Using a Finite Mixture Model of Heterogeneous Households to Delineate Housing Submarkets

Authors

Eric Belasco, Michael C. Farmer, and Clifford A. Lipscomb

Abstract

We use a finite mixture model to identify latent submarkets from household demographics that estimates a separate hedonic regression equation for each submarket. The method is a relatively robust empirical tool to extract submarkets from demographic information with far less effort than suspected. This method draws from latent class models to group observations in a straightforward data-driven manner. Additionally, the unique information about each submarket is easily derived and summarized. Results are also shown to more convincingly sort submarkets than a prior study in the same area that used more comprehensive data.

One concern in the application of hedonic pricing methods is the reliable detection of housing submarkets. There are several approaches to approximate submarket delineation. Some rely on a well-partitioned geographic area as a submarket indicator or on the type of domicile (apartment or house) to sort households into submarkets. Yet if the analyst is interested in the economic demand properties of hedonic attributes, submarkets need to differentiate households according to different preferences for hedonic amenities. Two recent articles in this journal illustrate the issue. First, Shultz and Schmitz (2009) note the considerable variation in the price effect of golf courses. Further, the authors note that residential composition is not as homogeneous as they initially thought. Second, Farmer and Lipscomb (2010) use within-neighborhood household diversity to explore the competition for different bundles of housing attributes between different types of households. They found households with distinctly different tastes offering similar prices for some houses in a single neighborhood. These two works illustrate the theoretical challenge facing real estate researchers interested in demand-side analyses: household heterogeneity requires matching diverse household tastes to particular houses. Operationally, we define submarkets as subsets of market agents whose preference rankings over the stock of real estate are similar, meaning that a group (submarket) orders houses from most preferred to least preferred in a similar way that is distinct enough from another group (submarket) that has a

different ranking from least preferred to most preferred. Empirically, this can be a very time-consuming process; so other strategies have been developed.

In the absence of very specific household level demographic and attitudinal data, researchers such as Osland (2010) use geographically-weighted regression and spatial Durbin models to account for spatial heterogeneity. The current work follows up on prior tests that elicit and analyze a small set of household information directly from a small survey sample (Lipscomb and Farmer, 2005). That work was robust in submarket identification; but unnecessarily data intensive. The current work introduces a more flexible econometric method to detect submarkets that could plausibly be implemented as it greatly reduces the costs to ascribe much more detail in submarket identification at the individual unit level.

The returns to responsible submarket isolation are not trivial. First, explanatory power to predict housing price variation in the dependent variable has been shown to improve dramatically with submarket delineation (Goodman and Thibodeau, 2007). Also, proper aggregability of households into submarkets, which can be characterized by a representative household for each submarket, is required for consistent attribute coefficient estimation (Palmquist, 2004). Put another way, grouping types for hedonic price analysis into those with similar preference rankings is essential, especially if the goal is to conduct welfare analyses of, say, an environmental attribute change using hedonic price estimates (Sieg, Smith, Banzhaf, and Walsh, 2002; Banzhaf and Walsh, 2008).

In a review of the housing submarket literature, we find no clear consensus on how to delineate housing submarkets. Some researchers focus on (1) the characteristics of the housing stock itself to delineate submarkets (Ugarte, Goicoa, and Militino, 2004), (2) the characteristics of the neighborhoods to delineate submarkets (Tiebout, 1956; Sieg Smith, Banzhaf, and Walsh, 2002), (3) spatial stratification of large geographic areas (e.g., Abraham Goetzmann, and Wachter, 1994), (4) price (Goodman and Thibodeau, 2007), and (5) the characteristics of the residents who occupy houses to delineate submarkets (Lipscomb and Farmer, 2005). Almost all analysts acknowledge that the fifth category, which focuses on the characteristics of the residents themselves (the market consumers), is at the deep core of submarket identification (Palmquist, 2004). Yet, given the difficulty in observing or extracting information on individuals by housing unit, analysts opt for observed secondary data that has, a priori, a strong theoretical reason to be correlated with individual characteristics of homeowners. Those proxies are both practical and responsibly justified theoretically. Advances that delineate nuances of attribute value, as those suggested by Shin, Saginor, and Van Zandt (2011) or the quasi-Bayesian home appraisal technique by Zurada, Levitan, and Guan (2011), illustrate the advantages of leveraging high resolution data collected at only a few locations. From these high resolution data, the authors can make robust inferences without comprehensive data coverage. In that same spirit, we suggest that relatively comprehensive information on housing characteristics buttressed by a few observations on a sample of the households themselves may more easily and more efficiently extract hedonic value and delineate equally nuanced submarket divisions.

The burden to extract information from some samples of individual households is heavy and in some cases may not be necessary. For example, in some markets a sharp distinction may divide submarkets by simple neighborhood location if all residents in that area have relatively homogeneous preferences. Similarly, if those in an area owning single-family homes have similar housing preferences as a group and differ from those in apartments who are themselves like other apartment dwellers, submarkets based on the housing characteristic of single-family unit or apartment may delineate well those two submarkets (Watkins, 2001). At other times location may not be such a good submarket predictor. Houses that sell for the same price, even if in different locations, may be the key determinant of submarket distinctions (Goodman and Thibodeau, 2007). Yet in responding to all three, Lipscomb and Farmer (2005) found three very distinct types to be living in the same small neighborhood. These distinctions were not predicted solely based on renter-owner segregation. Similarly, Farmer and Lipscomb (2010) found that members from these three submarkets with very distinct, different housing preferences at times competed for the same units, offering virtually identical prices for those dwellings; yet, those different types bidding for the same dwelling showed far more commonality with residents occupying differently priced houses than those with whom they may find themselves in a bidding war for a given unit in a very similar price range.

So, we may not know what set of distinctions will robustly demarcate submarkets in a given city or region; yet we still may not need the expensive and comprehensive interview and survey of households conducted by Lipscomb and Farmer (2005). Therefore, a technique could be highly valued if it showed the promise to leverage the information from a far more limited set of surveys and yet continued to sort submarkets with the same accuracy of Lipscomb and Farmer (2005), whether those submarkets were delineated by location, home characteristics or price, or by some combination of the three. With the rise of mixed-use designs and the trend toward urbanization in some areas, multiple market segments can inhabit the same areas, making geographic separation itself a poor indicator of submarkets in at least some high growth, high development areas. If different submarkets co-exist side by side in the same area, perhaps paying similar prices for housing units, current methods might aggregate very different types of households into a single homogeneous preference model, such as a single hedonic price equation. Potentially, a very modest extraction of household characteristics at the individual unit level can produce the sharper submarket delineation that is more economically meaningful. Delineated market segments may be highly correlated with physical characteristics or spatial patterns that can form the basis for expanded generalization. Proof of that concept is the purpose of this work.

Identification of such diversity based on direct household data in a single neighborhood has been shown to considerably improve the precision of hedonic estimates in a single neighborhood (Lipscomb and Farmer, 2005). Neighborhood residents rank both individual houses and individual neighborhoods from best to

worst in clearly different ways. With submarkets existing in each neighborhood in different proportions, at times competing for the same houses while ranking houses and neighborhoods from best to worst differently, it is clear that greater precision in household diversity is needed to make nuanced evaluations of the net benefits of amenity changes, such as zoning policies or public green space additions.

In this paper, we propose a method that is relatively robust in delineating housing submarkets based directly on the characteristics of the residents who occupy houses. We extend the work to identify the submarkets of Lipscomb and Farmer (2005) by using a method that requires far less effort to tease out submarket (micro-market) effects and appears to provide a clearer picture of submarket dispersion across space. The result of our work is a housing submarket delineation technique that is calibrated, in this exercise, within a single neighborhood to illustrate the advantages. Yet the method is applicable to much larger hedonic studies, even at a regional level. Critically, we are able to replicate the submarket identification and housing characteristic values found in Lipscomb and Farmer (2005) with far less effort and seemingly greater precision.

In summary, we hypothesize that there exists two endogenously-related phenomena that we look to determine simultaneously. First, there exist a finite number of distinct submarkets when purchasing a home. Second, each submarket values housing attributes differently. To simultaneously characterize these two phenomena, we use latent class analysis in the form of a finite mixture model. A finite mixture model can be thought of as a mechanism used to (1) combine estimating latent class membership though traditional discrete choice modeling and (2) utilize maximum likelihood estimation that is based on latent class membership and independent variables. While the technical details of estimation are discussed later in the text, we assume that a finite number of discrete classes exist where individuals are relatively homogenous within each class.

Data and Model

Data

To estimate the finite mixture model, we use data from an 820-dwelling neighborhood in Atlanta, Georgia, well-situated away from other residential areas by a major interstate highway on the east, a new 137-acre mixed-use development to the north, a commercial corridor to the west, and the Georgia Institute of Technology to the south. The data include survey questions administered to Home Park neighborhood residents near the Georgia Tech campus, publicly available Multiple Listing Service (MLS) data, and geographical information system (GIS) data.

Survey subjects were given an envelope that contained a short 20-question multiple-choice survey, as well as a pencil and a self-addressed, stamped envelope.

Two waves of survey deliveries resulted in 400 observations by August 2002. Then, in September 2002, we collected data on 50 non-respondents to test for non-response bias. Finding none, house value was determined by either a recent arms-length transaction recorded in the local MLS or by a reported appraisal from a refinance as during this period virtually every household had refinanced or pursued the process. MLS listings back 20 years were acquired, reporting home characteristics at the latest sale. (Note: Over 60% of the homes in this neighborhood sold within the last five years prior to survey administration.) Also, 12% of the respondents reported being issued a building permit to complete work on their homes in the three years prior to survey administration. County records were used for the remainder of the data. Renters were asked most of the same questions. We estimated house value based on the reported monthly rent times the local rental multiplier (126), which was obtained by asking local realtors and by observing actual sales of multi-family homes. The exceptional coverage here would not be required in a larger geographic area study as our concern was a sufficient number of data points to conduct our analyses. The effective response rate for the survey was 51%. Variable names, descriptions, and summary statistics are shown in Exhibit 1.

Exhibit 1 shows that the average house in the Home Park neighborhood has 1,332 heated square feet. With so many renters in the neighborhood, it is interesting to see that 77% of respondents live adjacent to renters and 20% of respondents live adjacent to undergraduate students; this suggests that Home Park is very different from the typical single-family neighborhoods analyzed in most hedonic studies. With such a large percentage of renters and students living in the area, our estimation methods must reflect this reality. As we document later, we account for the likelihood that certain respondents have preferences that significantly differ from the preferences of other respondent households in the same neighborhood.

Model

We employ a finite mixture model to sort respondents into endogenously determined latent submarkets. To predict home prices, the finite mixture model is:

$$h(P_i|x_i,\ \beta_j,\ \pi_j) = \sum_{j=1}^{m} \pi(z_i) f(P_i|x_i,\ \beta_j). \qquad (1)$$

A mixing model, $\pi(z_i)$, is used to assign each observation a percentage chance of belonging to each latent submarket, while $f(.)$ is a submarket-specific conditional hedonic regression. In essence, the predicted hedonic value associated with each home is the weighted average of predicted values across each submarket weighted by the probability of residing within each submarket. We also define d_i =

Exhibit 1 | Descriptive Statistics (n = 400)

Variable	Description	Mean	Std. Error	Min.	Max.
Price	Sales price of a house (in dollars)	148,100.00	69,400.00	54,000.00	312,500.00
Sqft	Square footage in house	1,332.77	530.99	496.00	3,816.00
BldgPerm	1 = if you have been issued a building permit in last 3 years; 0 otherwise	0.12	0.33	0.00	1.00
Condo	1 = if home is a condominium; 0 otherwise	0.06	0.24	0.00	1.00
Topo	1 = if dwelling is above street level; 0 otherwise	0.39	0.49	0.00	1.00
NHP	1 = if respondent lives north of 14th Street; 0 otherwise				
RentAdj	1 = if renters live in adjacent property	0.77	0.42	0.00	1.00
UgradAdj	1 = if undergraduates live in adjacent property; 0 otherwise	0.20	0.40	0.00	1.00
HoimpAdj	1 = if neighbors have made home improvements in last 3 years; 0 otherwise	0.33	0.47	0.00	1.00
Rent	1 = if some part of house is rented; 0 otherwise	0.65	0.48	0.00	1.00
Age	Age (in years) of the survey respondent	33.44	14.44	17.00	91.00
Student	1 = if you are a current student; 0 otherwise	0.45	0.50	0.00	1.00
Educ	Respondent's education level (0 = Did not complete high school; 5 = Has a graduate degree)	3.25	1.38	0.00	5.00
Educ_iss	1 = if education is the most important national issue besides national security; 0 otherwise	0.32	0.47	0.00	1.00
Tenure	Number of years lived at current location	5.09	10.22	0.01	70.00

(d_{i1}, d_{i2}, ... , d_{im}) to be binary variables that indicate the inclusions of individual i into each latent group. These variables are treated as missing data and are incorporated into the likelihood function based on a logistic function, which is conditional on factors that do not directly influence the sales price of a dwelling

(such as *Age*, *Student*, *Rent*, *Educ*, *Educ_iss*, or *Tenure* as described in Exhibit 1). In the case where d is fully observable or participants in each submarket are known, this can be written as:

$$d_{ij} = \frac{e^{\gamma_j z_i}}{1 + \sum_{j=2}^{m} e^{\gamma_j z_i}}. \quad (2)$$

In equation (2), $j = 1,2,...m$, where m is the number of identified latent classes (or submarkets) and $e^{\gamma_1 z_i} = 1$. It is worth mentioning that under the scenario where submarkets are directly observable, the estimation of equation (2) would be similar to a conditional logit model using maximum likelihood estimation. The point of departure here is that types are not observable, and therefore the characteristics of each type are also unknown. The optimal number of latent classes is found using the Akaike Information Criteria (AIC) statistic to compare competing models that allow for a different number of latent classes. The AIC statistic was originally derived in Akaike (1974) and is commonly used in maximum likelihood estimation to assess the ability of competing models to fit data while penalizing models for additional parameters. A general AIC statistic can be written as $AIC = 2k - 2LogL$, where k is the number of estimated parameters and $LogL$ is the log-likelihood value using the estimated parameters. The model that minimizes AIC is preferred. In this study, $m = 3$ minimizes AIC.

A study by Lipscomb and Farmer (2005) also identified three distinct submarkets (latent classes) using the same data and a different iterative sorting process. The main difference between the current work and that of Lipscomb and Farmer (2005) is in the use of demographic information to identify latent submarkets. Lipscomb and Farmer (2005) use a two-step process where the first step identifies potential submarkets using seemingly unrelated regression (SUR) analysis and the second step uses successive iterations of the SUR model coupled with principal components analysis. The appeal of this method is the endogenous determination of types by the iterative search process. However, the major limitations of this procedure are that each household is assigned to a single submarket and the technical separation of the type identification and the subsequent regression estimation procedure.

In the current work, the finite mixture model unites both steps into a single optimization process and allows for only demographic factors to be used when estimating submarkets. This is consistent with product differentiation theory (Rosen, 1974). This mixture model approach allows an individual household to be assigned to each class under a probability distribution, retaining the information regarding the relative uncertainty that a given respondent attaches to a particular submarket. Households that clearly belong to a given submarket carry a higher (probability) weight than households that exhibit characteristics similar to several

submarkets. This feature of the finite mixture model is important as it adds efficiency to the exercise. In future work, it will be important as researchers consider its use in larger real estate markets and submarket studies that use a broader unit of analysis, such as the Census block group or tract. The finite mixture process is explained below.

Because submarket identification (the variable d) is unknown, an expectation maximization (EM) algorithm is used to estimate the likelihood of class identification simultaneously with the estimation of hedonic regression parameters, which are conditional on class identification. The EM algorithm was originally developed in Dempster, Laird, and Rubin (1977) and is now commonly used to estimate maximum likelihood parameters when the likelihood function is based on latent variables or missing data points. The estimated log-likelihood function can be written as:

$$LogL = \sum_{i=1}^{n} \sum_{j=1}^{m} \hat{d}_{ij} \log[f_j(P_i|x_i, \beta_j)] + \hat{d}_{ij} \log[\hat{\pi}_j], \tag{3}$$

where:

$$\hat{d}_{ij} = \frac{\hat{\pi}_j f_j(P_i|x_i, \hat{\beta}_j)}{\sum_{j=1}^{m} \hat{\pi}_j f_j(P_i|x_i, \hat{\beta}_j)}, \tag{4}$$

such that $\hat{\pi}_j = \Sigma_{i=1}^{n} \hat{d}_{ij}$. Given that $\hat{d}_{ij}$ is the estimated probability that individual i is identified with a latent class (or submarket) j, $\hat{d}_{ij}$ becomes the dependent variable in equation (2) in order to evaluate the impact of demographic sorting variables on belonging to a particular submarket. Given the first part of the specification in the likelihood function, shown in equation (3), it is helpful to illustrate the difference between this method and that of a simple hedonic regression model. The modified hedonic regression can be shown as:

$$y_{ij} = d_{ij}(\beta_j X_i) + \varepsilon_{ij}, \tag{5}$$

where y is the log of *Price* and X includes typical hedonic covariates, including *log*(*Sqft*), *BldgPerm*, *Condo*, *Topo*, *NHP*, *RentAdj*, *UgradAdj*, and *HoimpAdj* for $i = 1$ to 400 and $j = 1$ to 3. Notice that there are two unknown components that are simultaneously determined: (1) the probability that each individual belongs to

each submarket ($\hat{d}_{ij}$); and (2) the parameter estimates that are unique to each submarket (β_j). Since each of these components influences the other, the EM algorithm is used. Preferences are assumed to be heterogeneous across the population, but homogeneous within each latent submarket to be aggregated into a single hedonic regression for each submarket. Estimated regression parameters use a unique percentage weight for each observation (e.g., $d_{ij} = .20$ for Type A; $d_{ij} = .55$ for Type B; and $d_{ij} = .25$ for Type C); this shows how each estimate is conditional on its latent submarket. The method explicitly allows for different submarkets to possess different parameter estimates so that each type maintains different values associated with traditional hedonic pricing variables. For example, the three types identified in this study are shown to consist of established households, students, and young professionals. Given the different characteristics of these types of households, it is reasonable to expect that these types are likely to be willing to pay different premiums for square footage or extra bedrooms (Farmer and Lipscomb, 2010).

Since types cannot be predicted with certainty, the prediction for observation i using the finite mixture model ($\hat{y}_i$) depends on the predicted probability of being type j, $\hat{d}_j$, and the associated $\hat{\beta}_j$ for each type. To illustrate, from equation (5), we compute the predicted dependent variable values based on:

$$\hat{y}_{ij} = \sum_{j=1}^{3} \hat{d}_{ij}(\hat{\beta}_j X_i). \tag{6}$$

The ability of this model to fit the in-sample data can then be assessed by using the standard R^2 metric, where $R^2 = 1 - \Sigma_{i=1}^{n} (\hat{y}_i - y_i)^2 / \Sigma_{i=1}^{n} (y_i - \overline{y}_i)^2$.

Results

Submarket Identification

The results corresponding to latent class identification are reported in Exhibit 2. Parameter estimates in Exhibit 2 are interpreted similarly to parameters estimated in a multinomial logit model where independent variables do not vary across alternative but only across individuals. As shown in Cameron and Trivedi (2005, p. 502), marginal effects for this type of model can be derived using:

$$\frac{d\hat{d}_{ij}}{dx_i} = \hat{d}_{ij}(\beta_j - \overline{\beta}_i), \tag{7}$$

Exhibit 2 | Estimation Results and Marginal Impacts from Latent Classification

	Submarket 2			Submarket 3		
Variable	Estimate	Std. Error	Med. Marginal Response	Estimate	Std. Error	Med. Marginal Response
Int	−1.044	0.594		0.550	0.602	
RENT	−0.224	0.287	−0.007	−0.120	0.291	<0.001
Age	0.028*	0.012	0.001	0.006	0.013	<0.001
Student	−0.034	0.294	0.004	−0.975*	0.302	−0.054
Educ	0.105	0.081	0.006	−0.173*	0.083	−0.010
Educ_iss	0.496*	0.231	−0.001	0.867*	0.242	0.040
Tenure	0.016	0.016	<0.001	0.039*	0.016	0.002

where β_j is the estimated set of parameters for type j and $\overline{\beta}_i$ is the weighted average of the estimated parameter estimates, weighted by $\hat{d}_i$. As is typically used, we report median marginal responses across all individuals in Exhibit 2. Marginal responses are interpreted as the relative change in the probability of being in Submarket 2 or 3 relative to being in Submarket 1. For example, an increase in age by one year has a slight positive influence on the probability that an individual is identified in Submarket 2 (0.001) or Submarket 3 (<0.001); or Submarket 1 households tend to be younger. As a further example, if an individual denotes that some part of the home is rented, then they are 0.7 percentage points less likely to reside in Submarket 2, relative to Submarket 1. It is also clear from the results that a higher level of education attainment is positively associated with membership in Submarket 2 and negatively associated with Submarket 3, relative to Submarket 1. Finally, using statistically significant differences, Submarket 3 households are less likely to be students; they are less educated; they more often consider education issues to be the more important social issue; and they have lived in the area much longer.

While the signs of almost all parameter estimates in Exhibit 2 are consistent for Submarkets 2 and 3, their relative magnitudes vary. Based on these results, it is easy to see that many of these variables would be expected to have statistically significant impacts on submarket identification; that is, we have identified households with different willingness to pay values for amenities.

To better understand these parameter estimates in considering how submarkets differ in their preferences, it is helpful to look at the characteristics of each submarket. For illustration purposes, Exhibit 3 presents the final mean characteristics of each submarket after they have been sorted statistically. The table provides a clearer picture of the statistical sorting.

Exhibit 3 | Comparison of Demographic Summary Statistics

	Submarket 1		Submarket 2		Submarket 3	
Variable	FM	LF	FM	LF	FM	LF
n	144	70	169	184	87	146
Rent	0.90	0.76	0.54	0.83	0.43	0.36
Age	26.13	30.79	36.44	30.81	39.72	38.03
Student	0.74	0.40	0.31	0.58	0.25	0.32
Educ	3.12	3.27	3.48	3.14	3.01	3.37
Educ_is	0.39	0.33	0.29	0.33	0.26	0.30
Tenure	2.65	2.73	5.16	3.68	9.01	8.01

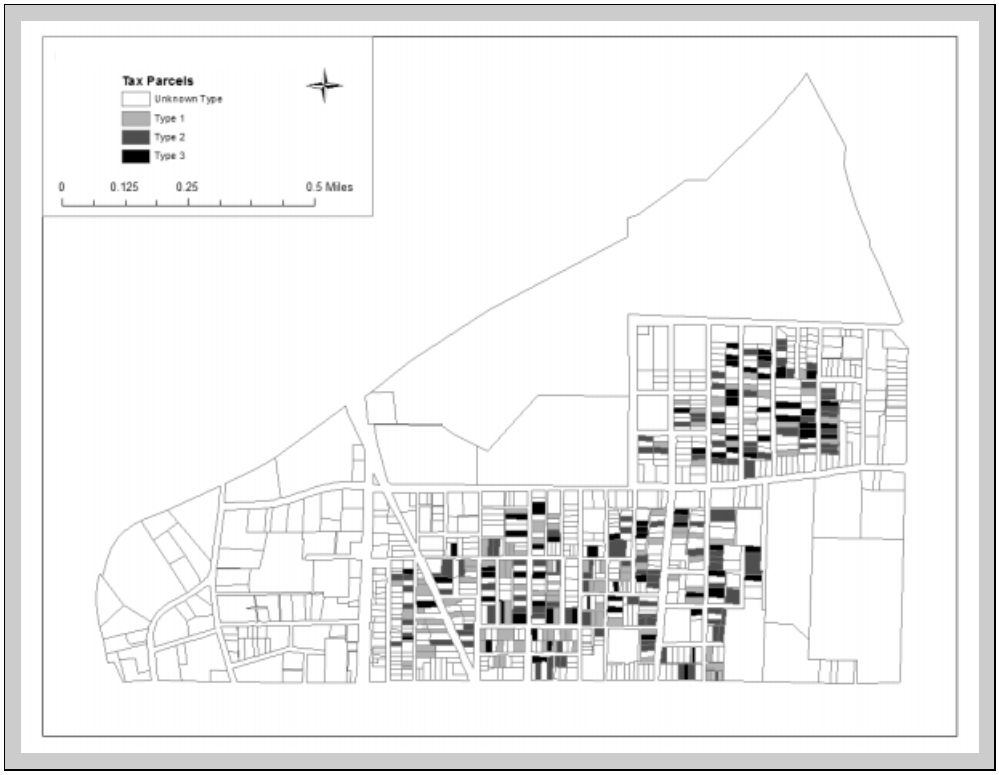
Note: FM results are from the Finite Mixture model presented in this paper and LF results are from Lipscomb and Farmer (2005).

Exhibit 3 is for descriptive purposes: an individual household is identified with a particular submarket if $\hat{d}_{ij} > \hat{d}_{ih}$ for all $j \neq h$ where $\hat{d}_{ij} = e^{\hat{\gamma}_j z_i}/1 + \Sigma_{j=2}^{m} e^{\hat{\gamma}_j z_i}$. Based on these results, Submarket 1 appears to be most closely characterized as the ''student'' submarket. These households are substantially younger than the other groups, include a very high percentage of renters, and have a shorter tenure at their dwelling.

Submarket 2 can be characterized as ''young professionals.'' While more established and older than Submarket 1 households, they are still younger than Submarket 3 households and have lived in their dwelling for less time than Submarket 3 households. Submarket 3 comprises more ''established households'' who, on average, have a longer tenure in the neighborhood. These households are older, are most likely to own their homes, are somewhat less educated, and are the least likely to cite education as the most important social issue. These results do not claim that the finite mixture model results are superior to that of Lipscomb and Farmer (2005), but do result in more distinct groups under the finite mixture model, which comes along with the added parsimony, lower computational burden, and theoretical consistency of the model.

It is worth mentioning at this point that submarket identification of an individual household is not an absolute designation as it was in Lipscomb and Farmer (2005). Results in Exhibit 2 and the final regression results below identify each household as having some probability of being in each of the three groupings. Exhibit 4 shows the spatial distribution of households among the three groupings when each household is classified either as Type 1, Type 2, or Type 3. For example, a household is classified as Type 1 when the probability of being in Type 1 is higher than that household's probability of being classified as Type 2 or Type 3.

Exhibit 4 | Highest Probability Types



The dispersal of household types across the Home Park neighborhood in Exhibit 4 mimics prior results from Lipscomb and Farmer (2005). Noting the limits of analyzing a single neighborhood, this result provides a strong *prima facia* case that counters the dominance of location per se as the primary driver of household location; there is no obvious or persistent clustering by type across the neighborhood. The ability to distinguish among submarkets at a higher resolution, using a tool more conducive to multi-neighborhood submarket delineation, suggests a more flexible submarket identification process that closely follows core demand theory. The extension of this technique to larger geographic scales is the obvious next step for future research on this topic.

Hedonic Estimation by Submarket

Next, we discuss the results from the hedonic regression results for each identified submarket. Each household is partially identified into a group according to the probabilities $\hat{d}_{ij}$. Exhibit 5 reports submarket-specific parameter estimates and compares those estimates to the aggregated OLS model, which does not make submarket distinctions.

Exhibit 5 | Estimation Results by Latent Class

Variable	Submarket 1	Submarket 2	Submarket 3	OLS (All)
Constant	11.651*	9.241*	8.277*	9.799*
	(0.030)	(0.378)	(0.707)	(0.399)
log(Sqft)	0.013*	0.295*	0.550*	0.282*
	(0.004)	(0.053)	(0.098)	(0.055)
BldgPerm	0.007	0.100*	0.337*	0.103
	(0.004)	(0.050)	(0.118)	(0.060)
Condo	−0.057*	−0.272*	0.120	−0.051
	(0.007)	(0.078)	(0.138)	(0.087)
Topo2	0.001	0.028*	0.039*	0.018*
	(0.001)	(0.007)	(0.013)	(0.008)
NHP	0.004	0.054	0.119	0.112*
	(0.003)	(0.039)	(0.070)	(0.041)
RentAdj	−0.016*	−0.154*	−0.329*	−0.248*
	(0.005)	(0.040)	(0.078)	(0.049)
UgradAdj	0.018*	0.149*	−0.324*	−0.037
	(0.003)	(0.058)	(0.093)	(0.051)
HoimpAdj	0.007	1.064*	−0.280*	0.345*
	(0.004)	(0.041)	(0.077)	(0.046)
Sigma	0.016	0.2045	0.3624	0.3809
R^2	0.837	0.301		

Note: The dependent variable is log(*Price*).

As shown in Exhibit 5, square footage has a different marginal impact on price based on the submarket. In this context, an additional 10% added to square footage adds nearly 5.5% to the house value for an established household but only 0.1% in housing value for the student household submarket. For a single neighborhood, the level of difference in value for square footage is arguably high, suggesting multiple markets.

All submarkets also appear to have a negative response to living adjacent to renters (*RentAdj*); yet this response is much stronger for established households than it is for students. In similar fashion, this is also seen as students and young professionals have a significantly positive response to living adjacent to undergraduates. Established households on the other hand have a significantly negative response to living next to undergraduates. This provides much cleaner differentiation than the aggregate OLS model that predicts living next to an undergraduate does not have a statistically significant impact on sales price. Once we control for submarkets, however, the result is much clearer that different types have different preferences.

A characteristic of this neighborhood is that it exhibits a high degree of local growth, as well as within-neighborhood remodeling and reconstruction (demolition, then rebuilding). How and why different types of improvement or redevelopment occur along with the effects on, or caused by, local submarkets is an important part of the information set for planners, zoning officials, and utility providers; and this process is much better explained by the market sorting above.

Home renovations/additions to one's dwelling (*BldgPerm*) appear to have substantially positive value gains for the established households, but much smaller impacts for the two other submarkets. It is worth noting that Submarket 2 still has a statistically significant reaction to recent renovations while Submarket 1 shows no significant price response to renovation to their unit (through *BldgPerm*). The distinction appears to be the difference between renovation and redevelopment.

Renovations in the homes adjacent to a particular dwelling (*HoimpAdj*) interestingly have a negative impact for established households. Many of those improvements are complete redevelopments of highly depreciated, single-family residences into multi-unit or multi-bedroom student houses. Those units are new and yield a high price, but are not occurring everywhere. Being adjacent to a home improvement, mostly redeveloped parcels for high-valued multi-bedroom student houses, has a positive impact on home prices of young professionals nearby and an insignificant impact for nearby students themselves. Those redeveloped parcels have an especially large marginal impact on the lower-valued single-family homes and condominiums owned by young professionals in the immediate neighborhood. These distinctions make sense. Established households take advantage of higher quality improvable property enclaves by renovating an existing structure. New development for higher-end student housing in the vicinity of owned units by young professionals takes the form of total redevelopment of well placed but highly depreciated single-family units into larger multi-bedroom houses.

As is clear with many of the parameter estimates, the marginal values associated with amenities tend to vary across submarkets, demonstrating the effect of each submarket on the overall pricing of homes throughout the neighborhood. This entire nuance on home improvement versus redevelopment is confounded in the aggregated OLS regression, which suggests both changes may be good for all units. Yet the weight, on average, in the OLS emphasizes redevelopment per se (0.345 for *HoimpAdj*, which is significant versus 0.103 for *BldgPerm*, which is not significant). This loss of distinction by over-aggregating could unwittingly drive out one important submarket—the most established households in the neighborhood—if an aggressive policy intervention favored redevelopment over renovation.

Making these differentiations in preferences among submarkets highlights the reason why the coefficient of determination (R^2) also improves dramatically from 0.301 in the OLS estimation to 0.837 with the finite mixture model. The relatively

low sigma values, which are computed in the same step as the parameter estimates using normal maximum likelihood estimation, reveal that there is relatively efficient estimation across the submarkets. It is also worth mentioning that the variance in all regressions, based on submarket delineation, is far lower than the variance in the aggregated OLS regression. This also suggests that the submarket delineation minimizes the deviations in outliers; or seemingly unusual properties find a more coherent explanation in the context of several housing submarkets than could be identified in an aggregated OLS model.

Comparison to Prior Approaches

A final result is that the delineation of a proto-typical student, or young professional, or established household, as economic actors in this housing market appear to be much sharper than prior submarket delineation. Exhibit 5 shows a more plausible division among the three submarkets in this neighborhood compared to prior work.

Starting at the lower end, the length of time in the neighborhood for those expressing economic preferences of ''students'' is 2.65 years, about the median time in school for a college student at an engineering school. The stay for young professionals is longer at 5 years, but established households show a mean tenure in the neighborhood of 9 years. On educational attainment, students show some post-secondary education while young professionals show more education (generally having completed their educations or completing higher degrees). Established households show the lowest education, with some adult family members having completed post-secondary degrees.

Finally in Submarket 1, 75% report being students, 90% are renters, their average age is 26, and they believe education is the highest social concern (40%, compared to 26% and 29% for the representative household in the other two submarkets). Young professionals are older on average (36), about half of them rent (50%), and they also believe education is the highest social concern. Established households are older still (almost 40 years old on average), about 31% rent their dwellings, and they express the least concern for education as the most important social issue.

Casual inspection shows that these are much crisper distinctions than the prior sorting by Lipscomb and Farmer (2005). In that work, Submarket 1 (students) was smaller, in part because students occupied the most and the least expensive properties in the neighborhood. By sorting each observation exclusively into one submarket, that work obscured the blended nature of many young professionals, who may still be in school pursuing professional degrees or entering the homeownership phase. In that work, the age for ''student renters'' was close to 30 years old, almost identical to young professionals; in the current model, students have an average age of 26.13 while young professionals are predicted to be 36.44 years old, on average. Submarket 2 households compete for some units

with Submarket 1 and Submarket 3, which demands more precision if the distinction rewarding mixed development in this neighborhood, renovation, and redevelopment is to be made. The finite mixture model, with a more fluid assignment of each individual as a distribution over their degree of likeness to each latent submarket class, arguably provides a much better market representation. The treatment of Submarket 2 in particular resolves certain anomalies found initially in characterizing the young professionals.

Conclusion

Our research focuses on a single neighborhood but has widespread implications for larger hedonic studies. These results suggest the method posed may be more successful in helping the analyst ''scale up'' her housing market analysis to larger geographic areas without surrendering efficiency and consistency in valuing particular housing amenities by each submarket or latent class. We also learned that modest demographic and attitudinal information on households must be collected to sort households using the finite mixture model. While a survey is still needed, the number of questions that require a valid response to be able to estimate these kinds of models is relatively small (less than ten). The overall gain in R^2 improves from 0.301 to 0.837 by the sorting of the sample into these submarkets, suggesting a substantial improvement in model fit.

Using a finite mixture model, we find diverse agents who exhibit very different patterns in their revealed preferences for location, dwelling characteristics, and adjacency factors. While these agents exhibit heterogeneous decisions across submarkets, they appear to be relatively homogenous within each latent submarket. The patterns found are distinct enough to violate a smooth continuity in the distribution of heterogeneity that is required by randomized parameters, at least for our sample size ($n = 400$). After sorting respondents with a finite mixture model, we find that truly different welfare implications can be drawn for each submarket, implications that would be overlooked in the absence of submarket delineation.

Certainly the relative computational ease of the finite mixture model compared to the coupling of principal components analysis and the SUR estimator is very appealing. The computational ease comes from the fact that the finite mixture model can be estimated more parsimoniously through the use of an EM algorithm, while the PCA with SUR estimation is completed in two separate steps where class identification is made with complete certainty. In fact, considering the computing advances that permit bootstrapping parameter estimates and standard errors (e.g., Neill, Hassenzahl, and Assane, 2007) or more robust results that conduct Monte Carlo estimation with an EM estimator or Gibbs sampler to approximate the distribution of submarkets, where submarkets may correlate in very highly particular ways with spatial and housing characteristics (LeSage, 2009; LeSage, et al., 2010; Paulson, Hart, and Hayes, 2010), the demands on the analyst may not be that burdensome.

Still, the demands are greater than the current state-of-the-art in hedonic modeling, but it may be worth the effort for many purposes. The reasons for the added effort depend on the purpose of the hedonic study. As planners try to assess social gains or losses, determining the likelihood that a given submarket is likely to enter or exit a given geographic area due to a proposed policy or land use change is very important for valid welfare evaluation (Palmquist, 2004). For policy analysts, determining the correct number of statistically distinct submarkets within the framework suggested here helps them to avoid policy prescriptions that may provide perverse incentives for one type of household at the expense of other household types. With the analytic extensions suggested here in this proof of concept, the gains may well warrant the potentially modest extra effort.

Finally, we note that with reasonable extensions in Bayesian statistics, such as the advance of the EM estimator, the seemingly strong correlations between household type and structural features or other amenities can be captured over a much larger space from a relatively small household survey. Critically, using the same finite mixture model technique directly on the housing unit characteristics does not differentiate submarkets according to preferences; so the initial step to isolate different household preferences appears in this exercise to be needed to deliver a reliable submarket identification. An advantage of the mixture of demographic and attitudinal variables in this approach is that an abbreviated survey reduced to four or five basic demographic variables and one attitudinal variable may be all that is required. The corresponding increase in response rates would further bolster these results.

References

Akaike, H. A New Look at the Statistical Model Identification. *IEEE Transactions on Automatic Control*, 1974, 19:6, 716–23.

Abraham, J.M., W.N. Goetzmann, and S.M. Wachter. Homogeneous Groupings in Metropolitan Housing Markets. *Journal of Housing Economics*, 1994, 3:3, 186–206.

Banzhaf, H.S. and R.P. Walsh. Do People Vote with Their Feet? An Empirical Test of Tiebout. *American Economic Review*, 2008, 98:3, 843–63.

Cameron, C. and P.K. Trivedi. MICROECONOMETRICS: Methods and Applications. New York: Cambridge University Press, 2005.

Dempster, A.P., N.M. Laird, and D.B. Rubin. Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 1977, 39:1, 1–38.

Farmer, M.C. and C.A. Lipscomb. Using Quantile Regression to Reveal Hedonic Submarket Competition, *Journal of Real Estate Research*, 2010, 32:4, 435–60.

Goodman, A.C. and T.G. Thibodeau. The Spatial Proximity of Metropolitan Area Housing Submarkets. *Real Estate Economics*, 2007, 35:2, 209–32.

LeSage, J.P. Spatial Probit Models, Estimation and Interpretation. Keynote Address. Third World Conference of the Spatial Econometrics Association, Barcelona, Spain, 2009.

LeSage, J.P., R.K. Pace, N. Lam, R. Campanella, and X. Liu. New Orleans Business Recovery in the Aftermath of Hurricane Katrina. Paper presented at Southern Regional Science Association (SRSA). Alexandria, VA. March 18, 2010.

Lipscomb, C.A. and M.C. Farmer. Household Diversity and Market Segmentation Within a Single Neighborhood. *Annals of Regional Science*, 2005, 39, 791–810.

Neill, H.R., D.M. Hassenzahl, and D.D. Assane. Estimating the Effect of Air Quality: Spatial Versus Traditional Hedonic Price Models. *Southern Economic Journal*, 2007, 73: 4, 1088–1111.

Osland, L. An Application of Spatial Econometrics in Relation to Hedonic House Price Modeling. *Journal of Real Estate Research*, 2010, 32:3, 289–320.

Palmquist, R.B. Property Value Models. In K.G. Mäler and J. Vincent (eds.), *Handbook of Environmental Economics*. Volume 2, North-Holland, 2004.

Paulson, N.D., C.E. Hart, and D.J. Hayes. A Spatial Bayesian Approach to Weather Derivatives. *Agricultural Finance Review*, 2010, 70:1, 79–96.

Rosen, S. Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. *Journal of Political Economy*, 1974, 82, 34–55.

Shin, W-J., J. Saginor, and S. Van Zandt. Evaluation of Subdivision Characteristics on Single Family Housing Value Using Hierarchical Linear Modeling. *Journal of Real Estate Research*, 2012, 33:3, 317–48.

Shultz, S.D. and N.J. Schmitz. Augmenting Housing Sales Data to Improve Hedonic Estimates of Golf Courses. *Journal of Real Estate Research*, 2009, 31:1, 63–79.

Sieg, H., V.K. Smith, H.S. Banzhaf, and R. Walsh. Interjurisdictional Housing Prices in Locational Equilibrium. *Journal of Urban Economics*, 2002, 52: 131–53.

Tiebout, C. A Pure Theory of Local Expenditures. *Journal of Political Economy*, 1956, 64: 5, 416–24.

Ugarte, M.D., T. Goicoa, and A.F. Militino. Searching for Housing Submarkets Using Mixtures of Linear Models. In J.P. LeSage and R.K. Pace (eds.), *Spatial and Spatiotemporal Econometrics*. Oxford: Elsevier Ltd., 2004.

Watkins, C.A. The Definition and Identification of Housing Submarkets. *Environment and Planning A*, 2001, 33:12, 2235–53.

Zurada, J., A.S. Levitan, and J. Guan. A Comparison of Regression and Artificial Intelligence Methods in a Mass Appraisal Context. *Journal of Real Estate Research*, 2011, 33:3, 349–88.

Eric Belasco, Montana State University, Bozeman, MT 59717-2920 or eric.belasco@montana.edu.

Michael C. Farmer, Texas Tech University, Lubbock, TX 79409 or michael.farmer@ttu.edu.

Clifford A. Lipscomb, Greenfield Advisors LLC, Atlanta, GA 30339 or cliff@greenfieldadvisors.com.