Report

Spatial Dependence, Housing Submarkets and House Price Prediction

BOURASSA, Steven C., CANTONI, Eva, HOESLI, Martin E.

Abstract

This paper compares alternative methods of controlling for the spatial dependence of house prices in a mass appraisal context. Explicit modeling of the error structure is characterized as a relatively fluid approach to defining housing submarkets. This approach allows the relevant submarket to vary from house to house and for transactions involving other dwellings in each submarket to have varying impacts depending on distance. We conclude that "for our Auckland, New Zealand, data" the gains in accuracy from including submarket variables in an ordinary least squares specification are greater than any benefits from using geostatistical or lattice methods. This conclusion is of practical importance, as a hedonic model with submarket dummy variables is substantially easier to implement than spatial statistical methods.

FACULTE DES SCIENCES ECONOMIQUES ET SOCIALES
HAUTES ETUDES COMMERCIALES

Spatial Dependence, Housing Submarkets, and House Price Prediction

Steven C. BOURASSA
Eva CANTONI
Martin HOESLI

HEC GENÈVE
UNIVERSITÉ DE GENÈVE

Spatial Dependence, Housing Submarkets, and House Price Prediction

Steven C. Bourassa

KHC Real Estate Research Professor,
School of Urban and Public Affairs, University of Louisville, 426 W. Bloom Street,
Louisville, Kentucky 40208, phone: (502) 852 5720, fax: (502) 852 4558,
email: steven.bourassa@louisville.edu

Eva Cantoni

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

and

Martin Hoesli

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

Spatial Dependence, Housing Submarkets, and House Price Prediction

Abstract

This paper compares alternative methods of controlling for the spatial dependence of house prices in a mass appraisal context. Explicit modeling of the error structure is characterized as a relatively fluid approach to defining housing submarkets. This approach allows the relevant submarket to vary from house to house and for transactions involving other dwellings in each submarket to have varying impacts depending on distance. We conclude that – for our Auckland, New Zealand, data – the gains in accuracy from including submarket variables in an ordinary least squares specification are greater than any benefits from using geostatistical or lattice methods. This conclusion is of practical importance, as a hedonic model with submarket dummy variables is substantially easier to implement than spatial statistical methods.

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

1. Introduction

House prices are customarily modeled using hedonic regression models whereby the price is explained by structural and locational attributes. As with all regression models, errors should be independent from one another, else parameter estimates will be inefficient and confidence intervals will be incorrect. The independence assumption is unlikely to be valid in a standard ordinary least squares (OLS) context, as house price residuals have been shown to exhibit spatial dependence in spite of efforts to model locational effects accurately (Pace, Barry and Sirmans, 1998). This obviously creates problems as such models are used for house price index construction (Can and Megbolugbe, 1997) and also for mass appraisal (Basu and Thibodeau, 1998; Bourassa, Hoesli and Peng, 2003). Basu and Thibodeau (1998), for instance, argue that spatial

dependence exists because nearby properties will often have similar structural features (they were often developed at the same time) and also share locational amenities. More generally, LeSage and Pace (2004) discuss both theoretical and statistical reasons that would explain why data from several fields of study would be prone to spatial dependence.

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

$$Y = \mu(X) + \varepsilon, \tag{1}$$

where Y is a vector of transaction prices, X is a matrix of values for residential property characteristics, and ε is an error term. One approach is to model $\mu(X)$ so that residuals over space do not exhibit any pattern. This usually implies incorporating geographical coordinates or other spatial indicators as regressors, parametrically or even nonparametrically (Colwell, 1998; Clapp, 2003; Fik, Ling and Mulligan, 2003). One such approach includes as a regressor the weighted average of recent sale prices for nearby properties (Can and Megbolugbe, 1997). An alternative approach is to model ε, that is, to assume not only that $E(\varepsilon) = 0$, but also that $E(\varepsilon\varepsilon') = \Omega$, which is a matrix with at least some nonzero off-diagonal elements. Ripley (1981) and Cressie (1993) provide a discussion of relevant spatial statistical methods for modeling ε. These include geostatistical models such as those applied in real estate by Dubin (1998) or Basu and Thibodeau (1998) and the lattice models that have been refined and applied by Pace and his colleagues (e.g., Pace and Barry, 1997a).

Theoretically speaking, the assumptions behind the two classes of spatial statistical models differ in terms of the definition of the domain over which spatial locations are permitted to vary (see Cressie, 1993, pp. 8-9, for details). In the case of lattice models, which include simultaneous autoregressive (SAR) and conditional autoregressive (CAR) variants, locations are restricted to a discrete set of points. In contrast, geostatistical models permit an infinite number of locations within a given geographical area. This has implications for the way predictions based on each type of model take into account spatial information. Given their constraints, the lattice models seem less suited than the geostatistical models for ex-sample prediction purposes. Whether this is of practical relevance is an empirical question that we will address here.

Our focus is to compare the utility of spatial statistical methods relative to each other and to simpler OLS methods in a mass appraisal context. Given our emphasis on practical applications of hedonic price models, we limit our analysis of spatial statistical models to those that can be implemented readily using existing software packages. This means that the spatial methods used here are well developed and established in the statistical literature. Our OLS models that are most directly comparable to the spatial statistical models rely on valuer-defined geographical submarkets (or neighborhoods) within which house values are considered to be interdependent.

For testing purposes, we use a sample of 4,880 residential sales in Auckland, New Zealand. For each method, 100 random samples that each contain 80% of the transactions are generated to estimate the predictive ability of each technique for the 20% ex-samples. We estimate four geostatistical models, one each based on exponential and spherical variograms and then robust versions of the same models. We then estimate two lattice models, SAR and CAR. We compare predictions from these models with each other and with the predictions from two OLS models. The predictions from one OLS model are adjusted by the unweighted average residuals for valuer-defined submarkets, while the predictions from the other OLS model are not. Finally, we add a set of submarket dummy variables to each of the models (OLS, geostatistical, and lattice) in order to assess the impacts of simple controls for neighborhood effects on the accuracy of predictions.[1] We are particularly interested in comparing the predictions based on the OLS model that incorporates submarket dummies with those based on the spatial statistical models without submarket dummies.

Most previous research has either focused on a limited subset of the available spatial techniques or used a small sample of properties. Using a small sample from Baltimore, Dubin (1988) compares ex-sample predictions using OLS and a geostatistical technique and concludes that the geostatistical approach is superior even when some neighborhood (census block group) characteristics are included as explanatory variables. Basu and Thibodeau (1998) compare the predictive ability of OLS and one geostatistical technique, concluding that the latter is superior for six of eight regions in Dallas. Dubin, Pace and Thibodeau (1999) compare regression coefficients across OLS and four different spatial methods (including both geostatistical and lattice models) using a small simulated

example for which the true parameters are known. Most of the spatial models performed better than OLS with respect to parameter estimation. Militino, Ugarte and García-Reinaldos (2004) apply several models – including CAR, SAR, and geostatistical – to 293 transactions from Pamplona, Spain, but do not attempt ex-sample predictions.

A recent study by Case et al. (2004) is in some ways similar to the present one. They apply OLS and several spatial statistical methods to a very large sample of about 50,000 transactions from Fairfax County, Virginia, using out-of-sample prediction accuracy for comparison purposes. In their final results, two of the three spatial methods produced more accurate ex-sample predictions than an OLS model that included median residuals for a small number of nearest neighbors. Although these authors estimated an OLS model with neighborhood (census tract) dummy variables, they did not then use that model for prediction purposes. Also, unlike the present paper, they performed their predictions using only one split of the data. This means that their results may depend on the particular split.

The present article is structured as follows. We first discuss the relationship between the ideas of spatial dependence and housing submarkets. These ideas are very closely related; thinking in terms of housing submarkets is helpful in conceptualizing the problem that spatial dependence models seek to rectify. Section 3 contains a presentation of the spatial statistical methods that are used in the paper, while section 4 outlines our research design. We discuss our empirical analysis in the following section. Section 6 concludes the paper.

2. Spatial Dependence and Housing Submarkets

The concepts of spatial dependence and housing submarkets are closely related. The submarket concept relies on the idea of substitutability. Substitutes are pairs of goods for which an increase in the price of one leads to an increase in the demand for the other. Pairs of goods with similar characteristics are likely to be substitutes. In equilibrium, prices equalize across substitutes. Within housing submarkets, prices of houses are similar because submarkets contain close substitutes. Implicit prices of the characteristics of houses are similar for the same reason.

Spatial dependence or autocorrelation refers to the existence of covariance in the errors in the context of hedonic price estimation for residential property markets. Given the similarities in the prices of housing characteristics within a submarket, errors are more likely to be correlated within submarkets than across submarkets. Therefore, controlling for submarkets in hedonic equations can substantially reduce estimation errors. This can be accomplished in a variety of ways. Simple methods include incorporating a series of dummy variables for the submarkets, estimating a separate equation for each submarket, or adjusting predicted values using the errors within each submarket.[2]

Controlling for submarkets in hedonic price equations assumes either that one has a predefined set of submarkets or that one is going to use some method to define them. Predefined submarkets are typically geographical areas, such as those defined by real estate agents (e.g., Palm, 1978) or by valuers (e.g., Bourassa, Hoesli and Peng, 2003). Alternatively, submarkets can be defined in terms of the characteristics of dwellings, neighborhoods, or census units. Statistical techniques, such as principal components and cluster analysis, can be used to combine similar dwellings or neighborhoods into submarkets, which may or may not be geographical areas (e.g., Bourassa et al., 1999). Ugarte, Goicoa and Militino (2004) demonstrate the use of mixture models which both estimate hedonic equations and classify transactions into submarkets which are not geographical areas. However, there is some evidence to suggest that geographical submarkets are more meaningful and therefore useful for improving prediction accuracy (Bourassa, Hoesli and Peng, 2003).

Spatial statistical methods allow for a more fluid concept of submarkets than is permitted by the fixed definitions based on geographical areas or housing or neighborhood characteristics. In effect, methods such as the lattice or geostatistical approaches applied here allow for the relevant submarket to vary from property to property. The relationships between the focal property in a submarket and nearby properties are captured in a matrix of weights in the case of lattice models or by a distance function based on a fitted variogram (or semivariogram) in the case of geostatistical models. This more fluid approach to modeling the relationships among

properties would seem *a priori* to allow for more effective reduction of prediction errors due to spatial dependence.

It is useful in this context to consider Can's (1992) distinction between *adjacency* and *neighborhood* effects. The lattice and geostatistical methods focus on adjacency effects, or the external effects of nearby properties on the property in question. The simpler methods mentioned above, such as controlling for location within a relatively homogeneous geographical area defined by valuers for appraisal purposes, imply a focus on neighborhood effects. Thus our empirical question is whether adjacency or neighborhood effects predominate. In other words, is it more important to account for each property's situation within the boundaries of relatively homogeneous neighborhoods that are recognized as such in a particular market or to account for the relationships between each property and its neighbors? The results will depend, of course, on how well the neighborhoods are defined. Our sense is that the classification created by Auckland valuers is based on relatively careful consideration of property characteristics and prices.

3. Alternative Methods for Modeling Spatial Dependence

In this section, we present two modeling approaches for spatial data that we use in this paper: lattice and geostatistical models. In a nutshell, the lattice approach models the covariance matrix of the errors parametrically, whereas the geostatistical approach builds the covariance matrix indirectly through a parametric variogram. Moreover, the underlying assumptions of the two approaches differ (see the discussion in Section 1). We refer the reader to Ripley (1981) and Cressie (1993) for a more complete and detailed description of the statistical aspects of these models.

3.1. Lattice Models

We assume that the data are issued from equation (1), with $E(\varepsilon) = 0$ and $E(\varepsilon\varepsilon') = \Omega$. Lattice models assume $\mu(X) = X\beta$ and parameterize the covariance function of the error term of the model by assuming either that $\Omega^{-1} = \sigma^2(I - \phi C)$ (CAR models) or that

$\Omega^{-1} = \sigma^2 (I - \alpha D)'(I - \alpha D)$ (SAR models), where C and D represent spatial weight matrices that specify the dependence among observations. These matrices satisfy the conditions that their rows sum to 1, that their diagonal is 0 (an observation does not affect its own prediction), and that $\phi\tau_i^C < 1$ and $\alpha\tau_i^D < 1$, where τ_i^C and τ_i^D are the n eigenvalues of matrices C and D, respectively, with $i = 1,\ldots,n$ (see Haining, 1990, p. 82). Many choices for specifying the weight matrices are available in the literature (see Getis and Aldstadt, 2004, for a review), but some of them show only small practical differences, such as the results in Militino, Ugarte and García-Reinaldos (2004).

The estimates of the parameters α (or ϕ) and β are then obtained by maximizing the log-likelihood

$$\ln L = const. + \frac{1}{2}\ln\left|\Omega^{-1}\right| - \frac{1}{2}\left((Y - X\beta)'\Omega^{-1}(Y - X\beta)\right). \qquad (2)$$

The most important computational issue here is the evaluation of the log-determinant ($\ln\left|\Omega^{-1}\right|$), which is infeasible by standard methods for large sample sizes, given that Ω is of size n by n. Pace and Barry (1997a, 1997b) have derived approximations to these terms that are implemented in their Matlab code.[3] In this paper, we use their code to fit CAR and SAR models, with the Delauney spatial weight matrix (Cressie, 1993, p. 374).

Predictions are computed simply as $\hat{Y} = X\hat{\beta}$. Ripley (1981, p. 90) gives a formula to compute fitted values (that is predictions for in-sample observations) for SAR models (see also Pace and Gilley, 1997, 1998). This formula borrows strength from the information provided by neighboring observations through the spatial weight matrix D. We are interested in ex-sample predictions, therefore ruling out the use of Ripley's formula.

Pace et al. (1998) use spatial and temporal weight matrices for nearby and recent transactions to improve ex-sample predictions. Their approach, however, has not been developed to the point that it is included in available statistical software. Moreover, our data would allow for only a very simplified approximation of their method given that we do not have a temporal dimension. Consequently, we do not attempt to implement their method here.

3.2. Geostatistical Models

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

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

$$2\gamma(h) = Var(Y(s+h) - Y(s)), \tag{3}$$

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

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

$$2\hat{\gamma}(h) = \frac{1}{|N(h)|} \sum_{N(h)} \left(Y(s_i) - Y(s_j)\right)^2, \tag{4}$$

where $N(h) = \{(i, j): s_i - s_j = h\}$ and $|N(h)|$ is the number of distinct elements of $N(h)$. For a given distance h, this variogram estimator is a variance estimator over all pairs of observations that are at a distance h apart. Note that when data are irregularly spaced, the variogram is usually smoothed by summing over pairs of points that lie in a tolerance region. $\hat{\gamma}(h)$ is an unbiased estimator of $\gamma(h)$, but is badly affected in presence of outliers because of the $(\cdot)^2$ term in the sum. Therefore, Cressie and Hawkins (1980) have defined a more robust estimator:

$$2\tilde{\gamma}(h) = \left\{ \frac{1}{|N(h)|} \sum_{N(h)} \left|Y(s_i) - Y(s_j)\right|^{1/2} \right\}^4 \Bigg/ \left(0.457 + \frac{0.494}{|N(h)|}\right), \tag{5}$$

which achieves robustness through $|Y(s_i) - Y(s_j)|$. In the presence of outlying observations this estimator is more stable.

The second step of the procedure consists of fitting a parametric model to the empirical variogram (either classical or robust). The most popular variogram models include the exponential variogram defined by

$$\gamma(h;\vartheta) = \begin{Bmatrix} 0 & \text{if } h = 0 \\ c_0 + c_e\left(1 - \exp(-h/a_e)\right) & \text{if } h \neq 0 \end{Bmatrix}, \tag{6}$$

where $\vartheta = (c_0, c_e, a_e)'$ with $c_0 \geq 0$, $c_e \geq 0$ and $a_e \geq 0$, and the spherical variogram defined by

$$\gamma(h;\vartheta) = \begin{Bmatrix} 0 & \text{if } h = 0 \\ c_0 + c_s\left\{(3/2)(h/a_s) - (1/2)(h/a_s)^3\right\} & \text{if } 0 < h \leq a_s \\ c_0 + c_s & \text{if } h > a_s \end{Bmatrix}, \tag{7}$$

where $\vartheta = (c_0, c_s, a_s)'$, with $c_0 \geq 0$, $c_s \geq 0$ and $a_s \geq 0$. The parameter c_0 is the limit of $\gamma(h)$ when $h \to 0$ and is called the "nugget effect". The other parameters in ϑ control the functional form of $\gamma(h;\vartheta)$ (see Cressie, 1993, pp. 61-63, for details). The parametric variograms can be fitted to data by several procedures, which include – among others – (restricted) maximum likelihood and generalized least squares.

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

$$\hat{Y}(s_0) = \lambda' Y = \sum_{i=1}^{n} \lambda_i Y(s_i), \tag{8}$$

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

$$E(Y(s_0) - \sum_{i=1}^{n} \lambda_i Y(s_i))^2 . \tag{9}$$

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

When the process is not stationary, a preliminary step can be performed to capture the trend. For instance, one can first fit a regression model and then compute the variogram

on the residuals of this regression. The final predictions are then computed by adding the kriging predictions on residuals to the fitted values of the regression. This can introduce some bias, and a GLS iterative procedure could be constructed based on the covariance matrix resulting from a variogram model (see Basu and Thibodeau, 1998). The bias is of order $1/n$, making this issue of little concern in our case given that our sample is fairly large.

A local version of Dubin's kriging model could in theory produce better results than the kriging methods described above (Case et al., 2004). In contrast to the standard kriging approach, the local method involves estimating a separate variogram for each property, using a subsample of the data containing neighboring properties. We have experimented with the local method and concluded that it is unstable with respect to our data; in many cases it works well, but in some cases it does not work at all. This may have two causes. First, the very flat empirical variograms indicate that we do not have much correlation. This holds true whether we define the relevant neighborhoods in terms of radial distance from each property or in terms of the number of nearby properties included in the subsample. Second, the properties are not spread uniformly around the city, leading the procedure to crash when the coverage in a given area is sparse.

To implement the geostatistics approach, we used the S+Spatial Toolbox of the commercial software Splus. Several other free or commercial software packages are available.

4. Research Design

The main objective of the paper is to contrast the out-of-sample accuracy in house price predictions of several alternative specifications. We consider eight different techniques: two OLS approaches, four types of geostatistical models, and two types of lattice models. One of the two OLS methods involves adjusting the predictions by the unweighted average residuals for each valuer-defined submarket; the other OLS method does not make this adjustment. The four geostatistical models involve estimation of exponential and spherical variograms as well as robust versions of those. The two types of lattice models are the conditional and simultaneous autoregressive estimators. Each of

the eight models is estimated with and without a set of dummy variables for valuer-defined submarkets.

For each model, we estimate hedonic regressions using 100 randomly selected samples of our data each containing 80% of the observations. Hence the methods are compared based on the same 100 samples of data. For each of the 100 splits, out-of-sample predictions are generated for the remaining 20% of data. We then calculate error statistics and the proportion of predictions that are within 10% and 20% of the sale prices. This is done for each of the 100 samples and the medians of the error statistics and prediction accuracy proportions are calculated for each model. These form the basis for our comparisons.

The source of data for this study is the official database of all real estate transactions in New Zealand. We use data pertaining to detached dwellings only. We focus on sales in the City of Auckland in 1996. A total of 4,880 transactions were retained for the analysis.[4] The database contains the date of sale, the sale price, and such information as: exact location, floor area, age, wall material and condition, and quality of the principal structure. The land area is provided for 76% of the units. The units for which no land area is provided are generally "cross-leased", which means that the land is owned collectively by the owners of the dwellings on that site. The collective owners lease a fraction of the land to each individual owner for a "peppercorn", or nominal, rent. For all such cross-leased dwellings, we set land area equal to zero and set a dummy variable equal to one. Supplementary information used for mass appraisal purposes is also available. These data include important characteristics such as water views, and the quality of landscaping and of the neighborhood. We use the sale price net of the value of any chattels as the dependent variable in our hedonic models.

The data were supplemented with the distance between each property and the central business district (CBD). In addition, we use geographical areas defined by official valuers as our spatial submarkets. These areas, referred to as "sales groups", were defined for mass appraisal purposes and are considered to be relatively homogeneous submarkets. For estimation purposes, we combined three sales groups located in or near

the CBD because they had relatively few transactions; these form the default category. Finally, quarterly time dummy variables are included in each model.

Some variables were transformed before entering into the estimations. We use the natural logarithms of the dependent variable, house price, as well as both land and floor area and distance to the CBD.[5] In addition, both age and age squared are included in the model as the relation between house value and age is expected to follow a U-shaped curve. Table 1 contains the means of the raw independent and dependent variables used in the analyses.

[Table 1 here]

5. Empirical Analysis

Table 2 contains examples of hedonic regression results using OLS with and without submarket dummy variables for a random sample of 80% of the data. The adjusted R^2 statistic increases from 0.722 to 0.798 when the 33 submarket dummy variables are added to the model. The results are consistent with expectations. The logarithms of land and floor area are positively related to sale price, as is the square of the age of the property. Age itself is negatively related to the sale price. The quality and condition of the properties are also important. The logarithm of distance to the CBD is negatively related to sale price and is highly significant. The sale price is approximately 10% higher for properties with a water view. Good landscaping, the number of attached garages, and to a lesser extent, a driveway, significantly affect dwelling prices. The quality of the neighborhood is very important, and higher quality levels are associated with higher prices. In the model with submarket dummy variables, the estimated coefficients and standard errors imply significant differences across valuer-defined areas. When such variables are included in the model, there is a decline in the percentage price impact of being in better neighborhoods. This would be expected as submarket variables will capture part of the variation in neighborhood quality.

[Table 2 here]

We perform two exploratory analyses to determine whether spatial dependence exists in our data. First, we depict the error structure of the OLS regression that does not

include submarket variables by constructing a semivariogram (Figure 1). We also investigate the median of the residuals for various *x* and *y* coordinates, respectively (Figure 2). For this purpose, we divide the city into a grid of 19 cells from west to east by 13 cells from south to north (the west-east dimension is greater than the south-north dimension in Auckland). The semivariogram shows that there is covariance in the error structure and that this declines with distance (the semivariogram increases with distance). Figure 2 shows that the median of residuals is not constant across geographical areas. In particular, residuals tend to be negative to the west and south and positive to the east and north. Hence, spatial dependence clearly exists in the error structure of the OLS model.

[Figures 1 and 2 here]

The next step is to determine how best to account for such dependence to obtain more accurate house price predictions. Table 3 reports comparative statistics for the various models: medians of the average absolute errors, average absolute relative errors, average squared errors, and percentages of predictions within 10% and 20% of the actual price. Figure 3 displays the median and distribution of the proportion of predictions within 10% and 20% of sale price. Results for six methods are shown: OLS with and without adjustment by the average residuals in submarkets, the exponential and robust exponential variogram models, and the CAR and SAR models. Predictions with the spherical models are not reported as they are very similar to predictions with the exponential variogram model. Both parts of Figure 3 depict boxplots of results with and without submarket dummy variables.[6]

[Table 3 and Figure 3]

Referring to Table 3, the OLS model without submarket dummy variables exhibits higher average absolute errors than the geostatistical models. A simple adjustment to the OLS predictions using the unweighted average residuals for submarkets reduces the absolute errors, but these remain higher than with the geostatistical models. The CAR and SAR methods yield larger errors than the unadjusted OLS predictions. When submarket variables are added to the OLS model, the absolute errors are lower than those for the geostatistical models without the submarket dummies. However, the average squared errors for the OLS models containing submarket dummy variables are comparable to those with the geostatistical models, indicating that there is a greater

occurrence of large errors with the OLS models. The average absolute errors are very similar when submarket dummy variables are included in all of the models.

The median of predictions within 10% of actual price across the 100 splits is 39.8% for the OLS model without submarket dummy variables.[7] The geostatistical models yield significantly more accurate predictions (median of approximately 44%). Consistent with the absolute error statistics, a simple adjustment to the OLS predictions using the average residuals in submarkets yields predictions that are only marginally less accurate than the predictions generated using the geostatistical models. Also, the CAR and SAR methods produce predictions that are worse than the unadjusted OLS predictions. This is because the predictions are computed as $\ln Y = X\hat{\beta}$. While $\hat{\beta}$ is estimated more efficiently than with OLS, we are unable to take into account the spatial weight matrix when making ex-sample predictions. This implies that the lattice models as they are now implemented in existing software are not well suited for mass appraisal purposes. Similar results are obtained for predictions within 20% of actual price. Some 73.8% of adjusted OLS predictions are within 20%, which is only marginally less than with the geostatistical models (approximately 75.5%). Again, CAR and SAR models yield the least accurate predictions.

When submarket variables are added to the OLS model, the median of predictions within 10% of actual price rises from 39.8% to 46.8% and that of predictions within 20% from 68.7% to 77.9%.[8] Comparing these results with those for the spatial statistical models that exclude submarket dummy variables leads to the conclusion that the geographical subdivisions used by appraisers improve the accuracy of house price predictions more than do the spatial models. In other words, the use of geographical submarkets appears to be more important in predicting house prices than the more fluid approach which permits "submarkets" to vary from house to house. This conclusion is of practical importance, as a hedonic model with submarket dummy variables is substantially easier to implement than spatial statistical methods. Nevertheless, adding submarket variables to the geostatistical and lattice models results in considerable improvement in predictive accuracy relative to the same models estimated without those variables. In comparison with the OLS model that includes submarket variables, the geostatistical models yield slightly higher percentages of predictions within the 10% and

20% limits. It is likely that the geostatistical estimations that incorporate submarket variables are superior to the corresponding OLS model because, to use Can's (1992) terminology, the former capture adjacency effects as well as neighborhood effects. The CAR and SAR models, in contrast, offer little or no increase in prediction accuracy relative to the OLS model.

The good performance of the OLS models could be a consequence of the quality of the data. One possibility is that the predictions benefited from the extensive set of property characteristics available in the transactions data. This does not appear to be the case, however, as our results were largely unchanged when we re-estimated the models after removing a number of property characteristics.[9] Another possibility is that the characteristics are measured more accurately than typical. Testing this conjecture would require data from another jurisdiction. Another, more likely, possibility is that the good results obtained for the OLS predictions with submarket dummy variables are due to the fact that the submarkets have been carefully outlined and hence capture much of the spatial dependence in house prices in Auckland.

6. Conclusions

The price of a house is related to the prices of adjacent properties. If a hedonic model cannot perfectly capture the effects of location, then the residuals of adjacent properties will be correlated. The aim of this paper is to consider how best to take into account this spatial dependence in a mass appraisal context. We investigate whether spatial statistical models that can be estimated using existing software perform better than an OLS model with neighborhood dummy variables. The comparison is therefore about whether the structure of the errors has to be modeled or whether neighborhood variables can be used. This is also an issue of ease of use as the latter approach is simpler to implement.

We use a rich database of over 4,800 residential sales in Auckland, New Zealand. Two variations each of two OLS, four geostatistical, and two lattice models are considered. Our results suggest that the geostatistical methods perform better than the simple OLS model, but that a simple adjustment of predictions using the average residuals in neighborhoods (submarkets) is almost as good. When submarket dummy

variables are added to the OLS model, the predictions are more accurate than the predictions generated with the geostatistical methods. The lattice models perform poorly, in most cases worse than the unadjusted OLS predictions. However, when submarket dummy variables are added to the geostatistical models, they perform better than the augmented OLS models. When the lattice models include submarket dummies, they perform about the same as the corresponding OLS models.

We conclude that, relative to a simple OLS model, the benefits from incorporating submarket dummy variables are greater than the benefits from using more complicated techniques that attempt to model the structure of errors. For example, the percentage of ex-sample predictions within 10% of the actual sale price increases from 39.8% for the simple OLS model to 46.8% for the OLS model that includes submarket variables. In contrast, the geostatistical models increase the accuracy to about 44%, while the lattice models have a lower accuracy rate than the simple OLS model. This suggests that, for our data at least, the valuer-defined geographical submarkets are more useful in a mass appraisal context than the more fluid concept of submarkets implied by formal modeling of the spatial dependence of residuals. This appears to differ from Case et al.'s (2004) conclusions, although our methods are not the same as theirs and, as noted above, they do not report predictions from their OLS estimation with census tract dummies.

This work could be expanded in a number of ways. First, we could compare our results with those obtained using various methods that focus on measuring the impact of location more effectively in $\mu(X)$. Our best OLS model yields a proportion of predictions within 10% of house values that is just shy of 47%. Fik, Ling and Mulligan (2003) show that when x and y coordinates are interacted with a limited set of independent variables in a hedonic equation of residential units in Tucson, the proportion of predictions within 10% of actual price is 65%. It would be interesting to apply their method to the Auckland data used here. Second, the models that we consider could be estimated for a city where property attributes are measured less accurately than is the case in Auckland. Finally, it might be useful to apply our models to another city where the submarkets used for mass appraisal purposes are defined less carefully. It may be that spatial statistical methods yield better forecasts in such an environment.

Acknowledgments

We thank John Clapp, Xavier de Luna, and two anonymous referees for useful comments.

Notes

[1] An alternative approach would be to incorporate variables measuring neighborhood characteristics (as in Dubin, 1988, for example). Such data would typically be available for small areas defined for census purposes. However, census areas are less likely to correspond to housing submarkets than are the valuer-defined areas used here.

[2] Bourassa, Hoesli and Peng (2003) show that the latter method, although quite simple, results in significant improvements in the accuracy of predictions based on a market-wide hedonic equation; thus we test for its impact here.

[3] This code is available at http://www.spatial-statistics.com. Other general results on sparse matrices exist; see, for example, Bai and Golub (1997) and Reusken (2002).

[4] A sale was removed from the sample if it fell into one of the following categories: (a) the property had a land area larger than 0.25 hectares (this excluded properties that may have been sold primarily for redevelopment purposes); (b) the property had a floor area either less than 30 square meters (probably due to an error in data entry); (c) the transaction was flagged as not being "arm's length"; or (d) the property was located on an island.

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

[6] The body of each boxplot is constructed from the first to the third quartiles, while the whiskers are set at ±1.5 times the inter-quartile range from the median. However, if no observation exists at that distance, the whisker is set at the closest observation towards the body of the boxplot. If there are observations outside of the whiskers, each of these is depicted by a line. Within the body of each boxplot, the median over the 100 splits appears as a bar, while the 95% confidence interval is depicted by the indented and unshaded area in the center of the plot.

[7] For comparison purposes, Fik, Ling and Mulligan (2003) note that Freddie Mac prefers to have at least 50% of predictions within 10% of the actual values.

[8] The adjustment using average residuals in submarkets is ineffective when submarket dummy variables are included in the model because the average of residuals for each submarket equals zero.

[9] In the list of variables in Table 2, we deleted variables from "Walls in good condition" to "Average quality of the principal structure" and from "Water view" to "Very good neighborhood."

References

Bai, Z. and G. H. Golub (1997). "Bounds for the Trace of the Inverse and the Determinant of Symmetric Positive Definite Matrices," *Annals of Numerical Mathematics*, 4, 29-38.

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

Bourassa, S. C., F. Hamelink, M. Hoesli and B. D. MacGregor (1999). "Defining Housing Submarkets," *Journal of Housing Economics*, 8(2), 160-183.

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

Can, A. (1992). "Specification and Estimation of Hedonic Housing Price Models," *Regional Science and Urban Economics*, 22(3), 453-474.

Can, A. and I. Megbolugbe (1997). "Spatial Dependence and House Price Index Construction," *Journal of Real Estate Finance and Economics*, 14(1/2), 203-222.

Case, B., J. Clapp, R. Dubin and M. Rodriguez (2004). "Modeling Spatial and Temporal House Price Patterns: A Comparison of Four Models," *Journal of Real Estate Finance and Economics*, 29(2), 167-191.

Clapp, J. M. (2003). "A Semiparametric Method for Valuing Residential Locations: Application to Automated Valuation," *Journal of Real Estate Finance and Economics*, 27(3), 303-320.

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

Cressie, N. (1993). *Statistics for Spatial Data*, Wiley: New York.

Cressie, N. and D. M. Hawkins (1980). "Robust Estimation of the Variogram, I," *Journal of the International Association for Mathematical Geology*, 12(2), 115-125.

Dubin, R. A. (1988). "Estimation of Regression Coefficients in the Presence of Spatially Autocorrelated Error Terms," *Review of Economics and Statistics*, 70(3), 466-474.

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

Dubin, R., R. K. Pace and T. G. Thibodeau (1999). "Spatial Autoregression Techniques for Real Estate Data," *Journal of Real Estate Literature*, 7(1), 79-95.

Fik, T. J., D. C. Ling and G. F. Mulligan (2003). "Modeling Spatial Variation in Housing Prices: A Variable Interaction Approach," *Real Estate Economics*, 31(4), 623-646.

Getis, A. and J. Aldstadt (2004). "Constructing the Spatial Weights Matrix Using a Local Statistic," *Geographical Analysis*, 36(2), 90-104.

Haining, R. (1990). *Spatial Data Analysis for the Social and Environmental Sciences*, Cambridge: Cambridge University Press.

LeSage, J. P. and R. K. Pace (2004). "Introduction," in *Spatial and Spatiotemporal Econometrics*, LeSage, J. P. and R. K. Pace (Eds), Advances in Econometrics, Volume 18, Oxford: Elsevier, 1-32.

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

Militino, A. F., M. D. Ugarte and L. García-Reinaldos (2004). "Alternative Models for Describing Spatial Dependence among Dwelling Selling Prices," *Journal of Real Estate Finance and Economics*, 29(2), 193-209.

Pace, R. K. and R. Barry (1997a). "Quick Computation of Regressions with a Spatially Autoregressive Dependent Variable," *Geographical Analysis*, 29(3), 232-247.

Pace, R. K. and R. Barry (1997b). "Sparse Spatial Autoregressions," *Statistics and Probability Letters*, 33(2), 291-297.

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

Pace, R. K., R. Barry and C. F. Sirmans (1998). "Spatial Statistics and Real Estate," *Journal of Real Estate Finance and Economics*, 17(1), 5-13.

Pace, R. K. and O. W. Gilley (1997). "Using the Spatial Configuration of the Data to Improve Estimation," *Journal of Real Estate Finance and Economics*, 14(3), 333-340.

Pace, R. K. and O. W. Gilley (1998). "Generalizing the OLS and Grid Estimators," *Real Estate Economics*, 26(2), 331-347.

Palm, R. (1978). “Spatial Segmentation of the Urban Housing Market,” *Economic Geography*, 54(3), 210-221.

Reusken, A. (2002). “Approximation of the Determinant of Large Sparse Symmetric Positive Definite Matrices,” *SIAM Journal on Matrix Analysis and Applications*, 23(3), 799-812.

Ripley, B. (1981). *Spatial Statistics*, New York: Wiley.

Ugarte, M. D., T. Goicoa and A. F. Militino (2004). “Searching for Housing Submarkets using Mixtures of Linear Models,” in *Spatial and Spatiotemporal Econometrics*, LeSage, J. P. and R. K. Pace (Eds), Advances in Econometrics, Volume 18, Oxford: Elsevier, 259-276.

Table 1. Sample statistics

Variable	Mean
Net sale price (NZ$)	328,398
Age of dwelling	46.45
Land area (10s of square meters)	55.42
Cross-leased	0.24
Floor area (square meters)	144.42
Wall condition (proportion)	
Good	0.39
Average	0.58
Bad	0.03
Roof material (proportion)	
Tile	0.41
Metal	0.55
Other	0.04
Wall material (proportion)	
Wood	0.63
Brick	0.13
Fibrolite	0.06
Other	0.18
Quality of the principal structure (proportion)	
Superior	0.19
Average	0.76
Poor	0.05
Distance to CBD (km)	6.79
Water view (proportion)	0.09
Modernization (proportion)	0.26
Landscaping (proportion)	
Good	0.16
Average	0.79
Poor	0.05
Driveway	0.85
Quality of the neighborhood (proportion)	
Very good	0.03
Good	0.20
Average	0.68
Poor	0.09
Number of attached garages	0.75
Sample size	*4,880*

Table 2. Examples of results for OLS estimations without and with submarket dummy variables

Variables	Estimates for equation without submarket dummies	Estimates for equation with submarket dummies (estimates for submarket dummies are preceded by the sales group number)	
Intercept	12.816**	11.344**	(4) -0.206**
Log of floor area	0.690**	0.457**	(5) -0.053
Log of land area	1.898**	2.560**	(6) -0.009
Cross-leased or strata-titled	0.099**	0.128**	(7) 0.226**
Age of dwelling	-0.003**	-0.004**	(8) 0.031
Age of dwelling squared	4.720×10^{-5}**	5.096×10^{-5}**	(9) 0.176**
Walls in good condition	0.088**	0.083**	(10) 0.165**
Walls in average condition	0.063**	0.052**	(12) -0.062
Dwelling with a tile roof	-0.030	-0.028	(13) 0.103**
Dwelling with a metal roof	-0.068**	-0.041*	(14) 0.214**
Dwelling with wooden walls	-0.015	-0.006	(15) -0.059
Dwelling with brick walls	-0.056**	-0.019	(16) -0.379**
Dwelling with fibrolite walls	-0.102**	-0.049**	(17) -0.237**
Superior quality of the principal structure	0.231**	0.124**	(18) -0.410**
Average quality of the principal structure	0.094**	0.050**	(19) -0.322**
Log of distance to the CBD	-0.172**	-0.137**	(22) -0.647**
Quarter 2	0.008	0.014	(23) -0.117**
Quarter 3	-0.020*	-0.019*	(24) -0.006
Quarter 4	0.015	0.017	(25) 0.177**
Water view	0.103**	0.079**	(26) 0.181**
Modernization	0.034**	0.029**	(27) -0.096*
Average landscaping	0.026	0.013	(28) -0.315**
Good landscaping	0.077**	0.060**	(29) -0.220**
Driveway	0.019	0.010	(30) -0.141**
Average neighborhood	0.098**	0.021	(31) -0.119**
Good neighborhood	0.231**	0.067**	(32) -0.130**
Very good neighborhood	0.323**	0.205**	(33) -0.189**
Number of attached garages	0.039**	0.036**	(34) -0.010
			(35) -0.074
			(37) -0.233**
			(38) -0.268**
			(39) -0.225**
			(53) -0.274**
Adjusted R^2	*0.722*	*0.798*	

Note: The symbols * and ** denote significance at the 5% and 1% levels, respectively.

Table 3. Median error and prediction accuracy statistics

Statistic	OLS	Adjusted OLS	Expo-nential	Robust expo-nential	CAR	SAR
Average absolute error						
Without submarkets	58,521	53,168	51,586	51,377	61,915	60,534
With submarkets	49,901	49,901	47,963	47,464	49,948	49,074
Average absolute relative error						
Without submarkets	17.8%	16.3%	15.9%	15.8%	18.4%	18.1%
With submarkets	15.0%	15.0%	14.5%	14.3%	15.0%	14.6%
Average squared error (millions)						
Without submarkets	9,214	7,696	7,496	7,472	10,718	10,197
With submarkets	7,514	7,514	7,056	6,942	7,628	7,373
Predictions within 10%						
Without submarkets	39.8%	43.3%	44.0%	44.2%	38.9%	39.2%
With submarkets	46.8%	46.8%	49.0%	49.3%	47.4%	48.0%
Predictions within 20%						
Without submarkets	68.7%	73.8%	75.5%	75.6%	66.8%	67.7%
With submarkets	77.9%	77.9%	79.3%	79.7%	77.7%	78.7%

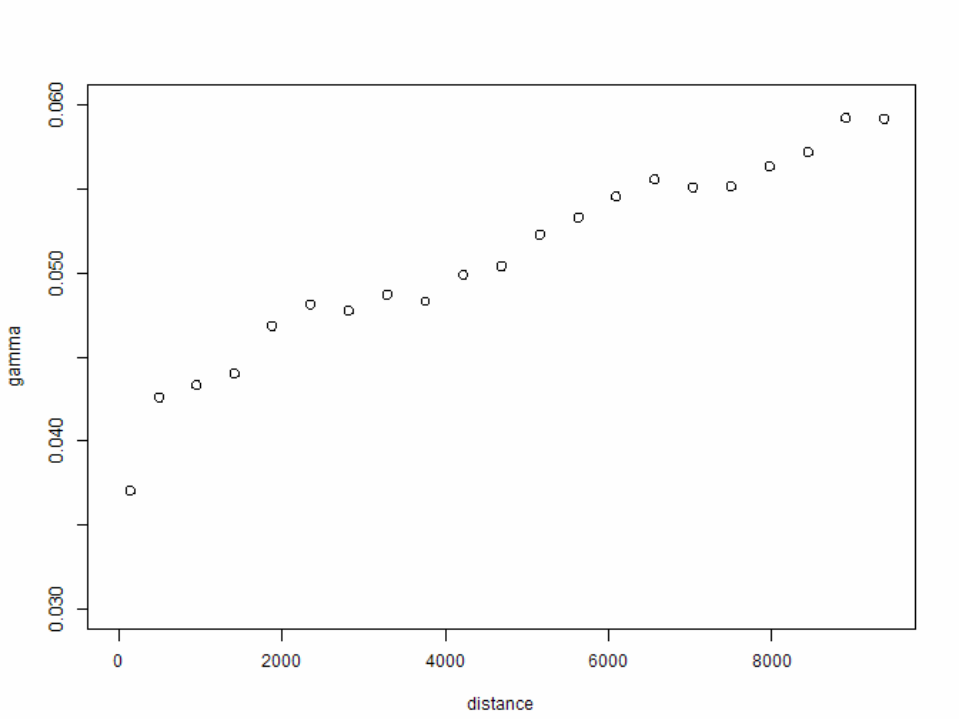

Figure 1. Empirical semivariogram

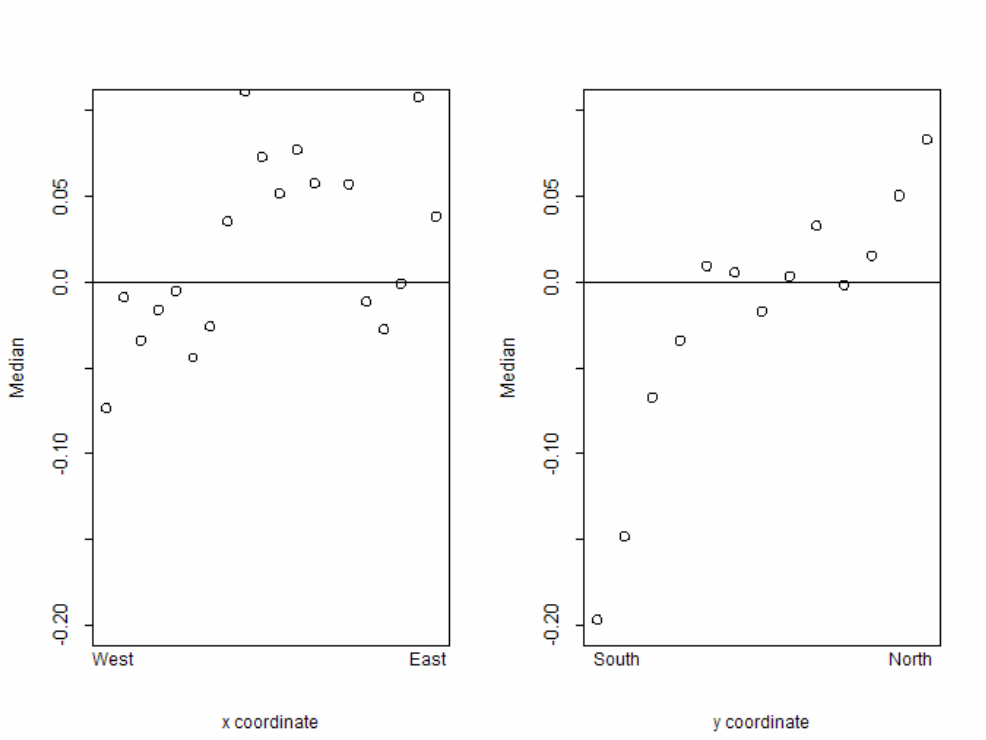

Figure 2. Median values of residuals

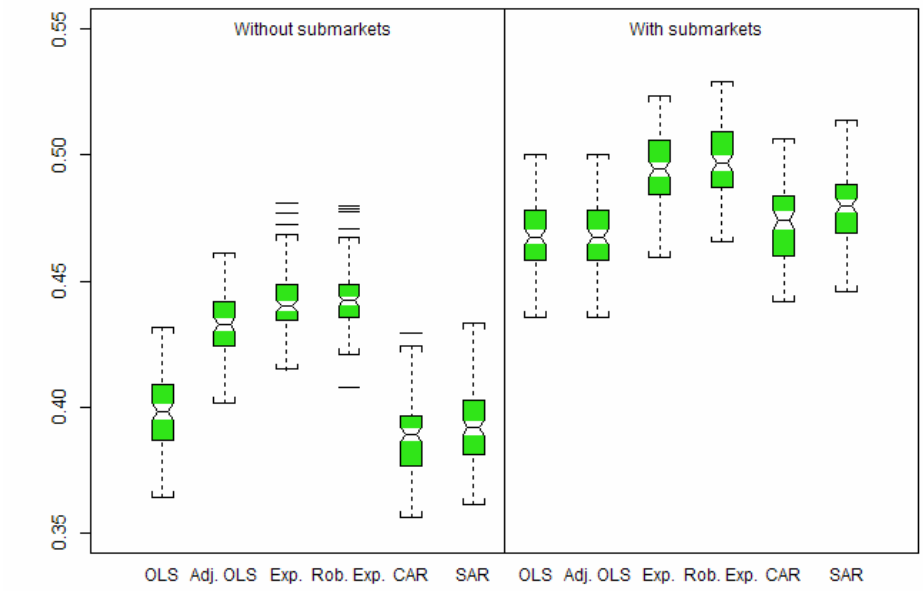

Figure 3a. Proportion of predictions within 10% of actual value

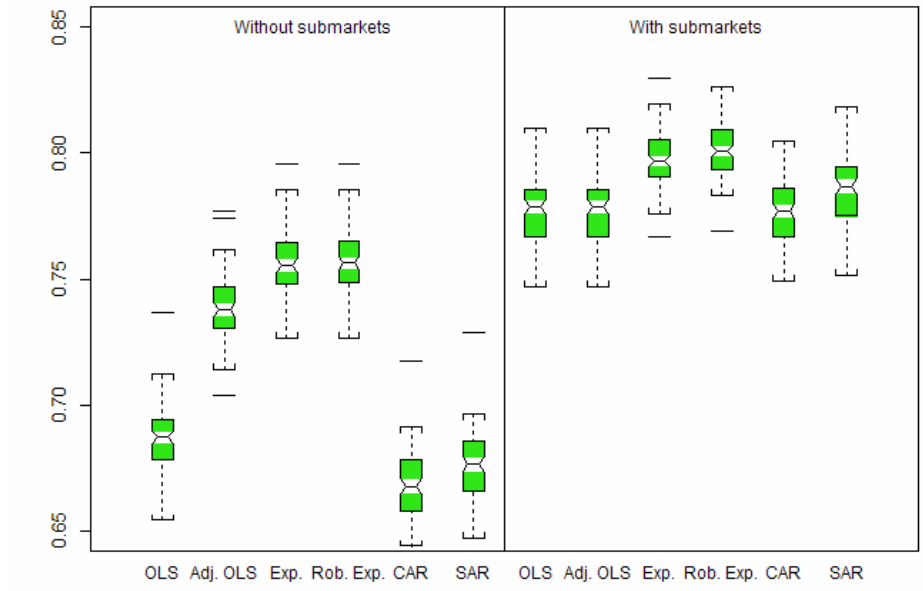

Figure 3b. Proportion of predictions within 20% of actual value