The Comparable Sales Method as the Basis for a Property Tax Valuations System and its Relationship and Comparison to Geostatistical Valuation Models

Richard A. Borst

Keywords: Comparable Sales, geostatistical, spatial autocorrelation, weight matrix

Abstract

There is an identifiable theoretical relationship between the comparable sales method (CSM) of valuation as practiced by mass appraisers and the recent developments in geostatistical valuation models. The CSM is shown to be a special case of a spatially lagged weight matrix model. There is a less formal but clear relationship with Geographically Weighted Regression as well. The predictive accuracy of CSM is compared to several Ordinary Least Squares Model configurations, and results obtained from Geographically Weighted Regression via empirical studies on diverse datasets. An example of a comparable sales weighting scheme as practiced by mass appraisers is provided. In addition, particular interest is focused on how well each method is able to model the spatial variations in property values. This is done by examining the local and global spatial autocorrelation in residual errors of the predicted values.

Introduction

The comparable sales method of valuation as implemented in a mass appraisal setting has gained widespread use in North America. As to whether it is a “best practice” is a matter for discussion, and even debate. This paper takes the position that is certainly among the best, if not the best method for mass appraisal. Other candidates for a “best practice” include the use of a well structured linear or nonlinear model calibrated for an entire jurisdiction, market segmentation models, the “response surface” method, and a collection of advanced modelling techniques. The purpose of this paper is to compare a selected subset of these techniques to the CSM.

The topics contained herein in their order of presentation are:

- A description of CSM as implemented in a mass appraisal environment.
- An introduction to weight matrix models
- The relationship of a spatial lag weight matrix model to the comparable sales model
- Geographically Weighted Regression (GWR) Models
- Relationship of GWR to CSM
- An empirical comparison of several selected methods
- Conclusions

Comparable Sales Method

The main processing steps within the method include finding the n most comparable sales properties (comps); computing an adjusted sale price for each comp; weighting these estimates according to their similarity to the subject; and summing the weighted comp estimates to get the final estimate

Consider that finding the most comparable sale is equivalent to finding the least dissimilar. The actual dissimilarity measure can be based on physical separation, and differences in physical characteristics, date of sale, and the neighbourhood to which the comparable sale belongs. The expression by which “comparability is judged was documented by Gipe (1975). It takes on the form

$$D_{jk} = (\sum_{i=1}^{n} [DW_i(X_{ij} - S_{ik})]^2)^{1/2} \qquad (1.1)$$

where D_{jk} is the dissimilarity measure, often referred to as "Distance" between subject property j and sale property k, DW_i is a weighting factor assigned to characteristic i, X_{ij} is the value if the ith variable for the subject property j and S_{ik} is the value of the i th variable for the k th sale.

Initially consider estimating the value of a subject based on one comparable sale. Cannaday (1989) describes the method of adjustment of a comparable sale by using the difference in the MRA estimates of the subject and the comparable. This adjustment process can be expressed as:

$$Estimate\ of\ S_k\ based\ on\ C_j\ =\ ASP(C_j)+(ESP(S_k)\text{ - }ESP(C_j)) \tag{1.2}$$

where S_k refers to subject k, C_j to comparable sale j, $ASP(C_j)$ is the actual selling price of C_i, and $ESP(S_k)$ refers to the estimated selling price of subject k. Equation (1.2) can be rewritten as:

$$Estimate\ of\ S_k\ based\ on\ C_j\ =\ ESP(S_k)+(ASP(C_j)\text{ - }ESP(C_j)) \tag{1.3}$$

where the expression $(ASP(C_j)\text{ - }ESP(C_j))$ is the residual error of the estimate for comparable sale j. Generalizing (1.3) to the case of valuing a subject by using several comparable sales results in

$$\begin{gathered}CSM(S_k)=\sum_{j=1}^{n}CW_{jk}[ESP(S_k)+(ASP(C_j)-ESP(C_j))]\ \ with\\ \sum_{\mathrm{j=1}}^{\mathrm{n}}CW_{jk}=1\end{gathered} \tag{1.4}$$

where $CSM(S_k)$ refers to the comparable sales method, CW_{jk} is a "comparability weight applied to each comparable sale property. Noting that $\sum_{j=1}^{n}CW_jESP(S_k)=ESP(S_k)$, (1.4) can be rewritten as:

$$CSM(S_k)=ESP(S_k)+\sum_{j=1}^{n}CW_{jk}(ASP(C_j)-ESP(C_j))\ (1.5)$$

In matrix notation this expression can be rewritten as:

$$CSM(S)=X\hat{\beta}+CW(Y-X\hat{\beta}) \tag{1.6}$$

where $\hat{\beta}$ are the estimates of β obtained by OLS methods. The formulation of (1.6) can be rewritten in the compact form:

$$\tilde{Y}=\hat{Y}+CW\hat{\varepsilon} \tag{1.7}$$

where the substitution of variables is $CSM(S)=\tilde{Y}$, $\hat{Y}=X\hat{\beta}$, and $W\hat{\varepsilon}$ represents the weighted residual errors from OLS. However, it is the formulation of (1.6) that will be useful in the comparison to the weight matrix methods.

Thompson (2006) describes a weighting mechanism in common use for developing CW_{jk} combining the n comparable sales estimates expressed as

$$CW_{jk}=1/[(D_{\max}/2)^2+D_{jk}{}^2+(2D_{\max}P_{jk})^2] \tag{1.8}$$

where

$CW_{jk}\ =\ comparability\ weight\ of\ jth\ sale\ to\ the\ kth\ subject$

$D_{\max}\ =\ \max imum\ acceptable\ comparability\ dis\tan ce$

$D_{jk}\ =\ actual\ comparaiblity\ dis\tan ce\ between\ jth\ sale\ and\ kth\ subject$

$P_{jk}\ =\ fractional\ adjustment\ to\ the\ jth\ sale\ for\ the\ kth\ subject$

Table 1 illustrates the method for $n = 5$. For purposes of simplicity of exposition j ranges from 1 to 5, and k is not specifically defined.

Sale Number	**1**	**2**	**3**	**4**	**5**
Price	45,000	30,000	50,000	25,000	40,000
Adjusted Selling Price	45,000	40,000	50,000	40,000	30,000
Adjustment to Price	0	10,000	0	15,000	-10,000
P_{jk}	0	0.333	0.000	0.600	-0.250
D_{jk}	10	60	70	80	120
$(D_{\max}/2)^2$ with $D_{\max} = 100$	2,500	2,500	2,500	2,500	2,500
D_{jk}^2	100	3,600	4,900	6,400	14,400
$(2D_{\max} P_{jk})^2$	0	4444	0	14400	2500
Total Denominator	2,600	10,544	7,400	23,300	19,400
$CW_{jk} \times 10,000 \quad \sum CW_{jk} \times 10,000 = 7.091$	3.846	0.948	1.351	0.429	0.515
Normalized W_{jk}	0.542	0.134	0.191	0.061	0.073
Weighted Contribution	24,410	5,350	9,529	2,421	2,181
Weighted Estimate	43,891				

This methodology is in widespread use in North America. Representative references to its use are found in Underwood and Moesch (1982), Thompson and Gordon (1987), McCluskey and Borst (1998) and Todora and Whiterell (2002). Additional analytical foundations for the method are described in Colwell et al (1983). Vandell (1991) provides the theoretical background to provide decisions about how many comparables to select, what the comparable selection criteria should be, and how proper weights for each adjusted value estimate can be determined such that the final value estimate is both unbiased and of minimum variance. Kang and Reichert (1991) compared MRA to the comparable sales method on data sets with a variety of characteristics. Gau et al, (1992) extended Vandell's method. Lai and Wang (1996) compare the accuracy of the minimum variance grid[1] method to MRA value estimates. They establish the theoretical basis upon which it is shown that the grid method should have a lower variance than MRA estimates. They also show a means by which confidence intervals can be computed. Pace and Gilley (1998) describe the grid estimator in the context of a spatial autoregression formulation.

Weight Matrix Model

In this discussion W is an $n \ x \ n$ matrix, the elements of which represent the "strength" of the connections between each property and all other properties. There are two variants of the weight matrix model. In one the matrix is used to model a spatial process directly. These models are referred to as spatial lag models. The other uses a weight matrix to model the error term. Only the former is addressed herein because it has a clear relationship to the comparable sales method.

The simplest form of the spatially lagged weight matrix model is also referred to as the autoregressive model, meaning that the dependent variable is related to itself in a specific way. The logic behind the autoregressive model is that there are "spill over" effects in which the sales prices of nearby properties affect the value of a given property more than those that are farther away. Stated another way, values at close-by locations are more correlated than values at locations that are far apart. This autoregressive model has the following formulation:

$$Y = \rho WY + \varepsilon \qquad (9)$$

[1] The "grid method" or "grid estimator" is a common description given to the comparable sales analysis technique.

where ρ the coefficient of autocorrelation variable is, W is a matrix of weights with elements W_{ij} that specify the strength of the relationship between properties i and j, and ε is the error term. In other words, the value of a single observation of the dependent variable, Y_i, is a weighted average of its neighbouring properties Y_j (with $i \neq j$). Models of this from are often referred to as lattice models because the weight matrix is developed for a specific set of points which can be from a regular or irregular arrangement of the points on the two-dimensional plane. Specification of spatially autoregressive model in hedonic form was reported by Can (1990), (1992). The model is expressed as

$$Y = X\beta + \rho WY + \varepsilon \ . \qquad (10)$$

It is often referred to as a mixed regressive, autoregressive model because it is a combination of the basic hedonic model and the autoregressive model forms. Pace and Gilley (1997) use a somewhat different model specification of a weight matrix spatially autoregressive model. In the equation

$$Y = X\beta + \rho W(Y - X\beta) + \varepsilon \qquad (11)$$

the term $\rho W(Y - X\beta)$ represents a weighted average of the errors on nearby properties. As before, W is a matrix of weights W_{ij} with the added properties that $W_{ii} = 0$, the rows of W sum to 1 , $0 \leq \rho < 1$ and ε □ $N(0,\sigma^2 I)$. This specification is referred to as a simultaneous autoregressive specification (SARS) model with log-likelihood function

$$L(\rho,\beta,\sigma^2) = \frac{1}{2}\ln|B| - \frac{1}{2}\left[n\ln(2\pi\sigma^{-2}) + \sigma^{-2}(Y - X\beta)'B(Y - X\beta)\right] \qquad (12)$$

where $\beta = (I - \rho W)^{'}(I - \rho W)$. The maximum likelihood (ML) estimation method efficiently estimates the model asymptotically provided that the model assumptions are true. The study in the cited reference compares OLS to SARS on an often cited set of data originally presented by Harrison and Rubinfeld (1978). The SARS method has a lower (44% reduction) error and more intuitively satisfying coefficients.

Pace and Gilley (1998) show how a linear combination of an OLS estimate and comparable sales estimate form the basis of a simultaneous autoregressive (SAR) model.[2] The formulation is the same as that in (11). The parameters of this model were estimated by two methods, Estimated Generalized Least Squares (EGLS), and Maximum Likelihood (ML) as mentioned above. A comparison of the two estimating methods was performed on the same sample data used by Can (1992) It was found that the EGLS SARS model was superior in performance relative to OLS and the Grid estimator. In the study of 563 properties the OLS had 27.9% and grid estimator had 11.7% greater median absolute error than the EGLS SARS. They observe that although the EGLS SARS is superior, the grid estimator provided evidence of its utility in the presence of spatial information. It was called a “poor man’s” spatial autoregression.

Finding solutions by ML involves finding the log determinant of a matrix. When the number of observations is large, and the weighting function is such that the weights are non-zero for a relatively small number of observations in the vicinity of a specific observation, the matrices become large (NxN). Direct methods for finding the log determinant become computationally intense. Barry and Pace (1999) provide a Monte Carlo simulation method that has been shown to be efficient on matrices of size $1,000,000x1,000,000$. Such techniques are a necessity for ML techniques to be useful for prediction applications such as assessment. Kelejian and Prucha (1998) describe the theoretical background for a generalized spatial two-stage least squares procedure for estimating a model of the type described in this section. They state that it is computationally simple procedure for estimating models with both a spatially lagged independent variables and error terms. It is fair to say that

[2] The article points to Cressie (1993) pp. 402-410, as the support to the SAR designation.

calibration of such models goes well beyond the skills, expertise and software tools needed to calibrate OLS models.

Relationship of Comparable Sales Model to Weight Matrix Model

Based on the presentation in Pace and Gilley (1998), consider the following combination of the comparable sales estimator and the OLS estimator:

$$\breve{Y} = \alpha X\hat{\beta} + (1-\alpha)[X\hat{\beta} + W(Y - X\hat{\beta})] \qquad (13)$$

this can be rewritten as

$$\tilde{Y} = X\hat{\beta} + (1-\alpha)[W(Y - X\hat{\beta})] \qquad (14)$$

or

$$\tilde{Y} = X\hat{\beta} + \rho^{*}[W(Y - X\hat{\beta})] \qquad (15)$$

The analogy to equation (11) is evident. It is presented again here for ease of comparison

$$Y = X\beta + \rho W(Y - X\beta) + \varepsilon \qquad (16)$$

The difference being that the parameters of (16) have yet to be estimated and require advanced calibration methods, while those in (15) could have been estimated by OLS with systematic variation in ρ^{*} providing an estimate of its optimum value.

Geographically Weighted Regression

In a typical application of Multiple Regression Analysis (MRA) one equation is calibrated on a given set of sales, each of which is weighted equally. There are variants of this technique that weight the points in the sample data set differently, such as Robust Regression, which seeks to down-weight or eliminate outliers in the data set. However, there is still only one equation developed for the entire data set, and there is no inherent spatial component to the weighting method. GWR, on the other hand is a computationally intensive technique that weights each point in the dataset based on its location. The introduction to the concepts involved in GWR often includes description of moving window regression. In this case the sample points within a fixed distance of a given point are included in the regression, all with equal weight, and all others are excluded. This was the case with Dubin's procedure in Case et al (2004). She took the 200-300 points nearest the sample point an included them in the regression model. This procedure was repeated for each data point in the sample set. GWR operates in a similar fashion; however the points do not receive equal weight. Instead, the weight is a function of location, and diminishes with the distance from the regression point.

Figure 1 Example Spatial Weighting Kernel illustrates the concept of a weighting function based on the coordinates of the regression point and the data points near it. The peak of the surface is the regression point; any sample points under the surface would receive the weight based on the height of the surface at that point.

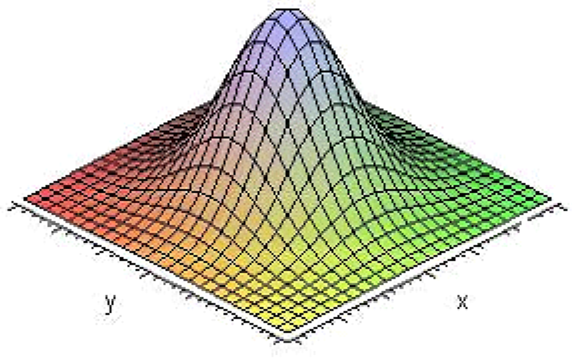

Figure 1 Example Spatial Weighting Kernel

Two dimensional representations of the kernel density function are used to illustrate the concept of fixed and variable bandwidth weighting. Figure 2 Spatial Kernel shows the spatial dimension along the X axis, and the weighting function is represented by the vertical height of the curve. The height of the curve at point j, given by w_{ij}, is the weight applied to point j when point i is the regression point and d_{ij} is the distance between the regression point i and data point j. The bandwidth of the spatial kernel, h, is a parameter that affects how the weight is computed as the distance between the regression point and other sample points increases.[3] It can be fixed or variable. This becomes an important consideration when the sample points are not regularly spaced. A fixed bandwidth could result in there being insufficient points considered in the regression if the bandwidth is too small.

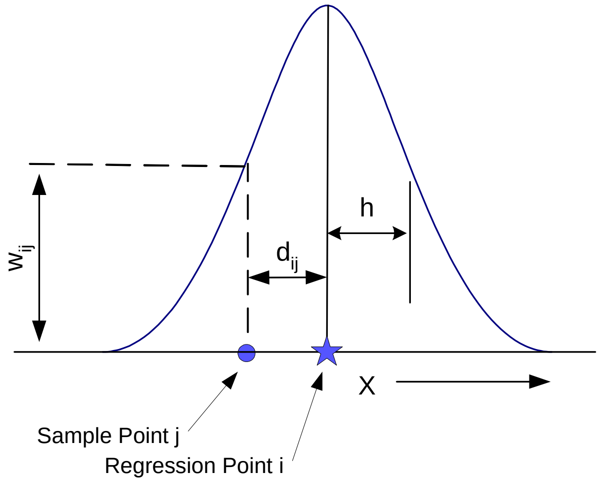

Figure 2 Spatial Kernel

Figure 3 Variable Bandwidth Kernels shows two kernels. In the left part of the figure, the ten data points are more closely spaced than the ten points on the right. In this hypothetical example it can be seen that by appropriately adjusting the bandwidth, ten data points can be considered in each case.

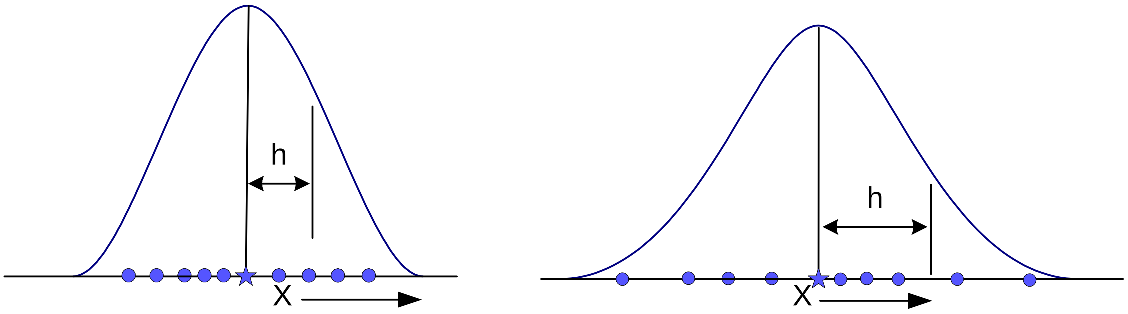

Figure 3 Variable Bandwidth Kernels

The adaptive kernel is more appropriate for application with real property transaction data. This can be reasoned by consideration of the variations in the spatial density of property locations. The number of points contained in a fixed bandwidth does vary considerably in all three of the counties of the study data, thus all GWR estimation reported herein was performed with an adaptive bandwidth kernel.

Relationship of GWR to Comp Sales

The relationship between GWR and CSM lends itself better to a narrative description. In GWR a localized model is developed by weighting points based on their distance from the regression point.

[3] In the Gaussian formulation of the kernel, the weighting function is given by the distance decay function $w_{ij} = \exp[-1/2(d_{ij}/h)^2]$.

More distant points receive less weight than those that are closely proximal to the regression point. The regression process itself accounts for the differences in the independent variables for each property included in the regression. CSM achieves its "localization" in two ways. First, physical distance can be incorporated into the dissimilarity measure. This places emphasis on finding nearby properties. However, it also allows for weighting on physical characteristics as well. This allows for finding properties more like the subject, for example in story height, living area, and number of bedrooms. The choice is only limited by the characteristics available in the database of property descriptors. Furthermore, the variables used in the selection process need not have been used in the OLS model used to adjust the comparable sales weights.

Response Surface Methods

The term "Location Value Response Surface" (LVRS) emerged from the assessment community starting in the early 1980's. Most often it refers to the application of a location based correction factor to a base value determined via multiple regression analysis or similar multivariate technique without explicit incorporation of location in the model structure. In one of the first reports of its use, O'Connor (1982) described a methodology for developing a location based correction that was in turn applied to a multiplicative form of a valuation model. A model using various functions of x, y coordinates including the distance to certain "Value Influence Centres" (VICS) as independent variables, and the relative value of a typical home at 86 points in the study area[4] was developed as the "Response Surface". A separate model using building and land characteristics as independent variables and selling price as the dependent variable in a multiplicative form was developed to arrive at a to-be-corrected value estimate. A final value estimate is the product of the base value and the response surface correction computed for the exact location of the property being valued. The LVRS technique evolved over time. O'Connor and Eichenbaum (1988) and Eichenbaum (1988) provide additional insight as to how the VICS are incorporated into the model.

By 1999 the terminology for some authors in the assessment community had changed from LVRS to Response Surface Analysis (RSA) and GIS technology had advanced in capability and accessibility. That is, its migration to desktop computers had made it available to a larger number of users. Ward, et al (1999) detailed the use of GIS to develop a surface of normalized sale price per square foot of living area. The normalized factor derived from the surface was utilized as an independent variable in the CAMA model. McCluskey et al (2000) investigate alternative approaches which specifically model the spatial distribution of house prices with the objective of developing location adjustment factors. These approaches were based on the development of surface response techniques such as inverse distance weighting and universal Kriging. The results generated from the surfaces created were then calibrated within a model structure including other descriptive detail about the sale properties. Ward et al (2002) provide an extensive study of the use of Global Response Surface (GRSA) tools to develop location factors for use as an independent variable in a nonlinear model formulation. Instead of using price per square foot (meter) or price as the dependent variable in the response surface, the concept of a "z-score" ($z = (x - \mu) / \sigma$) which allows for combining data having different fundamental means (μ) and standard deviations (σ)distributions. The advantage cited is that, for example, vacant property sales can be combined with sales of improved properties to form a combined dataset. Similarly, z-scores can be used to combine size with condition on a meaningful scale. The latter allows for developing surfaces on unsold properties based on their physical descriptions. Korbo et al (2003) use a similar technique to develop and use location adjustments for agricultural land in Saskatchewan.

Empirical Results

We compare the following model/methods on datasets from three U.S. counties.

- OLS regression – global model
- OLS regression – segmented models
- OLS regression – segments as binary variables in global model

[4] Lucas County Ohio

- A response surface model
- A combined segmented and response surface model
- GWR regression
- CSM

The Study Data

Sales transaction and property descriptive data have been acquired from three rather different counties in the U.S. They are Catawba County, North Carolina, Sarasota County, Florida, and Fairfax County, Virginia. Table 1 Brief Comparison of Subject Counties provides a high level summary of the three jurisdictions. Catawba County is a relatively small, low population density jurisdiction. Sarasota County has a higher population density, but of importance is the much larger amount of "water area". The number of persons per household is much lower than either Catawba or Fairfax indicative of an older population. This is consistent with the theme of Florida as a place to retire. It has a coastline on the Gulf of Mexico, and considerable inland water resources. It has more than triple the number of housing units of Catawba. Fairfax County is a high population density suburban community bordering on Washington, DC. It has a much higher average income and more than six times as many housing units than does Catawba.

County	State	Pop (1990)	Housing Units	Land Area sq KM	Water Area sq KM	Pop Density	Per Hshld	2003 Median Income	Pop 2004 Estimate	Pop Incr.	2004 Pop Density
Catawba	North Carolina	118,412	49,192	1036	35.069	114.302	2.4071	$40,112	149,466	0.262	144.2779
Sarasota	Florida	277,776	157,055	1481	397.61	187.58	1.7687	$42,306	355,477	0.28	240.0506
Fairfax	Virginia	818,584	307,966	1025	29.109	798.976	2.658	$82,481	1,003,157	0.225	979.1282

Table 1 Brief Comparison of Subject Counties

There are differences in the housing stock as well. The major differences are shown in Table 2. Each data set was pared down from the original files obtained from the jurisdiction. Some reasonability checks on the relation between the selling price and the property characteristics were employed as well restricting the study to exclude condominiums and town homes.

Variable	County	Count	Mean	St. Dev	Std. Error	Minimum	Maximum	Range
Price	Catawba	7107	159,941	107,103	1270	22,000	1,800,000	1,778,000
	Sarasota	25276	295,036	358,283	2254	13,000	11,000,000	10,987,000
	Fairfax	21700	648,346	317,537	2156	175,000	11,700,000	11,525,000
Sale Date	Catawba	7107	Feb-03			Jan-00	Aug-05	71 months
	Sarasota	25276	Nov-03			Jan-02	Oct-05	43 months
	Fairfax	21700	Jan-05			Jan-04	Jan-06	25 months
Living Area	Catawba	7107	1737	764	8.97	384	8024	7640
	Sarasota	25276	1730	737	4.64	260	10867	10607
	Fairfax	21700	2193	1096	7.44	448	10341	9893
Lot Size (Sq. Ft.)	Catawba	7107	33,343	93,357	1096	3485	3,858,109	3,854,624
	Sarasota	25276	16,699	77,578	488	1283	10,890,000	10,888,720
	Fairfax	21700	19,407	36,055	245	871	2,583,979	2,583,108
Age	Catawba	7107	25.8	23.45	0.275	1	187	186
	Sarasota	25276	25.8	16.64	0.105	2	110	108
	Fairfax	21700	30.4	17.57	0.113	1	136	135

Table 2 Summary Statistics by County

There are several observations about the data that highlight the differences in the housing stock of the three counties:

- Sarasota has a mean selling price nearly double that of Catawba, while Fairfax is approximately four times as high.
- The selling prices start at $175,000 in Fairfax!
- The range of prices for both Sarasota and Fairfax is over six times as large as that of Catawba.
- The sale date range is quite different among the counties
- The homes of Fairfax are somewhat older and larger than that of the other two counties.
- The lot sizes in Catawba are considerably larger than those of the other two counties, consistent with the lower population density.

Histograms of “Price” illustrate the price differential among the three counties. All three histograms were trimmed at $1,000,000 to make the comparison more direct.

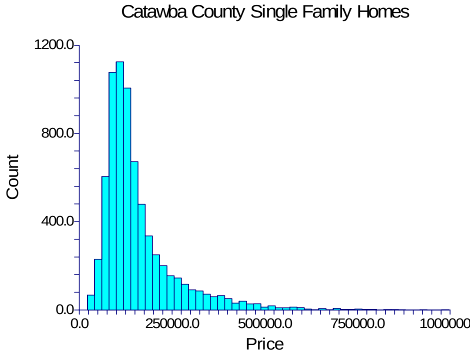

Figure 4 Catawba County Sales

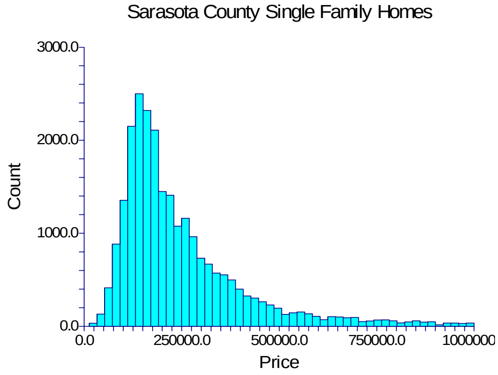

Figure 5 Sarasota County Sales

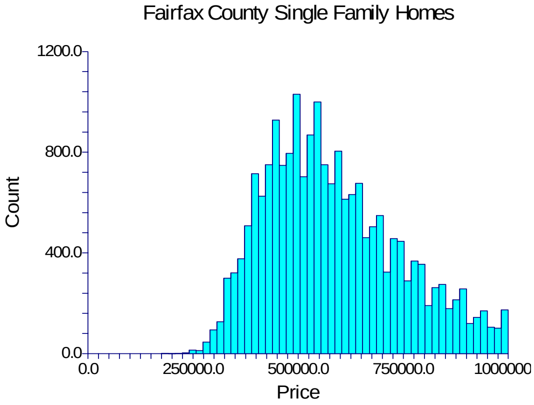

Figure 6 Fairfax County Sales

The variables available for use in the model specification and calibration process are somewhat different in each county. The one thing in common is that they are obtained from an assessment database, and are quite rich in information.

OLS Models

The development of a baseline global model and a segmented markets model is described in a paper presented at the IPTI'S 9TH International Conference 2006.[5] In brief, a GWR model is developed for each of the three counties. That model is used to value a "market basket value" (MBV) home at each of the regression points in the data. This allows for the formation of a surface, the variation of which represents the composite change in the calibrated model across the study area. Figure 7 Fairfax County MBV Surface provides an example surface. Submarkets are said to exist when the equation describing a candidate submarket is different from a reference equation. This principle is used in forming the submarkets.

[5] Copies of this paper are available upon request. Contact richborst@msn.com.

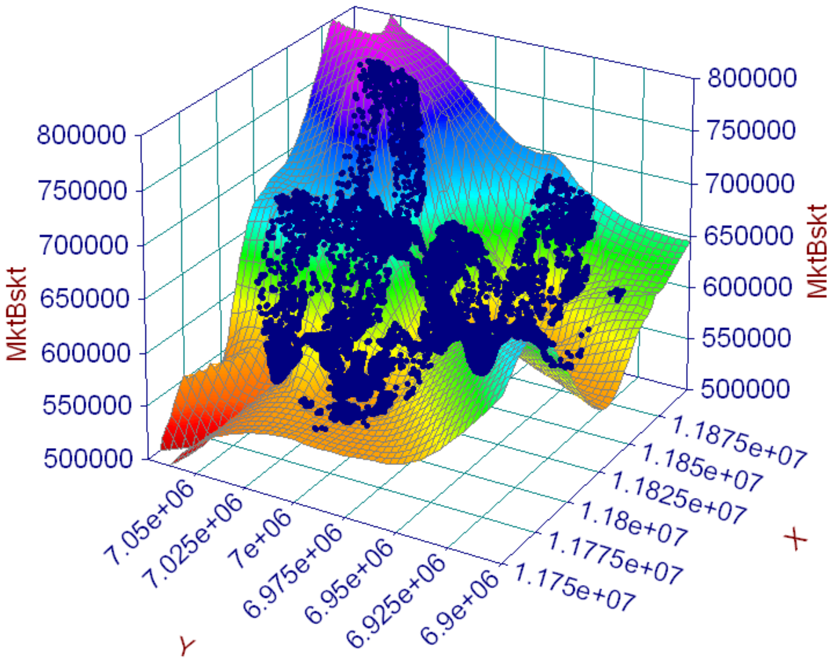

Figure 7 Fairfax County MBV Surface

Goodness of Variance Fit optimization, Smith (1986), is used to partition the MBV's into submarkets. The optimal number of submarkets is found by a variety of techniques including COD[6] minimization, Akaike Information Criterion minimization and spatial autocorrelation in residual errors minimization. Here we present the results for the global model and for the best segmented market models. The global model is based on structural and land characteristics and does no contain specific location information. Similarly, the segmented models contain no explicit location information. The segmentation process is the means by which locational effects are addressed. Three model structures were evaluated for each county – linear, semi-log and log-linear. The model structure with the best performance was chosen individually by county.

OLS Models with Binary Market Segments

An alternative use of the segments identified by the GWR/MBV process is to extend the baseline model by incorporating them as binary (dummy) variables. There are several possibilities for the outcome of such models. If their performance is equal to or superior to that of the segmented models, it could be concluded that all of the spatial variation can be accounted for in one model, thus obviating the need for segmented models. This would reduce the segmentation process to one of finding useful "super neighbourhoods". If, on the other hand, the performance of the extended baseline model is inferior to the segmented model, then further support for the fundamental need for market segmentation has been provided

GWR Models

GWR modelling is a computationally intense process. Processing times increase dramatically with the number of points in the dataset and the number of independent variables considered in the regression. To make the best comparison considering this limitation, all data points were used in each calibration run, but the number of independent variables were chosen based on the principle of parsimony. Multivariate variable selection routines[7] were used to select the number of variables. With reference to Figure 8 Variable Selection Curve - Catawba County, it can be seen that there is little change in R-Squared after ten variables enter the model. Thus, in the case of Catawba county ten variables were used in GWR.

[6] Coefficient of Dispersion is the average absolute deviation about the median estimated value/actual sale value. It is most commonly used by the assessment community.

[7] The variable selection process used herein is one described by McHenry (1978)

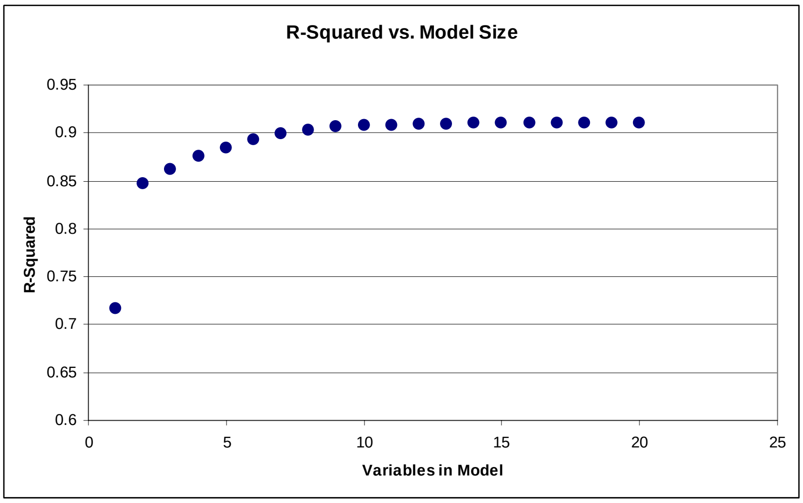

Figure 8 Variable Selection Curve - Catawba County

Response Surface Models

Ward et al (2002) used the *z-score* surface as an independent variable in an OLS model formulation. Instead of a *z-score*, we use the MBV surface value normalized to its mean value as an independent variable. The fundamental difference is in the fact that the MBV is based on all the information in the independent variables.

Combined Segmented and Response Surface Models

All segmented models were recalibrated using the MBV factor as an additional independent variable in the model.

CSM

The procedure outlined in the section, Comparable Sales Method was utilized to compute values on all properties in each of the three counties. The number of comparable properties considered was set to five. The selection of comparable weights is usually done from both statistical and aesthetic viewpoints. An example will illustrate. In this example eleven variables were considered in the comparability weighting scheme. A certain penalty weight is first selected for each variable. In general it is preferable to have the comparables drawn from the same neighborhood. They are more likely to share the same market influences, and are also more likely to be appreciated by a non-appraiser viewing the evidence supporting the value. As a starting point, a weight of 150 points is set as the "penalty" for leaving a neighborhood. The weight for the remaining variables is set within this context. As an example, the question is asked how much smaller or larger a house can be in a given neighborhood before it would be preferable to look outside the neighborhood. A representative set of weights is provided inTable 3 Representative Set of Weights. The last variable in the list is Parcel ID. A very large weight is place on this variable to insure that the subject property (all of which are also sale properties) is not included as its own comparable. This would give an artificially low error in the predicted value.

Variable Name	**Type**	**Weight**
NBHD Group	Binary	250
NBHD	Binary	150
Reverse Month of Sale	Continuous	1
Story Height	Continuous	10
Number of Beds	Continuous	15
Age	Continuous	1
Pool Area	Continuous	0.05
Garage Area	Continuous	0.05
Xcoord	Continuous	0.01
Ycoord	Continuous	0.01
SFLA	Continuous	0.1
PARID	Special	1000

Table 3 Representative Set of Weights

Model Performance Results

The results of the studies described above are summarized in the following tables and charts. Note that the global (baseline) models had no location influence factors in the model structure. They are based on physical characteristics and land size. The relatively high COD's are not presented as representative of a good modelling effort, but rather they serve as a baseline of comparison for the methods described herein to incorporate location into the model. They are simply reference points to show the improvement to be gained by the several methods described herein. The following summary of the methods tested in preparing this paper are restated for convenience of reference: The numbering of each item is cross referenced in the table headings.

1. Global OLS Model of best form
   a. Linear - Catawba
   b. Semi-log – not selected as "best"
   c. Log-linear – "best" in Fairfax and Sarasota
2. Segmented Model with Best Number of Segments derived via market basket value from GWR regression surface. In all three cases ten segments were used.
3. Best Global Model using Market Segments as Binary Variables
4. Global Model with normalized Market Basket Value as an Independent Variable
5. Best Segmented Model with normalized Market Basket Value as Independent Variable
6. Global GWR Model – same form as OLS models
7. Comp Sales Method using best predictions from segmented model as adjustment factors

Table 5 provides the performance of each model as measured by the Coefficient of Dispersion. Figures 9 and 10 present the same information in graphical form. The figures use a ratio of each model's performance to the baseline COD.

	Number of Sales in Model	Best Global Model	Best Segmented Model	Best Binary Model	MBV as Independent Variable in Global Model	MBV and Best Segmented Model	GWR Global	Comps Method
		(1)	(2)	(3)	(4)	(5)	(6)	(7)
Catawba	7,107	11.05%	10.21%	11.00%	11.08%	10.55%	9.97%	9.75%
Fairfax	19,983	8.94%	6.95%	7.22%	7.05%	6.88%	7.30%	6.52%
Sarasota	24,616	18.99%	14.36%	15.68%	16.09%	14.09%	13.15%	12.92%

Table 4 Comparison of COD Among All Tested Methods

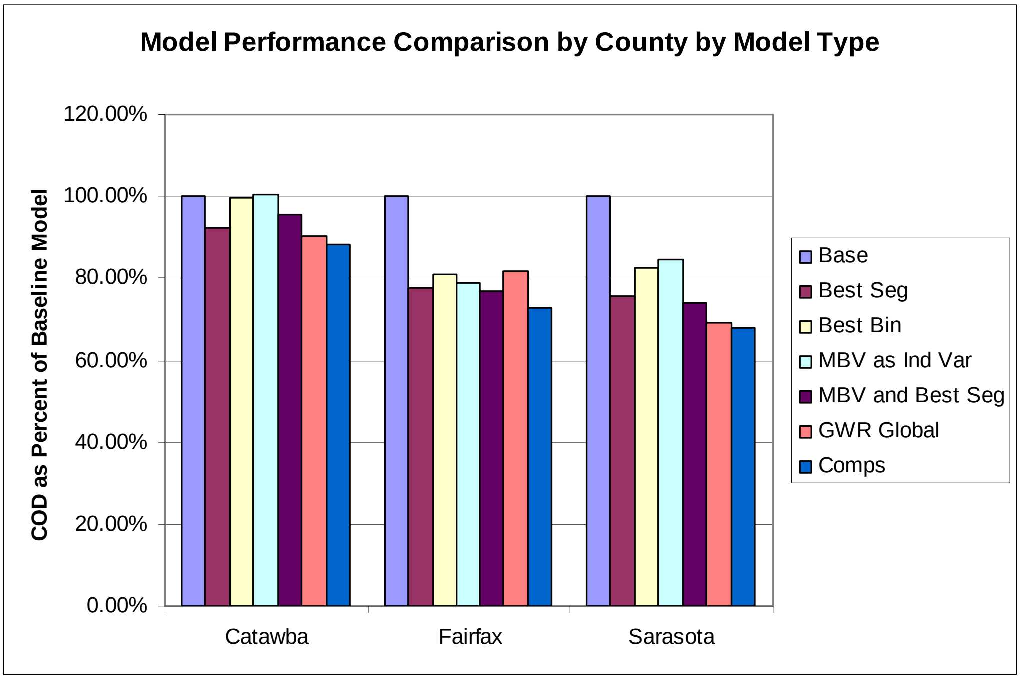

Figure 9 Relative Performance of the Methods by County

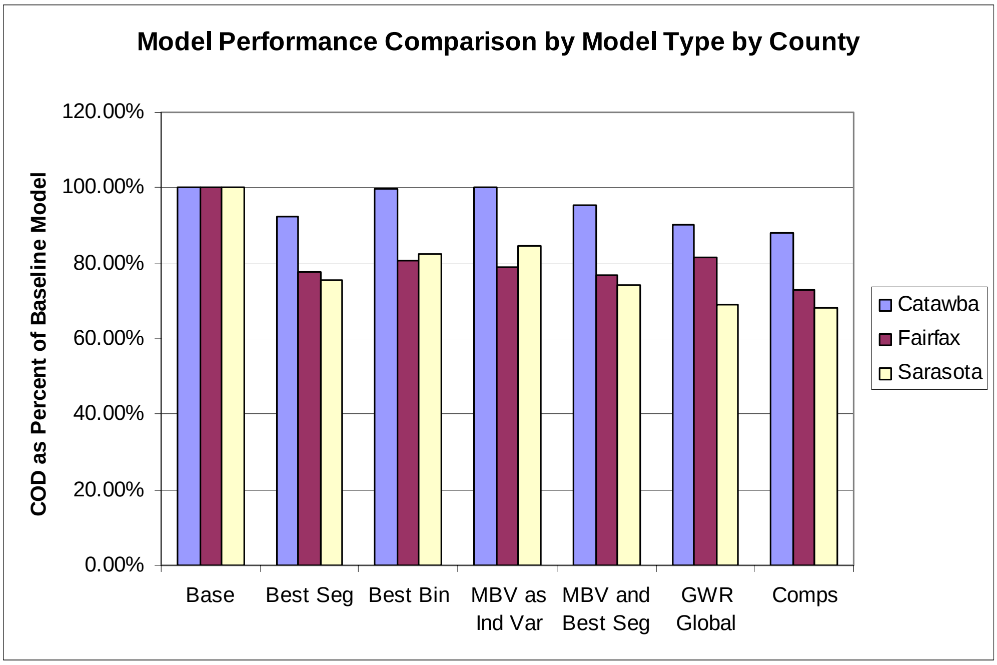

Figure 10 Relative Performance of the Methods by County by Model Type

Spatial Autocorrelation in the Residuals

If there are statically significant spatial patterns in the residual errors of a particular model it is possible that they can be detected by a measure called Moran’s I which tests for global spatial autocorrelation in group-level data. It is a weighted correlation coefficient used to detect departures from spatial randomness. The formula for Moran’s I is

$$I = n\frac{\sum_{i=1}^{n}\sum_{j=1}^{n} w_{ij}(y_i-\mu)(y_i-\mu)}{\left(\sum_{i=1}^{n}(y_i-\mu)^2\right)\left(\sum\sum_{i\neq j} w_{ij}\right)} \qquad (17)$$

where n is the number of observations, w_{ij} are the spatial weights between observations, y_i is the value at location i, and μ is the mean value of y. It ranges between values of -1 to 1. Departures from randomness indicate spatial patterns, such as clusters or patterns such as geographic trends. When values in nearby areas are similar, Moran's I will be large and positive. When values are dissimilar, Moran's I will be negative.

To illustrate the concept we present the spatial autocorrelation in the residual error ratio as a function of the number of market segments in the “segmented markets model” and for the CSM. With reference to Table 5 Three County Comparison of Moran's I three observations are immediate. The first is that the spatial correlation coefficients in Segment 1 are quite different among the three counties. The second is that they all improve with the number of segments. The third is that the CSM method’s reduction in spatial autocorrelation is dramatic.

	Moran's I - 1% Significance Test		
Segments	**Catawba**	**Fairfax**	**Sarasota**
1	0.178	0.416	0.3
2	0.155	0.345	0.269
3	0.179	0.306	0.277
4	0.16	0.266	0.25
5	0.142	0.25	0.217
6	0.138	0.244	0.201
7	0.133	0.247	0.206
8	0.138	0.238	0.202
9	0.138	0.24	0.19
10	0.136	0.235	0.19
CSM	0.012	0.132	0.020

Table 5 Three County Comparison of Moran's I

In Figure 11 Three County Comparison in Spatial Autocorrelation in Residual Error Ratios the same information is provided in the form of a graphic plot. In that figure is can clearly be seen that the CSM method makes a dramatic reduction in spatial autocorrelation.

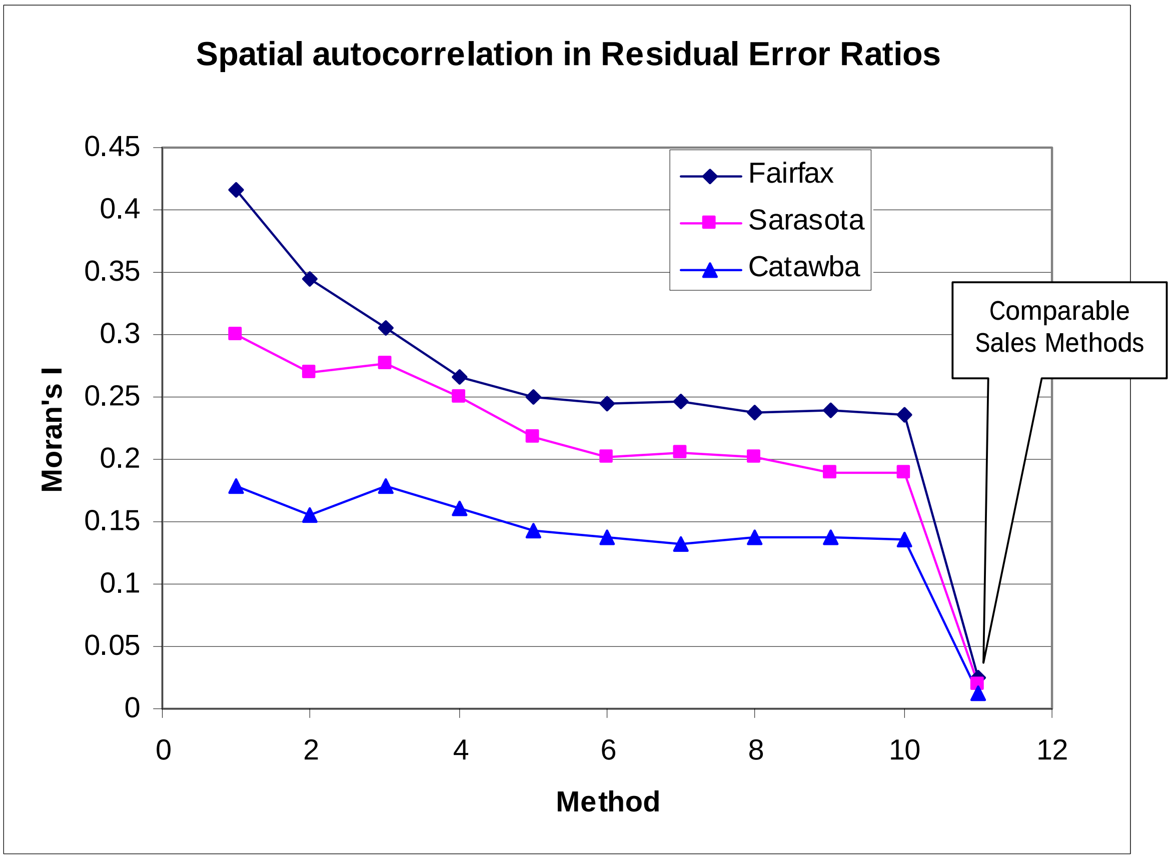

Figure 11 Three County Comparison in Spatial Autocorrelation in Residual Error Ratios

Conclusion and Observations

The comparable sales method has been demonstrated to be superior in predictive performance to all other methods presented herein. The second best performance was not as clear. In two of the counties, the GWR model was second in terms of prediction accuracy. In one, Fairfax, the segmented model with a location factor included as an independent variable was second best. We note that the

GWR method is computationally intense and is two orders of magnitude slower than traditional OLS techniques. This is not viewed as a major disadvantage because it is expected that the performance of GWR will improve over time with advances in software and hardware. We were not able to test the mixed regressive spatially autoregressive model at the time of submission of this paper. The reason is simple. There is no readily available user friendly software for this purpose. All the spatial statistics software to date does not include this type model. Again, this is expected to change over time and we hope to be able to report on its performance in the near future.

The dramatic reduction in spatial autocorrelation in the residual error provided by CSM is further evidence of its claim to be the “best practice”.

References

Barry, Ronald Paul, Pace, R. Kelley, 1999, Monte Carlo Estimates of the Log Determinant of Large Sparse Matrices, Linear Algebra and its Applications, 289, 41-54

Can, Ayse, 1992, Specification and Estimation of Hedonic Housing Price Models, Regional Science and Urban Economics 22, 453-474

Cannaday, Roger E., 1989, How Should You Estimate and Provide market Support for Adjustments in Single Family Appraisals, Real Estate Appraiser and Analyst, 55:4, 43-54

Case, Bradford, Clapp, John, Dubin, Robin, Rodriguez, Mauricio, 2004, Modeling Spatial and Temporal House Price Patterns: A Comparison of Four Models, Journal of Real Estate Finance and Economics, 29:2, 167-191, 2004

Colwell, Peter, F. , Cannaday, Roger E., Wu, Chunchi, 1983, The Analytical Foundations of Adjustment Grid Methods, Journal of the American Real Estate and Urban Economics Association, 11:1, 11-29

Eichenbaum, Jack, 1988, Incorporating Location into Computer Assisted Valuation: Progress in New York City., Paper presented at the World Congress III Computer-Assisted Valuation and Land Information systems Assisted Valuation, Cambridge, MA

Gau, George W., Lai, Tsong-Yue, Wang, Ko, 1992, Optimal Comparable Selection and Weighting in Real Property Valuation: An Extension, AREUEA Journal 20:1, 107-123

Gipe, George W., 1974, Analysis of Residuals to Improve Multiple Regression Equations, Assessors Journal, 9:1, 9-16

Harrison, D. and Rubinfeld, D. L., 1978, Hedonic Housing Prices and the Demand for Clean Air, Journal of Environmental Economics and Management, 5, 81-102 as cited in Pace and Gilley 1997.

Kang, Han-Bin; Reichert, Alan K., 1991, An Empirical Analysis of Hedonic Regression and Grid-Adjustment Techniques in Real Estate Appraisal, AREUEA Journal, 19:1 70-91

Kelejian, Harry H., Prucha, Ingmar R., 1998, A Generalized Spatial Two-Stage Least Squares Procedure for Estimating a Spatial Autoregressive Model with Autoregressive Disturbances, Journal of Real Estate Finance and Economics, 17:1, 99-121

Korbo, Bradley G., Rizvi, Syed, Ghebre, Kefala, Merritt, Garth, 2003, Location Adjustments for Agricultural Land Using the Geostatistical Capabilities of a GIS, Assessment Journal, Fall 2003

Lai, Tsong-Yue, Wang, Ko, 1996, Comparing the Accuracy of the Minimum-Variance Grid Method to Multiple Regression in Appraised Value Estimates, Real Estate Economics, 24:4, 531-549

McCluskey, William J., Borst, Richard A., 1998, Application of Hybrid Intelligent Appraisal Techniques Within The Field of Comparable Sales Analysis, 1998 IAAO Conference Proceedings: Today’s Vision Tomorrow’s Reality

McCluskey, William J., Deddis, William G., Lamont, Ian, Borst, Richard A., 2000, The application of surface generated interpolation models for the prediction of residential property values, Journal of Property Investment and Finance 2002, 18: 2

O’Connor, Patrick M., 1982, Making one MRA Model Behave Appropriately Across the Whole of Large County, Paper presented at the First World Congress on Computer Assisted Valuation, Cambridge, MA 1982

O’Connor, Patrick M., Eichenbaum, Jack, 1988, Location Value Response Surfaces: The Geometry of Advanced Mass Appraisal, IAAO Property Tax Journal, 277-298

Pace, R. Kelley, Gilley, Otis W., 1997, Using the Spatial Configuration of the Data to Improve Estimation, Journal of Real Estate Finance and Economics, 14:3, 333-340

Pace, R. Kelley,., Gilley, Otis W, 1998, Generalizing the OLS and Grid Estimators, Real Estate Economics; Summer 1998, 26, 2 331-347

Smith, Richard A., 1986, Comparing Traditional Methods for Selecting Class Intervals on Choropleth maps, Professional Geographer, 38(1), 1986, 62-67

Thompson, John F, 2006, Comparable Sales Weighting Formulae, Private communication

Thompson, John F., Gordon, Jack F., 1987, Constrained Regression Modeling and the Multiple Regression Analysis-Comparable Sales Approach, Property Tax Journal v. 6 no. 4

Todora, Jim, Whiterell, David , 2002, Automating the Sales Comparison Approach, Assessment Journal, Jan/Feb. 25-33

Underwood, William E., Moesch, James R., 1982, The Second Generation of CAMA in New York State., Paper presented at the First World Congress on Computer Assisted Valuation, Cambridge, MA

Vandell, K. D., 1991, Optimal Comparables Selection and Weighting in Real Property Valuation, Real Estate Economics, 19, 2 213-239

Ward, Richard D., Guilford, Jason, Jones, Brian, Pratt, Debbie and German, 2002, Piecing Together Location: Three Studies by the Lucas County Research and Development Staff, Assessment Journal, Sept/Oct. 15-48

Ward, Richard D., Weaver, James R., and German, Jerome C., 1999, Improving CAMA models using geographic information systems response surface analysis location factors, Assessment Journal, Jan/Feb