuracy, although the standard errors of the three measures have implications for forecasting efficiency. Thus, the value of APE is further scrutinized. Table 4 presents the distribution of APE cumulative percentile for the eight forecasting models. The cumulative frequency percentile is presented for 1%, 5%, 10 %, 25%, … 99%. These cumulative frequency percentiles are presented to evaluate their prediction accuracy under the framework of the automated valuation model (AVM). The AVMs are computer algorithms that provide real estate market analysis and estimates of housing values (Moore, 2005). The criterion used in the real estate industry that adopts AVM requires that at least 50% of the predicted house prices should be within 10% of the observed sale prices (Goodman Thibodeau, 2003). All five submarket models and the two forecast-combining algorithms meet the AVM's 10% threshold criterion while the no-market-segmentation model fails the threshold criterion. With a trivial difference, the median deviation of the predicted values based on the expert-defined submarket model, high school districts, simple average and the encompass combining algorithm from the true housing prices were notably different from and significantly smaller than other set of submarket models (in the range of 8.6 to 8.8 in the former set to 9.1 to 10.1 in the latter set). The numbers indicate that the predicted values of 50% of the transactions were within an interval enclosed by their true sale prices plus and minus 8.6% (or 8.8%). Again, this confirms the better

performance of the expert-defined submarket model, high school districts and the two combining algorithm than those of the rest.

Overall, our result is consistent with the findings of Goodman and Thebodeau (2003) that models with spatially disaggregated submarkets perform better in terms of housing price forecasting and Bourassa *et al.* (2003) that the expert defined submarket structure has better prediction accuracy than those generated from simple statistical technique (eg, *k*-means clustering). Our result also shows that forecasting accuracy and efficiency of expert defined submarket does not significantly differ from that in submarkets based on high school district criteria and some other commonly used forecasting combining algorithms, although it shows advantage over models based submarkets identified through more complicated statistical procedures.

8. Conclusion

Notwithstanding the widespread application of the hedonic model in housing studies, with several notable exceptions, little research has been performed on the housing-price forecasting accuracy of these models. This paper compares the forecasting accuracy of the hedonic model applied with various submarket structures. These submarkets were either imposed or derived from intuition or from the data using different clustering techniques. Using multiple measures of forecasting accuracy, we compare a single-market hedonic model with five spatially disaggregated submarket models as well as two forecast-combining algorithms.

Although this study has not provided a clear choice for the “best” housing-price forecasting model, it has demonstrated that the forecasting accuracy of the hedonic price model improves when *a priori* expert knowledge, school district, and combining information conveyed in different modeling strategies are used to define housing submarkets. Particularly, the

forecasting accuracy of the expert-defined submarket structure, using *a priori* information from realtors who deal with the local real estate market on daily basis, and the high school district reflecting school quality perform significantly better than the models that use systematic clustering techniques to define submarkets. This result implies that the boundaries drawn with expert knowledge and school quality serve better for housing market segmentation than the boundaries drawn with clustering or predefined geographic units.

Using a more comprehensive set of submarket structures that are derived from a variety of delineation criteria, we have obtained results that are consistent with Bourassa *et al.*'s (2003) finding where prediction accuracy of housing price using real estate appraiser's defined submarket is better than statistically defined submarket. Furthermore, we confirm that school district and combining information conveyed in different modeling strategies are on par with the expert-defined submarket structure. This implies that under the circumstance where the expert defined submarkets are not easily available or are not in consensus, the housing price forecasting based on submarkets by high school district may serve as well as expert defined submarket.

References

Ahlburg, A. "A commentary on error measures: error measures and the choice of a forecast method." *International Journal of Forecasting* 8 (1992): 99-100.

Armostrong, J. S. , Lusk, E.J. Gardner, E.S. Jr. , Geurts, M. D. , Lopes, L. L., Markland, R. E. , McLaughlin, R. L. , Newbold, P. , Pack, D. J., Andersen, A. , Carbone, R. , Fildes, R. , Newton, H. J. , Parzen, E. , Winker, R. I. , and Makridakis, S. "Commentary on the Makridakis time series competition (M-Competition)". *Journal of Forecasting* 2 (1983): 259-311.

Baker, D. "The housing bubble fact sheet, Center for Economic and Policy Research, Issue Brief, (2005).

Bin, O. , S. Polasky. "Effects of flood hazards on property values: evidence before and after hurricane flood." *Land Economics* 80 (2004): 490-500.

Bogart, W. T., B.A. Cromwell. "How much more is a good school district worth?", *National Tax Journal* 50 (1997)**:**280-305.

Bourassa, S. C., F. Hamelink, M. Hoesli, and B. D. Macgregir. "Defining housing submarkets." *Journal of Housing Economics* 8 (1999): 160-83.

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

Brasington, D. "Which measures of school quality does the housing market value?" *Journal of Real Estate Research* 18 (1999): 395-413.

Campbell, P. R. Evaluating forecast error in state population projections using census 2000 counts, Population Division Working Paper Series NO. 57, Population Division, U. S. Bureau of the Census, Washington D. C. 2002.

Chen, W. Y., and K. K. Lian. "A comparison of forecasting models for Asian equity markets." *Sunway Academic Journal* 2 (2005): 1-12.

Chen, Z. and Y. Yang. Assessing forecast accuracy measures, Department of Statistics, Iowa State University, Preprint Series: 2004-10, http://seabiscuit.stat.iastate.edu/departmental/preprint/articles/2004-10.pdf, accessed September, 2006.

Clemen, R. T. "Combining forecasts: a review and annotated bibliography." *International. Journal of Forecasting* 5 (1989): 559-83.

Dale-Johnson, D. "An alternative approach to housing market segmentation using hedonic price data" Journal *of Urban Economics* 11 (1983): 311-32.

Day, B. "Submarket identification in property markets: A hedonic housing price model for Glasgow." Working Paper, The Centre for Social and Economic Research on the Global Environment, School of Environmental Science, and University of East Anglia, 2003.

Dowell, D. and J. D. Landis "Land use controls and housing costs: an examination of San Francisco bay area communities." *AREUEA Journal*10 (1982) 67-93.

Down, A. "Have housing prices risen faster in Portland than Elsewhere?" *Housing Policy Debate*, 13 (2002): 7-31.

ESRI, Environment and Scientific Research Institute Maps and Data. (2006) Available at http://www.esri.com/

Fletcher, M., J. Mangan and E. Raeburn. "Comparing hedonic models for estimating and forecasting house prices." *Property Management* 22 (2004): 189-200.

Fletcher, M., P. Gallimore, and J. Mangan. "The modelling of housing submarkets." *Journal of Property Investment and Finance*, 18 (2000): 473-87.

GEOLYTICS. The 2000 Long Form Census CD. Available at: http://www.geolytics.com/USCensus,Census-2000-Long-Form, Products.asp.
Goodman, A. C., and T. G. Thibodeau. "The spatially proximity of metropolitan area housing submarkets." Forthcoming in *Real Estate Economics*, 2006.
Goodman, A. C., T. G.Thibodeau. "Housing market segmentation." *Journal of Housing Economics* 7 (1998): 121-43.
Goodman, A. C., T.G. Thibodeau. "Housing market segmentation and hedonic prediction accuracy." *Journal of Housing Economics* 12 (2003):181-201.
Grangler, C. W., and R. Manathan. "Improved methods for combining forecasts." *Journal of Forecasting* 3 (1984): 197-204.
Haurin, D. R. and D. M. Brasington. "The impact of school quality on real house prices: inter- and intrametropolitan effects." *Journal of Housing Economics* 5 (1996): 351-68.
Hayes, K. J. and L. L. Taylor. "Neighborhood school characteristics: what signals quality to homebuyers?" *Federal Reserve Bank of Dallas Economic Review Fourth Quarter* (1996): 2-9.
Grigsby, W. , M. Baratz, G. Galster and D. Maclennan. "The dynamics of neighborhood change and decline, progress in planning." 28 (1987): 1-76.
Iwata, S., H. Murao, and Q. Wang. "Nonparametric assessment of the effects of neighborhood land uses on the residential house values." Advances in Econometrics: Applying Kernel and Nonparametric Estimation to Economic Topics 14, 2000.
Johnson, D. D. "An alternative approach to housing market segmentation using hedonic price data." *Journal of Urban Economics* 11 (1982): 311-32.
Katz, L. and K. Rosen. "The interjurisdictional effects of growth controls on housing prices." *Journal of Land Economics* 30 (1987): 149-60.
KGIS. (2006) Knoxville, Knox County, Knoxville Utilities Board Geographic Information System, "Knox net Where." Available at http://www.kgis.org/KnoxNetWhere.
Knoxville Area Association of Realtors® Internet Data Exchange Program. (2006), Available at: http://public.kaarmls.com/
Koehler, A. B. "The asymmetry of the APE measure and other comments on the M3-competition." *International Journal of Forecasting* 17 (2001): 570-74.
Kunst, R. M., Jumah, A. Toward a theory of evaluating predictive accuracy, Economics Serie 162, ISSN: 1605-7996, the Department of Economics and Finance, Institute for Advanced Studies, Vienna, Austria 2004.
Lutzenhiser, M. and N. R. Netusil. "The effect of open space type on a home's sale price: Portland, Oregon." *Contemporary Economic Policy* 19 (2001): 291-98.
Mahan, B. L., S. Polasky and R. M. Adams. "Valuing urban wetlands: a property price approach." *Land Economics* 76 (2000) 100-13.
Makkridakis, S. and M. Hibon. "The M3-Competition: results, conclusions and implications." *International Journal of Forecasting* 16 (2000): 451-76.
Makridakis, S. "Accuracy measures: theoretical and practical concerns" *International Journal of Forecasting* 9 (1993): 527-29.
Michaels, R. G. and V. K. Smith. "Market segmentation and valuing amenities with hedonic models: the case of hazardous waste sites." *Journal of Urban Economics* 28 (1990): 223-42.
Moore, J. W. "Performance comparison of automated valuation models." *Journal of Property Tax Assessment & Administration* 3 (2005): 43-60.

NAR, National Association of Realtors. (2006) Existing-Home Sales Slip in April, (http://www.realtor.org/PublicAffairsWeb.nsf/Pages/EHS06April) Accessed on June 8, 2006.

Palms, R. "Spatial segmentation of the urban housing market." *Economic Geography* 54 (1978): 210-21.

Philips, J. and E. Goodstein. "Growth management and housing prices: the case of Portland, Oregon." *Contemporary Economic Policy* 18 (2000): 334-44.

Schnare, A. B. and R. J. Struyk. "Segmentation in urban housing markets." *Journal of Urban Economics* 3 (1976): 146-66.

Small, K. and S. Song. "'Wasteful' commuting: a resolution." *The Journal of Political Economy* 100 (1992) 4: 888-98.

Song, S. "Determinants of bargaining outcome in single-family housing transactions: an empirical examination." *Urban Studies* 32 (1995): 605-14.

Song, S. "Does generalizing density functions better explain urban commuting?, Some evidence from the Los Angeles region." *Applied Economics Letters*, 2, pp. 148-50.

Song, S. "Home buyer's characteristics and selling prices." *Applied Economics Letters* 5 (1998): 11-14.

SPSS. (2006) The SPSS TwoStep cluster component. a scalable component enabling more efficient costumer segmentation. White Paper-Technical Report. www.spss.com.

STATA, Statistical Software for Professionals. (2006) Available online at wwww.stata.com.

TACIR, Tennessee Advisory Commission on Intergovernmental Relations. Growth policy, annexation, and incorporation, under public chapter 1101 of 1998: a guide for community leaders (2006). Available at: http://www.state.tn.us/tacir/PDF_FILES/Growth_Policy/Annexation98.pdf.

Notes

1. The map is available at
http://public.kaarmls.com/property/searchDetails.asp?AreaID=1&pt=1

2. The rural areas include land to be preserved for farming, recreation, and other non-urban uses. The land within the UGB is reasonably compact but adequate to accommodate the entire city's expected growth for the next 20 years. PGAs are large enough to accommodate growth expected to occur in unincorporated areas over the next 20 years (MPC, 2001).

Table 1. Variable Name, Definition, and Descriptive Statistics

Variable	Definition	Pooled sample	Estimation sample	Validation sample
Dependent Variable				
House price	Housing sales price	114,598.30	114,683.80	113,834.83
		(41,306.09)	(41,345.23)	(40,957.83)
Structural Variable				
Finished area	Total finished structure square footage	1,793.16	1,794.89	1,780.74
		(716.54)	(7,211.25)	(6,742.44)
Age	Year house was built subtracted from 2005	27.78	27.79	27.70
		(20.18)	(20.21)	(19.89)
Lot size	Lot square footage	25,483.30	25,723.69	23,335.35
		(67,896.71)	(70,205.51)	(41,916.41)
Stories	Count of stories	1.30	1.30	1.29
		(0.46)	(0.46)	(0.45)
Bedroom	Count of Bedroom	3.03	3.03	3.05
		(0.57)	(0.57)	(0.57)
Fireplace	Dummy variable for fireplace (1 if fireplace 0 otherwise)	0.71	0.71	0.70
		(0.55)	(0.55)	(0.54)
Garage	Dummy variable for garage (1 if garage 0 otherwise)	0.63	0.64	0.61
		(0.48)	(0.47)	(0.48)
Brick	Dummy variable for all brick exterior walls (1 if all brick 0 otherwise)	0.21	0.21	0.21
		(0.41)	(0.41)	(0.40)
Pool	Dummy variable for pool (1 if pool 0 otherwise)	0.04	0.04	0.04
		(0.21)	(0.21)	(0.21)
Condition	Dummy variable for condition of structure (1 if excellent, very good, and good 0 otherwise)	0.75	0.75	0.76
		(0.42)	(0.42)	(0.42)
Quality	Dummy variable for quality of construction (1 if excellent, very good, and good 0 otherwise)	0.29	0.29	0.29
		(0.45)	(0.45)	(0.45)
Neighborhood Variable				
Population density	Population density for census-block group in 2000	2.36	2.36	2.34
		(1.73)	(1.76)	(1.52)

Per capita income	Per capita income for census-block group in 2000	24,278.09	24,276.32	24,293.89
		(8,358.67)	(8,337.08)	(8,551.47)
Travel time to work	Average travel time to work for census-block group in 2000 (minuets)	22.67	22.68	22.58
		(3.35)	(3.36)	(3.29)
Vacancy rate	Vacancy rate for census-block group in 2000, which is unoccupied housing units in 2000.	0.06	0.06	0.06
		(0.02)	(0.02)	(0.02)
Unemployment rate	Unemployment rate for census-block group in 2000	0.03	0.03	0.03
		(0.02)	(0.02)	(0.02)
Rural area	Dummy variable for rural area (1 if census-block group with only rural type houses 0 otherwise)	0.06	0.06	0.06
		(0.25)	(0.25)	(0.24)
Rural-urban interface	Dummy variable for rural-urban interface (1 if census-block group with mixed urban and rural type houses 0 otherwise)	0.22	0.22	0.23
		(0.42)	(0.41)	(0.42)
Urban growth boundary	Dummy variable for urban growth boundary (1 if urban growth boundary 0 otherwise)	0.08	0.09	0.07
		(0.28)	(0.28)	(0.26)
Planned growth area	Dummy variable for planned growth area (1 if planned growth area 0 otherwise)	0.46	0.46	0.47
		(0.49)	(0.49)	(0.49)
Bearden	Dummy variable for Bearden High School District (1 if Bearden, 0 otherwise)	0.16	0.16	0.16
		(0.36)	(0.36)	(0.36)
Carter	Dummy variable for Carter High School District (1 if Carter, 0 otherwise)	0.03	0.03	0.03
		(0.18)	(0.18)	(0.18)
Central	Dummy variable for Central High School District (1 if Central, 0 otherwise)	0.09	0.09	0.09
		(0.29)	(0.29)	(0.29)
Doyle	Dummy variable for Doyle High School District (1 if Doyle, 0 otherwise)	0.06	0.06	0.07
		(0.25)	(0.25)	(0.25)
Fulton	Dummy variable for Fulton High School District (1 if Fulton, 0 otherwise)	0.04	0.04	0.04
		(0.20)	(0.20)	(0.19)
Gibbs	Dummy variable for Gibbs High School District (1 if Gibbs, 0 otherwise)	0.06	0.06	0.06
		(0.24)	(0.24)	(0.24)
Halls	Dummy variable for Halls High School District (1 if Halls, 0 otherwise)	0.06	0.06	0.06
		(0.24)	(0.24)	(0.24)
Karns	Dummy variable for Karns High School District (1 if Karns, 0 otherwise)	0.16	0.16	0.16
		(0.37)	(0.37)	(0.37)

Powell	Dummy variable for Powell High School District (1 if Powell, 0 otherwise)	0.07 (0.26)	0.07 (0.26)	0.06 (0.24)
Austin	Dummy variable for Austin High School District (1 if West, 0 otherwise)	0.00 (0.08)	0.00 (0.08)	0.00 (0.08)
Farragut	Dummy variable for Town of Farragut & Farragut High School District (1 if Farragut, 0 otherwise)	0.11 (0.32)	0.11 (0.32)	0.13 (0.34)
Distance Variable				
Downtown	Distance to downtown Knoxville (feet)	44,991.79 (19,009.86)	45,020.19 (19,002.96)	44,737.99 (19,074.67)
Greenway	Distance to nearest greenway (feet)	7,970.83 (5,724.16)	7,974.29 (5,733.96)	7,939.91 (5,898.05)
Water	Distance to nearest stream, lake, and river (feet)	8,905.03 (6,023.85)	8,904.56 (6,037.94)	8,909.24 (5,898.05)
Sidewalk	Distance to nearest sidewalk (feet)	3,470.90 (4,938.17)	3,467.48 (4,945.06)	3,501.43 (4,877.43)
Golf	Distance to nearest golf course (feet)	11,056.63 (5,080.07)	11,048.84 (5,089.06)	11,126.25 (4,999.83)
Railroad	Distance to nearest railroad (feet)	7,182.90 (5,539.48)	7,220.32 (5,559.69)	6,848.57 (5,345.37)
Park Size	Area of nearest park (square feet)	16,55,444.00 (60,95,089.00)	16,57,784.00 (62,14,606.00)	1634383.00 (4890652.00)
Time Variable				
1999 Dummy	Dummy variable for year of sale (1 if 1999 0 otherwise)	0.19 (0.39)	0.19 (0.39)	0.18 (0.38)
2000	Dummy variable for year of sale (1 if 2000 0 otherwise)	0.19 (0.39)	0.19 (0.39)	0.19 (0.39)
2001	Dummy variable for year of sale (1 if 2001 0 otherwise)	0.22 (0.41)	0.22 (0.41)	0.22 (0.41)
2002	Dummy variable for year of sale (1 if 2002 0 otherwise)	0.23 (0.42)	0.23 (0.42)	0.22 (0.41)
February	Dummy variable for month of sale (1 if February 0 otherwise)	0.07 (0.25)	0.07 (0.25)	0.06 (0.24)
March	Dummy variable for month of sale (1 if March 0	0.08	0.08	0.09

	otherwise)	(0.28)	(0.28)	(0.29)
April	Dummy variable for month of sale (1 if April 0 otherwise)	0.08 (0.28)	0.08 (0.28)	0.09 (0.29)
May	Dummy variable for month of sale (1 if May 0 otherwise)	0.10 (0.30)	0.10 (0.30)	0.09 (0.29)
June	Dummy variable for month of sale (1 if June 0 otherwise)	0.09 (0.29)	0.10 (0.30)	0.09 (0.29)
July	Dummy variable for month of sale (1 if July 0 otherwise)	0.09 (0.29)	0.09 (0.29)	0.08 (0.28)
August	Dummy variable for month of sale (1 if August 0 otherwise)	0.09 (0.29)	0.09 (0.29)	0.10 (0.31)
September	Dummy variable for month of sale (1 if September 0 otherwise)	0.07 (0.26)	0.07 (0.26)	0.07 (0.26)
October	Dummy variable for month of sale (1 if October 0 otherwise)	0.08 (0.27)	0.08 (0.27)	0.08 (0.27)
November	Dummy variable for month of sale (1 if November 0 otherwise)	0.07 (0.27)	0.07 (0.27)	0.08 (0.27)
December	Dummy variable for month of sale (1 if December 0 otherwise)	0.06 (0.25)	0.06 (0.25)	0.06 (0.24)

Note: Number in parenthesis is standard deviation.

Table 2. Distribution of Houses by Submarkets

Number of Submarket	Methods of Delineating Submarkets					
	No market segmentation	K-means clustering	Two-step clustering with Price	Two-step clustering without price	High school districts	Expert defined submarkets
1	18425	1179	4912	5885	1853	1586
2		622	3940	8374	1274	644
3		10416	7097	4166	1395	508
4		5843	2476		3078	537
5		365			1135	219
6					1192	1246
7					804	1138
8					2204	1073
9					1737	653
10					627	168
11					3004	588
12					122	1610
13						3640
14						4815

Table 3. Summary Statistics for Prediction Sample Residuals (N=1842)

	No market segmentation	K-means clustering	Two-step clustering with price	Two-step clustering without price	High school districts	Expert defined submarkets	Simple average	Encompass combining algorithm
Error								
Mean	-$1229.97	-$1372.43	-$1581.85	-$1343.25	-$1479.59	-$1830.99	-$1521.62	-$1506.15
Std. Dev.	$25596.89	$24895.12	$23992.51	$24820.64	$23745.06	$23456.26	$23517.75	$23239.58
APE (%)								
Mean	15.51	15.20	14.81	15.06	14.42	14.38	14.38	14.12
Std. Dev.	24.48	23.76	20.04	23.91	20.06	21.20	22.16	21.41
PPE (%)								
Mean	-5.60	-5.44	-5.30	-5.40	-5.10	-5.30	-5.30	-4.90
Std. Dev.	28.5	27.70	26.00	27.70	25.90	25.10	25.90	25.20

Table 4. APE (%) Distribution Summary Statistics

	No market segmentation	K-means clustering	Two-step clustering with price	Two-step clustering without Price	High school districts	Expert defined submarkets	Simple average	Encompass combining algorithm
1%	0.20	0.10	0.10	0.10	0.10	0.10	0.10	0.10
5%	0.90	1.00	0.70	0.70	0.80	0.80	0.70	0.70
10%	1.90	1.80	1.50	1.70	1.70	1.60	1.50	1.40
25%	4.50	4.70	4.40	4.30	4.10	4.10	3.90	3.80
50%	10.10	9.50	9.10	9.30	8.80	8.60	8.80	8.70
75%	18.50	17.20	17.50	17.80	16.80	17.60	16.80	16.80
90%	29.60	29.40	28.50	28.60	29.30	28.50	28.20	28.30
95%	40.30	41.90	43.50	41.90	41.70	42.10	40.30	41.60
99%	147.00	115.30	123.00	111.10	100.00	103.00	109.30	106.80

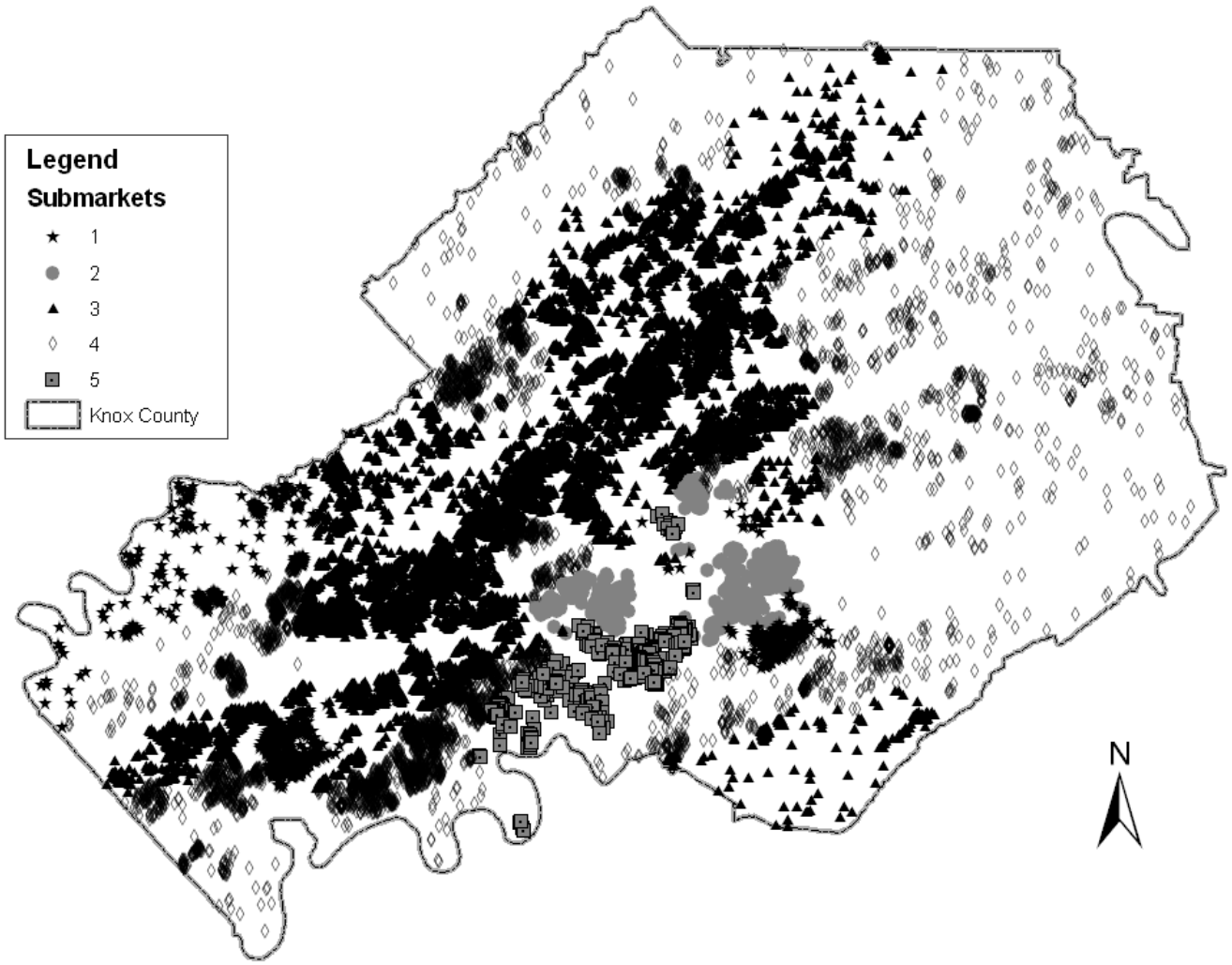

Figure 1: Submarkets derived from k-means clustering on census tract data

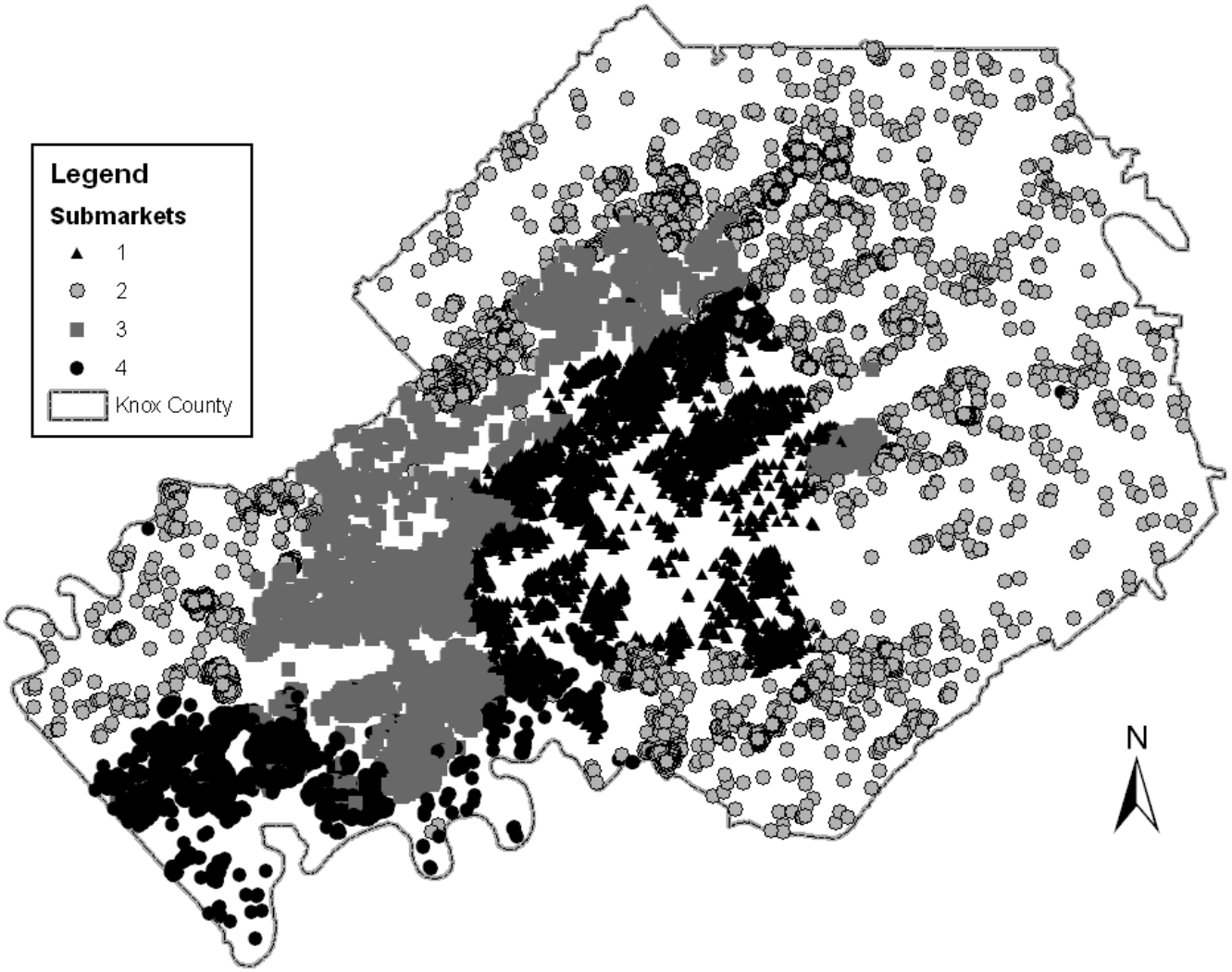

Figure 2: Submarkets derived from individual house data using two-step clustering with price

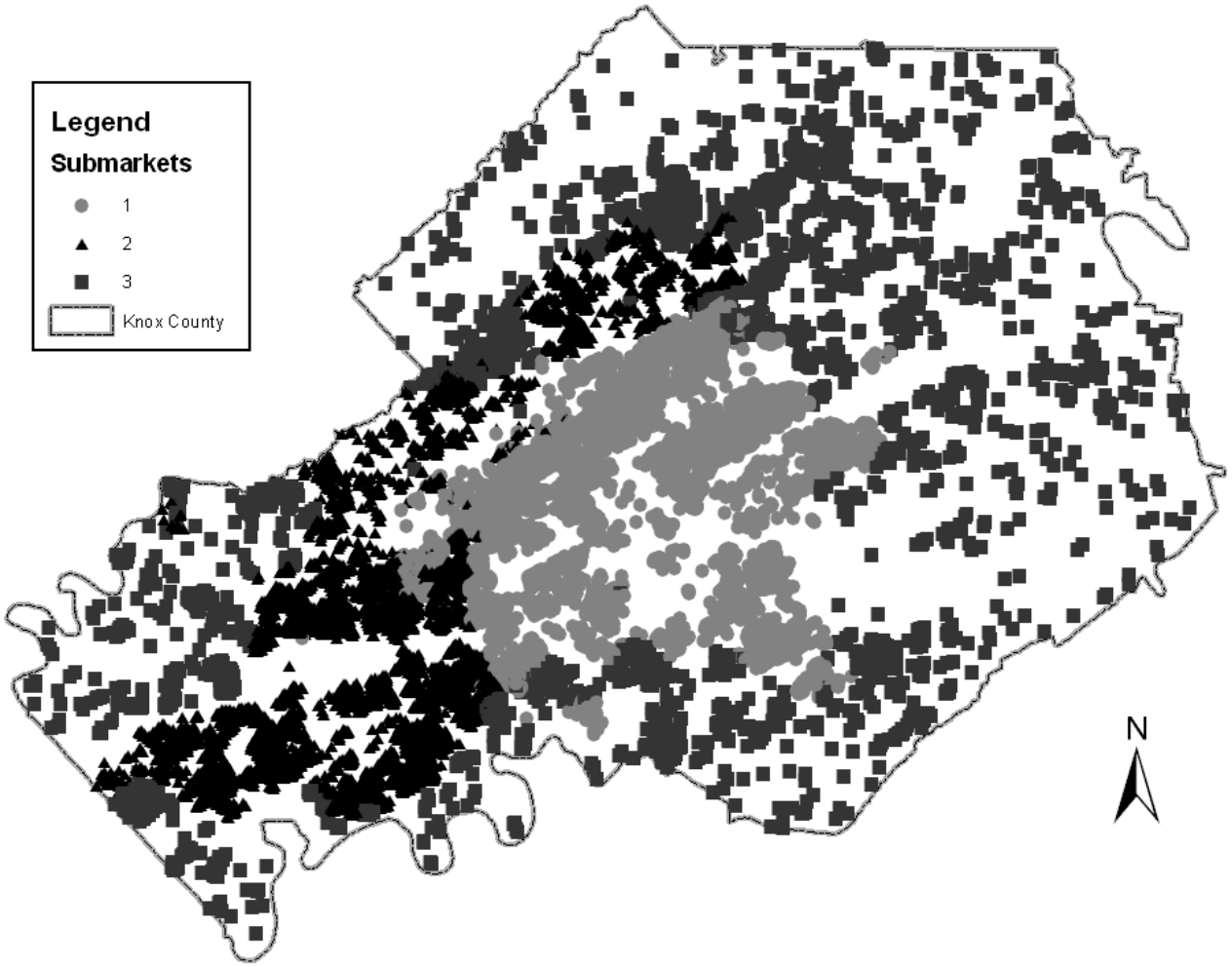

Figure 3: Submarkets derived from individual house data using two-step clustering without price

Figure 4. High school districts

Figure 5: Expert-defined submarkets