Journal of Real Estate Finance and Economics, 29:2, 167–191, 2004

Modeling Spatial and Temporal House Price Patterns: A Comparison of Four Models

BRADFORD CASE
Board of Governors of the Federal Reserve System, Washington, DC 20551
E-mail: bradford.case@frb.gov

JOHN CLAPP
Centre for Real Estate, University of Connecticut, Storrs, CN 06269-1041, U.S.A.
E-mail: John.Clapp@business.uconn.edu

ROBIN DUBIN
Department of Economics, Weatherhead School of Management, Case Western Reserve University, Cleveland, OH 44106
E-mail: rad4@case.edu

MAURICIO RODRIGUEZ
TCU Box 298530, Fort Worth, TX 76129
E-mail: M.Rodriguez@tcu.edu

Abstract

This research reports results from a competition on modeling spatial and temporal components of house prices. A large, well-documented database was prepared and made available to anyone wishing to join the competition. To prevent data snooping, out-of-sample observations were withheld; they were deposited with one individual who did not enter the competition, but had the responsibility of calculating out-of-sample statistics for results submitted by the others. The competition turned into a cooperative effort, resulting in enhancements to previous methods including: a localized version of Dubin's kriging model, a kriging version of Clapp's local regression model, and a local application of Case's earlier work on dividing a geographic housing market into districts. The results indicate the importance of nearest neighbor transactions for out-of-sample predictions: spatial trend analysis and census tract variables do not perform nearly as well as neighboring residuals.

Key Words: kriging, out-of-sample prediction, data snooping, local polynomial regression, smoothing regression, semiparametric models, cluster analysis, nearest neighbors, hedonic models

1. Introduction

It is well known that real estate values are highly dependent on their locational and market characteristics. Over the past decade, technological advances have led to a significant increase in the amount of available data with spatial attributes that allow investigators to more readily account for these important factors in their pricing models.[1] Indeed, over the past several years, researchers have made great strides accounting for spatial and temporal factors in real estate pricing models.[2] However, it is not at all obvious how best to account for space and time in empirical investigations. This research contributes to the existing

literature by comparing alternative approaches to modeling spatial and temporal house prices.

A competition was initiated to compare the results produced by alternative modeling approaches. A large, well-documented database was prepared and made available to anyone wishing to join the competition. Out-of-sample observations were withheld from entrants in the competition and made available to one individual who did not enter the competition, but had the responsibility of calculating out-of-sample statistics for results submitted by others. The competition turned into a cooperative effort, with participants sharing models and programs.

All results produced by the models employed by the entrants of this competition were superior to the results that were obtained from using ordinary least squares (OLS). The competition was the catalyst for enhancements of previously employed methods such as: a localized version of Dubin's kriging model, a kriging version of Clapp's local regression model, and a local application of Case's method for dividing a geographic housing market into districts.

The next section describes the data made available to the researchers who entered this competition. Section 3 describes OLS models and the models employed by each of the three entrants in this competition. Section 4 reports the results obtained from OLS and from the models submitted by each entrant. The final section contains concluding remarks.

2. The data made available to the entrants in this competition

Mo Rodriguez and Kelley Pace developed the data for this study with a database provided by the Fairfax County Assessors' Office.[3] They started with more than 60,000 transactions of single-family properties that sold from the first quarter in 1967 through the second quarter of 1991. They filtered and enhanced the data in order to accommodate a range of spatial models.[4] After filtering, all transactions had data within acceptable ranges for sales price, property characteristics, and date of sale. The spatial distribution of the transactions is displayed in Figure 1. The apparent spatial clustering of houses by size sets the stage for the spatial modeling in Section 3.

The data were subsequently divided into in-sample observations and out-of-sample observations. The in-sample data covered the period from (including) the first quarter of 1967 to the second quarter of 1991. The out-of-sample data (for which entrants did not have sales price) covered the time period from the first quarter of 1972 through the fourth quarter of 1991. The purpose of this was to allow a burn-in period (1967Q1 to 1971Q4) and an out-of-sample forecasting period for time series models. Models with no time series properties, such as the simple OLS model described in the next section, would use in-sample data starting in 1972Q1 and the subset of 5,000 out-of-sample transactions from 1972Q1 to 1991Q2.[5] To summarize, Table 1 shows the sample sizes available for each type of model.

The four sub-samples did not differ greatly with respect to mean numbers of rooms, bedrooms, and bathrooms. Likewise, the ranges of characteristics, and their standard deviations, were similar across the four sub-samples.[6]

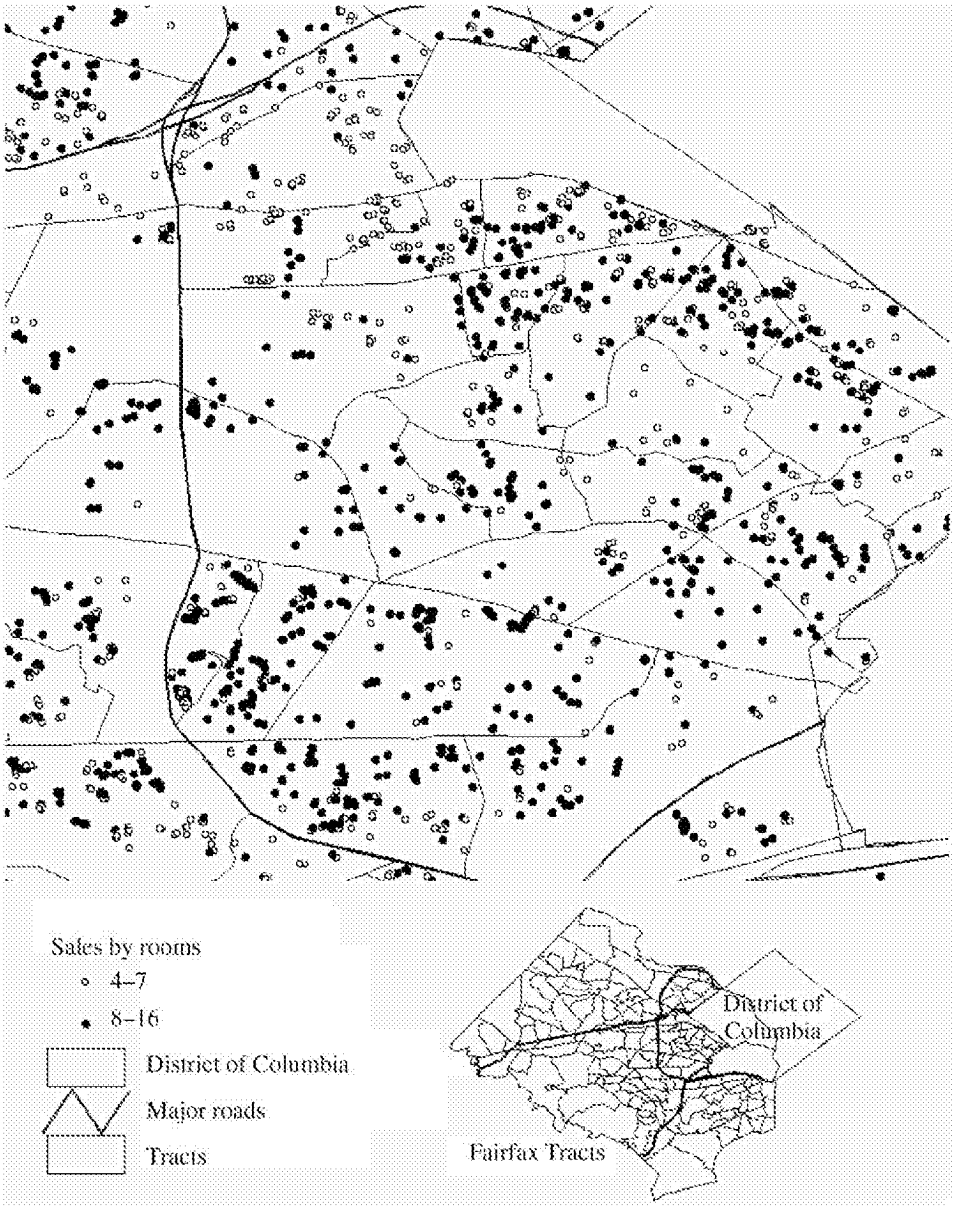

Figure 1. Inner Fairfax County houses by number of rooms. The larger houses (dark dots indicate eight or more rooms) tend to cluster. There are no sales available for Fairfax City.

Table 1. Number of observations: Number of transactions passing the described filters.

	Number of Observations	
	In-Sample	Out-of-Sample
Cross sectional model	49,511	5,000
Time series model*	51,190	7,177

Note: *The in-sample TS data included 1,679 observations during the burn-in period; out-of-sample TS data included 2,177 sales during the last two quarters of 1991.

A geographic information system (GIS) was used to match street addresses of the transactions with census tract boundaries; a 1990 census tract number and a set of relevant census information were assigned to each transaction. Table 2 provides a detailed listing of the variables that were available to the entrants.

To facilitate spatial analysis, the entrants were also provided with observation numbers identifying the nearest neighbor for the first to the 15th nearest neighbors to each of the sample observations and the corresponding nearest neighbor distances for both the sample and out-of-sample sets.[7] The number 15 may be viewed as too small, but this was given as part of the data provided for the competition.

Table 2. Listing of variables available to entrants.

Obs.#	Integers from 1 to 51,190 for in-sample (1–7,177 for out-sample)
Price	Sales price in dollars (withheld from out-sample to be predicted)
Date	Date of sale (decimal date format)
PriorPrice	Previous sales price in dollars (when available, otherwise 0)
PriorDate	Previous date of sale (decimal date format when available, otherwise 0)
LandArea	Land area in square feet
Rooms	Number of rooms in the house
Beds	Number of bedrooms in the house
Baths	Number of bathrooms in the house
HalfBaths	Number of half bathrooms in the house
Fire	Number of fireplaces in the house
YearBuilt	The year the house was built
Lon	Longitude of house in decimal degrees
Lat	Latitude of house in decimal degrees
censusID	Number giving the state, county, and census tract ID
CenLon	Longitude of census tract centroid
CenLat	Latitude of census tract centroid
Pop	Population
HH	Number of households
Black	Black population
Hispanic	Hispanic population
BA_edu	Number with BA degrees
Grad_edu	Number with graduate education
MedHHinc	Median household income

3. The models employed in this competition

3.1. Ordinary least squared models

The purpose of this section is to describe the baseline OLS regression models developed by Clapp. Since the overall purpose is out-of-sampling prediction, relatively little emphasis will be given to expected signs of variables. The OLS model is essentially a cross-sectional model (i.e., no lagged variables are included). Thus, the burn-in period (1967–1971) was excluded from the sample.

To account for spatial factors in the OLS model, Clapp used latitude, longitude, each of these two variables squared and latitude times longitude.[8] These variables are part of the geography literature known as "trend surface analysis." As described below, some of our OLS regressions included dummies for each census tract with more than 10 observations out-of-sample. In essence, this allows for variation in neighborhood land values, with neighborhood boundaries described by census tract boundaries.[9]

Table 3 presents the basic OLS model with annual time dummy variables, with the 1991 dummy (YY91) omitted. Clapp controlled for heterogeneous property characteristics with: total rooms, bedrooms, bathrooms, fireplaces, age of the structure and age squared. Sales prices were logged, a standard transformation in the hedonic modeling literature.

A good fit is obtained; quarterly time dummy variables would not substantially improve on this. The price index shows plausible annual rates of increase, with a leveling out (and small decline) during the recession and early recovery years of 1989–1991. From 1972 to the middle of 1991, standardized prices in Fairfax increased by 163 percent, continuously compounded.

All of the trend surface variables are statistically significant, indicating some spatial trend over Fairfax County. The introduction of cubed terms was not feasible as it caused perfect co-linearity (inability to invert the $X'X$ matrix).

Coefficients on building characteristics have plausible signs and magnitudes. For example, an additional room increases house price by 5.3 percent. The larger coefficients on bathrooms and fireplaces probably reflect high implicit prices on these variables. Over this time period, the number of bathrooms and fireplaces included in new construction has risen.[10] This suggests that developers responded to high relative implicit prices for these characteristics. These large coefficients may also reflect the omission of interior square footage from the data: Bathrooms and fireplaces are positively correlated with square footage.

Building age has a plausible quadratic form. It has been documented that age is generally associated with depreciation and obsolescence; however, beyond some point, only those houses with the best locations and the highest construction quality survive (see Clapp and Giaccotto, 1998; Goodman and Thibodeau, 1995). Neighborhood (Census tract) characteristics show plausible signs and magnitudes. The negative sign on percent of households with a graduate education (pctGRAD) suggests some model misspecification.[11]

Census tract dummies were introduced to try to improve the goodness-of-fit statistics. In fact, the coefficients on these dummies constitute a form of trend surface analysis. They estimate a pattern of neighborhood land values.

Table 3. Basic OLS model with trend surface analysis ($N = 49{,}511$).

Total SS:	15,071.763	Degrees of freedom:	49,472
R-squared:	0.853	Rbar-squared:	0.853
Residual SS:	2,219.236	Std error of est:	0.212
F(38,49472):	7,539.819	Probability of *F*:	0.000

Variable	Estimate	Std Error	*t*-value	Prob > |*t*|
Constant	− 1,968.99	603.63	− 3.26	0.001
YY72	− 1.6349	0.0100	− 163.12	0
YY73	− 1.4575	0.0100	− 145.54	0
YY74	− 1.3162	0.0106	− 124.35	0
YY75	− 1.2738	0.0093	− 136.34	0
YY76	− 1.2346	0.0080	− 154.85	0
YY77	− 1.1558	0.0074	− 156.82	0
YY78	− 1.0393	0.0070	− 148.59	0
YY79	− 0.9135	0.0069	− 132.22	0
YY80	− 0.7951	0.0074	− 108.05	0
YY81	− 0.7038	0.0076	− 92.13	0
YY82	− 0.6852	0.0079	− 86.77	0
YY83	− 0.6576	0.0064	− 102.53	0
YY84	− 0.5846	0.0061	− 95.20	0
YY85	− 0.4980	0.0058	− 85.56	0
YY86	− 0.3949	0.0056	− 70.75	0
YY87	− 0.2443	0.0056	− 44.00	0
YY88	− 0.0746	0.0055	− 13.64	0
YY89	0.0312	0.0055	5.63	0
YY90	0.0288	0.0057	5.06	0
Rooms	0.0531	0.0011	46.77	0
Beds	− 0.0219	0.0021	− 10.53	0
Baths	0.1275	0.0020	62.62	0
HalfBaths	0.0997	0.0022	46.12	0
Fire	0.1341	0.0017	81.16	0
Lon	126.6179	15.1319	8.37	0
Lat	354.5654	16.5485	21.43	0
Lonsq	1.2375	0.1246	9.94	0
Latsq	− 2.9000	0.2198	− 13.19	0
LatLon	1.6738	0.2404	6.96	0
Age	− 0.0137	0.0002	− 57.15	0
Agesq	0.0002	0.0000	33.85	0
InLandArea	0.1642	0.0013	127.06	0
InMedHHinc	0.0931	0.0084	11.13	0
InHH	− 0.0210	0.0020	− 10.28	0
PctBlack	0.0007	0.0002	3.37	0.001
PctHispanic	0.0025	0.0002	10.36	0
PctBA_edu	0.0022	0.0003	6.32	0
PctGRAD_edu	− 0.0037	0.0004	− 9.89	0

Notes: ln(Sales Price) = dependent variable. The 1991 time dummy variable (YY91) is omitted. Introduction of latitude and longitude cubed was impossible due to collinearity.

Implementation of this idea was made difficult by the fact that some census tracts have very few houses or house sales. Since the purpose is forecasting out-of-sample, we started with the holdout sample and formed dummies for all tracts with more than 10 transactions. This produced 99 dummy variables. In-sample dummy variables were calculated for these same 99 tracts. More than 86 percent of the tract dummy variables had significant coefficients at the 5 percent level or better. Thus, a complex trend surface pattern does characterize Fairfax County.

When tract dummies are used, the significance of other spatial variables is reduced, and signs may change (see Table 4 versus Table 3). For example, the coefficients on longitude and latitude change dramatically, but their effect would need to be added to the effect of tract dummies in order to evaluate the spatial pattern of house prices. Also, most of the census tract demographic variables have signs that are difficult to interpret in Table 4, and all show reduced economic and statistical significance relative to the results reported in Table 3.

Because of the difficulty in interpreting results, we did not pursue the tract dummy regressions. However, the improved R^2s (Table 4 compared to Table 3) motivates the more elaborate spatial models discussed below. Clearly, there is a complex spatial pattern that should be modeled.

3.2. Clapp's local regression model

The local regression model (LRM) uses large databases to produce neighborhood price indices. LRM is shown to provide estimates of a surface of constant-quality-house values as a function of time. Thus, a constant-quality-house price index over time can be estimated at any point in space.

The LRM starts with a standard cross-sectional hedonic model, except that a flexible function is introduced for the value of space and time. The log of sales price ($\ln SP_i$) is regressed on a vector of housing characteristics ($\mathbf{X}_i$) and a 3D vector of latitude, longitude, and time ($\mathbf{Z}_i$), for a given house, i, sold at time t:

$$Y_{it} = \ln SP_{it} = \mathbf{X}_i B_1 + f(\mathbf{Z}_i) + \gamma_{it}, \tag{1}$$

where ε_{it} is an i.i.d. noise term that is assumed to be normally distributed for the purposes of OLS hypothesis testing.

Robinson (1988) and Stock (1989) propose a two-step algorithm for estimating a model of the form of equation (1): OLS estimates the parameters on $\mathbf{X}$ and nonparametric smoothing methods estimate $f(\mathbf{Z}_i)$. Robinson achieves orthogonality between the parametric and nonparametric parts by using conditional variables.[12] Specifically, he uses a nonparametric procedure to estimate:

$$E(Y|\mathbf{Z}) = \text{smooth}(Y|\mathbf{Z}) \tag{2}$$

$$E(X_k|\mathbf{Z}) = \text{smooth}(X_k|\mathbf{Z}) \tag{3}$$

Table 4. OLS model with trend surface analysis and tract dummies ($N = 49{,}511$).

Total SS:	15,071.763	Degrees of freedom:	49,373
R-squared:	0.872	Rbar-squared:	0.872
Residual SS:	1,929.923	Std error of est:	0.198
F(137,49373):	2,454.059	Probability of *F*:	0.000

Variable	Estimate	Std Error	*t*-value	Prob > |*t*|
Constant	− 11,807.91	1578.21	− 7.48	0.00
YY72	− 1.688	0.009	− 178.46	0.00
YY73	− 1.509	0.009	− 159.95	0.00
YY74	− 1.359	0.010	− 136.62	0.00
YY75	− 1.317	0.009	− 149.64	0.00
YY76	− 1.267	0.008	− 168.84	0.00
YY77	− 1.192	0.007	− 171.34	0.00
YY78	− 1.076	0.007	− 162.90	0.00
YY79	− 0.942	0.007	− 144.81	0.00
YY80	− 0.825	0.007	− 119.21	0.00
YY81	− 0.728	0.007	− 101.37	0.00
YY82	− 0.707	0.007	− 95.37	0.00
YY83	− 0.678	0.006	− 112.57	0.00
YY84	− 0.605	0.006	− 104.93	0.00
YY85	− 0.512	0.005	− 93.83	0.00
YY86	− 0.402	0.005	− 76.89	0.00
YY87	− 0.250	0.005	− 48.13	0.00
YY88	− 0.076	0.005	− 14.81	0.00
YY89	0.033	0.005	6.31	0.00
YY90	0.029	0.005	5.38	0.00
Rooms	0.047	0.001	43.31	0.00
Beds	− 0.010	0.002	− 5.07	0.00
Baths	0.112	0.002	57.93	0.00
HalfBaths	0.088	0.002	42.59	0.00
Fire	0.120	0.002	75.80	0.00
Lon	− 121.877	37.717	− 3.23	0.00
Lat	366.703	37.840	9.69	0.00
Lonsq	− 0.879	0.291	− 3.02	0.00
Latsq	− 5.067	0.449	− 11.29	0.00
LatLon	− 0.349	0.599	− 0.58	0.56
Age	− 0.017	0.000	− 69.24	0.00
Agesq	0.00024	0.00001	39.56	0.00
InLandArea	0.162	0.001	124.01	0.00
InMedHHinc	0.073	0.019	3.84	0.00
InHH	− 0.007	0.004	− 1.73	0.08
pctBlack	− 0.003	0.000	− 7.77	0.00
PctHispanic	0.007	0.001	9.79	0.00
PctBA_edu	− 0.003	0.001	− 5.11	0.00
PctGRAD_edu	− 0.003	0.001	− 4.26	0.00
Tract dummy1	− 0.149	0.016	− 9.61	0.00
Tract dummy2	0.084	0.016	5.21	0.00

Notes: ln(Sales Price) = dependent variable. Ninety-nine tract dummies were in the regression: 86 percent were statistically significant at the 5 percent level or lower. Only the first two tract coefficients are presented.

where $k = 1, \ldots, K$ and K is the number of explanatory variables, not including a constant term; the subscript i has been suppressed on all variables. A suitable nonparametric procedure (the "smooth") will be discussed below.

To consistently estimate the linear portion of equation (1), compute the vectors $Y^* = Y - \mathrm{E}(Y|\mathbf{Z})$ and $X_k^* = X_k - E(X_k|\mathbf{Z})$. The parameters are estimated by an OLS regression of Y^* on $\mathbf{X}^*$, where $\mathbf{X}^*$ has K columns and a constant is included in the regression. Finally, partial residuals are calculated using equation (4):

$$\hat{\eta} = Y - X\hat{\beta}_1. \tag{4}$$

Note that Y and X are original variables: Y^* and $\mathbf{X}^*$ are used only to get consistent estimators for β_1. Once $\hat{\eta}$ has been calculated, $f(\mathbf{Z})$ can be estimated:[13]

$$f(\mathbf{Z}) \equiv \mathrm{smooth}((\hat{\eta})|Z). \tag{5}$$

Smoothing (equations (2), (3) and (5)) was done with local polynomial regression (LPR) as opposed to the older Nadaraya–Watson (NW) method used by McMillen (1994). LPR has superior theoretical properties, especially at boundary points (see Wand and Jones, 1995; Fan and Gijbels, 1996, for accessible treatments of LPR and its relationship to the NW method).

Local polynomial regression now takes the form of equation (6)[14]:

$$\eta_{it} = \gamma_0 + (\mathbf{Z}_i - \mathbf{z}_0)\gamma_1 + (\mathbf{Z}_i - \mathbf{z}_0)^2\gamma_2 + \cdots + (\mathbf{Z}_i - \mathbf{z}_0)^p\gamma_p + \xi_{it}, \tag{6}$$

where the $\gamma_j (j = 1, \ldots, p)$ have three elements, the coefficients on latitude, longitude and time; γ_0 is a scalar. Note that, when $\mathbf{Z}_i$ equal $\mathbf{z}_0$ then equation (6) reduces to γ_0, the parameter of interest. Thus, LPR fits a surface to the η-values conditional on the values of $\mathbf{z}$ given by $\mathbf{z}_0$: for example, $\mathbf{z}$ is a cube of equally spaced latitude, longitude and time points that span the data; the level of η is estimated conditional on each knot of the grid.

Kernel weights, $K_h(\cdot)$, are multiplied by each observation when estimating equation (6) by weighted least squares. The kernel function gives less weight to transactions more distant in space; hence the term "local" polynomial (of degree p) regression. $K_h(\cdot)$, is defined such that greater distances—larger values for $\| \mathbf{Z}_i - \mathbf{z}_0 \|$—imply lower values for $K_h(\cdot)$, a product kernel[15]:

$$K_h(\mathbf{W}_i - \mathbf{w}_0) = K_{hw1}(W_{i1} - w_{01})K_{hw2}(W_{i2} - w_{02})K_{ht}(t - t_0). \tag{7}$$

Following common practice, normal pdfs will be used as the kernel functions, $K_{hw1}(\cdot), K_{hw2}(\cdot)$ and $K_{ht}(\cdot)$. In general, the kernels can be thought of as probability densities with standard deviations ("bandwidths") equal to hw_1, hw_2 and h_t.

With any nonparametric smoothing method, selection of bandwidth (analogous to the standard deviation) is of critical importance. The LRM uses a cross-validation approach to find an optimal balance between bias (bandwidth too large) and variance

(bandwidth too small). More detail on the LRM is given in Clapp et al. (2002) and in Clapp (2002).[16]

As discussed in the next section, the first round of results (Round 1) showed that Clapp's LRM method did not perform as well as Case's and Dubin's models in terms of out-of-sample statistics. Analysis showed that this is due to their inclusion of nearest neighbors. Case included the estimated disturbance term (i.e., the residuals) from the first five nearest-neighboring sales directly in the second stage of his OLS model. Similarly, Dubin's procedure incorporates the effect of nearby houses in both the estimation of the regression coefficients and in the prediction.

The revised LRM (employed in Round 2) was estimated in steps as follows: (1) The LRM was estimated from equations (1)–(7); (2), residuals were calculated for the first 15 nearest neighbors; (3) these residuals were entered as explanatory variables into a second stage estimate of the LRM.[17] Coefficients were calculated for each of the first five nearest neighbors, then for the median of the 6th to 10th and for the median of the 11th to 15th. Finally, this procedure was repeated for the subset of nearest neighbors that sold within 3 years of the subject property sale.

3.3. Dubin's maximum likelihood estimation of the hedonic regression

The model that Dubin estimated is $Y = X\beta + u,\ u \sim N(0, \sigma^2 K)$. K is assumed to have a constant main diagonal and be positive definite. This is a standard hedonic formulation, except that the errors are allowed to be correlated.

The covariance matrix of the error terms is of dimension $(N \times N)$, where N is the number of observations. Even given the assumptions, K has $N(N-1)/2$ unknown elements. In order to make the estimation feasible, a correlation function, or correlogram, must be assumed. Here Dubin uses a negative exponential correlogram, shown in equation (6).

$$\begin{aligned} K_{ij} &= b_1 \exp\left(-\frac{d_{ij}}{b_2}\right) && \text{if } i \neq j, \\ &= 1 && \text{if } i = j, \end{aligned} \tag{8}$$

where d_{ij} is the (straight-line) distance separating houses i and j; b_1 and b_2 are parameters to be estimated. Dubin used maximum likelihood estimation to simultaneously estimate β, σ^2, b_1 and b_2. The log-likelihood function is shown in equation (7).

$$L = -\frac{N}{2}\ln \sigma^2 - \frac{1}{2}\ln|K| - \frac{1}{2\sigma^2}(Y - X\beta)'K^{-1}(Y - X\beta). \tag{9}$$

Note that the ML estimate of β is

$$\tilde{\beta} = (X'K^{-1}X)^{-1}X'K^{-1}Y \tag{10}$$

and thus the spatial autocorrelation affects the estimated regression coefficients.

One advantage of this procedure is that it addresses the problem of autocorrelation, which should improve the efficiency of the estimators. The second is that the estimated correlation function can be used to provide an estimate of the error term, which can be added to the standard $\tilde{X}\beta$, to improve the prediction. The process of predicting the error terms is known as kriging (see Ripley, 1981; Dubin, 1992, for discussions). Briefly, in kriging the predicted error is a weighted sum of the residuals from the estimation data. Thus, $\hat{e} = \lambda' e$, where λ are the weights and e are the regression residuals. The optimal set of weights can be shown to be $\lambda = \tilde{K}^{-1}\tilde{k}$, where $\tilde{K}$ is the correlation matrix for the estimation data and $\tilde{k}$ is the correlation vector between the point to be predicted and the estimation data. The tildes denote that K and k are obtained from the estimated correlogram.

Rather than estimating one equation for the entire study area, Dubin chose to estimate a new equation for each prediction point. Thus, for each of the 5,000 observations to be predicted, Dubin selected, from among the set of houses that sold up to 3 years prior to the prediction house, the 200–300 closest observations from the estimation data. These observations were then used to estimate the model described above. This approach has two major advantages. First, the estimation dataset is quite large, containing 51,190 observations. Evaluating the likelihood function involves inverting K. Without using sparse matrix techniques (and the negative exponential correlation matrix is not sparse), it is computationally intensive to invert such a large matrix. Second, it is not reasonable to expect that the implicit prices and the spatial relationships are constant over the entire study area. Table 5 shows summary statistics for some of the estimated parameters.

Examination of Table 5 shows that the spatial autocorrelation parameters, b_1 and b_2, vary considerably across the study area. These parameters can be used to determine the geographic extent of the spatial autocorrelation. Using the means reported in the table in conjunction with equation (6), it can be seen that the spatial autocorrelation falls to 0.1 at an average of 0.33 miles from the prediction house. There is, however, considerable variation in this distance, known as the range of the correlogram, over the study area. RADIUS shows how far the most distant estimation house is from the prediction house. This distance is determined by the requirement that the estimation sample consist of at least 200 houses. It is larger in less dense areas and on the periphery. The range of the

Table 5. Summary statistics for selected round one parameters.

	Mean	Std Dev.	Min.	Max.
b1	0.62	0.21	0	0.99
b2	0.18	0.19	0.01	2.31
RADIUS	1.77	1.16	0.41	10.0
VDIFF	60.71	40.58	0.0	291.09

Note: The parameters for 5,000 predictions are summarized here. VDIFF is absolute value of twice the difference between the value of the likelihood function with and without the spatial autocorrelation term. Radius is the distance in miles from the prediction house to the most distant house used for estimation.

correlogram is generally much smaller than the radius, which has an average value of 1.8 miles.

VDIFF is absolute value of twice the difference between the value of the likelihood function with and without the spatial autocorrelation term. That is, it is the likelihood ratio test for the significance of spatial autocorrelation. This statistic is significant when its value exceeds six. The last row of Table 5 shows that almost all of the LR statistics are highly significant. Only 4 percent are insignificant.

For the first round, Dubin used the following independent variables: age, lot size, number of rooms, bedrooms, bathrooms, fireplaces and half baths; no interaction terms or census variables were used. The dependent variable is the log of sales price. Dubin controlled for temporal effects (within her three-year sales window), by including the date of sale, to account for inflation. Finally, Dubin accounted for spatial effects by including a linear geographic trend (i.e., the longitude and latitude) in an attempt to achieve stationarity.

For the second round, Dubin made two changes. First, the bottom end of the acceptable sample size was increased to 225. The second change was to include the census variable Median Income. Including census variables posed a problem in that some of the estimation areas were contained entirely in one census tract. For these observations, median income was not included (there were only 23 of these). These changes did not improve the predictive ability of the model. Dubin also experimented with other functional forms for the correlogram; these models performed considerably worse than the negative exponential (these results are not shown but are available from Dubin).

3.4. Case's hedonic price model with homogeneous districts and nearest-neighbor residuals

Case employed a hedonic model that includes homogeneous within-county districts created on the basis of cluster analysis. This modeling technique allows for differences in the parameters of the hedonic price function across local neighborhoods. This approach is broadly similar to the local smoothing aspect of Clapp's model and the 5,000 local models estimated by Dubin.

The construction of homogeneous within-county districts is a local application of the clustering technique described in Case (2000) with an application to international commercial real estate markets, and in Case (2001) with an application to U.S. regional house price appreciation rates. In brief, first a hedonic price model is estimated separately for each of the 123 census tracts in Fairfax County that had at least 160 in-sample property transactions. The hedonic price models employed available property characteristics data (age in quadratic, lot size, and numbers of rooms, bedrooms, bathrooms half baths, and fireplaces) as well as annual dummies for year of transaction; no census tract data were used as the models were estimated at the census tract level. The estimated parameters of these hedonic price models (including quarterly time series) are then used to cluster census tracts into groups with similar parameter sets.

The k-means cluster algorithm starts with an initial allocation of candidates to clusters, so the final clustering may be sensitive to the particular initial allocation. To remove the

influence of this effect, the clustering is performed 10,000 times with randomly chosen initial cluster allocations. This yields a cluster association frequency, which is the number of times (out of 10,000) that each pair of census tracts ended up in the same final cluster. The cluster association frequency, then, reflects the degree of similarity between parameter estimate sets for each of the 7,503 pairs of census tracts.

The cluster association frequency is then regressed on other variables representing similarities among census tracts: the distance between each pair of census tracts, and a separate cluster association frequency reflecting the degree of similarity in tract-level demographic information (median income, household size, percents black and Hispanic, and percents with college and graduate degrees):

$$F_{ij}^{H} = \alpha + \beta F_{ij}^{D} + \gamma D_{ij} + \varepsilon_{it}, \tag{11}$$

where F_{ij}^{H} is the number of times that census tracts i and j clustered together based on the parameters of the hedonic price models; F_{ij}^{D} is the number of times that the same census tracts clustered together based on demographic information, D_{ij} is the distance between the centroids of census tracts i and j, ε_{ij} is the disturbance term, and α, β, and γ are parameters to be estimated. Finally, the predicted cluster association frequency from this regression is then used to form final clusters.

The clustering technique also results in a test statistic that identifies the optimal number of clusters, which is 12 in this application. The districts, shown in Figure 2, comprise contiguous groups of census tracts with two minor exceptions.

In addition to the use of within-county districts, Case's first-round model included residuals (in percent deviations) from the five nearest neighbors as regressors in a second-stage estimation. This technique makes use of implicit information on the value of unobserved property and neighborhood attributes. (Note that nearest neighbors need not be in the same district as the subject property.)

The neighboring disturbances are likely correlated with X variables, so the parameter estimates produced by this model are likely to be inconsistent. Inconsistent estimators are acceptable here because we focus on ex-sample predicted values. If the nearest neighbor residuals improve out-of-sample prediction, as they do here, then one can search for consistently estimated models that have similar predictive accuracy.

After determining the within-county districts and estimating the residuals from a first-stage model, the hedonic price models are then re-estimated separately for each district:

$$\ln(P_{lt}) = \alpha^{k} + \beta_1^{k}\mathrm{AGE}_{lt} + \beta_2^{k}\mathrm{AGE}_{lt}^{2} + \sum_{m}\beta_m^{k}X_{lmt} + \sum_{n=1}^{T}\gamma_n^{k}D_{ln} + \sum_{q=1}^{5}\delta_q^{k}\hat{\varepsilon}_{ltq} + \varepsilon_{lt}, \tag{12}$$

where $\ln(P_l)$ is the natural logarithm of the transaction price of property l at time t; AGE_{lt} is the age of property l at time t; D_{ln} is a dummy variable that is one if property l sold at time n (i.e., $n=t$), otherwise zero; X_{lmt} denotes one of m property characteristics for property l at time t; $\hat{\varepsilon}_{ltq}$ is the estimated qth nearest neighbor residual from the first-stage regression; and α^{k}, β_m^{k}, γ_q^{k}, and δ_q^{k} are parameters to be estimated for district k.

The first-round results indicated that the use of nearest-neighbor residuals is much more

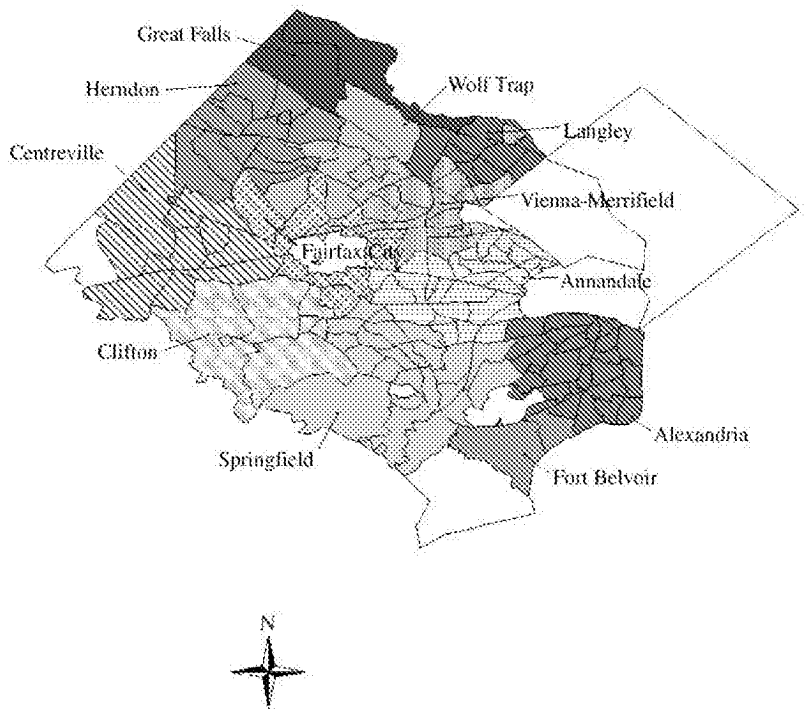

Figure 2. Districts produced from the clustering algorithms.

important than the use of homogenous within-county districts, so for Round 2 Case expanded the number of nearest-neighbors to nine. In addition, he employed the correction for transformation bias suggested by Miller (1983), $\hat{P}_{it} = e^{X\hat{\beta}+\hat{\sigma}^2/2}$. As Goldberger (1968) and Thibodeau (1989) point out, taking the antilog of the predicted value from a regression in which the dependent variable is in logs yields an asymptotically unbiased estimate for the median of the price distribution, but an downward-biased estimate for the mean. While this correction does not eliminate this problem, the remaining bias is probably negligible.[18] As the next section shows, the effect of the inclusion of additional nearest-neighbor residuals and the use of this correction for transformation bias in Round 2 was to reduce the mean-based measures of the prediction error, while increasing slightly the median-based measures of the prediction error.

The use of homogeneous within-county districts can be interpreted as analogous to one aspect of Dubin's method, the use of data on 200–300 nearest neighboring properties to estimate coefficients separately for each subject property. Both procedures are designed to allow parameters to vary locally while preserving an adequate number of observations for model estimation. The 12 districts defined above are substantially larger than the geographic area included for a given property under Dubin's procedure. For example, the Annandale district includes more than 10,000 property transactions in 30 census tracts.

Case justified the use of nearest-neighbor residuals as a means of recovering data on the

market value of unobserved property and neighborhood attributes. In this sense the use of nearest-neighbor residuals is analogous to another aspect of Dubin's method, the use of kriging to incorporate spatial correlations into predicted values. On the other hand, the residuals may reflect local variation in the "true" parameters of the hedonic price model that remain even after parameters are allowed to vary at the district level—an interpretation that makes the use of nearest-neighbor residuals more analogous to the use of nearest neighbors in model estimation.

4. Results

4.1. Results from Round 1

Table 6 provides the results from Round 1 of the competition. The OLS mean out-of-sample prediction error is 4,465 and the MSE is 2,528,392,028. The standard deviation of the prediction error is 50,283.

Clapp's LRM results exhibit better out-of-sample predictions than the OLS-based results. The mean prediction error is lower than its OLS counterpart (2,968 versus 4,465) and the standard deviation of the prediction error for the LRM model is also lower (45,882 versus 50,283). However, the LRM results did not predict as well as the models employed by the other entrants.

Case's initial model employing homogeneous within-County districts and nearest-neighbor residuals produced the lowest standard deviation of the prediction error among Round 1 models (36,657). However, Dubin's model with a different equation for each prediction point resulted in the lowest mean prediction error among the results submitted for Round 1 (619).

Table 6. Results from Round 1: statistics for predicted sales under alternative methodologies.

	OLS	Clapp	Case	Dubin
Common subset of data (Obs = 5,000)				
Mean prediction error	4,465	2,968	2,522	619
Median prediction error	(1,471)	(1,133)	243	(589)
Standard deviation of prediction error	50,283	45,882	36,657	42,173
Root mean squared prediction error	50,476	45,973	36,740	42,173
Root median squared prediction error	16,704	15,454	10,939	11,356
Mean absolute value prediction error	29,034	26,624	19,928	21,593
Median absolute value prediction error	16,704	15,454	10,939	11,356
Mean absolute value percent error	16.62%	15.44%	11.90%	12.54%
Median absolute value percent error	12.27%	11.31%	8.07%	8.34%

Note: Since none of the models had explicit time series properties, the results for the 7,177 observations (including 2,177 for 1991Q3 and 1991Q4) are not reported herein.

4.2. *Results from Round 2*

Cooperation among the entrants led to improvements in the predictions submitted for Round 2. Results from the OLS and LRM models were improved by including residual data for nearest neighbors in the models (i.e., the model was estimated, residuals calculated, and then the model was re-estimated with these residuals included). Of course, the coefficients from these models are not consistent and not efficient, but that does not matter for our purpose of out-of-sample forecasting.

There was a strong statistical significance of all the nearest neighbor coefficients including coefficients on the 11th to 15th neighbors (see Appendix Table A.5). Moreover, these coefficients display a plausible declining pattern, not dissimilar to the negative exponential pattern required by Dubin's model.[19] This indicates that researchers may need to include data associated with more than 15 nearest neighbors in order to fully exploit this relationship. Dubin's model allows for many nearby observations; it is likely that this characteristic enables her model to perform very well out-of-sample.

Table 7 shows out-of-sample results for OLS and for Clapp's LRM models based on including the nearest-neighbor residual (actual for 1–5, median for 6th to 10th and 11th to 15th residuals) and results based on including residuals for the nearest observations within three years (i.e., similar to Dubin's model).

The second round OLS results were an improvement over the first round. Table 7 shows the mean prediction error went down to 3,049 (using the median residual) and 2,743 (using residuals from prior three years) from the previous 4,465 reported for Round 1 in Table 6.

The results reported for Case in Table 7 also made use of nearest-neighbor data, expanding the number of nearest neighbors from five to nine. In addition, in Round 2 Case employed the correction for transformation bias suggested by Miller (1983), which improved mean-based measures of the prediction error at the cost of increased prediction error according to median-based measures. In general, however, results from Case's models were better predictors—even using median-based measures—as a result of including

Table 7. Results from Round 2: Statistics for predicted sales under alternative methodologies.

	OLS-Median	OLS-3yr	Clapp-Median	Clapp-3yr	Case	Dubin
Common subset of data (Obs = 5,000)						
Mean prediction error	3,049	2,743	3,477	3,344	366	744
Median prediction error	152	(272)	893	841	(1,094)	(550)
Standard deviation of prediction error	38,655	39,247	38,370	38,699	35,859	42,412
Root mean squared prediction error	38,771	39,339	38,524	38,839	35,857	42,414
Root median squared prediction error	11,520	11,768	11,892	11,644	11,079	11,500
Mean absolute value prediction error	21,037	21,150	21,085	21,091	19,484	21,695
Median absolute value prediction error	11,520	11,768	11,892	11,644	11,079	11,500
Mean absolute value percent error	12.48%	12.56%	12.39%	12.40%	11.84%	12.58%
Median absolute value percent error	8.53%	8.43%	8.66%	8.59%	8.04%	8.34%

Note: Since none of the models had explicit time series properties, the results for the 7,177 observations (including 2,177 for 1991Q3 and 1991Q4) are not reported herein.

additional nearest-neighbor residuals. The mean prediction error for Case's models went down from 2,522 to 366 and MSE declined by nearly 5 percent (see Table 7 versus Table 6).

For Round 2, Dubin's increase to the bottom end of the acceptable sample size and inclusion of the census variable Median Income did not improve upon the results submitted for Round 1. The mean prediction error increased from 619 to 744, while the mean absolute value percent error increased from 12.54 percent to 12.58 percent. The median absolute value percent error did not change from 8.34 percent. Increasing the sample size did not work because the new observations were outside the range of the autocorrelation function.[20] Median Income may not have varied enough in the small geographical areas used for each local regression.

Case estimated several alternative versions of his model (results not shown, but available) to determine the relative importance of nearest-neighbor residuals and homogeneous within-county districts. The results suggested that the use of nearest-neighbor residuals results in a much greater improvement in the predictive power of the model than does the use of homogeneous within-county districts. For example, the standard deviation of the prediction error is about 19 percent larger for a model without nearest-neighbor residuals than for the model with them, whereas the difference between district-level models and one county-level model is less than one percent. This suggests that the greater payoff to future research is likely to be in the improvement of methods for incorporating geographic correlation rather than in the improvement of methods for defining discrete local areas.

Case also estimated versions of his model in which transaction price and date information for earlier transactions of the same properties were included in the data set used to estimate the hedonic price model in each within-county district:

$$\ln(P_{lt}) = \alpha^k + \beta_1^k \mathrm{AGE}_{lt} + \beta_2^k \mathrm{AGE}_{lt}^2 + \sum_m \beta_m^k X_{lm} + \sum_{n=1}^{T} \gamma_n^k D_{ln} + \sum_{q=1}^{5} \delta_q^k \hat{\varepsilon}_{lt} + \varphi^k I(\mathrm{prior})_{lt} + \varepsilon_{lt}, \tag{13}$$

where $I(\mathrm{prior})$ indicates whether observation lt is for a prior transaction ($I=1$) or for the most recent transaction ($I=0$).

Interestingly, the empirical results (not shown) indicate that this actually degraded the predictive power of the model, even with the inclusion of a dummy variable indicating whether each observation represented the most recent or a prior transaction. For example, the standard deviation of the prediction error increased by six percent when prior transactions were included.[21] These results are important because many property tax data sets maintain the most recent property characteristics data rather than the property characteristics that were valid at the time of the most recent transaction (much less at the time of any earlier transactions).

One implication of this is that hedonic price models estimated from recently-transacting properties may be better than models estimated from properties transacting earlier when the data are maintained in this way, because the property characteristics data

are more likely to represent the actual property characteristics at the time of the sale. This tends to validate Dubin's restriction of the data included in her models to the nearest properties transacting within three years of the subject property. It also suggests that nearest-neighbor residuals included as regressors should be restricted temporally in a similar way, but empirically this restriction does not seem to have improved the estimates (results not shown).

5. Conclusions

All the first round results from this competition were better than the results produced by OLS, even though OLS included trend surface analysis and a number of controls for neighborhood characteristics. We conclude that a well-specified OLS model ignores important spatial structure in the data. The main question from Round 1: What common factor in the Case and Dubin models enabled them to strongly outperform OLS and the Clapp model?

These two Round 1 models both include some form of nearest-neighbor variables. Case used the five nearest neighbors and Dubin's approach incorporates the effect of over 200 nearby houses. Round 2 included the 15 nearest-neighbor residuals as explanatory variables in the OLS and Clapp models. These variables enabled these models to perform almost as well as the other submitted models. We conclude that the information on the value of unobserved property and neighborhood attributes contained in nearest-neighbor residuals should be included in the model if the purpose is to predict out-of-sample.

Dubin's introduction of a Census variable and increase of the minimal sample size for Round 2 did not improve the results submitted for Round 1. As expected, Case's Round 2 model that employed Miller's (1983) correction for transformation bias reduced the mean prediction error, and had a negative impact on the median prediction error. Future research with the Dubin and Clapp models, or any model using $\exp(\hat{\log}\, x)$, should correct for bias.

The models might be compared on dimensions other than out-of-sample prediction. The OLS and Case models use less econometric theory and more generally accessible software. Case's cluster analysis adds explanatory power and reveals interesting spatial patterns at the expense of the substantial computational effort required to estimate optimal clusters. But none of his models provide consistent and efficient estimators of model parameters.[22] The same criticism applies to Clapp's second round models, and both rounds of his models require substantial computational effort with software (e.g., Xplore or Stata) that perform nonparametric smoothing.[23]

Dubin's kriging model is well established in the spatial literature; software is available from Dubin (Gauss programs); Anselin's SpaceStat estimates lattice models that use nearest neighbors.[24] Dubin's model produces consistent and efficient estimators given widely accepted assumptions, but it is also computationally intensive. The innovation in this paper is to estimate her kriging model locally, with one model estimated for each out-of-sample point. The local estimates of Dubin's parameters in Table 5 might be

useful for further spatial analysis. For example, the estimated parameters b_1 and b_2 indicate the spatial extent of neighboring influence. It would be interesting to plot these on a map and test their spatial patterns relative to patterns revealed by the Clapp and Case methods.

Local regression models have received considerable attention in this paper. Future research should examine proximity in characteristic space as an alternative to proximity in physical space. Methods used by real estate appraisers motivate this extension of our research.[25] An appraiser would probably reject a sale of a 4,000 square foot house as a useful comparable when valuing a 1,500 square foot house, even if the sale were quite close in space and time.

Appendix

Table A.1. Descriptive statistics, in-sample Fairfax sales 49,511 sales from 1972Q1–1991Q2.

Variable	Mean	Std Dev.	Variance	Minimum	Maximum
lnSP	11.9121	0.5517	0.3044	9.9523	13.8105
YY72	0.0118	0.1081	0.0117	0	1
YY73	0.0118	0.1081	0.0117	0	1
YY74	0.0103	0.1008	0.0102	0	1
YY75	0.0142	0.1182	0.0140	0	1
YY76	0.0227	0.1489	0.0222	0	1
YY77	0.0293	0.1687	0.0284	0	1
YY78	0.0353	0.1847	0.0341	0	1
YY79	0.0366	0.1878	0.0353	0	1
YY80	0.0289	0.1676	0.0281	0	1
YY81	0.0255	0.1576	0.0248	0	1
YY82	0.0228	0.1494	0.0223	0	1
YY83	0.0493	0.2165	0.0469	0	1
YY84	0.0605	0.2384	0.0568	0	1
YY85	0.0798	0.271	0.0734	0	1
YY86	0.1040	0.3052	0.0932	0	1
YY87	0.1068	0.3088	0.0954	0	1
YY88	0.1171	0.3215	0.1034	0	1
YY89	0.1044	0.3058	0.0935	0	1
YY90	0.0875	0.2825	0.0798	0	1
YY91	0.0414	0.1992	0.0397	0	1
Rooms	7.8093	1.5663	2.4534	3	18
Beds	3.5935	0.7416	0.5500	1	8
Baths	2.1444	0.6282	0.3946	1	6
HalfBaths	0.8902	0.5395	0.2911	0	4
Fire	1.035	0.7108	0.5053	0	4
Lon	−77.2324	0.0936	0.0088	−77.4913	−77.0434
Lat	38.8493	0.0730	0.0053	38.6325	39.0436
Lonsq	5,964.8549	14.4573	209.0141	5,935.6804	6,004.896
Latsq	1,509.2705	5.6688	32.1355	1,492.4732	1,524.404
LatLon	−3,000.4258	8.066	65.0604	−3,019.511	−2,980.03

Table A.1. (continued)

Variable	Mean	Std Dev.	Variance	Minimum	Maximum
Age	7.5982	9.5684	91.5533	0	85
Agesq	149.2846	347.8768	12,1018.2979	0	7,225
InLandArea	8.9596	1.1191	1.2524	6.4568	12.2891
InMedHHinc	10.9965	0.2628	0.0691	10.1994	11.7259
InHH	7.4597	0.5114	0.2615	1.0986	8.6002
PctBlack	6.1468	6.4306	41.3532	0	96.2917
PctHispanic	6.6719	6.5676	43.1333	0	47.115
PctBA_edu	19.0731	4.3562	18.9764	1.0928	66.129
PctGRAD_edu	14.1765	5.9631	35.5581	0.8957	30.1959

Table A.2. Subset of descriptive statistics, in-sample Fairfax sales 51,190 sales from 1967Q1–1991Q2.

Variable	Mean	Std Dev.	Variance	Minimum	Maximum
Rooms	7.8188	1.5615	2.4383	3	18
Beds	3.6028	0.7425	0.5513	1	8
Baths	2.1465	0.6275	0.3938	1	6

Table A.3. Descriptive statistics, out-of-sample Fairfax sales 5,000 sales from 1972Q1–1991Q2.

Variable LnSP	Mean Na	Std Dev. Na	Variance Na	Minimum Na	Maximum Na
YY72	0.0116	0.1071	0.0115	0	1
YY73	0.0152	0.1224	0.0150	0	1
YY74	0.0124	0.1107	0.0122	0	1
YY75	0.0174	0.1308	0.0171	0	1
YY76	0.020	0.1400	0.0196	0	1
YY77	0.0252	0.1567	0.0246	0	1
YY78	0.0376	0.1902	0.0362	0	1
YY79	0.0392	0.1941	0.0377	0	1
YY80	0.0292	0.1684	0.0284	0	1
YY81	0.0304	0.1717	0.0295	0	1

Table A.3. (*continued*)

Variable LnSP	Mean Na	Std Dev. Na	Variance Na	Minimum Na	Maximum Na
YY82	0.0194	0.1379	0.0190	0	1
YY83	0.0502	0.2184	0.0477	0	1
YY84	0.0620	0.2412	0.0582	0	1
YY85	0.0800	0.2713	0.0736	0	1
YY86	0.1034	0.3045	0.0927	0	1
YY87	0.1108	0.3139	0.0985	0	1
YY88	0.1114	0.3147	0.0990	0	1
YY89	0.1012	0.3016	0.0910	0	1
YY90	0.0834	0.2765	0.0765	0	1
YY91	0.0400	0.1960	0.0384	0	1
Rooms	7.8194	1.5562	2.4217	4	15
Beds	3.5912	0.7390	0.5462	1	7
Baths	2.1404	0.6092	0.3712	1	6
HalfBaths	0.8764	0.5407	0.2924	0	4
Fire	1.0174	0.6800	0.4624	0	4
Lon	−77.2318	0.0930	0.0087	−77.4756	−77.0451
Lat	38.8493	0.0727	0.0053	38.6358	39.0435
Lonsq	5,964.767	14.3701	206.4986	5,935.943	6,002.467
Latsq	1,509.273	5.6498	31.9202	1,492.728	1,524.395
LatLon	−3,000.41	8.0305	64.4886	−3,018.29	−2,980.22
Age	7.7574	9.6010	92.1798	0	76
Agesq	152.3386	354.9114	125,962.11	0	5776
InLandArea	8.9639	1.1013	1.2129	6.4583	12.0977
InMedHHinc	11.003	0.2588	0.0670	10.1994	11.7259
InHH	7.4515	0.5205	0.2710	2.9957	8.6002
Pctblack	6.0541	6.0936	37.1320	0	41.951
Pcthisp	6.4235	6.1631	37.9832	0	47.115
PctBA	19.2328	4.5350	20.5666	1.9573	66.129
PctGRAD	14.2889	5.9210	35.0587	2.1570	30.1959

Table A.4. Subset of descriptive statistics, out-of-sample Fairfax sales 7,177 sales from 1972Q1–1991Q4.

Variable	Mean	Std Dev.	Variance	Minimum	Maximum
Rooms	7.81	1.57	2.46	4	16
Beds	3.58	0.74	0.54	1	7
Baths	2.15	0.62	0.38	1	6

Table A.5. OLS with lagged neighboring residuals.

	Basic OLS		OLS-spatial lags NN		LRM residuals on spatially lagged		
Variable	Estimate	*t*-value	Estimate	*t*-value	Variable	Estimate	*t*-value
CONSTANT	−1,968.99	−3.3	−947.18	−1.9			
YY72	−1.6349	−163.1	−1.68	−204.8			
YY73	−1.4575	−145.5	−1.50	−181.1			
YY74	−1.3162	−124.4	−1.35	−154.6			
YY75	−1.2738	−136.3	−1.31	−170.7			
YY76	−1.2346	−154.9	−1.25	−190.1			
YY77	−1.1558	−156.8	−1.18	−194.4			
YY78	−1.0393	−148.6	−1.07	−186.7			
YY79	−0.9135	−132.2	−0.93	−163.9			
YY80	−0.7951	−108.0	−0.81	−134.9			
YY81	−0.7038	−92.1	−0.72	−115.1			
YY82	−0.6852	−86.8	−0.70	−108.3			
YY83	−0.6576	−102.5	−0.67	−127.3			
YY84	−0.5846	−95.2	−0.60	−119.1			
YY85	−0.4980	−85.6	−0.51	−107.3			
YY86	−0.3949	−70.8	−0.40	−86.3			
YY87	−0.2443	−44.0	−0.24	−52.7			
YY88	−0.0746	−13.6	−0.07	−15.4			
YY89	0.0312	5.6	0.04	7.8			
YY90	0.0288	5.1	0.03	6.9			
Rooms	0.0531	46.8	0.04	39.7			
Beds	−0.0219	−10.5	−0.0021	−1.2			
Baths	0.1275	62.6	0.0970	57.6			
HalfBaths	0.0997	46.1	0.0789	44.3			
Fire	0.1341	81.2	0.1039	75.8			
Lon	126.6179	8.4	179.0715	14.3			
Lat	354.5654	21.4	406.3981	29.9			
Lonsq	1.2375	9.9	1.6057	15.6			
Latsq	−2.9000	−13.2	−3.4514	−19.1			
LatLon	1.6738	7.0	1.7898	9.1			
Age	−0.0137	−57.2	−0.0150	−76.5			
Agesq	0.0002	33.9	0.0002	36.6			
InLandArea	0.1642	127.1	0.1747	164.2			
InMedHHinc	0.0931	11.1	0.0953	13.9			
InHH	−0.0210	−10.3	−0.0232	−13.8			
PctBlack	0.0007	3.4	0.0005	2.7			
PctHispanic	0.0025	10.4	0.0024	11.9			
PctBA_edu	0.0022	6.3	0.0020	7.1			
PctGRAD_edu	−0.0037	−9.9	−0.0042	−13.6			
errNN1			0.1741	38.4	LRMerrNN1	0.1613	34.7
errNN2			0.1272	28.5	LRMerrNN2	0.1133	24.7
errNN3			0.1065	23.4	LRMerrNN3	0.0910	19.5
errNN4			0.0937	20.9	LRMerrNN4	0.0846	18.4
errNN5			0.0733	16.5	LRMerrNN5	0.0636	13.9
errNN6–10			0.2733	33.8	LRMerrNN6–10	0.2278	27.1
errNN11–15			0.1812	20.3	LRMerrNN11–15	0.1508	16.1
R-squared	0.853		0.9050		0.269		
Num Obs	49,511		46,979		46,979		

Notes: Model 1 with ln(Sales Price) as the dependent variable was estimated and residuals saved. For each observation in the next (OLS) model, the 15 nearest neighbor residuals were included as regressors: The median of the 6th to 10th and 11th to 15th residuals was used. A similar process was followed for the residuals from the LRM model. The last column of the table indicates the part of the LRM residual that could be explained by nearest neighbor LRM residuals.

Acknowledgments

Professor Rodriguez received research support from the Charles Tandy American Enterprise Center at Texas Christian University. Professor Clapp received support from the Center for Real Estate, University of Connecticut. We thank Matt Klena, Kelley Pace, Tom Thibodeau and an anonymous referee for comments on earlier drafts.

Notes

1. For example, spatial coordinates for census data are widely available as are data with zip + 4 coordinates. Geographic information systems (GIS) and global positioning systems (GPS) make it feasible to add spatial coordinates to existing databases. In addition, some spatial statistical software is now also readily available. For a description of some GIS and spatial statistical software, see http://www.ai-geostats.org/.
2. For example, see Pfeifer and Bodily (1990), Goetzmann and Spiegel (1997), Raman (2002) and Pace et al. (1998a).
3. Fairfax County, Virginia is a suburb of Washington, DC. The data source excludes the City of Fairfax, which is independent of Fairfax County. This is a relatively affluent county. The median family income in Fairfax County was $72,000 in 1997, compared to $41,434 in the state of Virginia and $37,005 for the United States as a whole. Interstate highway access is excellent in the county, and a subway system links about one-third of the county to Washington DC.
4. All possible observations were filtered using the following rules: House price > $20,000 and < $1,000,000, total rooms < 20, number of fireplaces < five, year sold ≥ 1967, and land area ≤ 5 acres. Additional filters to eliminate potential data errors were employed as follows: Non-negative age, non-negative year built, positive number of baths, total number of rooms greater than number of baths, positive number of bedrooms, number of bedrooms greater than or equal to number of baths, total number of rooms greater than number of bedrooms, date sold greater than prior date sold, total rooms greater than one, land value greater than zero, and improvement value greater than zero. Common filters were used so that differences in results would be driven by alternative modeling approaches as opposed to different filters that could have been employed.
5. None of the current authors estimated a time series model. This part of the competition, together with a repeat sales model (repeats are a subset of these data), is completely open to any new contestants.
6. Sample statistics for the in-sample and out-sample data sets are provided in appendix tables at the end of this paper. Since none of the models submitted by entrants had explicit time series properties, the results for the 7,177 observations (including 2,177 for 1991Q3 and 1991Q4) are not reported herein.
7. The insight that nearest neighbor observations are important in spatial modeling is largely drawn from Professor Kelley Pace's previous work with spatial and temporal models (see Pace et al. (1998b) for an overview of this field). We thank Professor Pace for identifying the 15 nearest neighbors and providing us with that information for this research. We did not study the potential impact of using a higher number of nearest neighbors.
8. We wanted the OLS model to perform as well as possible in terms of out-of-sample mean squared error.
9. A referee pointed out that it would be highly desirable to account for municipal boundaries and for school district boundaries. We agree, but those data were not available for this study.
10. The out-of-sample data indicates the average number of fireplaces increased by about 20 percent for homes built in the 1980s versus those built in the 1970s (from 0.91 to 1.1 fireplaces on average). The number of half baths increased by about 11 percent (from 0.91 to 1.01 half baths on average) and the number of full baths increased by about 4.5 percent (from 2.09 to 2.19 full baths on average) over the same time period.
11. The economic significance of this variable is small: A 10 percent increase in the percent of the neighborhood with graduate education decreases house prices by only 4 percent. Likewise, all of the census tract variables have small economic effects as measured by elasticity (for number of households and median family income) and by the impact of a 10 percent increase (all other variables).

12. A step-by-step exposition of this procedure, showing its relationship to OLS, is given in Hu (1999).
13. Note that the constant is absorbed into the LPR part of the model. This is arbitrary; it has no effect on the results.
14. The exponents in equations (6) and (7) are taken element-by-element.
15. Product kernels are standard in the literature; a different bandwidth, h, is used for each dimension. Locally adaptive bandwidths are allowed by increasing bandwidth at each point until at least 20 observations are within one bandwidth, provided that this increase is necessary.
16. Both papers can be found at http://www.sba.uconn.edu/users/johnc/currentresearch.html
17. Where neighboring residuals were lost for particular observations due to truncation of the time period, zeros were entered. Reuse of the residuals produces inconsistent estimators, but the focus here is ex-sample prediction, where pragmatism rules. See discussion in Section 3.4.
18. The unbiased predicted value (Neyman and Scott, 1960), is

$$\hat{P}_{it} = e^{X\hat{\beta} + \hat{\sigma}^2/2(1 - \frac{1}{2(T-K-1)}X'(X'X)^{-1}X)}$$

which differs from Case's predicted value only by the factor

$$1 - \frac{1}{T-K-1} - \mathrm{tr}X'(X'X)^{-1}X \approx 1.$$

19. Note that the coefficients on the 5th to 10th and on the 11th to 15th lags should be divided by 5.
20. Our results appear to bracket the optimal number of nearest neighbors. Fifteen is probably too small, but the 225 used by Dubin in the second round is probably too large. However, the optimal number may depend on the model used.
21. The details for these results are not reported herein, but are available upon request from Case.
22. Future research can determine if spatial autoregressive (SAR) techniques might be used to solve this econometric difficulty and preserve good out-of-sample performance.
23. Clapp's model is capable of producing interesting and significant spatial and temporal patterns of house prices: see Clapp (2002).
24. Software to estimate similar space-time models is available from James LeSage's spatial econometrics toolbox at www.spatial-econometrics.com and Kelley Pace's toolbox at www.spatial-statistics.com.
25. See Pace et al. (2002) for a discussion on the level of accuracy an appraisal system should possess.

References

Case, B. (2000). "Co-Movements in International Commercial Real Estate Markets," Working paper, Federal Reserve Bank (August 2000).

Case, B. (2001). "Co-Movements in U.S. House Price Appreciation Patterns," Working paper, Federal Reserve Bank (January 2001).

Case, B. (2002). "Homogeneous Within-County Districts for Hedonic Price Modeling," Working paper, Federal Reserve Bank (May 2002).

Clapp, J. M. (2002). "A Simi Parametric Method for Estimating Local House Price Indices," Draft paper, University of Connecticut Center for Real Estate. Available at: *http://www.sba.uconn.edu/users/johnc/currentresearch.html*

Clapp, J. M., and C. Giaccotto. (1998). "Residential Hedonic Models: A Rational Expectations Approach to Age Effects," *Journal of Urban Economics* 44, 415–437.

Clapp, J. M., H.-J. Kim, and A. E. Gelfand. (2002). "Spatial Prediction of House Prices Using LPR and Bayesian Smoothing," *Real Estate Economics* 30(4), 505–532.

Dubin, R. (1992). "Spatial Autocorrelation and Neighborhood Quality," *Regional Science and Urban Economics* 22, 433–452.

Dubin, R., R. K. Pace, and T. G. Thibodeau. (1999). "Spatial Autoregression Techniques for Real Estate Data," *Journal of Real Estate Literature* 7, 79–95.

Fan, J., and I. Gijbels. (1996). *Local Polynomial Modeling and Its Applications*. New York: Chapman and Hall.

Goetzmann, W. N., and M. Spiegel. (1997). "A Spatial Model of Housing Returns and Neighborhood Substitutability," *Journal of Real Estate Finance and Economics* 14(1), 11–31.

Goldberger, A. S. (1968). "The Interpretation and Estimation of Cobb-Douglas Functions," *Econometrica* 36, 464–472.

Goodman, A. C., and T. Thibodeau. (1995). "Age Related Heteroskedasticity in Hedonic House Price Equations," *Journal of Housing Research* 6(1), 25–42.

Hu, D. (1999). "A Semiparametric Investigation of Lower-Income Home Mortgage Purchases in the Secondary Mortgage Market," Draft Paper, Zell/Lurie Real Estate Center, The Wharton School, University of Pennsylvania.

McMillen, D. P. (1994). "Vintage Growth and Population Density: An Empirical Investigation," *Journal of Urban Economics* 36, 333–352.

Miller, D. M. (1983). "Reducing Transformation Bias in Curve Fitting," Virginia Commonwealth University, Institute of Statistics Working Paper No. 7, November 1983.

Neyman, J., and E. L. Scott. (1960). "Correction for Bias Introduced by a Transformation of Variables," *Annals of Mathematical Statistics* 31, 643–655.

Pace, R. K., R. Barry, J. M. Clapp, and M. Rodriguez. (1998a). "Spatial Autoregressive Models Neighborhood Effects," *Journal of Real Estate Finance and Economics* 17, 15–33.

Pace, R. K., R. Barry, and C. F. Sirmans. (1998b). "Spatial Statistics and Real Estate," *Journal of Real Estate Finance and Economics* 17, 1–13.

Pace, R. K., C. F. Sirmans, and C. Slawson. (2002). "Are Appraisers Statisticians?" In *Real Estate Valuation Theory*, Research Issues in Real Estate Monograph Series, Vol. 8, American Real Estate Society, Dordrecht: Kluwer Academic Press.

Pfeifer, P. E., and S. Bodily. (1990). "A Test for Space-Time ARMA Modeling and Forecasting of Hotel Data," *Journal of Forecasting* 9, 255–272.

Ripley, B. (1981). *Spatial Statistics*. New York: Wiley, pp. 45–51.

Robinson, P. M. (1988). "Root-*N*-Consistent Semiparametric Regression," *Econometrica* 56, 931–954.

Raman, P. (2002), "Towards Robust House Price Index Estimation: An Empirical Investigation of San Diego, California," Draft paper, Fannie Mae.

Stock, J. H. (1989). "Nonparametric Policy Analysis," *Journal of the American Statistical Association* 84, 567–575.

Thibodeau, T. G. (1989). "Housing Price Indexes from the 1974–1983 SIMSA Annual Housing Surveys," *AREUEA Journal* 1, 100–117.

Wand, M. P., and M. C. Jones. (1995). *Kernel Smoothing* New York: Chapman and Hall.

Yatchew, A. (1998). "Nonparametric Regression Techniques in Economics," *Journal of Economic Literature* 36, 669–721.