Predicting House Prices with Spatial Dependence:

A Comparison of Alternative Methods

Steven C. Bourassa

School of Urban and Public Affairs, University of Louisville, 426 W. Bloom Street, Louisville, Kentucky 40208, and CEREBEM, Bordeaux Management School, Bordeaux, France, phone: (502) 852 5720, fax: (502) 852 4558,

email: steven.bourassa@louisville.edu

Eva Cantoni

Department of Econometrics, University of Geneva, 40 boulevard du Pont-d'Arve, CH-1211 Geneva 4, Switzerland, email: eva.cantoni@unige.ch

and

Martin Hoesli

HEC and SFI, University of Geneva, 40 boulevard du Pont-d'Arve, CH-1211 Geneva 4, Switzerland, University of Aberdeen Business School, Scotland, and *CEREBEM, Bordeaux Management School, Bordeaux, France,* email: martin.hoesli@unige.ch

Paper presented at the 15th Conference of the Pacific Rim Real Estate Society (PRRES), Sydney, Australia, 18th- 21st January 2009

Submission for the best conference paper award.

Predicting House Prices with Spatial Dependence:

A Comparison of Alternative Methods

Abstract

This paper compares alternative methods for taking spatial dependence into account in house price prediction. We select hedonic methods that have been reported in the literature to perform relatively well in terms of ex-sample prediction accuracy. Because differences in performance may be due to differences in data, we compare the methods using a single data set. The estimation methods include simple OLS, a two-stage process incorporating nearest neighbors' residuals in the second stage, geostatistical, and trend surface models. These models take into account submarkets by adding dummy variables or by estimating separate equations for each submarket. Submarkets are defined at different levels of aggregation. We conclude that a geostatistical model with disaggregated submarket variables performs best.

Key Words: spatial dependence, hedonic price models, geostatistical models, mass appraisal, housing submarkets

JEL Codes: C21, R31

Introduction

The hedonic method is increasingly being used for price index construction, mass appraisal, and other purposes. With respect to price index construction, the hedonic

method yields indices that are used for multiple purposes, such as tracking housing markets, analysis of real estate bubbles, or investment benchmarking. For mass appraisal, the estimates yielded by hedonic models are used as a basis for the taxation of properties, but in some countries also to assess the value of properties for mortgage underwriting and for performance analyses of real estate portfolios. The method is also well suited to assess the impacts of externalities, such as increased noise levels resulting for instance from the extension of an airport, on house values.

Caution, however, should be exercised when devising hedonic models. Appropriate variables must be selected carefully and measured accurately. And, as with all regression models, errors should be independent from one another, else parameter estimates will be inefficient and confidence intervals will be incorrect. Both theory and empirical research suggests that the independence assumption is unlikely to be valid in a standard ordinary least squares (OLS) context. Basu and Thibodeau (1998), for instance, argue that spatial dependence exists because nearby properties will often have similar structural features (they were often developed at the same time) and also share locational amenities. Consistent with theory, much empirical analysis has concluded that house price residuals are spatially dependent.

Multiple authors have analyzed alternative methods for constructing and estimating hedonic models with spatial dependence in the context of mass appraisal. For example, Dubin (1988) compared geostatistical and OLS techniques, as did Basu and Thibodeau (1998). Other efforts include: Can and Megbolugbe (1997), who investigate a spatial lag model; Pace and Gilley (1997), who develop lattice models; Fik, Ling, and Mulligan (2003), who explore a trend surface model; Thibodeau (2003), who considers

the importance of spatial disaggregation in a geostatistical model; and Case, Clapp, Dubin, and Rodriguez (2004), who compare various approaches.

One difficulty in comparing these studies is that they use different data, and their results may be data-dependent. One contribution of the present paper is to compare several methods using the same data set. We use a data set from Louisville, Kentucky, containing approximately 13,000 house sales for 1999. Our approach is similar to that of Case, Clapp, Dubin, and Rodriguez (2004), but we consider their best model (contributed by Case) in comparison to other methods that have performed well in other studies. We focus specifically on the best models from Thibodeau (2003), Fik, Ling, and Mulligan (2003, henceforth FLM), and Bourassa, Cantoni, and Hoesli (2007, henceforth BCH). Another contribution of the present paper is to perform each type of analysis using 100 random samples of the data for estimation purposes to insure that the results are not specific to a particular sample.

The estimation methods include simple OLS, a two-stage process incorporating nearest neighbors' residuals in the second stage (similar to Case), geostatistical (similar to Thibodeau and BCH), and trend surface models (similar to FLM). These models take into account submarkets by adding dummy variables (as in BCH) or by estimating separate equations for each submarket (as in Thibodeau). Submarkets are defined at different levels of aggregation, ranging from highly disaggregated (as in Thibodeau) to less disaggregated (as in Case).

We conclude that taking into account submarkets is important in achieving more accurate house price predictions. Highly disaggregated submarkets are more effective than less disaggregated ones. Our results further show the benefits of modeling spatial

dependence in the error term. Geostatistical methods seem more useful than the two-stage nearest neighbors' residual procedure. Our best result is for a geostatistical model with dummy variables for disaggregated submarkets.

The structure of the paper is as follows. The next section contains a review of techniques for modeling spatial dependence. The third section summarizes previous comparative research. The subsequent section contains a discussion of the research design, which is followed by a section on our empirical analyses. The final section contains some concluding remarks.

Modeling Spatial Dependence

Spatial dependence can be treated in two basic ways. We assume a general model,

$$Y = \mu(X) + \varepsilon , \qquad (1)$$

where Y is a vector of transaction prices, X is a matrix of values for residential property characteristics, and ε is an error term. The first approach is to model $\mu(X)$ so that residuals over space do not exhibit any pattern. This may involve incorporating geographical coordinates (Colwell, 1998; Pavlov, 2000; Clapp, 2003). For example, FLM use data from Tucson, Arizona, to estimate what is in effect a three-dimensional or trend surface of property values based on a small number of property characteristics: land area, floor area, and age. Their OLS model includes these characteristics, plus x- and y-coordinates and submarket dummies. In addition, squares and cubes of these variables

and various interactive terms are included. A stepwise regression procedure eliminates collinear variables. They argue that this method captures the spatial dependence in the data and also results in substantial improvement in prediction accuracy. This approach is related to the local regression model of Clapp (2003), which is one of the models considered by Case, Clapp, Dubin, and Rodriguez (2004), and other methods that have come out of the geographical literature (see, e.g., Geniaux and Napoléone, 2008).

Another commonly used method is to add spatial indicators such as dummy variables for submarkets, which can be defined as geographical areas or non-contiguous groups of dwellings having similar characteristics and/or hedonic prices. An alternative to the use of dummy variables is to estimate separate equations for each submarket, thus allowing both intercepts and slopes to vary across areas or groups. For example, Thibodeau (2003) and Goodman and Thibodeau (2007) combine census block groups into small areas with enough transactions to estimate separate hedonic equations. Case (in Case, Clapp, Dubin, and Rodriguez, 2004) uses cluster analysis based on hedonic prices and demographic characteristics for census tracts to identify submarkets; he then estimates separate hedonic equations for each submarket.

Another approach to modifying $\mu(X)$ is to consider spatial lags, which are neighboring properties' prices or residuals. One such method includes as a regressor the weighted average of recent sale prices for nearby properties (Can and Megbolugbe, 1997). A variation on that method adds to predicted house prices an average (possibly weighted) of nearby properties' residuals (Bourassa, Hoesli, and Peng, 2003). A more complicated two-step estimation procedure takes an average of neighboring properties' residuals from a first-stage estimation and adds that as a regressor in the second stage

(Case, Clapp, Dubin, and Rodriguez, 2004). The latter method is equivalent to the former method if the estimated coefficient for the residual in the second-stage equation equals one; otherwise, the two-step procedure should yield better results.

The second approach is to model ε, that is, to assume not only that $E(\varepsilon) = 0$, but also that $E(\varepsilon\varepsilon') = \Omega$, which is a matrix with at least some nonzero off-diagonal elements. This approach includes geostatistical models such as those applied by Dubin (1998) or Basu and Thibodeau (1998) and the lattice models that have been refined and applied by Pace and his colleagues (e.g., Pace and Barry, 1997). BCH give an overview of these methods. The assumptions behind the two classes of spatial statistical models differ in terms of the definition of the domain over which spatial locations are permitted to vary. In the case of lattice models, which include simultaneous autoregressive (SAR) and conditional autoregressive (CAR) variants, locations are restricted to the discrete set of points represented by the data used to estimate the model. In contrast, geostatistical models permit an infinite number of locations within a given geographical area. This has implications for the way predictions based on each type of model take into account spatial information.

The lattice approach models the covariance matrix of the errors parametrically, whereas the geostatistical approach builds the covariance matrix indirectly through a parametric variogram. Moreover, the underlying assumptions of the two approaches differ. Lattice models assume $\mu(X) = X\beta$ and parameterize the covariance function of the error term of the model by assuming either that $\Omega^{-1} = \sigma^2(I - \phi C)$ (CAR models) or that $\Omega^{-1} = \sigma^2(I - \alpha D)'(I - \alpha D)$ (SAR models), where C and D represent spatial weight matrices that specify the dependence among observations. Predictions are generally

computed simply as $\hat{Y} = X\hat{\beta}$, although methods are available for incorporating information from *D* when calculating fitted values. An example of the lattice approach is Pace, Barry, Clapp, and Rodriguez (1998), who use both spatial and temporal weight matrices for nearby and recent transactions.

In contrast to lattice models, geostatistical models are based on the assumption that the observed data at a location *s* is a realization of a random process $\{Y(s) : s \in F\}$, which is supposed to satisfy a second-order stationarity assumption, that is, for which $E(Y(s)) = \mu$ for all $s \in F$ (constant mean) and $Cov(Y(s_1), Y(s_2)) = C(s_1 - s_2)$ for all $s_1, s_2 \in F$, where $C(\cdot)$ is called the covariogram. In effect, the covariance between locations depends only on the distance between them.

The geostatistical approach attempts to model the covariance matrix through a procedure based on three steps: (1) computation of an empirical variogram; (2) parametric modeling of this variogram; and (3) kriging (that is, prediction). The only information needed to perform these three steps is the notion of variogram defined as a function of the distance h between locations:

$$2\gamma(h) = Var(Y(s+h) - Y(s)), \qquad (2)$$

where $\gamma(h)$ is called the semivariogram.

The classical and most popular estimator of the variogram is obtained by the method of moments and was first proposed by Matheron (1962):

$$2\hat{\gamma}(h) = \frac{1}{|N(h)|} \sum_{N(h)} \left(Y(s_i) - Y(s_j) \right)^2, \qquad (3)$$

where $N(h) = \{(i, j) : s_i - s_j = h\}$ and $|N(h)|$ is the number of distinct elements of $N(h)$. For a given distance h, this variogram estimator is a variance estimator over all pairs of observations that are at a distance h apart. Note that when data are irregularly spaced, the variogram is usually smoothed by summing over pairs of points that lie in a tolerance region. $\hat{\gamma}(h)$ is an unbiased estimator of $\gamma(h)$, but is known to be badly affected in presence of outliers. Therefore, Cressie and Hawkins (1980) have defined a more robust estimator:

$$2\tilde{\gamma}(h) = \left\{ \frac{1}{|N(h)|} \sum_{N(h)} \left| Y(s_i) - Y(s_j) \right|^{1/2} \right\}^4 \Bigg/ \left(0.457 + \frac{0.494}{|N(h)|} \right). \qquad (4)$$

In the presence of outlying observations this estimator is more stable.

The second step of the procedure consists of fitting a parametric model to the empirical variogram (either classical or robust). The most popular models include the exponential and spherical variograms. BCH conclude that these two variograms yield quite similar results in the same type of application as in this article, so we focus here on the exponential variogram, which is defined as

$$\gamma(h; \vartheta) = \begin{cases} 0 & \text{if } h = 0 \\ c_0 + c_e \left(1 - \exp(-h / a_e)\right) & \text{if } h \neq 0 \end{cases}, \qquad (5)$$

where $\vartheta = (c_0, c_e, a_e)'$ with $c_0 \geq 0$, $c_e \geq 0$ and $a_e \geq 0$. The parameter c_0 is the limit of $\gamma(h)$ when $h \to 0$ and is called the "nugget effect". The other parameters in ϑ control the functional form of $\gamma(h;\vartheta)$. The parametric variograms can be fitted to data by several procedures, which include—among others—(restricted) maximum likelihood and generalized least squares. BCH also conclude that the robust method is slightly superior to the classical method, so the exponential robust estimator is used here.

Given a fitted variogram, the procedure goes on to compute the prediction at a point s_0 as a linear combination of the responses, that is,

$$\hat{Y}(s_0) = \lambda' Y = \sum_{i=1}^{n} \lambda_i Y(s_i), \tag{6}$$

where $\lambda = (\lambda_1, \ldots, \lambda_n)'$ is obtained by minimizing the mean squared prediction error

$$E(Y(s_0) - \sum_{i=1}^{n} \lambda_i Y(s_i))^2 . \tag{7}$$

The solution for λ depends on $\gamma(s_0 - s_i)$ for all $i = 1, \ldots, n$, and on $\gamma(s_i - s_j)$ for all $1 \leq i, j \leq n$. $\hat{Y}(s_0)$ is the best linear unbiased predictor. The solution obtained is an exact interpolation at the sample points, that is, $\hat{Y}(s_i) = Y(s_i)$ for all $i = 1, \ldots, n$. Note in particular that the formula above allows the computation of predictions at both sampled and unsampled locations, thus avoiding the problem with lattice models.

Previous Comparative Research

Previous empirical comparisons of the ex-sample prediction accuracy of different methods include Bourassa, Hoesli, and Peng (2003), who compare a set of spatial submarkets defined by real estate appraisers with a set of aspatial submarkets created using factor and cluster analysis. They also consider the impacts of adjusting predictions by neighboring properties' residuals. Using data for Auckland, New Zealand, they find that the most accurate predictions are obtained with a citywide equation with spatial submarket dummy variables and adjustment by neighboring residuals. Separate submarket equations performed slightly worse or better than the citywide equation, depending on whether the predictions were or were not adjusted for neighboring residuals, respectively. A similar conclusion was reached by Fletcher, Gallimore, and Mangan (2000), who compared predictions from a model with postcode dummies for the Midlands region of the United Kingdom with separate equations for each postcode. They found that the former was slightly superior to the latter.

Goodman and Thibodeau (2003) compare predictions for three submarket definitions with those for a market-wide model for Dallas. The submarket models are defined based on zip codes, census tracts, and a hierarchical method described in Goodman and Thibodeau (1998). They conclude that each of the submarket definitions yields significantly better results than the market-wide model, but none of the submarket definitions dominates the others. Goodman and Thibodeau (2007) compare spatial submarkets consisting of adjacent census block groups with aspatial submarkets constructed based on dwelling size and price per square foot. Both submarket methods

produce significantly better predictions than obtained from the market-wide model, although neither clearly dominates the other.

Dubin's (1988) study uses data for Baltimore to compare predictions using OLS and a geostatistical technique. She concludes that the geostatistical approach is superior even when some neighborhood (census block group) characteristics are included as explanatory variables. Basu and Thibodeau (1998) compare the predictive ability of OLS and one geostatistical technique, concluding that the latter is superior for six of eight regions in Dallas.

Thibodeau (2003) applies OLS and geostatistical methods to data from Dallas to compare the prediction accuracy of various models. He estimates an overall market model for the Dallas area, the same model with dummy variables for municipalities, individual models for municipalities, and individual models for "neighborhoods". The neighborhoods were constructed by combining adjacent census block groups until there were at least 150 transactions in each estimation sample, with the estimation sample consisting of 90% of transactions in each neighborhood. His best results are for the individual neighborhood geostatistical estimations; however, at the neighborhood level, the accuracy obtained from geostatistical methods is only marginally greater than for the OLS estimations.

Using data for Auckland, New Zealand, BCH compare an OLS model that includes submarket dummy variables with geostatistical, CAR, and SAR models. They show that lattice methods perform poorly in a mass appraisal context in comparison to geostatistical approaches or even a simple OLS model that ignores spatial dependence; however, they do not consider the possibility of using neighboring properties' residuals to

improve prediction accuracy. Their best results are obtained by incorporating submarket variables into a geostatistical framework.

Neill, Hassenzahl, and Assane (2007) compare OLS models including locational and submarket (census tract) variables with geostatistical models in a study of price impacts of air quality variations in Las Vegas. They find that geostatistical predictions outperform OLS predictions in approximately 90% of ex-sample cases.

Case, Clapp, Dubin, and Rodriguez (2004) apply OLS and several spatial statistical methods to a large sample of transactions (from Fairfax County, Virginia), using out-of-sample prediction accuracy for comparison purposes. The methods include: ordinary least squares with latitude and longitude variables for trend surface analysis as well as census tract and time dummies; Clapp's local regression model which applies OLS techniques to housing characteristics and a nonparametric smoothing method to a three-dimensional vector of latitude, longitude, and time for each transaction; Dubin's geostatistical approach which estimates a separate equation for each prediction point using a subsample of the data; and Case's approach which forms submarkets by applying cluster analysis to census tracts and then uses a two-stage estimation procedure that incorporates nearest neighbors' residuals from the first stage as variables in the second stage. In a second round of estimations, all of the models were supplemented by nearest neighbor residuals. After adjusting for neighbors' residuals, the results were quite similar across different estimation methods but Case's results were marginally better than the others.

FLM compare four models using data from Tucson: a standard OLS hedonic model; the standard model with the addition of submarket dummy variables; a trend

surface model with latitude and longitude interacted with each other and the hedonic characteristics; and the same model with the addition of submarket dummies (also interacted). The second and third models are superior to the first and similar to each other with respect to prediction accuracy. Their best results are for the fourth model that includes both trend surface variables and submarket dummies.

Research Design

Our focus is on comparing methods that have performed relatively well in terms of ex-sample prediction accuracy. These methods include a geostatistical model, models taking into account nearest neighbors' residuals, and trend surface models. The primary criterion for comparison purposes is the percentage of ex-sample predictions within 10% of the transaction price. According to FLM, for example, Freddie Mac's criterion for evaluating automated valuation models is that at least 50% of predictions should be within 10% of the actual price. Using a geostatistical model estimated for individual submarkets in Dallas, Thibodeau (2003) obtains a 63.6% accuracy rate. BCH estimate a geostatistical model with submarket dummy variables for Auckland, New Zealand. They achieve an accuracy rate of 49.3%, which is just shy of the Freddie Mac threshold. The most comparable approach in Thibodeau (2003) yields an accuracy rate of 43.4%. The best results in Case, Clapp, Dubin, and Rodriguez (2004) are those estimated by Case for homogeneous districts obtained using cluster analysis. A two-stage process first estimates individual equations for each district and then uses nearest neighbor residuals as variables in a second stage. He does not report the percentage of predictions within 10% of the actual price; however, he does report that the mean of the absolute value

percentage error is 11.8% (the median is 8.0%). BCH report a comparable statistic of 14.3% for their geostatistical model with submarket dummies. For their best model with both trend surface and submarket variables for Tucson, FLM report that 65.0% of their predictions are within 10% of transaction price.

The differences in accuracy across different studies may be due to either methods or data. Consequently, we apply the methods described in the preceding paragraph to the same data to facilitate comparisons. This approach is the same as that followed by Case, Clapp, Dubin, and Rodriguez (2004), but we take their best method and compare it with other methods that they did not consider.

Our base model is a simple OLS estimation with no controls for spatial effects. We then re-estimate the model with the average of the 10 nearest neighbors' residuals from the first-stage estimation included as a variable in the second stage. We also estimate a geostatistical model using the robust exponential technique (following BCH). Then we define submarkets using methods similar to those used by Case and by Thibodeau. We use census block groups as the building blocks for constructing submarkets by combining adjacent blocks with similar median house values until each resulting "transaction group" has at least 200 transactions. These transaction groups are similar to the "neighborhoods" defined by Thibodeau.[1] We also combine these transaction groups using cluster analysis to form "clusters" that are similar to the "districts" defined by Case. We then estimate a set of equations with dummy variables for transaction groups or clusters, using OLS, the two-stage nearest neighbors' residuals method, and the geostatistical method. We also estimate separate equations for each transaction group and cluster, using OLS and, where possible, the nearest neighbors and

geostatistical approaches. Given results previously reported by BCH, we do not estimate a lattice model; however, it may be worthwhile to compare some versions of SAR or CAR models with geostatistical approaches in future research.

Finally, we apply the trend surface method of FLM, using the Case-style clusters as submarkets. Because the FLM method can generate a large number of variables, we use a small number of property characteristics: lot size, floor area, and age. We include the squares and cubes of these characteristics as well as of the *x*- and *y*-coordinates and their squares and cubes. The variables include dummies for the submarkets. All possible pairs of variables are also interacted subject to the restriction that the sum of the powers is three or less. A stepwise procedure is used to eliminate collinearity.

The procedure for defining the Case-style clusters is a somewhat simplified version of Case's approach. Similar to Case, we consider two sets of variables: mean property characteristics and hedonic price estimates. We use Ward's hierarchical clustering method, which appears to be less sensitive to the initial seeds than the *k*-means method used by Case.

To insure that the results are not an artifact of the particular estimation sample chosen, each model is estimated using 100 random samples of the data. For each market-wide model (with or without submarket dummies), we estimate hedonic regressions using 100 samples each containing approximately 74% (9,600) of the total of 12,982 observations. For each of the 100 splits, ex-sample predictions are generated for the remaining 26% of the data. When separate submarket equations are estimated, we use 100 random samples consisting of 74% of the transactions for each Case-style cluster or 160 transactions for each Thibodeau-style group. Again, predictions are made for the ex-

sample transactions. We calculate error statistics and the proportions of predictions that are within 10% and 20% of the sale prices and report the medians for each model. These form the basis for our comparisons.

We specify our hedonic models using the variables available in the local property tax assessment database. We do not construct any spatial variables—such as measures of distance to the central business district or neighborhood characteristics—to allow spatial relationships to be captured to the extent possible by either submarket variables or geostatistical techniques. To implement the geostatistical approach, we used the S+Spatial Toolbox of the commercial software Splus.

Empirical Analysis

Data

The house price data are from the official records of the Property Valuation Administrator for Jefferson County (Louisville), Kentucky.[2] These records include sale prices, as well as various property characteristics, for all real estate transactions. We use data for all single family houses that sold in 1999. Some transactions were deleted due to missing data or because they could not be geocoded to census block groups. Also, transactions were deleted for properties whose sale prices, land areas, or floor sizes seemed unrealistically low or high. In the case of land area, properties greater than one-half acre in size were deleted because they are more likely to have been sold for redevelopment purposes. This results in a sample of 12,982 transactions.

Some variables were transformed before entering into the estimations. We use the natural logarithm of the dependent variable, house price.[3] Both age and age squared

are included in the model as the relation between house value and age is expected to follow a U-shaped curve. Also, the square of land area is included along with land area to reflect the decreasing marginal return to land.

Means for the sale prices, property characteristics, and quarterly time dummies are reported in Exhibit 1 (Panel A), along with statistics for the cluster analyses based on the transaction group hedonic characteristics and prices (Panels B and C, respectively). The census block groups, transaction groups, and clusters are mapped in Exhibit 2.

[Exhibits 1 and 2 here]

Cluster Analysis

The hierarchical cluster analysis procedure is applied to hedonic characteristics and prices, respectively, to produce varying numbers of clusters. The cluster analysis of hedonic prices includes estimates for land area squared, age of house squared, and the intercept term, in addition to the other hedonic characteristics. For both cluster analyses, we include *x*- and *y*-coordinates as variables to help impose some contiguity constraints. The optimal number of clusters is chosen such that adding another cluster results in only a minimal improvement in the percentage of variance explained. This rule of thumb suggests that eight clusters are optimal, whether the clusters are based on hedonic characteristics or prices. Also, the cluster definitions (as shown in Exhibit 2) turn out to be the same for both sets of data.

Hedonic Models

Exhibit 3 reports the results of the hedonic regressions performed for the first of our 100 estimation samples. We report regression results for the OLS model without submarket dummy variables, for the OLS model with cluster dummy variables, and for the OLS model with transaction group dummy variables (the estimated coefficients for the dummy variables are omitted from the table). The R^2 for the model with no submarket variables is 0.70 and increases when either set of submarket variables is added, to 0.75 and 0.78, respectively.

[Exhibit 3 here]

In the equation without submarket dummies, all variables are significant at the 1% level and all but one (age squared) have the expected signs. Property values are positively related to land area, floor area, the number of bathrooms, the degree to which the basement is finished, whether the house has air-conditioning and a fireplace, and the number of garages. The marginal utility of land decreases with lot size. The coefficients on the amenity variables for central air-conditioning and fireplaces appear quite high. For instance, air-conditioning adds 18% to the value of a house in the OLS model without submarket variables. There are at least two explanations for the magnitude of these coefficients. First, houses with central air-conditioning and a fireplace are likely to be of higher quality and hence these variables may be picking up other effects such as the quality of construction. Second, it is likely that these variables are capturing quality differences across submarkets which are not controlled for given that submarket variables are not included in the first model. As a matter of fact, the coefficients on the central air-

conditioning and fireplace variables decrease substantially when submarket variables are added to the model.

The age variable has the expected negative sign. However, the coefficient on age squared is significantly negative in the OLS model with no submarket variables and positive, but not significant, in the model including variables for clusters. This may be because there is no control for location (in the first model) or because the control is incomplete (in the second model). In the third model, the coefficient on age squared has the anticipated positive sign and is significant. The magnitudes of some of the coefficients for the cluster and transaction group dummy variables (not reported here) highlight the fact that there are substantial differences in prices across the Louisville housing market.

Applying the FLM method to the first estimation sample yields an R^2 of 0.76. Because this method as applied to our data leaves large numbers of variables in the models (for example, 41 for the first sample), we do not report an example of regression results.

House Price Prediction

Exhibit 4 reports comparative statistics for the various models: medians of the average absolute errors, average absolute relative errors, and percentages of predictions within 10% and 20% of the actual price. Here we focus on the percentage within 10% criterion. Our simple OLS model without submarkets or spatial adjustments yields a median accuracy of 36.5%, whereas the figure is 42.3% when the geostatistical method is used to control for spatial effects. These results are broadly similar to those for the most

comparable models reported in Thibodeau (2003): 35.9% and 46.9%, respectively. For the model without submarket dummies, there is not much difference in accuracy between the nearest neighbors' residuals and geostatistical methods for controlling spatial effects.

[Exhibit 4 here]

Adding dummy variables for clusters leads to a significant improvement in accuracy performance from 36.5% to 40.3%. Spatial adjustment by means of the geostatistical method improves the results even further (to 45.5%), whereas the nearest neighbor two-stage method improves accuracy only marginally. Using multiple equations for the clusters leads to better accuracy (42.4%) than the single equation with dummy variables (40.3%). This is consistent with results reported in Thibodeau (2003). However, using the two-stage nearest neighbor residual adjustment lowers performance to 38.4%. The latter result pertains to the method that is closest to Case's. For our data, this method does not appear to be particularly effective, although this could have something to do with differences in the ways nearest neighbors' residuals are treated. Also, the geostatistical method does not work here due to lack of convergence when fitting the parametric exponential variogram to the empirical one. The spherical variogram suffers from the same problem, which is probably related to the nature of the data at hand.

Considering a larger number of submarkets leads to better results. For instance, the accuracy is 44.0% when dummy variables for the transaction groups are included in the OLS model. The overall best result is for the geostatistical model with transaction group dummies (47.4%). This is consistent with the results of Thibodeau (2003) and

BCH; however, Thibodeau does not report results for a single equation with neighborhood dummy variables, so we are unable to fully compare these results with his.

Multiple equations for the 60 transaction groups yield slightly worse results (43.4%) than a single equation with submarket dummy variables (44.0%). For the reasons highlighted above, we are unable to implement the geostatistical model for the 60 groups and hence cannot replicate Thibodeau's best performing model.[4] However, his geostatistical multiple equation results are only slightly better (less than 2 percentage points) than his OLS results, suggesting that we would obtain a similar improvement. This implies that the multiple equation approach would not yield our best results even if we could implement the geostatistical model.

The FLM approach yields the second worst results (37.5%). This is only 1 percentage point better than an OLS model with neither submarket dummies nor any spatial adjustments. Given our data, the FLM method is not particularly effective in taking into account spatial dependencies for mass appraisal purposes. It performs well for FLM's Tucson data, although we do not know whether alternative methods would produce even more accurate results.

Overall, the prediction accuracy results reported in this paper tend to be lower than in other studies. While our best result is 47.4%, Thibodeau's best result is 63.6% and FLM's best result is 65.0%. In contrast, BCH's best result is 49.3%. These differences may be due to variations across cities in unmeasured characteristics related to property condition that do not exhibit a clear spatial pattern and hence are not controlled for using spatial techniques. We speculate that these variations are related to the age of the housing stock because there is likely to be greater variability in condition the older

the stock. We note that the average ages of houses in Thibodeau's Dallas sample, FLM's Tucson sample, and Case's Fairfax sample are 33, 22, and 8 years, respectively. In comparison, the average ages in BCH's Auckland sample and our Louisville sample are 47 and 38 years, respectively.

Conclusions

Automated valuation models are used in many countries for tax appraisal and mortgage underwriting purposes. Automated valuation is typically implemented using hedonic regression models. An important issue in such models is controlling for spatial dependence. Various authors have analyzed alternative methods for doing so. The results vary substantially from one study to another, which could be due to either methods or data. To control for the impacts of data, we apply several methods that perform well in the literature to a single data set.

Taking into account submarkets is important in achieving more accurate price predictions. More specifically, increasing the number of submarkets improves the results, confirming a conclusion in Thibodeau (2003) and Goodman and Thibodeau (2007). Obviously, the level of disaggregation is constrained by the number of transactions available for model estimation purposes. We are not able to reach any firm conclusion about the relative merits of single equation versus multiple equation methods of controlling for submarket effects.

Our results show the benefits of modeling spatial dependence in the error term. Geostatistical methods seem more useful than the two-stage nearest neighbors' residual procedure. An OLS estimation that takes into account disaggregated submarkets is

slightly more effective than a geostatistical model with no consideration of submarkets. However, our best result is for a geostatistical model with dummy variables for relatively disaggregated submarkets.

Endnotes

[1] Thibodeau (2003) uses two years of data, whereas we use only one year. Hence, all else being equal, he should have smaller and possibly more homogenous areas.

[2] Jefferson County merged with the City of Louisville in 2003.

[3] The OLS predictions are calculated as $\exp(\widehat{\ln Y})$, although the correct transformation would be $\exp(\widehat{\ln Y} + 0.5\hat{\sigma}^2)$. Because we are unable to implement equivalent transformations for predictions based on geostatistical methods, we do not add $0.5\hat{\sigma}^2$ before taking the antilogs of the OLS predictions. Given the large sample size, this has only a trivial impact on the results.

[4] Given the small size of the areas, we did not attempt to use the two-stage nearest neighbor residual adjustment.

References

Basu, A., and T.G. Thibodeau. Analysis of Spatial Autocorrelation in House Prices. *Journal of Real Estate Finance and Economics*, 1998, 17:1, 61–85.

Bourassa, S.C., E. Cantoni, and M. Hoesli. Spatial Dependence, Housing Submarkets, and House Price Prediction. *Journal of Real Estate Finance and Economics*, 2007, 35:2, 143–60.

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

Can, A., and I. Megbolugbe. Spatial Dependence and House Price Index Construction. *Journal of Real Estate Finance and Economics*, 1997, 14:1/2, 203–22.

Case, B., J. Clapp, R. Dubin, and M. Rodriguez. Modeling Spatial and Temporal House Price Patterns: A Comparison of Four Models. *Journal of Real Estate Finance and Economics*, 2004, 29:2, 167–91.

Clapp, J.M. A Semiparametric Method for Valuing Residential Locations: Application to Automated Valuation. *Journal of Real Estate Finance and Economics*, 2003, 27:3, 303–20.

Colwell, P.F. A Primer on Piecewise Parabolic Multiple Regression Analysis via Estimations of Chicago CBD Land Prices. *Journal of Real Estate Finance and Economics*, 1998, 17:1, 87–97.

Cressie, N., and D.M. Hawkins. Robust Estimation of the Variogram, I. *Journal of the International Association for Mathematical Geology*, 1980, 12:2, 115–25.

Dubin, R.A. Estimation of Regression Coefficients in the Presence of Spatially Autocorrelated Error Terms. *Review of Economics and Statistics*, 1988, 70:3, 466–74.

Dubin, R.A. Predicting House Prices Using Multiple Listings Data. *Journal of Real Estate Finance and Economics*, 1998, 17:1, 35–59.

Fik, T.J., D.C. Ling, and G.F. Mulligan. Modeling Spatial Variation in Housing Prices: A Variable Interaction Approach. *Real Estate Economics*, 2003, 31:4, 623–46.

Fletcher, M., P. Gallimore, and J. Mangan. The Modelling of Housing Submarkets. *Journal of Property Investment and Finance*, 2000, 18:4, 473–87.

Geniaux, G., and C. Napoléone. Semi-Parametric Tools for Spatial Hedonic Models: An Introduction to Mixed Geographically Weighted Regression and Geoadditive Models. In A. Baranzini, J. Ramirez, C. Schaerer, and P. Thalmann, editors, *Hedonic Methods in Housing Markets: Pricing Environmental Amenities and Segregation*, New York: Springer, 2008, 101–27.

Goodman, A.C., and T.G. Thibodeau. Housing Market Segmentation. *Journal of Housing Economics*, 1998, 7:2, 121–43.

Goodman, A.C., and T.G. Thibodeau. Housing Market Segmentation and Hedonic Prediction Accuracy. *Journal of Housing Economics*, 2003, 12:3, 181–201.

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

Matheron, G. *Traité de Géostatistique Appliquée, Tome I.* Mémoires du Bureau de Recherches Géologiques et Minières, No. 14. Paris: Editions Technip, 1962.

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, 1088–111.

Pace, R.K., and R. Barry. Quick Computation of Regressions with a Spatially Autoregressive Dependent Variable. *Geographical Analysis*, 1997, 29:3, 232–47.

Pace, R.K., R. Barry, J.M. Clapp, and M. Rodriguez. Spatiotemporal Autoregressive Models of Neighborhood Effects. *Journal of Real Estate Finance and Economics*, 1998, 17:1, 15–33.

Pace, R.K., and O.W. Gilley. Using the Spatial Configuration of the Data to Improve Estimation. *Journal of Real Estate Finance and Economics*, 1997, 14:3, 333–40.

Pavlov, A.D. Space-Varying Regression Coefficients: A Semi-Parametric Approach Applied to Real Estate Markets. *Real Estate Economics*, 2000, 28:2, 249–83.

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

Acknowledgments

We thank Martye Scobee for assistance with the transactions data and Elizabeth Riesser for preparing the transactions groups and map. Helpful comments from two anonymous reviewers are greatly appreciated.

Exhibit 1 Sample Means

Variable	Mean	Standard Deviation	Minimum	Maximum
Panel A: Characteristics from Tax Assessment Data (n = 12,982)				
Sale price ($)	115,478	66,478	25,000	420,000
Land area (square feet)	9,277	3,783	2,178	21,780
Floor area (square feet)	1,468	618	432	5,180
1.5 or 2 bathrooms	0.38	—	0.00	1.00
2.5 or more bathrooms	0.22	—	0.00	1.00
Age of house	37.7	27.8	0.0	99.0
Partial basement	0.12	—	0.00	1.00
Full basement	0.51	—	0.00	1.00
Central air conditioning	0.77	—	0.00	1.00
Fireplace	0.52	—	0.00	1.00
Number of garages	1.11	0.87	0.00	2.00
2nd quarter	0.28	—	0.00	1.00
3rd quarter	0.28	—	0.00	1.00
4th quarter	0.23	—	0.00	1.00
Panel B: Characteristics of Transaction Groups (n = 60)				
Land area (square feet)	9,260	2,259	4,390	14,324
Floor area (square feet)	1,455	411	922	2,578
1.5 or 2 bathrooms	0.39	0.16	0.12	0.74
2.5 or more bathrooms	0.21	0.24	0.00	0.87
Age of house	37.9	22.0	4.2	84.2
Partial basement	0.12	0.08	0.00	0.29
Full basement	0.51	0.19	0.09	0.93
Central air conditioning	0.77	0.18	0.29	1.00
Fireplace	0.51	0.29	0.09	0.98
Number of garages	1.10	0.38	0.44	1.93
2nd quarter	0.28	0.03	0.22	0.34
3rd quarter	0.28	0.03	0.20	0.33
4th quarter	0.23	0.04	0.17	0.34
x-coordinate (feet)	1,226,600	30,140	1,171,365	1,283,376
y-coordinate (feet)	258,644	20,175	220,561	300,444

Exhibit 1 Sample Means (continued)

Variable	Mean	Standard Deviation	Minimum	Maximum
Panel C: Hedonic Prices for Transaction Groups (n = 60)				
Intercept	10.8	0.4	9.8	11.7
Land area	3.04×10^{-5}	5.50×10^{-5}	-7.32×10^{-5}	16.3×10^{-5}
Land area squared	-1.04×10^{-9}	2.41×10^{-9}	-8.67×10^{-9}	3.68×10^{-9}
Floor area	2.17×10^{-4}	0.93×10^{-4}	-1.30×10^{-4}	3.93×10^{-4}
1.5 or 2 bathrooms	0.09	0.12	-0.20	0.61
2.5 or more bathrooms	0.16	0.15	-0.34	0.67
Age of house	0.00	0.01	-0.02	0.03
Age of house squared	0.08×10^{-4}	1.41×10^{-4}	-3.71×10^{-4}	5.49×10^{-4}
Partial basement	0.12	0.09	-0.11	0.37
Full basement	0.14	0.06	-0.06	0.26
Central air conditioning	0.06	0.09	-0.33	0.20
Fireplace	0.07	0.05	-0.02	0.24
Number of garages	0.04	0.03	-0.09	0.13
2nd quarter	0.04	0.06	-0.14	0.17
3rd quarter	0.06	0.05	-0.01	0.21
4th quarter	0.06	0.06	-0.10	0.19

Notes: Default categories (not shown) are 1 or fewer bathrooms, no basement, and the 1st quarter. For the estimation results summarized in Panel C, the dependent variable is the natural logarithm of sale price.

Exhibit 2 Census Block Groups, Transaction Groups, and Clusters

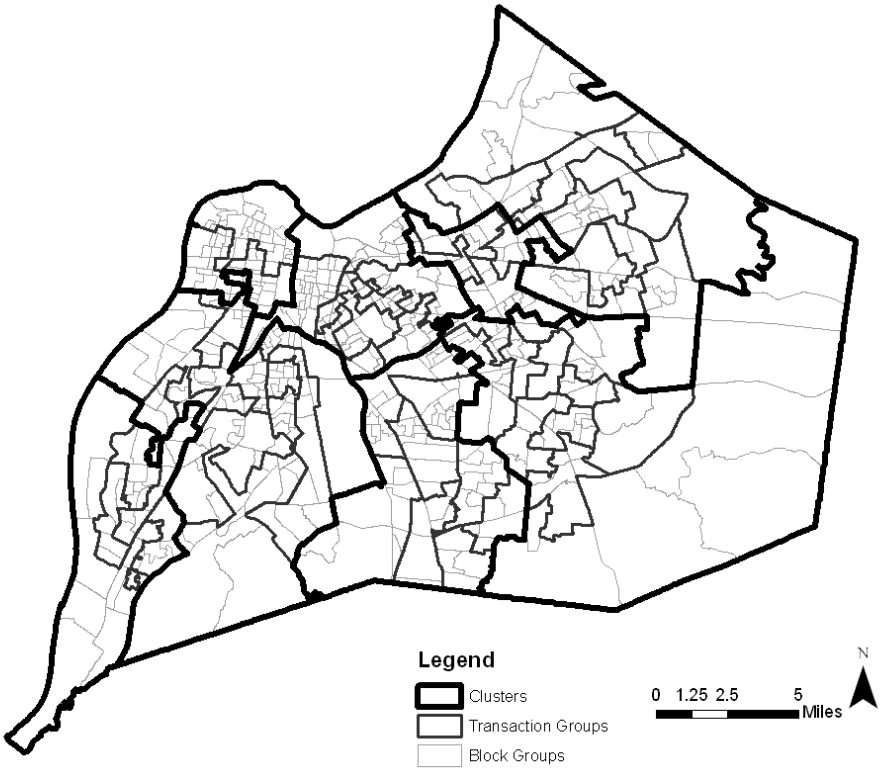

Exhibit 3 Sample OLS Estimations

Variables	No Submarket Variables	With 8 Clusters	With 60 Transaction Groups
Intercept	10.4**	10.7**	10.9**
Land area (square feet)	3.84×10^{-5}**	3.68×10^{-5}**	2.92×10^{-5}**
Land area squared	-1.20×10^{-9}**	-1.10×10^{-9}**	-0.90×10^{-9}**
Floor area (square feet)	3.16×10^{-4}**	2.67×10^{-4}**	2.36×10^{-4}**
Bathrooms (default is 1 or less)			
1.5 or 2	0.113**	0.083**	0.066**
2.5 or more	0.191**	0.147**	0.122**
Age of house	-0.003**	-0.005**	-0.007**
Age of house squared	-1.47×10^{-5}**	0.05×10^{-5}	2.24×10^{-5}**
Basement (default is no basement)			
Partial	0.129**	0.133**	0.123**
Full	0.156**	0.160**	0.137**
Central air conditioning	0.178**	0.125**	0.103**
Fireplace	0.146**	0.107**	0.084**
Number of garages	0.040**	0.041**	0.039**
Quarterly dummies (default is 1st quarter)			
2nd quarter	0.039**	0.043**	0.036**
3rd quarter	0.058**	0.061**	0.058**
4th quarter	0.056**	0.067**	0.063**
R^2	*0.697*	*0.753*	*0.784*

Notes: The dependent variable is the natural logarithm of sale price. These results are for the first random estimation sample (n = 9,600). The symbols * and ** denote significance at the 5% and 1% levels, respectively. The estimates for the submarket (cluster and transaction group) dummies are not reported.

Exhibit 4 Prediction Accuracy Statistics

Statistic	OLS	OLS with 10 Nearest Neighbor Residuals Variable	Geostatistical (Robust Exponential)
Median of average absolute error ($)			
Without submarkets	22,027	18,932	18,356
With 8 clusters			
Single equation	19,837	19,592	17,133
Multiple equations	19,081	20,621	—
With 60 transaction groups			
Single equation	18,145	17,600	16,678
Multiple equations	19,225	—	—
Trend surface	19,926	—	—
Median of average absolute relative error (%)			
Without submarkets	23.7	19.9	19.4
With 8 clusters			
Single equation	20.9	20.6	18.5
Multiple equations	20.1	21.6	—
With 60 transaction groups			
Single equation	19.1	18.6	18.0
Multiple equations	19.8	—	—
Trend surface	21.3	—	—
Median percentage of predictions within 10%			
Without submarkets	36.5	41.9	42.3
With 8 clusters			
Single equation	40.3	40.7	45.5
Multiple equations	42.4	38.4	—
With 60 transaction groups			
Single equation	44.0	45.4	47.4
Multiple equations	43.4	—	—
Trend surface	37.5	—	—

Exhibit 4 Prediction Accuracy Statistics (continued)

Statistic	OLS	OLS with 10 Nearest Neighbor Residuals Variable	Geostatistical (Robust Exponential)
Median percentage of predictions within 20%			
Without submarkets	63.4	70.6	71.1
With 8 clusters			
Single equation	69.0	69.3	74.0
Multiple equations	70.5	66.3	—
With 60 transaction groups			
Single equation	72.9	74.0	75.2
Multiple equations	70.8	—	—
Trend surface	66.4	—	—