The Spatial Proximity of Metropolitan Area Housing Submarkets

by

Allen C. Goodman*

and

Thomas G. Thibodeau**
(contact author)

August 2006

Forthcoming in ***Real Estate Economics***

* Professor of Economics
Department of Economics
Wayne State University
Detroit, Michigan 48202-3424
USA
Phone: 313.577.3235
FAX: 313.577.0149
Email: allen.goodman@wayne.edu

** Professor of Real Estate
Leeds School of Business
University of Colorado-Boulder
UCB 419
Boulder, CO 80309-0419
Phone: 303.735.4021
Email: tom.thibodeau@colorado.edu

The Spatial Proximity of Metropolitan Area Housing Submarkets

Abstract

An important question related to housing submarket construction is whether geographic areas must be spatially adjacent in order to be considered the same submarket. Housing consumers do not necessarily limit their search to spatially concentrated areas and may search similarly priced neighborhoods located throughout a metropolitan area when making housing consumption decisions. This paper examines two alternative procedures for delineating submarkets: one that combines adjacent census block groups into areas with enough transactions to estimate the parameters of a hedonic house price equation; and a second that permits spatial discontinuities in submarkets. The criterion used to evaluate the alternative techniques is the accuracy of hedonic house price predictions.

The empirical research is conducted using data obtained from the Dallas Central Appraisal District (DCAD). The DCAD provided information for every parcel of real property in Dallas County. As of January 1, 2003, there were approximately 500,000 single-family homes in the DCAD area and 44,000 transactions in the 2000:4-2002:4 period. We find that both submarket constructs significantly increase hedonic prediction accuracy over a standard pooled model, but that neither construct statistically dominates the other.

These results have important implications for empirically modeling submarkets within metropolitan area housing markets. Creating housing submarkets by combining spatially adjacent census block groups that lie within the same municipality and same independent school district is time consuming and costly. These results suggest that comparable increases in hedonic prediction accuracy can be achieved by delineating submarkets by dwelling size and median census block group per square foot transaction price.

Introduction

Understanding how metropolitan areas are partitioned into housing submarkets is important for several reasons. First, assigning properties to housing submarkets will likely increase the prediction accuracy of the statistical models that are used to estimate house prices. Second, identifying housing submarket boundaries within metropolitan areas will enable researchers to better model spatial and temporal variation in those prices. Third, an accurate assignment of properties to submarkets will improve lenders' and investors' abilities to price the risk associated with financing homeownership. Finally, providing submarket boundary information to housing consumers will reduce their search costs.

Analysts have examined numerous techniques for constructing housing submarkets. Some have used principal component analysis and statistical clustering techniques to group small geographic areas (e.g. census block groups, census tracts, zip code districts, or local government areas) into housing submarkets while others have developed procedures that explicitly model submarket boundaries. Goodman and Thibodeau (1998, 2003), for example, identify housing submarket boundaries using hierarchical models. Their implementation of the Bryk and Raudenbush (1992) technique assigns elementary school zones to housing submarkets depending on whether neighborhood public school quality is capitalized into neighborhood house prices.

Some submarket construction techniques focus on the supply side determinants of house prices and construct submarkets using characteristics of the housing stock (e.g. dwelling type, square feet of living area, dwelling age) and/or characteristics of the neighborhood (e.g. the quality of neighborhood schools, the quality of local police). Other submarket construction techniques focus on demand side determinants of house prices and form housing submarkets based on household incomes or other socioeconomic/demographic characteristics.

An important question related to housing submarket construction is whether geographic areas need be spatially adjacent to be considered the same submarket. Housing consumers do not necessarily limit their search to spatially concentrated areas when searching for housing. Fundamentally, most housing consumers are constrained by their incomes and may search

similarly priced neighborhoods located throughout a metropolitan area when making housing consumption decisions.

This paper empirically examines two alternative procedures for assigning single-family properties to submarkets. One combines spatially adjacent census block groups (located within the same municipality and same independent school district) into 372 areas with enough transactions to estimate the parameters of a hedonic house price equation. A second procedure permits spatial discontinuities by assigning properties to 325 submarkets based on dwelling size and on the average per square foot transaction price for the neighborhood. The empirical analysis is conducted using about 44,000 single-family transactions in the Dallas housing market over the 2000:4-2002:4 period. The criterion used to evaluate the alternative submarket constructions is the accuracy of hedonic house price predictions, over a 10% hold-out prediction sample. The alternative measures of hedonic house price prediction accuracy reported here are: (1) the average prediction error; (2) the mean absolute error; (3) the mean proportional error; (4) the mean squared error; and (5) the percent of the time that a predicted price is within ten (or fifteen or twenty) percent of an observed transaction price.

Literature Review

A housing (sub-) market is a geographic area where the price of housing (per unit of housing services) is constant. Identifying geographic areas with constant per unit housing prices is challenging because housing is a heterogeneous good, and the market value of a house (as estimated by its transaction price) is a function of the property's site, structural, neighborhood and location characteristics. Hedonic and other semi-parametric and non-parametric house price modeling techniques have been used to examine the influence that site and structural characteristics have on house price. Incorporating the influence that neighborhood and location characteristics have on house prices is more challenging.

Analysts have employed a variety of statistical techniques to measure and control for the influence that location has on house price. Kain and Quigley (1970) reduced services provided by 39 individual location characteristics to five factors using factor analysis. The indices include

the quality of adjacent parcels, the percent of the neighborhood dedicated to commercial uses, the amount of local commercial traffic, and numerous other potential externalities. Li and Brown (1980) separated the positive influence that accessibility has on residential real estate values from the negative effect that proximity to non-residential use has on residential property values. Proximity variables from Li and Brown include a corner grocery store, a neighborhood park, a school, a river, an ocean, conservation land, expressway interchange, or major thruway. Dubin and Sung (1990) group neighborhood characteristics into three broad categories: socioeconomic status of neighborhood residents (e.g. household income, education, occupation); quality of municipal services (e.g. education, public safety); and racial composition. They conclude that socioeconomic status and racial composition are more important than the quality of public services in determining house prices.

One way to control for the influences of neighborhood and location attributes on house prices is to group geographic areas with similar neighborhood and location characteristics into a "submarket". House price model parameters can then be estimated for all properties within these submarkets without having to measure explicitly the influence that the location attributes have on house prices. Eliminating (or significantly reducing) the influence that neighborhood and location attributes have on house prices enables analysts to focus on the site and structural determinants of house prices. In addition, analysts can examine the influence that location and neighborhood attributes have on house prices by modeling across submarket variation in house prices. The empirical challenge is to develop procedures that identify geographic areas sharing homogeneous location and neighborhood attributes.

Some of the neighborhood and location attributes that influence house prices may be nested. The quality of a neighborhood school, for example, is dependent upon, or nested within, the quality of the regional school district. Consequently, the value of a single-family detached house may depend on factors that are nested within a neighborhood, within a school district, and within a metropolitan area. Other attributes, such as ethnic areas, religious parishes, or housing types, may cross school or municipal boundaries, and will not necessarily be nested hierarchically or at all.

Goodman (1978) provides empirical support for geographically segmented housing markets. He

compared the hedonic coefficients for structural and neighborhood characteristics for five areas in metropolitan New Haven over a three-year period. He reported that hedonic coefficients for neighborhood characteristics were not constant over space and concluded that metropolitan markets were geographically segmented. Goodman and Dubin (1990) suggest both nested and non-nested tests for the optimal number and configuration of submarkets.

Dale-Johnson (1982) assigned properties to submarkets using factor analysis to reduce the influence that 13 neighborhood and location variables have on house prices. Maclennan and Tu (1996) used principal components to identify the most important neighborhood and location attributes and then defined submarkets using cluster analysis on the resulting factors.

Goodman and Thibodeau (1998) define economically meaningful submarket boundaries as geographic areas where: (1) the price of housing (per unit of service) is constant; and (2) individual housing characteristics are available for purchase. They examined housing market segmentation within metropolitan Dallas using hierarchical models (Bryk and Raudenbush, 1992) and single-family property transactions over the 1995:1 through 1997:1 period. They supplemented transaction data with information on elementary school student performance for public elementary schools and demonstrated the technique using data for the Carrollton-Farmers Branch Independent School District (CFBISD). Their results suggest that the metropolitan Dallas housing market is segmented by the quality of public education (as measured by student performance on standardized tests).

Goodman and Thibodeau (2003) subsequently applied the technique to all single-family properties in the Dallas County area and compared hierarchical model submarkets to two alternative housing submarket constructions: one that combined adjacent census tracts and a second that aggregated zip code districts. Using data for 28,000 single-family transactions for the 1995:1 through 1997:1 period, they examined hedonic house price prediction accuracy for the alternative housing submarket constructions. Their empirical results indicate spatial disaggregation yields significant gains in hedonic prediction accuracy. Orford (2000, 2002) also takes a multilevel approach to modeling the housing market in England.

Bourassa, Hamelink, Hoesli and MacGregor (1999) segment the Sydney and Melbourne, Australia housing markets by applying principal components and cluster analysis to a variety of neighborhood attributes. They report that three factors derived from twelve proximity and neighborhood attributes explain over 82 percent of the variance in house prices. They define housing submarkets by applying cluster analysis to these factors.

Bourassa, Hoesli and Peng (2003) and Thibodeau (2003) examine the effect that spatial disaggregation (e.g. employing submarkets) has on hedonic prediction accuracy. Bourassa, Hoesli and Peng (2003) examine two submarket constructions: (1) geographically concentrated "sales areas" used by local real estate appraisers in New Zealand; and (2) an aspatial submarket construction obtained by applying cluster analysis to the most influential factors generated from property, neighborhood and location attributes. They compared the hedonic house price predictions generated from these alternatives to a single equation for the entire city model. They concluded that while the statistically generated submarkets significantly increased hedonic house price prediction accuracy relative to the single equation model, the statistically generated submarkets did not outperform the "sales area" submarkets. Thibodeau (2003) constructed submarkets within Dallas County by combining adjacent census block groups located within the same municipality and the same independent school district. He compared the hedonic predictions from this model to a single Dallas County model and to a model that included dummy variables for municipality. He also reported significant increases in prediction accuracy associated with spatial disaggregation.

Watkins (2001) provides a detailed review of the alternative approaches that housing economists have employed for characterizing housing submarkets. Using transaction data for the Glasgow housing market, he examined three alternative approaches for delineating housing submarkets: (1) spatially stratified housing submarkets; (2) submarkets based on the similarity of structural characteristics; and (3) a hybrid definition that nests dwelling characteristics based submarkets within spatially defined submarkets. He concluded that the nested model provided the best empirical approach for delineating submarkets.

Some analysts have delineated within metropolitan area housing submarkets based on determinants of housing demand, while others have delineated submarkets based on supply-side variables. This paper proposes a method that delineates housing submarkets based on price.

Theory

From the earliest literature that explicitly recognized separate housing submarkets (Straszheim 1974, 1975), analysts have concentrated on the role of housing supply in the grouping of nearby units into submarkets. With the premise that similar units should be grouped together, it has been easiest to group nearby units, generally (although not always) within the same municipality. One can appeal to the premise that nearby units share similar neighborhood characteristics, either measured or unmeasured, and indeed the sale of nearby units may impact the sale price of units to be sold (labeled comparable properties by real estate appraisers). One can also look to the grouping of nearby units as a way of making an enormous problem slightly less enormous. Following Cliff et al (1975) and Goodman (1981), the number of different ways that m dwelling units can be grouped into k submarkets is:

$$a = a(f_1,\ldots,f_k) = \frac{m!}{\left[\prod_{i=1}^{k} f_i!\right][g_1!\,g_2!\,g_3!\ldots g_j!]} \tag{1}$$

in which f_i is the number of units in the i^{th} submarket, g_j is the number of submarkets which comprise j units in the analysis, and $A = \sum a$, where the summation is over all k-element partitions of m. A very restrictive continuity constraint that "lines up" the dwelling units with their nearest neighbors and allows only linear grouping, reduces the number of ways that m units can be grouped to:

$$A = \sum_{k=1}^{m} \frac{(m-1)!}{(k-1)!(m-k)!} = 2^{m-1}, \tag{2}$$

still a very large number.

All of the assumptions above, however, ignore the demand side of housing markets. Consider the traditional central place model, where consumers work downtown and live away from their jobs.[1] As noted in Figure 1, most models would locate consumers at locations relative to the Central Business District (CBD), where the locations are defined by income. If the income elasticity of land demand exceeds (is less than) the income elasticity of travel costs, higher income individuals will locate further from (closer to) the CBD.

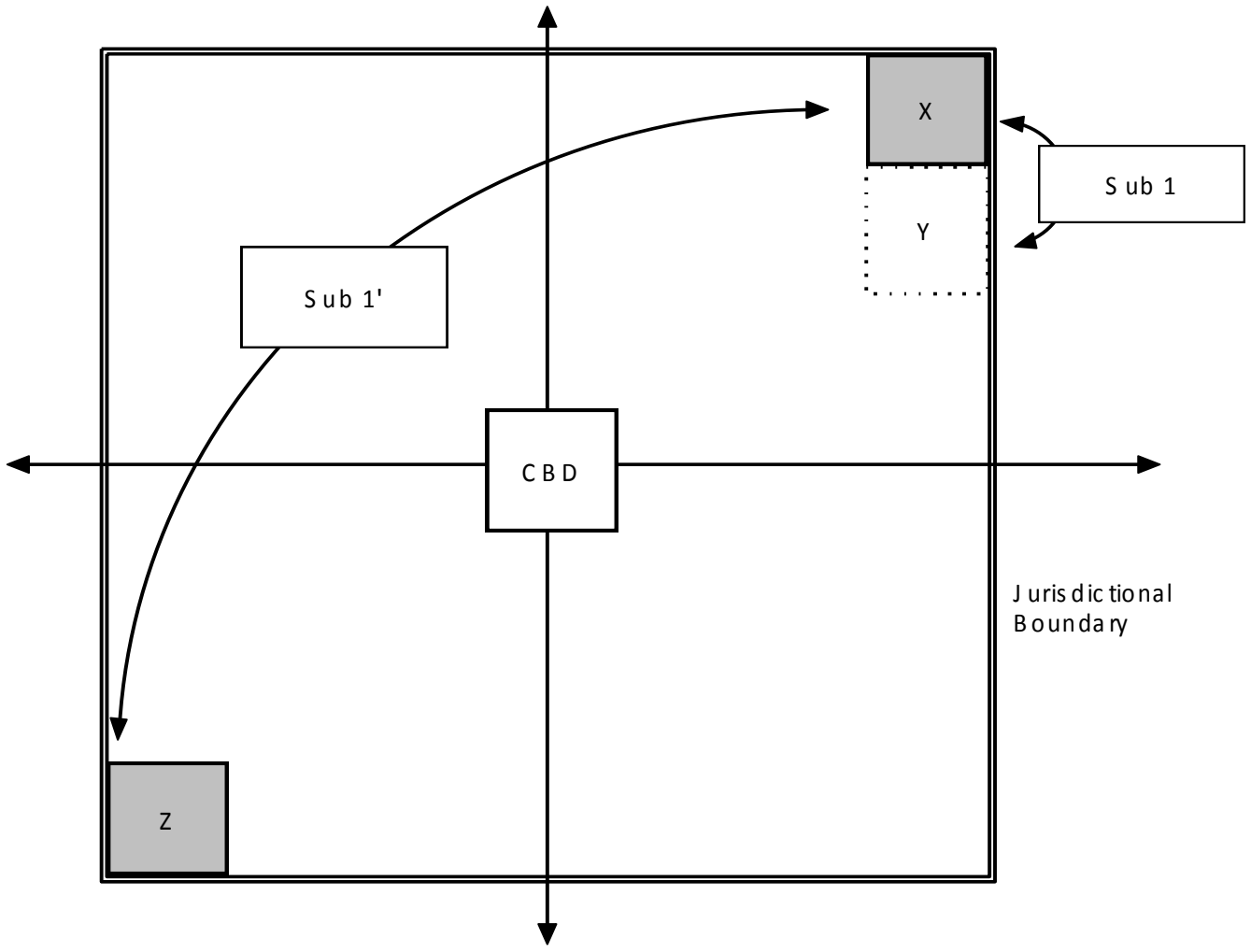

Figure 1 - Alternative Characterizations of Submarkets

[1] The central place model is provided for simplicity of exposition only. The same arguments will apply just as validly for areas with multiple workplace centers.

Consider dwelling unit *X* in Figure 1, at an arbitrary distance from the CBD. The researcher seeking to assign property *X* to a submarket might group *X* with dwelling *Y*, because dwelling *Y* is spatially "close." However, if lot sizes, house sizes, and municipal goods (even within the same municipality) are stratified by income, it could very well be that *X* is more appropriately grouped with unit *Z* in *Sub 1'*, which is the same distance from the CBD, but in the diametrically opposite direction, than with *Y* in *Sub 1* which is only close physically.[2]

What determines whether *X* should be grouped with *Y* or with *Z*? If a housing submarket is an area where the (per unit) price of housing is constant, then the house price should determine whether *X* is grouped with *Y* or with *Z*. If *X* is "priced" like *Z*, then *X* belongs in the same submarket as *Z*, even though *Z* is not "close" spatially.

Hedonic Estimation

This section describes the underlying hedonic regressions used to compare price delineated housing submarkets to spatially concentrated submarkets, using transaction data for Dallas, Texas. The spatially concentrated submarkets were constructed by combining adjacent census block groups located within the same municipality and the same independent school district. This grouping controls for two important neighborhood determinants of house price: public school quality and public safety. Goodman and Thibodeau (1998, 2003) have established that variation in school quality is capitalized in Dallas house prices. There is also substantial variation in the quality of municipal services. There are 25 separate municipalities within the Dallas Central Appraisal District (DCAD) area. Average police response times, for example, in Dallas County range from 25 minutes for the City of Dallas police to 2 minutes for police in Highland Park.

The main purpose of the paper is to empirically evaluate two alternative procedures for defining within-metropolitan area housing submarkets: the first alternative constructs housing submarkets by combining spatially adjacent census block groups within the same municipality and the same independent school district; the second alternative assigns properties to submarket based upon

[2] This point was first brought to Goodman's attention by Guy Orcutt, and later expounded by Stephen Mayo.

the temporally adjusted per square foot transaction price regardless of location. Fundamentally, the question is how well (sometimes unmeasured) spatial attributes get capitalized in the estimated coefficients for included structural characteristics. Naturally, the implementation of our test requires making empirical decisions that are subject to criticism. Alternative procedures for delineating spatial and aspatial submarkets should be considered.

We conduct our empirical investigation with just over 44,000 transactions. Forty-four thousand sales allow us to construct a significant number of spatially concentrated submarkets. Geographically small submarkets provide better control for (typically unmeasured) spatial attributes (including proximity externalities) compared to geographically large submarkets. Consequently, we construct as many submarkets as we think plausible given 44,000 transactions, and a hedonic specification (provided below) that estimates 25 unknown parameters.

The spatial submarkets were constructed by combining adjacent census block groups located in the same municipality and the same independent school district. Adjacent census block groups were combined until the submarket had about 120 transactions (using an econometric guideline suggesting 5 observations per estimated parameter to ensure parameter stability) available to estimate the parameters of the hedonic house price model. This procedure yielded 372 spatial submarkets.

The alternative housing submarket construction invokes demand criteria by assigning properties to submarkets based upon both dwelling size and on the average per square foot transaction price for the census block group. These submarkets were constructed in two steps. First, the distribution of census block group median per square foot transaction prices was divided into 100 segments. The census block groups (CBGs) with the lowest median per square foot transaction prices were assigned to the first percentile; the CBGs with the next to the lowest per square foot transaction prices assigned to the second percentile, etcetera. Second, properties within each per square foot price percentile were assigned to submarkets according to dwelling size (as measured by square feet of living area). Consequently the smaller properties in each collection of census block groups were separated from the larger properties holding CBG median per square foot transaction price roughly constant. This procedure yielded 325 housing submarkets. This

assignment *completely ignores* the spatial location of the property and could combine properties from different independent school districts and different municipalities. However, it is unlikely that a neighborhood with below average public schools would be combined with an area with superior public schools since school quality is capitalized in house price. Nevertheless, this assignment process is completely aspatial.

The empirical challenge in implementing this procedure using transactions that took place over a two year period is that Dallas house prices were not constant (in either nominal or real terms) over the 2000:4-2002:4 period. Furthermore, rates of house price appreciation varied spatially. Prior to constructing submarkets, the transactions were "marked to market" using a price index computed from hedonic house price equations. Separate hedonic equations were estimated for each municipality. In addition, for the large municipalities, separate house price indexes were estimated for low, median, and high priced housing.

The hedonic specification for marking property values to market and for evaluating the alternative submarket constructs includes numerous structural characteristics:

$$\begin{aligned}\ln(PRICE_{i,t}) = {} & \beta_0 + \beta_1*\ln(AREA) + \beta_2*\ln(SERVQ) + \beta_3*AGE + \\ & \beta_4*AGESQ + \beta_5*AGECUBE + \beta_6*BATHS + \\ & \beta_7*GHSYS + \beta_8*OHSYS + \beta_9*NACSYS + \beta_{10}*WACSYS + \\ & \beta_{11}*WETBAR + \beta_{12}*FIREPL0 + \beta_{13}*POOL + \\ & \beta_{14}*DTGAR + \beta_{15}*CARPORT + \beta_{16}*NOGAR + \\ & \sum_{t=1}^{T} \delta_t * SOLD_t + \mu_{i,t}, \end{aligned} \tag{3}$$

where

$PRICE_{i,t}$ = the transaction price of the i^{th} house sold in quarter t,

AREA = square feet of living area,

LNAREA = ln(AREA),

SERVQ = square feet of servant's quarters,

LNSERVQ = log (SERVQ) (ln (SERVQ)) = 0 if there are no servant's quarters),

DWELAGE = dwelling age,

AGE = dwelling age in decades,

AGESQ = AGE squared,

AGECUBE = AGE cubed,

BATHS = the number of bathrooms (two one-half bathrooms are counted as one full bath),

CHSYS = central heating system (the omitted heating system category),

GHSYS = dummy variable for (non-central) gas heating system,

OHSYS = dummy variable for other heating system--other heating systems include floor furnaces, wall heating systems, radiator heating systems, and no heating systems,

NACSYS = dummy variable for no air conditioning system,

WACSYS = dummy variable for window air conditioning system,

CACSYS = dummy variable for central air conditioning system (the omitted air conditioning category),

WETBAR = dummy variable for the presence of a wetbar,

FIREPL = dummy variable for the presence of at least one fireplace,

POOL = dummy variable equal to 1 if swimming pool present and zero otherwise,

ATGAR = dummy variable equal to 1 if the property has an attached garage and zero otherwise (the omitted category),

DTGAR = dummy variable equal to 1 if the property has a detached garage and zero otherwise,

CARPORT = dummy variable equal to 1 if the property has either an attached or a detached carport and zero otherwise,

NOGAR = a dummy variable equal to one if the property has no covered parking facility,

$SOLD_t$ = dummy variables for sale quarter, t = 2000:4 to 2002:3; the omitted sale quarter is 2002:4

Following Halvorsen and Palmquist (1980), the price index used to temporally adjust house prices to 2002:4 is e^{δ}.

Evaluating Alternative Submarket Definitions

To facilitate comparison of the alternative submarket delineation procedures, the sample of transactions was separated into an estimation subsample and a prediction subsample. The transactions in the estimation subsample are used to estimate parameters for the hedonic models defined by the alternative submarket delineations. The transactions in the prediction sample are excluded from the estimation sample and are used to evaluate prediction accuracy for the alternative submarket constructions. The same estimation and prediction subsamples are used for each alternative. Consequently, any variation in prediction accuracy cannot be attributed to differences in the underlying sample (although these particular results may be an artifact of the particular sample drawn). The estimation sample is a 90% random sample of all transactions. This sample was selected using a uniform random variable. The remaining observations are held out to form the prediction sample.

The alternative housing submarket definitions are evaluated using numerous statistical criteria: the mean absolute value of the prediction error, the mean percentage error, the mean squared error, and the percent of the time that a predicted price is within 10%, 15% and 20% of the observed price. The prediction accuracy threshold employed by the automated valuation model (AVM) industry is that at least 50 percent of the predicted house prices must be within ten percent of observed transaction prices.

We also evaluate the alternative definitions of housing submarkets using a non-nested *J*-test. Following Davidson and MacKinnon (1981), Goodman and Dubin (1990) employ the non-nested *J*-test to examine alternative definitions of submarkets. The non-nested *J*-test compares one specification (a particular set of regressors, functional form, or submarket definition) against an alternative when the alternative cannot be expressed as a restriction on the null hypothesis. In our case, the null hypothesis is that the spatially proximate submarket definition is the appropriate way to delineate submarkets and the alternative is that housing submarkets are more appropriately defined by dwelling size and census block group average per square foot prices. The two submarket formulations may be considered as the spatially proximate submarket formulation:

$$H_0: \mathbf{y} = \mathbf{X\beta} + \varepsilon_0, \tag{4}$$

and the per square foot formulation:

$$H_1: \mathbf{y} = \mathbf{Z\gamma} + \varepsilon_1, \tag{5}$$

H_1 cannot be written as a restriction on H_0, so conventionally nested *F*-tests of covariance are not appropriate.

One possibility for testing the restrictions involves an artificial nesting of the two models. Following Davidson and MacKinnon (1981) and Greene (2003), define $\mathbf{Z_1}$ as the set of $\mathbf{Z}$ that are not in $\mathbf{X}$, and $\mathbf{X_1}$ likewise with respect to $\mathbf{Z}$. A standard *F*-test can be carried out to test the hypothesis that in the augmented regression:

$$\mathbf{y} = \mathbf{X\beta} + \mathbf{Z_1\gamma_1} + \boldsymbol{\mu}_1, \tag{6}$$

the vector $\boldsymbol{\gamma}_1 = 0$, with the test then reversed (with $\mathbf{Z}$ as the null hypothesis). Greene notes that this compound model may have an "extremely large" number of regressors (in this problem the number of elements of $\mathbf{Z_1}$ will always equal the number of elements of $\mathbf{X}$ unless specific submarkets are identical). This is potentially troublesome if one is comparing more than two

alternative well-specified hedonic formulations, with large numbers of regressors.

The Davidson and MacKinnon *J*-test allows the researcher to test H_0 against the alternative H_1 with the *single* parameter α:

$$\mathbf{y} = (1 - \alpha)\, \mathbf{X\beta} + \alpha\, \hat{\mathbf{(Z\gamma)}} + \boldsymbol{\mu}, \tag{7}$$

and reversing the test with:

$$\mathbf{y} = (1 - \alpha')\, \mathbf{Z\gamma} + \alpha'\, \hat{\mathbf{(X\beta)}} + \boldsymbol{\mu}', \tag{8}$$

where: y is the (log of) the actual transaction price,

$\mathbf{X\beta}$ is the spatially proximate submarket model,

$\boldsymbol{Z\gamma}$ is the price per square foot model, and '^' denotes predicted value.

The test is H_0: $\alpha = 0$ vs. H_1: $\alpha \neq 0$. If the *t*-statistic is significant we reject H_0, which assumes that the alternative housing market constructions do not provide additional information. We compute similar test statistics with the per square foot submarket model as the null and with the spatially proximate submarket model as the alternative. For the spatially proximate submarket model to dominate, we must fail to reject the spatially proximate submarket null (i.e. the first *J* test must be insignificant), but we must reject similar hypotheses with the per square foot model as the null (the *J* tests must be significant).

To implement the *J*-test, we construct a block-diagonal design matrix. The block matrices, X_J, contain the regressors for submarket J. The design matrix includes the predicted house prices under the alternative submarket hypothesis, H_1, and $\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_N$ represent N vectors of hedonic coefficients (one vector of coefficients for each submarket). α is the scalar *J* test statistic with its accompanying confidence interval:

$$\begin{bmatrix} \mathbf{X_1} & 0 & 0 & ..... & 0 & \hat{Y}_{1,H_1} \\ 0 & \mathbf{X_2} & 0 & ..... & 0 & \hat{Y}_{2,H_1} \\ 0 & 0 & \mathbf{X_3} & ..... & 0 & \hat{Y}_{3,H_1} \\ ..... & & & & & \\ 0 & 0 & 0 & ..... & \mathbf{X_N} & \hat{Y}_{N,H_1} \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta_1} \\ \boldsymbol{\beta_2} \\ \boldsymbol{\beta_3} \\ .... \\ \boldsymbol{\beta_N} \\ \alpha \end{bmatrix} = \begin{bmatrix} Y_{1,H_0} \\ Y_{2,H_0} \\ Y_{3,H_0} \\ ... \\ Y_{N,H_0} \end{bmatrix} \quad (9)$$

The parameters are estimated twice: one under the null that spatially segmented markets is the appropriate submarket construct and a second time under the null that per square foot segmented markets is the appropriate submarket construct.

The *J*-test also provides an indirect demonstration of the benefits of combining estimators (Fair and Shiller, 1989, 1990). A hybrid predictor can be computed as a linear combination of the two alternatives:

$$y = (1 - \alpha)\, \hat{X\beta} + \alpha\, (\hat{Z\gamma}) + \mu, \quad (10)$$

The hybrid predictor will have a lower mean squared error when α is statistically significant.

The Data

The study data were obtained from the Dallas Central Appraisal District (DCAD). The DCAD assesses property value for tax purposes for all real property in Dallas County and in portions of adjacent counties. The characteristics of the 2002 DCAD single-family housing stock are summarized in Table 1. There were 502,541 single-family properties in the DCAD jurisdiction in 2002. The average single-family home had 1,778 square feet of living area and was 33.6 years old. Most properties have central heating and central air-conditioning systems. Just over ten percent of single-family homes in Dallas have swimming pools.

There were just over 44,000 sales of single-family properties between the fourth quarter of 2000 and the end of 2002. The mean (temporally unadjusted) transaction price was about $164,700.

The homes that sold were typically younger and larger than properties in the DCAD housing stock (Table 2).

Map 1 illustrates the locations of the municipalities within Dallas County and Table 3 provides information on the spatial distribution of single-family homes. The first four columns of Table 3 provide the number of properties in the metropolitan area, the percent of the Dallas County stock, the mean dwelling size and mean dwelling age for single-family homes for each municipality. The last six columns provide information on the single-family transactions for each municipality: the number and percent of sales, the means for square feet of living area and dwelling age, and the mean nominal (temporally unadjusted) transaction prices. Nearly 43% of the single-family housing stock and 35% of the sales are located in the City of Dallas. The oldest, largest, and most expensive homes are located in Highland Park. The youngest homes are in Coppell, a relatively new municipality located in the northwest corner of Dallas County. The least expensive homes are located in the southeast corner of the County (Wilmer and Hutchins). The properties that sold tend to be larger and younger than the average existing home in Dallas.

Census block groups were assigned to submarkets using contemporaneous (e.g. temporally adjusted) prices. Temporal adjustments were computed using estimated coefficients from municipality specific hedonic equations. Time adjustment factors were computed separately for low-priced, moderately-priced, and high-priced housing for the 15 largest municipalities. For the smaller municipalities, all properties within the city were marked to market using a city-wide average temporal price index. The average time adjusted price is about $169,000. The temporal adjustment indices derived from these equations are presented in Table 4. The index number for all places in 2002:4 is 1.0000. To estimate the 2002:4 market value for an Addison property that sold in 2000:4, for example, the observed transaction price was increased by 4.99%.

There is substantial variation in rates of house price appreciation: both across metropolitan areas and within a metropolitan area's distribution of house prices. In the portion of Carrollton located in Colin County, low-priced homes appreciated over 28% over the 2000:4-2002:4 period while the most expensive homes in the same area decreased in value over the same period. In the City of Dallas, low-priced homes appreciated 3.3% over the 2000:4-2002:4 period while the most

expensive homes appreciated at nearly twice that rate--6.5%. On average, house prices in the DCAD area increased by about 5% over the 2000:4-2002:4 period.

There are two separate issues here: (1) what determines house prices; and (2) what determines appreciation rates. This paper argues that housing markets could be established based on house prices (not appreciation rates). To evaluate this housing submarket construct against an alternative (spatial) construct, we need to control for temporal variation in house prices over our period of analysis. There are two ways to do this. One is to simply include dummy date of sale variables in the hedonic house price equations and not worry about spatial variation in appreciation rates. However, our empirical analysis of house price appreciation clearly indicates that appreciation rates vary substantially across metropolitan areas (and even within metropolitan areas by house price). Estimating price indices using dummy variables with data from multiple cities (e.g. the aspatial model) would not adequately control for temporal variation in house prices. An alternative assumption would be that house price appreciation rates for specific types of housing (for low, medium and high priced housing) are fairly constant for properties within a metropolitan area.

The alternative submarket constructions yield very different representations of housing submarkets. We computed the mean Euclidean distance between a transaction and the geographic center of the transaction's assigned submarket (as measured by the mean easting and northing for all transactions in the submarket). This produced 372 average distances for the spatial submarkets and 325 average distances for the aspatial submarkets. Table 5 reports the across submarket mean distances for these within submarket average distances. For the spatial submarket assignment, the mean distance between a transaction and the geographic center of the submarket is 0.85 kilometers (with a standard deviation of 0.88 kilometers). For the aspatial submarket definition, the average distance between a transaction and its geographic center is 10.88 kilometers (with a standard deviation of 4.82km).

The spatial submarket construct assigns all properties located in the City of Farmers Branch to one of five spatially concentrated Farmers Branch submarkets. The aspatial submarket construct assigns transactions to submarkets based on price and ignores location. Map 1 illustrates the

disparate locations of properties assigned to an aspatial submarket belonging to a particular property in one of the Farmers Branch submarkets. The aspatial construct assigned a subset of the Farmers Branch properties to six different municipalities located across northern Dallas County: Carrollton, Dallas, Farmers Branch, Garland, Irving and Richardson. A casual inspection of the map indicates that many of these properties separated by more than 30 kilometers.

There is significant variation in the distributions of transaction prices across submarket constructs. Table 5 shows the standard deviation for the distribution of transaction prices within each submarket for both submarket constructs. The mean standard deviation in (temporally adjusted) transaction prices for the spatial submarkets is $54,145 and the mean standard deviation for the aspatial submarkets is $36,924.

Results

Hedonic house price predictions were also computed using an all DCAD model to facilitate evaluation of the alternative submarket constructions. The estimated parameters for the all DCAD model (results available from the authors) explain 82% in the variation in the log of transaction price. Nearly all of the estimated coefficients are statistically significant at conventional levels and all the estimated coefficients have the expected signs.

The estimated coefficients from the hedonic equations for the three alternative housing submarket specifications (e.g. no submarkets, spatial submarkets, and aspatial dwelling size-per square foot submarkets) were used to predict 2002:4 transaction prices. The predicted prices were corrected for the finite sample bias that results from using a semi-log house price specification (see Thibodeau, 1992).

The hedonic prediction accuracy results are in Table 6. While the all-DCAD model explains over 80% of the variance in the log of transaction price, in part because there is considerable variance to explain, this model does not predict price very accurately. Less than 36% of the predicted prices are within ten percent of the observed transaction price, about half are within

15%. The all-DCAD model does not come close to satisfying the automated valuation models (AVMs) industry standard threshold for prediction accuracy.

The spatially concentrated submarkets produce a dramatic improvement in hedonic prediction accuracy. The mean absolute dollar error is reduced by over $ 15,000 – from $34,276 to 18,979. The percent of predicted prices that are within ten percent of observed prices increases from 36% to 66%! Over 86% of the predicted prices are within twenty percent of the observed price.

The aspatial submarket model has a lower mean and mean squared error, but slightly fewer predicted prices within ten, fifteen and twenty percent of the observed prices. The mean squared prediction error for the aspatial submarket model is 24.3% lower than the mean squared prediction error for the spatially concentrated submarket model.

Table 7 contains results for the non-nested *J*-test. The *J*-test statistics indicate that neither submarket construction statistically dominates the alternative. With spatially proximate submarkets the null hypothesis, the estimated coefficient for predicted prices from the (alternative) aspatial submarket model is 0.84. The standard error of the estimate is 0.0077. When the null is reversed, the estimated coefficient for predicted values for the (alternative) spatially proximate submarket model is 0.82 with a standard error of 0.0072. Both nulls are rejected at conventional levels. In economic terms, each alternative model provides additional information to the null for prediction purposes.

Can prediction accuracy be increased by combining models? We estimated the parameters of a hybrid model that minimizes the mean squared prediction error associated with taking a weighted average of the two estimators. The OLS parameters were computed without an intercept and with the constraint that the weights sum to one. The estimation results, in Table 8, indicate that Least Squares applies 80% weight to the aspatial submarket model and 20% to the spatially concentrated submarket model. The hybrid model reduces the mean absolute error to $18,400 (Table 6) and the mean squared prediction error, but the spatially concentrated model still has the highest percent of predictions within ten percent of observed transaction prices.

Conclusions

This research examined alternative procedures for delineating housing submarkets within metropolitan areas. The results indicate that delineating housing submarkets by dwelling size and per square foot house price perform *about* as well as spatially concentrated submarkets that control for variation in public school quality and the provision of public safety. In fact, the "winner" of the competition depends on how performance is measured. The spatially proximate submarket model yields more predictions within ten percent of observed prices, but the predictions from this model have a significantly higher mean squared prediction error.

These results have important implications for empirically modeling submarkets within metropolitan area housing markets. Creating housing submarkets by combining spatially adjacent census block groups that lie within the same municipality and same independent school district is time consuming and costly. These results suggest that comparable increases in hedonic prediction accuracy can be achieved by delineating submarkets by dwelling size and median census block group per square foot transaction price.

A version of this paper was presented at the January 2004 AREUEA meetings in San Diego, CA. We would like to acknowledge Chris Redfearn, Ed Coulson and two referees for providing comments on earlier drafts.

References

Bourassa, S.C., F. Hamelink, M. Hoesli and B.D. MacGregor. 1999. Defining Housing Submarkets. *Journal of Housing Economics* 8(June):160-183.

Bourassa, S. C., M. Hoesli and V.S. Peng. 2003. Do Housing Submarkets Really Matter? *Journal of Housing Economics* 12(1):12-28.

Bryk, A. S. and S.W. Raudenbush. 1992. *Hierarchical Linear Models: Applications and Data Analysis Methods*. Sage Publications. Newbury Park.

Cliff, A.D., P. Haggett, J.K. Ord, K.A. Bassett, and R.B. Davies. 1975. *Elements of Spatial Structure*. Cambridge: Cambridge University Press, Chapter 2.

Dale-Johnson, D. 1982. An Alternative Approach to Housing Market Segmentation Using Hedonic Price Data. *Journal of Urban Economics* 11:311-332.

Davidson, R. and J.G. MacKinnon. 1981. Several Tests for Model Specification in the Presence of Alternative Hypotheses. *Econometrica* 49(3):781-793.

Dubin, R.A., and C.H. Sung. 1990. Specification of Hedonic Regressions: Non-nested Tests on Measures of Neighborhood Quality. *Journal of Urban Economics* 27:97-110.

Fair, R.C. and R.J. Shiller. 1990. Comparing Information in Forecasts from Econometric Models. *American Economic Review* 80:375-389.

_____. 1989. Informational Content of Ex Ante Forecasts. *Review of Economics and Statistics* 71:325-331.

Goodman, A.C. 1978. Hedonic Prices, Price Indices, and Housing Markets. *Journal of Urban Economics* 5(4):471-484.

_____. 1981. Housing Submarkets within Urban Areas: Definitions and Evidence. *Journal of Regional Science* 21:175-185.

Goodman, A.C., and R.A. Dubin. 1990. Non-Nested Tests and Sample Stratification: Theory and a Hedonic Example. *Review of Economics and Statistics* 72 (February):168-73.

Goodman, A.C. and T.G. Thibodeau. 1998. Housing Market Segmentation. *Journal of Housing Economics* 7:121-143.

_____. 2003. Housing Market Segmentation and Hedonic Prediction Accuracy. *Journal of Housing Economics* 12(3):181-201.

Greene, W.H. 1993. *Econometric Analysis. Second Edition*. Macmillan Publishing Company.

Halvorsen, R. and R. Palmquist. 1980. The Interpretation of Dummy Variables in Semilogarithmic Equations. *American Economic Review* 70(3): 474-475.

Kain, J., and J.Quigley. 1970. Measuring the Value of Housing Quality. *Journal of the American Statistical Association* 65(330):532-48.

Li, M. and H.J. Brown. 1980. Micro-Neighborhood Externalities and Hedonic Housing Prices. *Land Economics* 56(2):125-141.

Maclennan, D. and Y. Tu. 1996. Economic Perspectives on the Structure of Local Housing Systems. *Housing Studies* 11:387-406.

Orford, S. 2000. Modelling Spatial Structures in Local Housing Market Dynamics: A Multilevel Perspective. *Urban Studies* 37 (9):1643-1671.

_____. 2002. Valuing Locational Externalities: a GIS and Multilevel Modeling Approach. *Environment and Planning B- Planning & Design* 29 (1):105-127.

Straszheim, M.R. 1974. Hedonic Estimation of Housing Market Prices: A Further Comment. *The Review of Economics and Statistics* 56 (3):404-406.

_____. 1975. *An Econometric Analysis of the Urban Housing Market.* New York: National Bureau of Economic Research.

Thibodeau, T.G. 2003. Marking Single-Family Property Values to Market. *Real Estate Economics* 31(1):1-22.

_____. 1992. *Residential Real Estate Prices: 1974-1983*. Blackstone Books: Studies in Urban and Resource Economics.

Watkins, C.A. 2001. The definition and identification of housing submarkets. *Environment and Planning A* 33 (12) (December): 2235-2253.

Table 1

Characteristics of the 2002 DCAD Single-Family Housing Stock

Variable	N	Mean	Std Dev	Minimum	Maximum
area	502,541	1778.20	847.56	500.00	10000.00
dwelage	502,541	33.61	18.80	0	75.00
BATHS	502,541	1.97	0.79	0	7.00
CHSYS	502,541	0.84	0.37	0	1.00
GHSYS	502,541	0.14	0.34	0	1.00
OHSYS	502,541	0.02	0.15	0	1.00
CACSYS	502,541	0.81	0.39	0	1.00
WACSYS	502,541	0.16	0.37	0	1.00
NACSYS	502,541	0.03	0.16	0	1.00
WETBAR	502,541	0.09	0.28	0	1.00
FIREPL	502,541	0.68	0.59	0	4.00
POOL	502,541	0.11	0.31	0	1.00
ATGAR	502,541	0.70	0.46	0	1.00
DTGAR	502,541	0.12	0.32	0	1.00
CARPORT	502,541	0.04	0.20	0	1.00
NOGAR	502,541	0.14	0.35	0	1.00

Table 2

Descriptive Statistics for 2000:4-2002:4 Single-Family Transactions

Variable	N	Mean	Std Dev	Minimum	Maximum
price	44,001	164695.67	131600.10	14000.00	2400000.00
tadjprice	44,001	169057.73	135450.49	15288.48	2591550.50
area	44,001	1896.50	762.92	518.00	7716.00
adjpsf	44,001	85.09	34.10	21.28	696.67
dwelage	44,001	27.70	18.33	0	75.00
BATHS	44,001	2.12	0.68	0	5.50
CHSYS	44,001	0.94	0.23	0	1.00
GHSYS	44,001	0.05	0.21	0	1.00
OHSYS	44,001	0.01	0.11	0	1.00
CACSYS	44,001	0.93	0.25	0	1.00
WACSYS	44,001	0.06	0.24	0	1.00
NACSYS	44,001	0.01	0.08	0	1.00
WETBAR	44,001	0.11	0.32	0	1.00
FIREPL	44,001	0.81	0.52	0	3.00
POOL	44,001	0.13	0.33	0	1.00
ATGAR	44,001	0.80	0.40	0	1.00
DTGAR	44,001	0.10	0.29	0	1.00
CARPORT	44,001	0.03	0.17	0	1.00
NOGAR	44,001	0.08	0.27	0	1.00
SQM01	44,001	0.12	0.32	0	1.00
SQM02	44,001	0.13	0.33	0	1.00
SQM03	44,001	0.10	0.31	0	1.00
SQM04	44,001	0.11	0.31	0	1.00
SQM05	44,001	0.14	0.34	0	1.00
SQM06	44,001	0.14	0.35	0	1.00
SQM07	44,001	0.10	0.30	0	1.00
SQM08	44,001	0.10	0.30	0	1.00

Note: tadjprice is the temporally adjusted price; adjpsf is the per square foot temporally adjusted price; and SQM01-SQM08 are dummy variables for sale quarter with SQM01 corresponding to 2000:4.