2000 V28 2: pp. 249–283
REAL ESTATE
ECONOMICS

Space-Varying Regression Coefficients: A Semi-parametric Approach Applied to Real Estate Markets

Andrey D. Pavlov*

This paper presents a method for estimating home values by non-parametrically incorporating the physical location of the properties. Specifically, I allow the parameters of the observed covariates to vary in space. This approach mitigates one of the biggest deficiencies inherent in hedonic pricing models—omitted variables. I demonstrate the advantages of the proposed method using real estate transaction data from Los Angeles County. The estimation finds a substantial spatial variation of the marginal values of the hedonic characteristics and provides an insight into the segmentation of the market. The proposed method is an extension of semi-parametric multi-dimensional *k*-nearest-neighbor smoothing. It alleviates a fundamental problem known as the curse of dimensionality by incorporating parametric components into a non-parametric estimation.

"Location, location, location" is a frequent explanation of home values. Take a Beverly Hills home and move it to Watts, and it loses most of its value. This cliché still finds its way into real estate valuation literature because location is used as a proxy for numerous unobserved variables. The standard hedonic approach to property valuation assumes that the price of a home is a function of the values markets place on its characteristics (physical and locational). If one knew the hedonic relationship's functional form and had measures of all relevant characteristics (*i.e.*, had a fully specified model), one would not need spatial modeling at all. One of the most significant drawbacks of hedonic valuation, however, is that we cannot observe or measure all relevant characteristics. This problem is typically addressed by using an artificial measure of location (*e.g.*, zip codes or cities) as a proxy for variations in the unobserved covariates. As demonstrated below, zip codes as a definition of neighborhoods can substantially be improved upon by letting the transaction data define neighborhoods. The method proposed in this chapter incorporates this idea. In particular, I do not specify neighborhoods *a priori*, nor do I assume a specific parametric relationship

*Concordia University, Montreal, PQ H3G 1M8, Canada or apavlov@anderson.ucla.edu.

between location and the value of a home (for example, a proportional shift in value estimated by an indicator variable in a linear regression). Instead, I allow the coefficients of a hedonic model to vary non-parametrically in space.

The proposed approach to estimating a spatially dependent relationship relies on the assumption that the unobserved variables on average change smoothly in space. In the case of residential real estate markets, this is quite plausible. Most of the unobserved characteristics are related to location (*e.g.*, pollution, crime, prestige) and typically do not change much over a short distance (*e.g.*, within a neighborhood).[1] Thus, even if I do not observe these covariates, the knowledge that they are nearly constant over small regions is valuable information, and the space-varying coefficients (SVC) method I propose takes advantage of it.

In the case of Los Angeles County home valuation, the SVC method outperforms the standard hedonic models in terms of cross-validation prediction error. I also find substantial spatial variability in the marginal value of physical characteristics. The SVC approach can further be applied to automated valuation systems. Such an application can quickly and inexpensively estimate the value of every home.

The SVC method can also be used to identify neighborhoods and to examine the influence of neighborhood characteristics on prices. Issues such as property taxation, government-related amenities and infrastructure projects require an estimate of the implicit prices of the characteristics. The proposed approach produces such estimates and explicitly models the spatial distribution of those prices.

In the remainder of this paper, I conceptually motivate the suggested application of the SVC method and review the relevant literature on real estate valuation. The actual estimation procedure is then presented and applied to real estate valuation using transaction data for Los Angeles County. Finally, I compare the proposed application with alternative valuation models and conclude with some generalizations and suggestions for future research.

[1] There may be some discontinuities or house-specific omitted variables. This issue is partially addressed on p. 251.

Motivation

A common way to model real estate markets is to use linear hedonic functions[2]:

$$z_i = a + bs_i + \varepsilon_i, \qquad i = 1, \ldots, N \tag{1}$$

where z is the transaction value (or natural logarithm of the value) of a house, s is the size of the house measured as square feet of the living area, and N is the total number of observations. For expositional purposes I assume that the only observable characteristic is the size of the living area in square feet. The actual estimation described below also includes the number of bathrooms and the number of bedrooms and can easily be extended to include more observed covariates.

A number of other variables influences the value of the house as well as the marginal value of the size of the property. For instance, assume the complete model is

$$z_i = a + bs_i + cl + dls_i + \varepsilon_i, \tag{2}$$

where l is an unobserved variable and c and d are parameters.[3] Since l is unobserved, the model translates into

$$z_i = (a + cl) + (b + dl)s_i + \varepsilon_i \tag{3}$$

As already discussed, l will likely depend on physical location, and is assumed not to change much for short distances. Even if I do not observe l directly, I can benefit from the relationship between l and physical location by allowing the coefficients to depend on that location:

$$z_i = p(x_i, y_i) + q(x_i, y_i)s_i + \varepsilon_i \tag{4}$$

where (x_i, y_i) denotes the coordinates of each observation.

A key assumption imbedded in the above model is that the omitted variables influence not only the intercept, but also the marginal value of the size of the house. This may be justified because, for example, some of the omitted

---

[2] The next section gives background for this model and compares it with some alternatives.

[3] l can be thought of as a vector of unobserved variables. This will not change the model and will have no effect on the empirical estimation.

variables may be related to the quality of construction. An additional square foot of marble floor, for instance, has higher value than an additional square foot of carpeted floor. Since, on average, the variation of quality between neighborhoods is greater than within them, I conjecture that the marginal value of size (as well as many other physical characteristics) will vary in space.

Using the method described below, I estimate p and q as non-parametric functions of the coordinates (x, y). This model can automatically include as many covariates as I observe without increasing the dimensionality of the non-parametric estimation.

Assuming linearity with respect to size may seem restrictive. The theory of hedonic pricing does not prescribe a specific functional form. It suggests, however, that the value of a house is a monotonically increasing function of size. Thus, linear approximation with respect to size is not unreasonable. Since I have no idea how the coefficients vary in physical space, non-parametric modeling of the location is justified.

In theory, one can perform the above estimation by directly applying some non-parametric method to the unknown function:

$$z = f(x, y, s) + \varepsilon \tag{5}$$

This approach does not assume linearity with respect to size, but it presents two substantial difficulties. First, the choice of an appropriate metric in a multi-dimensional space is not trivial. To illustrate the problem I examine the task of estimating home values as a function of distance from the ocean and size. Consider three houses A, B and C. Then assume A and B are close to the ocean, but C is far. On the other hand, A and C are of the same size, but B is smaller. To estimate the value of house A using non-parametric methods I need to take a weighted average of the values of houses B and C. To do this I need to determine which house, B or C, contains more information about house A. The answer, however, will depend on the preference functions of the potential homeowners. Without making further assumptions, this problem cannot be solved. To make matters worse, even if I define a metric, it will depend on the measurement units of the variables.

A typical approach to defining a metric in a multi-dimensional space is to use Euclidean distance. To make it invariant to measurement units, all variables are normalized to have the same variance. In the example above this would mean transforming distance from the ocean and size so that each variable has a sample variance of one. Even though this appears to be a

sensible procedure, a simple example illustrates why it will not solve the problem I discussed above. An argument can be made that the size of the house is really what contains the most information about the value. When valuing a house in this case, one should consider houses of similar size and pay little attention to the distance from the ocean. Through an appropriate specification of the utility functions, however, I can make an equally valid argument that the distance from the ocean contains the information and size is meaningless. In this case, I should consider houses with the same distance to the ocean and ignore the size. Thus, two researchers using the same data set will obtain substantially different results depending on the metric they choose. A solution is to use cross-validation to choose the optimal metric. This, however, amounts to minimizing a poorly behaved function in as many dimensions as the number of included variables, which may not be feasible.

Furthermore, normalizing all variables to have the same variance (*e.g.*, one) makes the metric extremely sensitive to outliers and undermines one of the biggest advantages of non-parametric estimation. If, for instance, I have a few extremely large houses, normalizing the variance to one will make all other houses appear to be of similar size.

The second problem of direct non-parametric estimation is known as the *curse of dimensionality.* Let me give a simple example. A possible procedure for estimating a surface in two dimensions is to create a 10×10 rectangular grid. In this case, each inner square will have eight neighbors. If I carry out this procedure in 5 dimensions, there will be a total of 10^5 hyper-rectangles, and each inner cell will have $3^5 - 1 = 242$ neighbors. With a sample size of 1,000 one would have to average over 60% or more of the observations to obtain reasonable estimates. Cleveland and Devlin (1988) recommend smoothing over at least 30% of the observations. Meese and Wallace (1991) estimate non-parametric hedonic functions in six dimensions. To obtain reasonable estimates, however, they average over 100% of the observations.

Frequently, the curse of dimensionality and the definition of a metric make direct non-parametric estimation in many dimensions infeasible. Furthermore, since I am targeting the problem of omitted variables rather than the issue of functional form, the deficiencies of the multi-dimensional non-parametric methods make a semi-parametric approach like the one I am proposing more appealing.

Real Estate Valuation Background

A common method used for estimating demand, constructing home value indexes or analyzing the effect of neighborhood characteristics is to estimate

hedonic price functions. Although conceptually attractive (Rosen 1974), hedonic modeling has two substantial drawbacks. The first is misspecification of the functional form. Meese and Wallace (1991) address this issue through direct non-parametric estimation.

The second drawback of hedonic modeling is its sensitivity to omitted variables. Two approaches have been developed to alleviate this problem. The first one uses repeated sales of the properties to estimate indexes (Case and Shiller 1987, 1990; Goetzmann and Spiegel 1997). This solves the problem of having appropriate controls (assuming properties have not changed between sales), but at a high price: valuable information is thrown away. The second approach is to reduce the sensitivity of the hedonic model to omitted variables. By incorporating physical location into a hedonic model, the SVC technique is an example of the second approach. However, the SVC method can be applied to repeat sales or hybrid models as well. For a detailed comparison of hedonic vs. repeat sales approaches see Case, Pollakowski and Wachter (1991).

Several basic methods have recently been used to incorporate location into hedonic models. A simple and widely used approach is to include indicator variables for zip codes (or some other *a priori* definition of neighborhood) and interaction effects between these zip codes and the physical characteristics of a house (Cauley 1997).

Other studies estimate a standard hedonic regression and then correct for spatially correlated errors (Dubin 1988, 1992; Lai and Wang 1996; Basu and Thibodeau 1997; Pace and Gilley 1997). Essentially, this allows the intercept to vary in space, but the other parameters remain fixed. Such models give us no insight into the spatial variation of the marginal value of the physical characteristics. The proposed method is more general in that it allows all parameters, and not just the intercept, to vary in space. As shown below, these estimates indeed suggest a substantial difference in marginal values across neighborhoods, and allowing all parameters to vary in space improves the out-of-sample performance of the model.

Yet another improvement to standard hedonic modeling is to include the prices of nearby properties in the regression equation (Anselin 1989; Can 1992; Can and Megbolugbe 1997). The drawback of this otherwise appealing approach is that these prices are not adjusted for the characteristics of the nearby properties. Since those characteristics are assumed to be priced correctly, this approach is reasonable if all of the included adjacent properties are very similar to the house in question in their observed characteristics. If

the house in question is away from the mean, however, the only meaningful information in nearby sales is the marginal value of the physical characteristics, not the overall value. The method I propose allows recent nearby sales to influence the estimation for a house, but it explicitly models the variation of the observed characteristics.

Estimation of the SVC Method

The following model is to be estimated:

$$z_i = p(x_i, y_i) + q(x_i, y_i)s_i + \varepsilon_i \tag{6}$$

where z is the transaction value, s is the size of the living area (the actual estimation also utilizes the number of bathrooms and the number of bedrooms), and p and q are coefficients which depend on the location of the house [as determined by its (x, y) coordinates]. The major assumption driving the model is that p and q are smooth in space.

The basic idea of non-parametric smoothing is to weight nearby observations more than remote observations. In particular, I estimate weighted least-squares regression with weights depending on the distance between observations. Formally, I estimate the parameters p and q at a particular location (x, y) by minimizing the weighted residual sum of squares:

$$\min_{p,q} \sum_i \{W_i(x, y) [z_i - p(x, y) - q(x, y)s_i]^2\} \tag{7}$$

where $W_i(x, y)$ are the weights. Notice that these weights depend on the location (x, y) at which the coefficients p and q are estimated. Since these weights will be different for different locations, the coefficient estimates will be different as well. Notice also that the weights depend on the physical location but not on the observed characteristics.

The solution of the above minimization problem in matrix notation is the standard weighted least-squares expression:

$$\begin{pmatrix} p_{(x,y)} \\ q_{(x,y)} \end{pmatrix} = (S' \; W \; S)^{-1} \; (S' \; W \; Z) \tag{8}$$

where S is an $N \times 2$ matrix with first column all ones and second column the sizes of the homes, Z is a column vector of the transaction values of the observations, and W is a diagonal matrix containing the weights. To estimate

the model with more than one independent variable, additional variables are included in the matrix S. Notice that this does not require any change to the weight matrix W.

The weighting scheme I utilize in this application is borrowed from the *k-nearest-neighbor* estimation. The k-nearest-neighbor estimate is a weighted average of a fixed number of observations closest to the house of interest. Thus, the relevant neighborhood of each house varies with the density of the observations. An important property of the k-nearest-neighbor weighting is that it keeps the number of observations that enter the estimation constant in space. This number can and will be varied to optimize the procedure, but this variation will be the same for all points in space.

The k-nearest-neighbor weight sequence $W_{ki}(x, y)$ is defined through the set of indexes

$$J_{x,y} = \{i : X_i, Y_i \text{ is one of the } k\text{-nearest observations to } x, y\} \tag{9}$$

With this set of indexes of neighboring observations the simplest k-nearest-neighbor weight sequence is

$$W_{ki}(x, y) = \begin{cases} 1/k & \text{if } i \in J_{x,y} \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

The above weighting scheme puts the same weight on all observations that fall within a specified neighborhood of the point of interest (x, y). More efficient weighting schemes have been developed by Stone (1977) that differentiate among the observations within the neighborhood of (x, y) and put more weight on nearby observations. In particular, one can use the kernel weights. To illustrate the proposed method I have chosen the parabolic shape of the weighting function (Epanechnikov 1969):

$$K(u) = 0.75(1 - u^2) \qquad I(|u| \leq 1) \tag{11}$$

This shape is commonly used in practice and enjoys some optimality properties (Hardle 1997). The argument u for observation i is typically given by

$$u_i = \frac{\begin{array}{c}\text{distance between observation} \\ i \text{ and the estimation point}\end{array}}{\begin{array}{l}\text{maximum distance between the} \\ \text{estimation point and all of its } k \text{ neighbors}\end{array}} \tag{12}$$

The k-nearest-neighbor weighting scheme was introduced by Loftsgaarden and Quesendberry (1965) for the purposes of density estimation. Lai (1977), Mack (1981), Devroye (1978) and Gyofty (1981) examined the asymptotic properties of the approximation and its rates of convergence to the true relationship. Cleveland and Delvin (1988) describe the *locally weighted regression*, which is essentially a variation of the k-nearest-neighbor smoothing.

A key parameter in the k-nearest-neighbor weighting scheme, and therefore in the whole estimation, is the number k of observations entering each estimation. This scale factor, also called the smoothing parameter or the search radius, determines how far the procedure will look for observations. A small k will force the procedure to consider only a few nearby observations, while a large k will allow distant observations to enter the estimation. If I knew the underlying shape of the function, I could determine the optimal smoothing parameter. Since the whole purpose of non-parametric estimation is to avoid specifying the functional form, there is no universal prescription for the choice of k. A common way to estimate the smoothing parameter for various non-parametric techniques is cross-validation. This procedure involves deleting one observation, estimating its value based on the remaining observations and calculating the square of the difference between the estimate and the true value. Iteration over all observations provides the total sum of cross-validation errors. The next step is to vary the smoothing parameter until this sum is minimized. The specific algorithm I have used is described in the Appendix.

Results Using Transaction Data from Los Angeles

The Data

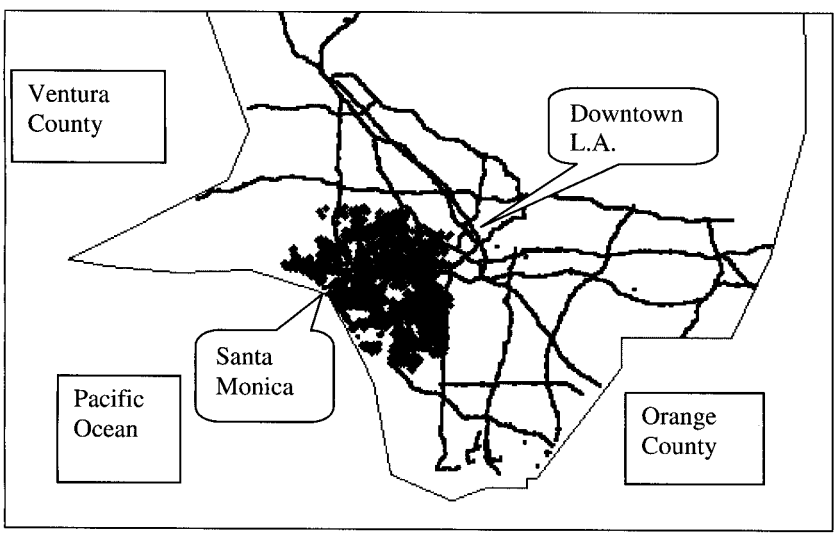
In this section I analyze the Los Angeles real estate market. As discussed below, this market is characterized by substantial variability of home values even within postal zip codes, which makes the application of the method proposed above particularly beneficial.

The dependant variable is the transaction value, and the independent variables are the size of the living area, the number of bedrooms, the number

of bathrooms and the location of the house. The area under consideration is the Los Angeles *West Side*, which includes Beverly Hills, Bel Air, West Los Angeles and Santa Monica. Figure 1 shows the location of the observations. I have screened the data for mistakes and outliers. The data set includes approximately 3,000 transactions that took place during the six-month period between April 1 and September 30, 1997. Transaction value (*i.e.*, sales price) is measured in dollars, size of the living area in square feet, bedrooms and bathrooms as the actual number, and the X, Y coordinates are given in Universal Transverse Mercator (Nad 27 for the U.S.) projection. This projection has been chosen for convenience and has no effect on the estimation.

The data set comes from DataQuick—a company specializing in assembling and providing publicly available data. The underlying data come from the Los Angeles County Recorder. The data set includes the addresses of the individual properties. These addresses were used to geocode each house (*i.e.*, assign X,Y coordinates). I use the TIGER files from the Bureau of Census to geocode the properties. These files contain detailed geographic data, including all streets, freeways, boundaries of administrative regions, etc. I was able to geocode approximately 80% of the transactions. Those constitute the sample of transactions that I use in my analysis.

Figure 1 ■ Location of the observations within Los Angeles County. Each dot represents a single transaction. The lines represent freeways.



To understand the characteristics of the underlying data, I provide summary statistics and figures. Table 1 reports the mean, median, standard deviation and coefficient of skewness of the included variables for the entire area. Figure 2 provides box plots for each variable. As expected, this is a high-priced area with a median home value of nearly $250,000. Furthermore, prices range between 50,000 and 2 million dollars, with a standard deviation of nearly 300,000. This wide dispersion of home values is not surprising, since the area under consideration includes the high-priced Bel Air and Beverly Hills areas as well as the South-Central region. Furthermore, as evident from the summary statistics and the box plots, the home values are highly skewed. This issue is partially addressed in the estimation by utilizing the natural logarithm of the value as a dependent variable. Nevertheless, this skewness and the large number of observations with unusually high transaction prices is an important feature of the data.

Since homes of similar values tend to cluster together in neighborhoods, an examination of the data for the entire region may be misleading. If indeed the local variation of home values is, as expected, substantially smaller than the overall dispersion for the entire area, procedures like the SVC can take advantage of this feature. The first step to examining the spatial properties of the data is to investigate the distribution of the typical prices in space. Figures 3 through 6 depict the spatial distributions of the home values and the independent variables. I have displayed each surface from two viewpoints—one directly above, and one at a 30-degree angle. The shading of the surfaces reinforces the shape. The major freeways in the area are also displayed.

Figures 3 through 6 use simple one-dimensional smoothing to depict the variables [*i.e.*, the only included independent variables are the (x, y) coordinates]. Thus, all figures are smoother than the actual observations and

Table 1 ■ Descriptive statistics of the included variables for the entire area.

Statistic	Value ($)	Size (ft^2)	No. of Bedrooms	No. of Bathrooms
Mean	343,932	1,738	2.85	1.88
Median	248,750	1,497	3	2
Standard deviation	294,092	860	0.89	0.97
Skewness	2.23	1.84	0.79	1.36

Value is the transaction price. Size is the size of the living area.

Figure 2 ■ Standard box plots for the four variables included in the model. As expected, the distribution of home values is highly skewed. To partially remedy this I have used the natural logarithm of the home values as a dependent variable.

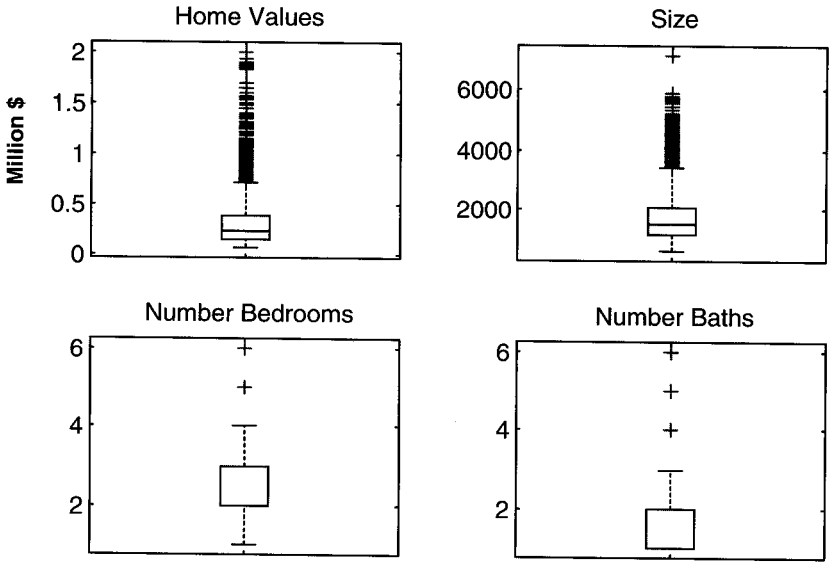


Figure 3 ■ Spatial distribution of observed transaction values. The plots are constructed using simple one-dimensional smoothing. The resulting surface is displayed from two viewpoints: directly above (left) and at a 30-degree angle (right). The shading reinforces the shape.

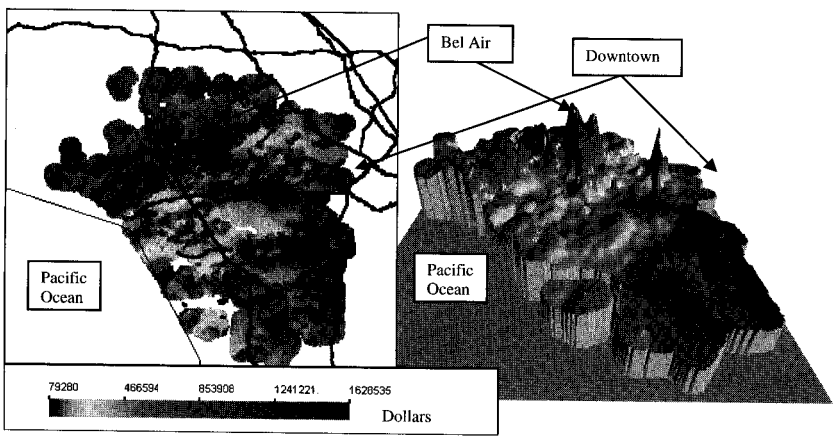
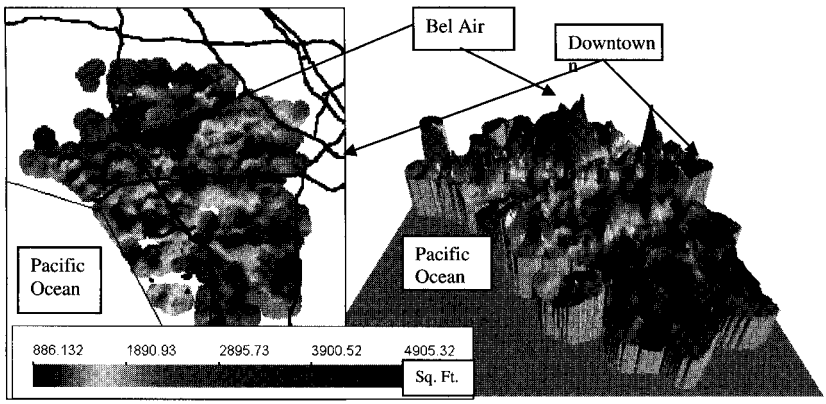


Figure 4 ■ Spatial distribution of the observed size of the living area. The plots are constructed using simple one-dimensional smoothing. The resulting surface is displayed from two viewpoints: directly above (left) and at a 30-degree angle (right). The shading reinforces the shape.



do not capture the local variations. Instead, they provide the mean value for each location. There are several features of the data that will be important in interpreting the results from the estimation presented below. For instance, the highest values are, as expected, in Bel Air. Notice, however, that the homes in Bel Air tend to be larger. Homes directly west of Bel Air (north Santa Monica and Pacific Palisades) are not as expensive, but are

Figure 5 ■ Spatial distribution of the observed number of bedrooms. The plots are constructed using simple one-dimensional smoothing. The resulting surface is displayed from two viewpoints: directly above (left) and at a 30-degree angle (right). The shading reinforces the shape.

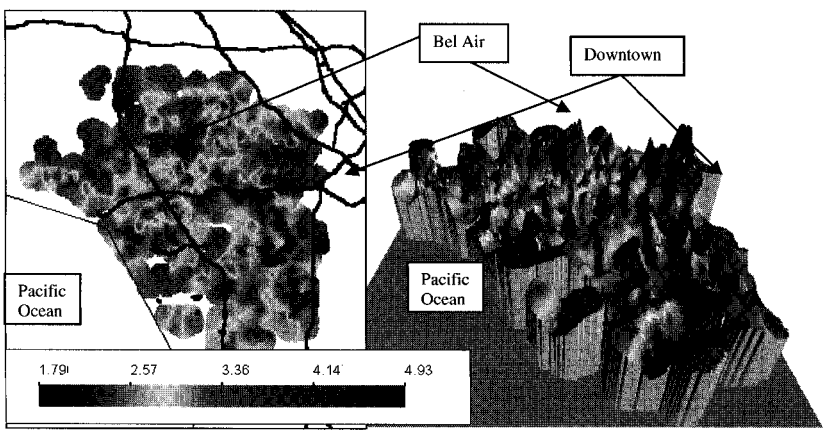
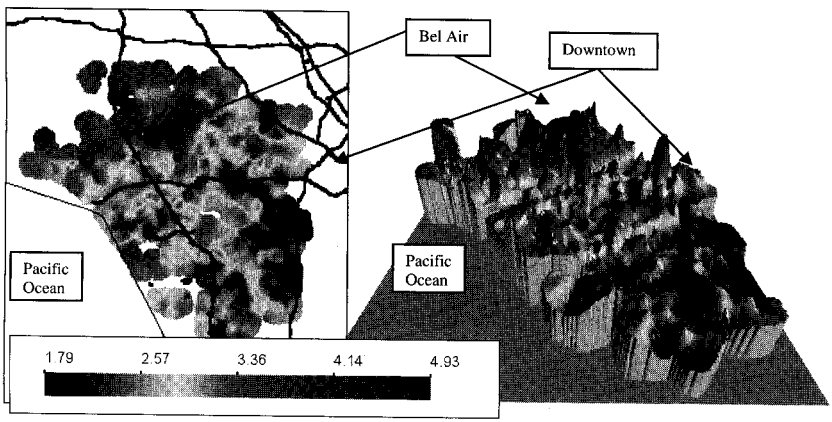


Figure 6 ■ Spatial distribution of the observed number of bathrooms. The plots are constructed using simple one-dimensional smoothing. The resulting surface is displayed from two viewpoints: directly above (left) and at a 30-degree angle (right). The shading reinforces the shape.



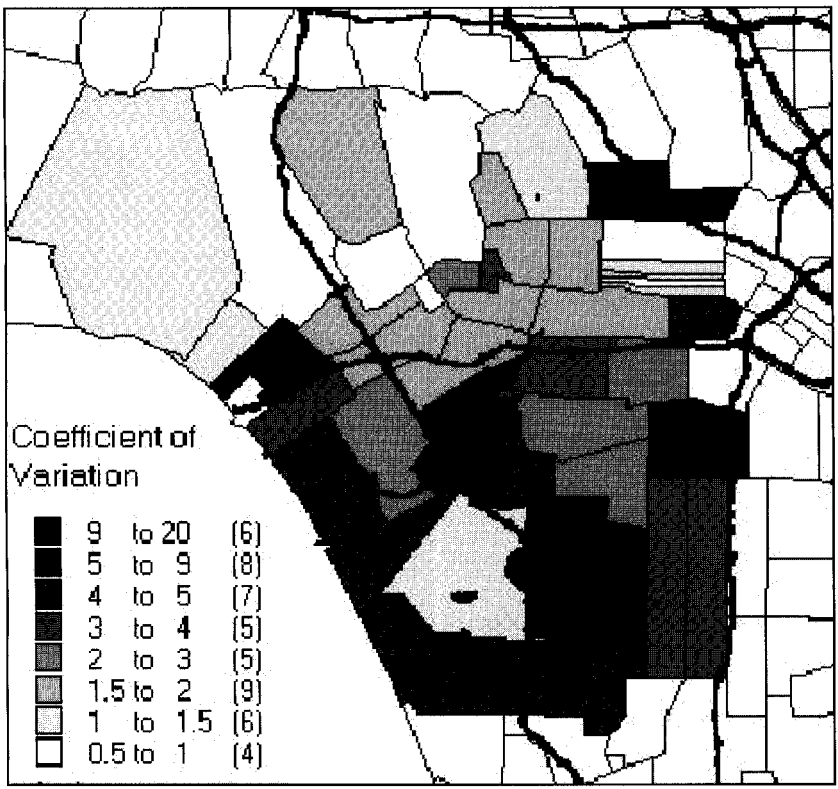
substantially smaller. Thus, just by observing the data, I am unable to determine whether the homes in Bel Air are expensive because they are larger or because of the neighborhood. This is important both for valuing homes that are away from the mean in physical characteristics and for valuing neighborhood amenities.

While Figure 3 provides the average price for each area, it is silent on the local variability of home values. To examine the local variability, I compute the coefficient of variation by postal zip code. Figure 7 presents these coefficients of variation. Notice that the home price variability relative to the mean within a zip code is substantial. For instance, 50% of the zip codes have coefficient of variation above 3. The standard deviation of home values is more than nine times the mean value for six zip codes.

Contrary to my expectations, the lower-priced areas have higher variability relative to the mean for those areas. For instance, ten out of the fourteen zip codes with coefficient of variation above 5 fall within the lower-priced South-Central region. On the other hand, almost all zip codes with coefficient of variation below 3 are within the high-priced areas. Notice, however, that the smallest coefficient of variation is 0.6, even for the very high-priced areas.

Some of the extremely large coefficients of variation may be due to a few outliers or mistakes in the data. To investigate this possibility, Figure 8

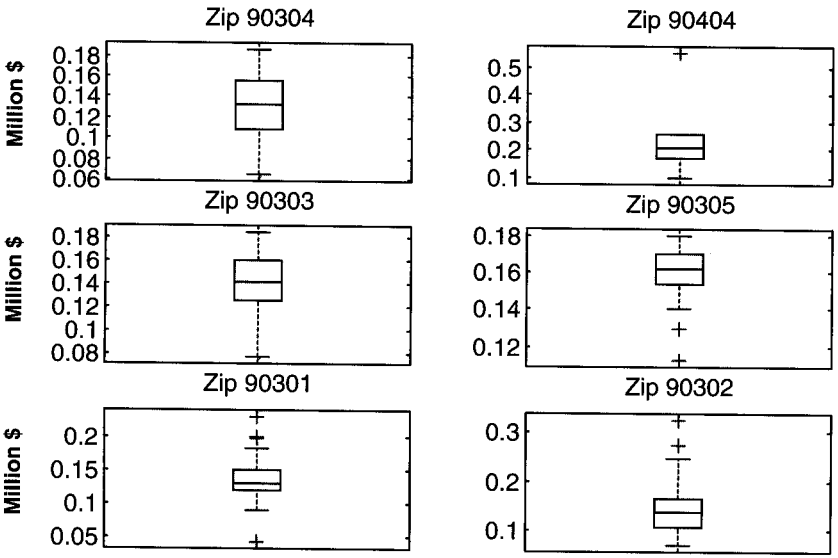
Figure 7 ■ Coefficient of variation of the home values by zip code.



provides box plots of the observations that fall within the six zip codes with largest coefficients of variation. Notice that while the observations within these zip codes are still highly skewed, the large variations are not driven by outliers and are instead characteristic of the data.

The high variability of home values evident from the above analysis suggests that the Los Angeles real estate market is extremely difficult to model. In other words, even models explaining as much as 80% of the variation will have substantial prediction error. Furthermore, the variability within zip codes is substantial, which suggests that the standard approaches of using pre-defined neighborhoods are unlikely to produce reasonable estimates. The individual location of the properties is likely to have a substantial effect on home values, and therefore location needs to be incorporated in a non-parametric fashion into the estimation.

Figure 8 ■ Box plots of the observations in the six postal zip codes with largest coefficients of variation. The main purpose of these plots is to show that the large coefficients of variation are not due to particular outliers but are rather characteristic of the data.



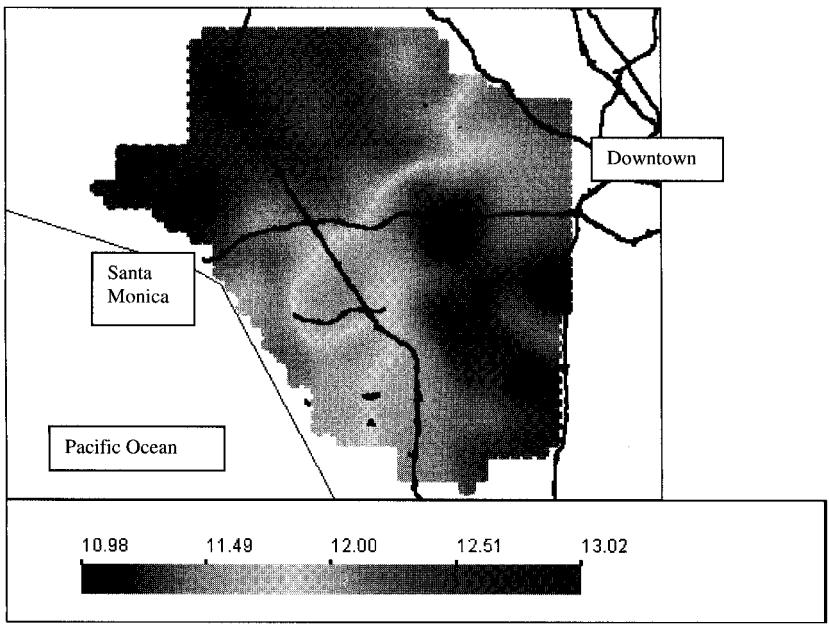
Results from the SVC Estimation Technique

To show the results from the estimation procedure I have constructed a fine grid.[4] On each node of the grid I estimate the parameters. I then display those estimates on a three dimensional graph with coordinates X, Y giving the physical location of the node, and the parameter of interest on the vertical axis.

Figure 9 depicts the spatial distribution of the intercept from the regression. Figures 10, 11 and 12 depict the marginal values of size of the living area, number of bedrooms and number of bathrooms, respectively. It comes as no surprise that expensive areas will have both high intercepts and high marginal values of the physical characteristics. These figures also indicate a substantial variation in all parameters, not just the intercept. This is consistent with the hypothesis that modeling the spatial distribution of the intercept alone is not sufficient.

[4] This is a regularly spaced rectangular grid. The distance between adjacent nodes is 300 meters.

Figure 9 ■ Spatial distribution of the intercept from the estimated hedonic regression $z_i = p(x_i,y_i) + q(x_i,y_i)s_i + \varepsilon_i$.



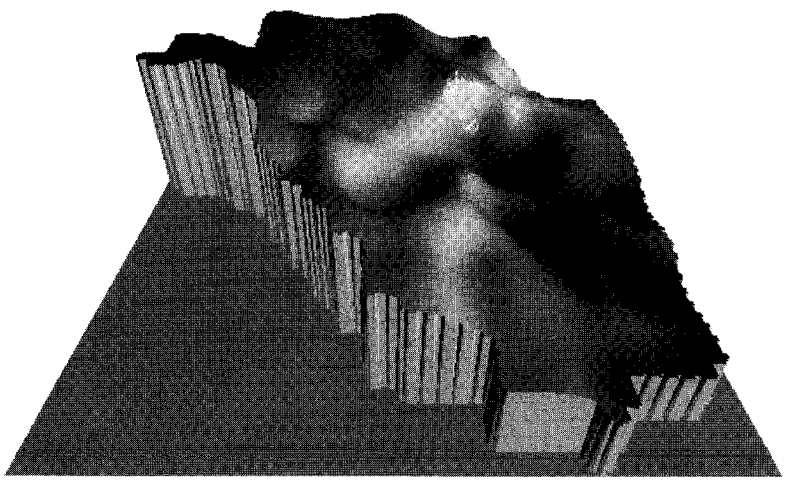
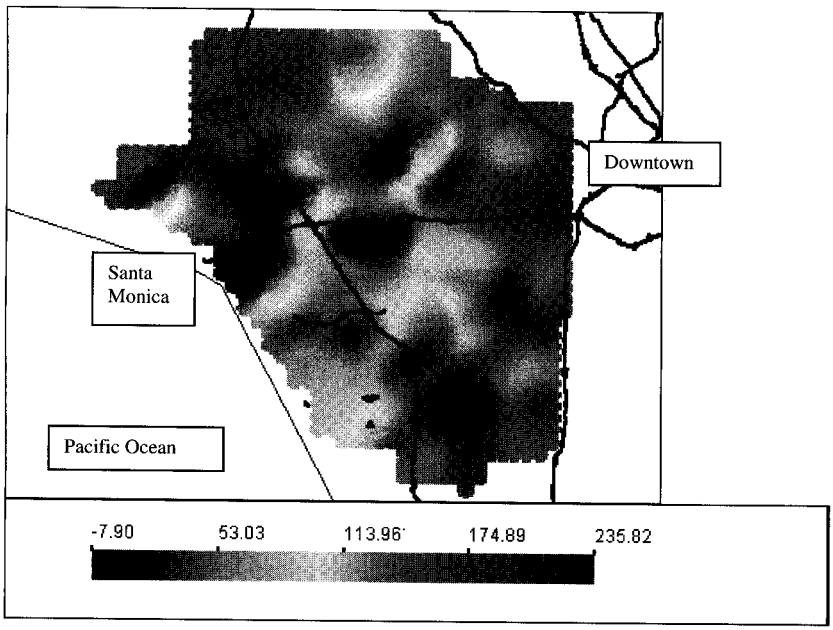


Figure 10 ■ Spatial distribution of the marginal value of size of the living area as estimated by the hedonic regression.



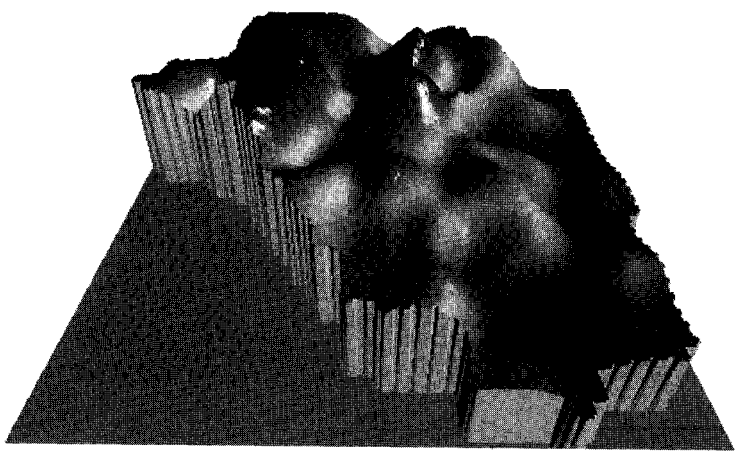


Figure 11 ■ Spatial distribution of the marginal value of bedrooms as estimated by the hedonic regression.

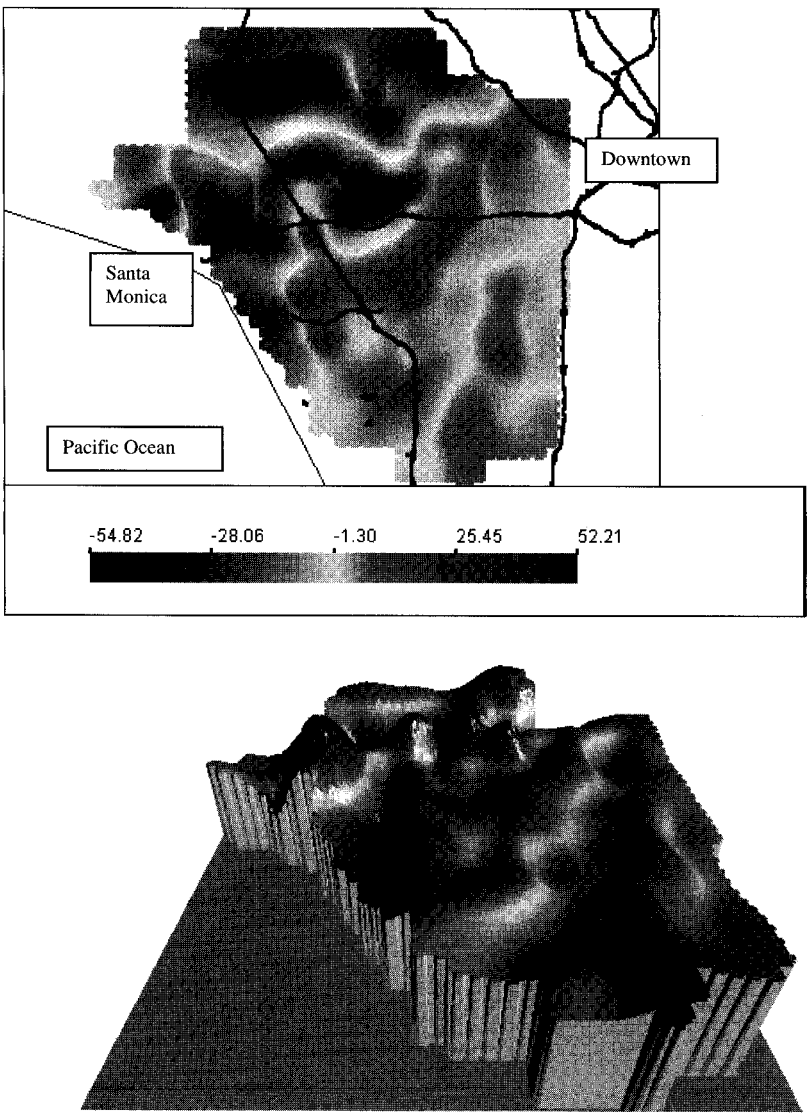
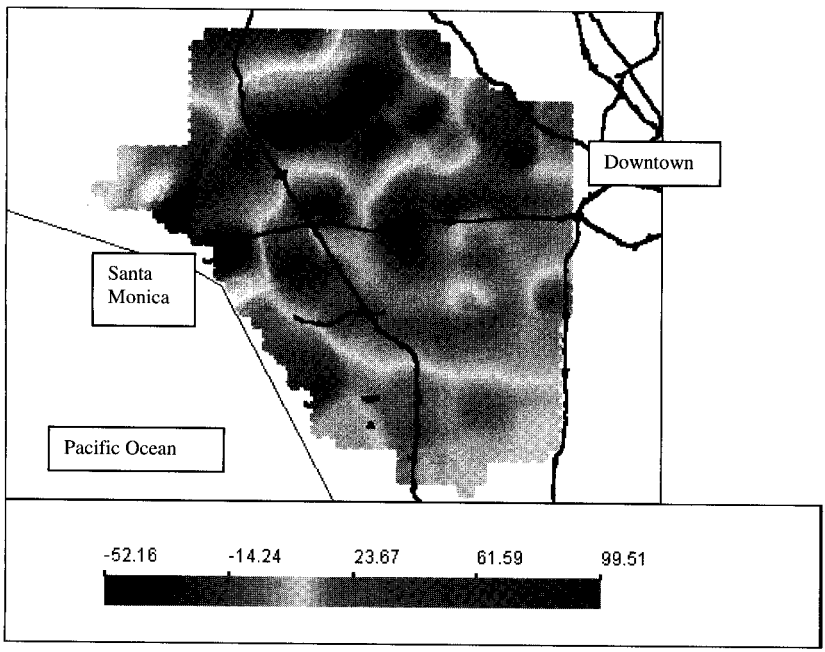
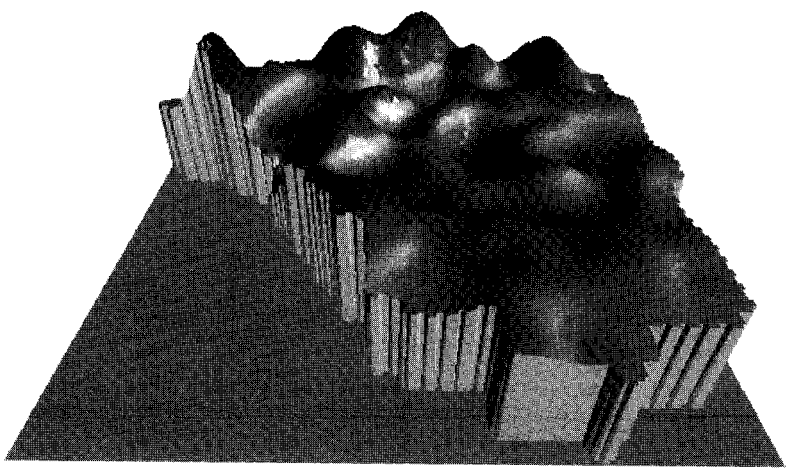


Figure 12 ■ Spatial distribution of the marginal value of bathrooms as estimated by the hedonic regression.





Using the parameter estimates at each node, I estimate the value of a typical house located on this node, using the mean values for the included variables. Figure 13 displays these estimated values. For people familiar with the areas, these maps present no surprise. For Los Angeles, we can clearly identify expensive areas like Bel Air and Pacific Palisades, as well as the low-valued South-Central region.

A comparison of the estimated value for a typical house in Figure 13 with the observed home values in Figure 3 gives an answer to the question of what drives the high prices in Bel Air—the fact that the houses are bigger or the quality of the amenities. As evident from Figure 13, the values for a typical home would actually be larger near the ocean directly west of Bel Air. In other words, the results suggest that if we could move a Bel Air home a few miles west, it would gain value.

Another way to interpret Figure 13 is that it depicts the unobserved neighborhood characteristics. Since controlling for the observed characteristics yields higher values near the ocean, the unobserved characteristics must be playing a role in determining these values. This is not surprising, since I would expect proximity to the ocean to have a big positive effect on values. On the other hand, the areas with the best unobserved amenities are not close to the downtown area or to any of the other major business districts. This suggests that values are driven more by the quality of the neighborhood amenities than by the proximity to employment opportunities. This is contrary to the central-business-district theory and its empirical confirmation by McMillen (1996), among others.

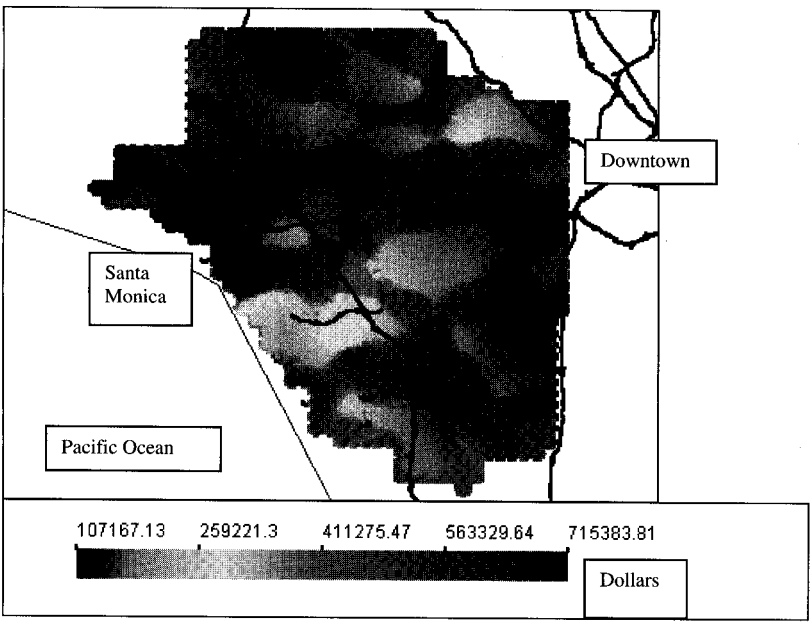
Residual Analysis

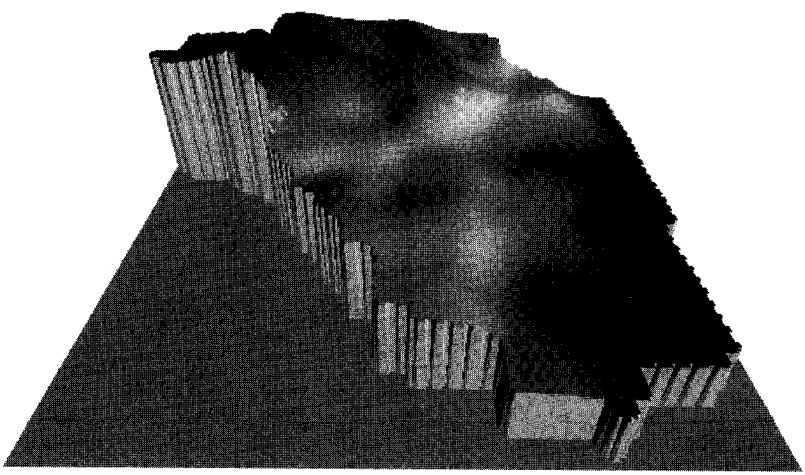
Let me now examine the residuals from the estimation technique and compare them with those from linear regression and several other alternative estimation methods. In addition to benchmarking the proposed method against various alternatives, the out-of-sample residual analysis presented below will, to some extent, remedy the lack of theoretical inferential framework for the method.[5]

As described above, the residuals from the proposed method are computed using cross-validation. Since cross-validation is used to choose the optimal

[5] Hastie and Tibshirani (1993) suggest an inferential approach for some general semi-parametric varying-coefficients models. While such an inferential framework is important from a theoretical perspective, it requires assumptions that are unlikely to hold in this application.

Figure 13 ■ Spatial distribution of the estimated value for a typical house. A typical house is defined to have the median value for the observed characteristics for the entire area. Thus, the observed spatial variation is due to the unobserved variables.





smoothing parameter, it does not represent a true out-of-sample performance measure for the SVC method. One solution is to use only half of the cross-validation errors to choose the smoothing parameter, and the other half to measure the out-of-sample performance. Due to the large number of observations in this application, however, the optimal smoothing parameter obtained using half of the cross-validation errors is virtually identical to the one obtained by using all errors. Furthermore, as will be seen below, the average cross-validation error is quite robust to the choice of smoothing parameter. Thus, the fact that I use the cross-validation errors to obtain the smoothing parameter does not undermine their ability to measure out-of-sample performance.

Figure 14 presents the average residuals for a wide range of the smoothing parameter k. The dashed line depicts the residuals from linear regression. This line is flat, since those residuals do not depend on the smoothing parameter. Effectively, the linear regression estimates are equivalent to applying the proposed method with $k = \infty$. Figure 15 plots the average distance to observations entering the estimation at each point in space as a function of the smoothing parameter.

Figure 14 allows us to make a number of observations. First, the average residuals do not change much over a wide range of smoothing parameters.

Figure 14 ■ Cross-validation residuals for Los Angeles County. The plot presents the root-mean-square cross-validation residuals from linear regression and from the SVC method for various values of the smoothing parameter k. The residuals from linear regression follow a straight line, since they do not depend on k. As k grows large, the mean residuals from the SVC method grow and eventually will approach the residuals from linear regression. As k gets small, on the other hand, very few observations enter the estimation at each point, and the model is overfitting the data. Thus, there is an optimal neighborhood size based on the trade-off between sampling error (small k) and bias of the estimates (large k).

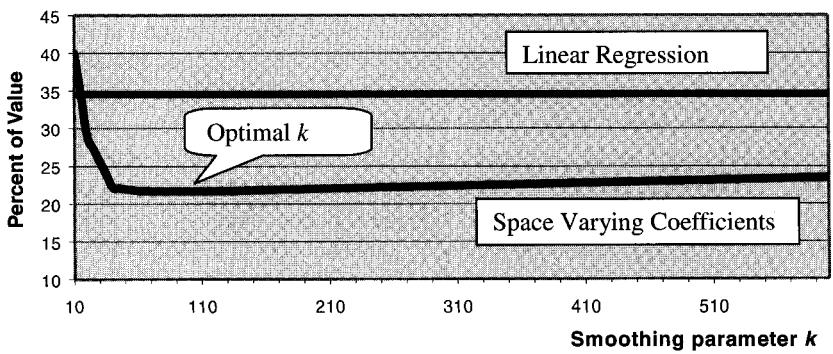
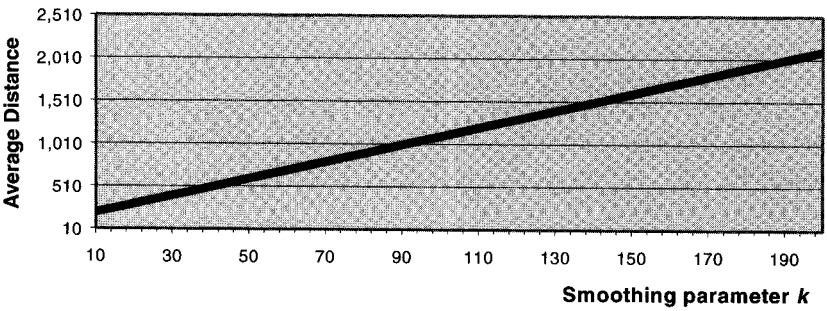


Figure 15 ■ Average distance to observations included in the estimation at each point for various values of the number of neighbors considered, k. Since the data are distributed fairly uniformly in space, the average distance is nearly a linear function of k.



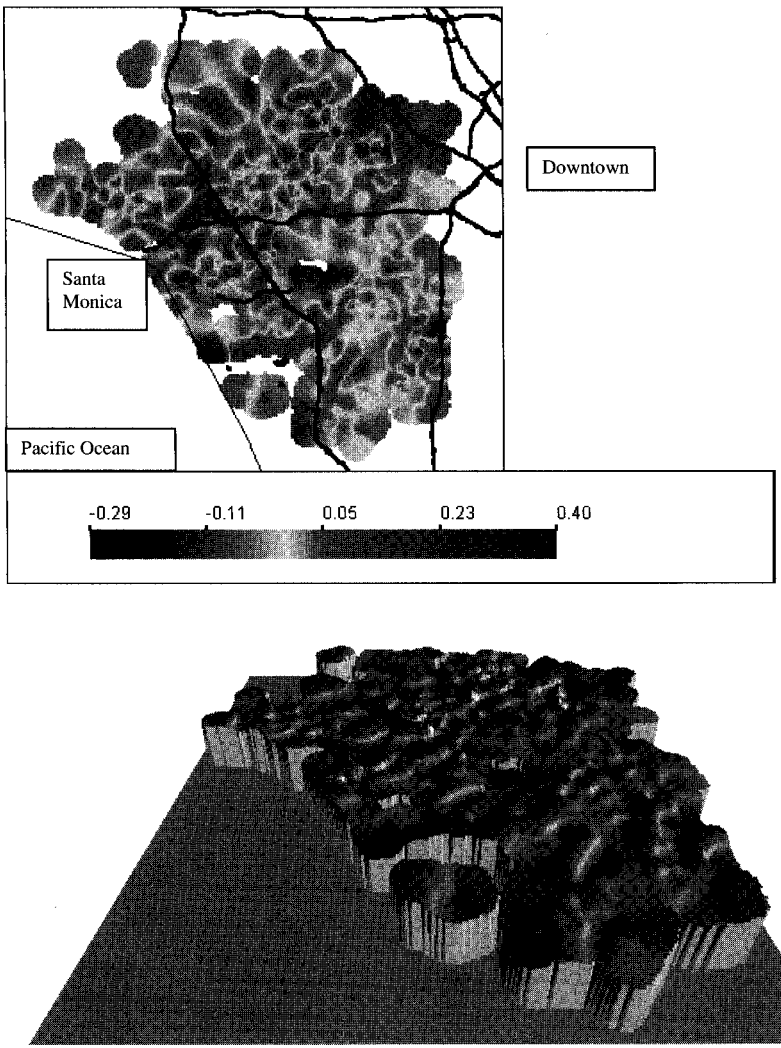
This means that only a rough estimate of the optimal k is necessary. As k grows large, the technique starts to smooth over very large regions, includes many observations and eventually approaches linear regression. If k gets too small, very few observations enter the estimation, the technique overfits the data, and the out-of-sample residuals start to grow. While the errors from too large k will never exceed linear regression, those from too small k may do so.

The parameter k controls the trade-off between bias and sampling error. To reduce the bias of the estimates I would aggregate over very small regions. According to the assumptions outlined in the introduction, only nearby observations are similar to the point of interest in the missing variables. On the other hand, the fewer the observations, the larger the sampling error. Hence, there is an optimal neighborhood size which minimizes the cross-validation errors.

It is important to see if the residuals are homogeneous in space. Figure 16 displays the cross-validation prediction errors in space. Notice that there is no spatial pattern in the cross-validation errors. This suggests that there are no systematic problems associated with the procedure. The spread of the residuals is not small (from -30% to 40%). This simply indicates that there are regions where values are harder to predict. This is caused by two factors. First, there are too few neighboring observations and the procedure includes remote houses. Second, the local variation is too large and constitutes a substantial departure from the smoothness assumption employed here.

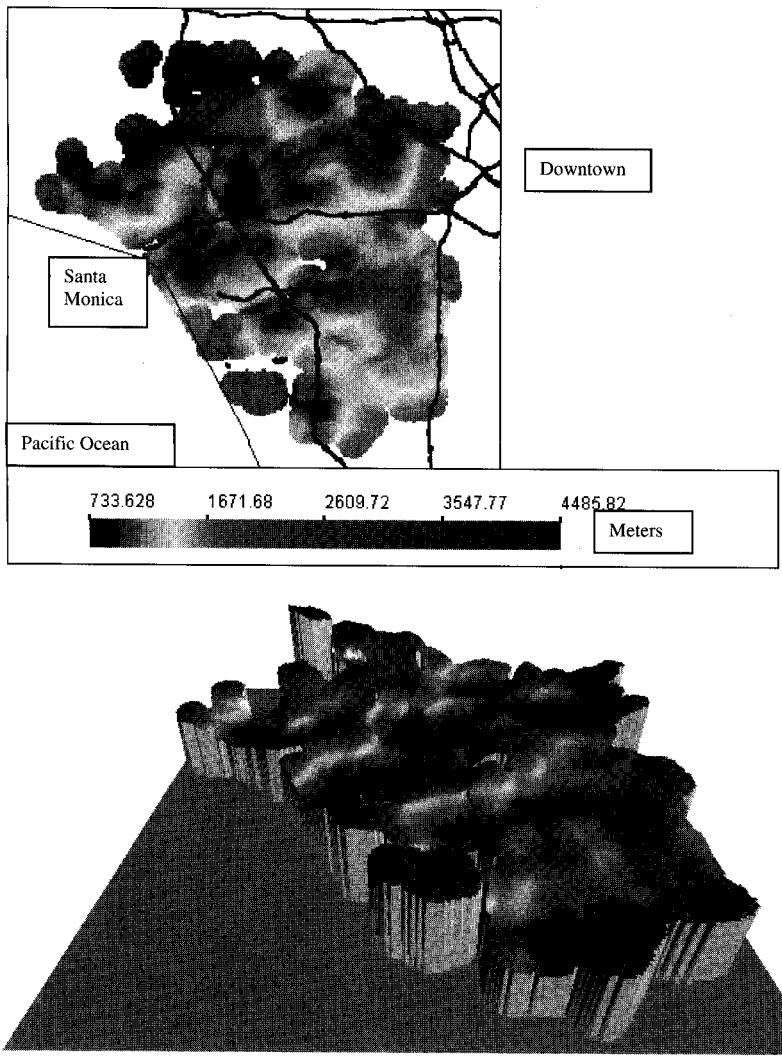
Figure 17 displays the average search radius in meters around each observation. Recall that the number of neighboring observations entering the

Figure 16 ■ Spatial distribution of the cross-validation errors. The absence of spatial pattern of those residuals suggests that the proposed method captures the spatial variation of home values.



estimation is kept constant, while the average distance to those observations varies with their density. The search radius is between 700 and 4,000 meters. Recall, however, that remote observations are weighted less than nearby observations. Thus, the observations within roughly half that distance receive most of the weight. The variation in the search radius is solely due to

Figure 17 ■ Spatial distribution of the average search radius. Large search radius indicates that the density of the observations is lower around that point and the procedure has to look farther to obtain enough neighboring observations.



variations in the density of the observations in space. In other words, the technique automatically adjusts to the density to obtain the best estimate given the surrounding observations.

Figure 18 provides a residual-vs.-predicted plot. Although some residuals are quite large, most residuals fall within −0.4 to 0.4, which is consistent with Figure 16. Furthermore, even though the residuals for more expensive homes are slightly larger, this plot does not indicate any apparent warning signs.

While the standard linear regression provides a useful benchmark, it is too restrictive in its assumptions to be a fair match for the proposed method and therefore cannot be used to evaluate its effectiveness. Several other methods used in previous research are closer competitors to the SVC method and provide a more meaningful evaluation. The most obvious ones are:

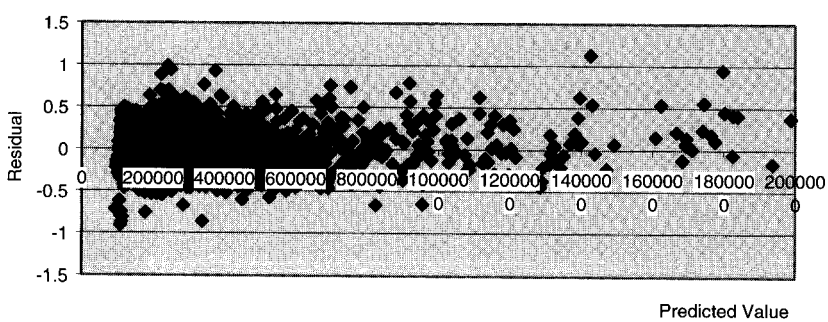
- Parameterize space, *i.e.*, use some *a priori* definition of neighborhoods. Here I use postal zip codes as a way of defining neighborhoods.
- Allow the parameters of the model to vary parametrically in space. I use linear and quadratic functions of the east–west and north–south coordinates.
- Vary only the intercept non-parametrically. In other words, estimate a standard linear model and then model the spatial distribution of the errors non-parametrically.

Below I present the implementation of each of the above three competitors and conclude that the SVC method outperforms those alternatives in terms of the average cross-validation errors.

Parameterizing Space

The first alternative to SVC is to specify neighborhoods exogenously. This can be accomplished by using city boundaries, school districts, census tracts,

Figure 18 ■ Estimation residuals vs. predicted values.



postal zip codes or any other definition of neighborhood. To illustrate this approach I use postal zip codes to define neighborhoods. Zip-code boundaries are usually along major streets, freeways, rivers and city boundaries and are thus a reasonable way to define neighborhoods. Furthermore, postal zip codes are routinely used by researchers and practitioners as a proxy for many missing variables.

For the estimation of home values, defining neighborhoods as postal zip codes means rewriting the hedonic model presented above the following way:

$$z_i = (zip_i) + q(zip_i)s_i + \varepsilon_i \tag{13}$$

where zip_i is the postal zip code for observation i. Thus, the parameters will be the same for all homes within a zip code and will change discretely between zip codes. This approach can be implemented either by including indicator variables for $N - 1$ zip codes and interactions between those indicator variables and other independent variables, or by simply estimating a separate regression for each zip code. I employ the latter technique.

Table 2 reports the cross-validation errors from the parameterization of space by zip codes along with the linear regression and the SVC estimations described above. While allowing the parameters of the linear regression to vary by zip code is a substantial improvement over standard linear regression (average residuals drop from 0.3450 to 0.2348), allowing the parameters to vary non-parametrically using the SVC method produces still lower average cross-validation errors.

Parametric Variation of the Parameters

A close competitor to the SVC approach when valuing real estate properties is to allow the parameters of the hedonic model to vary parametrically in space. Cassetti (1972, 1986) proposed the theoretical basis for this approach. Can (1990, 1992) and Can and Megbolugbe (1997) suggest an application of this approach to real estate markets.

In terms of the hedonic model discussed in this paper, the Casetti (1972) expansion method allows the parameters of the hedonic function to vary parametrically with a set of variables not included in the model. More specifically, I can rewrite the hedonic model the following way:

Table 2 ■ Comparison of alternative models.

Method	Root-Mean-Square Cross-Validation Error
Linear regression	.34
Parameterization with linear functions	.24
Parameterization with quadratic functions	.23
Varying only the intercept	.23
Linear regression with zip codes	.23
SVC (with optimal k = 80 observations)	.21
SVC + zip codes (with optimal k = 100, v = 450)	.19

The table compares the root-mean-square cross-validation errors from seven alternative approaches described in the paper:

- Standard linear regression of the transaction value on size, number of bedrooms and number of bathrooms.
- Linear parameterization—all coefficients are linear functions of the (x,y) coordinates.
- Quadratic parameterization—all coefficients are quadratic functions of the (x, y) coordinates.
- Varying only the intercept—the intercept of the regression is varied non-parametrically in space while all other coefficients remain fixed.
- Linear regression with postal zip codes allows the parameters of the linear regression to be different for each zip code. This is a way to parameterize space by defining neighborhoods *a priori.*
- Standard SVC—the proposed method which allows all parameters to vary non-parametrically in space.
- SVC + zip codes allows the proposed SVC method to take into account the postal zip code of the properties. This is a way to redefine the metric in space. The non-parametric features are fully retained.

$$z_i = p(m_i) + q(m_i)s_i + \varepsilon_i \tag{14}$$

where m_i is a vector of the new variables influencing the parameters of the model. In this case, the functions $p(m)$ and $q(m)$ are of known parametric form, usually linear or quadratic:

$$p(m) = c_p + d_p\, m + g_p\, m^2 \quad \textit{and} \quad q(m) = c_q + d_q m + g_q\, m^2 \tag{15}$$

Substituting the above functions into the hedonic model yields the interaction between the new variable m and the other included variables.

In the case examined here, the variable m are the (x, y) coordinates of each property. Thus, the above method translates into

$$z_i = p(x_i, y_i) + q(x_i, y_i)s_i + \varepsilon_i \tag{16}$$

with $p(x_i, y_i)$ and $q(x_i, y_i)$ now parametric functions of the following form (quadratic case only):

$$\begin{aligned} p(x, y) &= c_p + d_{px}\, x + g_{px}\, x^2 + d_{py}\, y + g_{py}\, y^2 + h_p\, xy \\ q(x, y) &= c_q + d_{qx}\, x + g_{qx}\, x^2 + d_{qy}\, y + g_{qy}\, y^2 + h_q\, xy \end{aligned} \tag{17}$$

Table 2 reports the root-mean-square cross-validation errors for the linear and quadratic parameterization of the coefficients p and q. While allowing the coefficients of the hedonic model to vary parametrically in space is a big improvement over the standard linear regression model, SVC still outperforms the parameterization approach in terms of cross-validation errors. This is not surprising considering how unpredictably those coefficients change in space and how important it is to model them non-parametrically.

Varying Only the Intercept in Space

Yet another competing method frequently used in practice is to model only the intercept of the hedonic model in space and keep all other parameters fixed. The implementation of this approach can take several forms. The most common is to estimate a standard hedonic regression and then model the spatial distribution of the errors. In terms of the motivation of the SVC method this would mean that the missing variables affect only the intercept of the hedonic model but do not interact with any of the other variables. This translates into the following specification:

$$z_i = p(x_i, y_i) + qs_i + \varepsilon_i \tag{18}$$

where the parameter q does not change in space.

Again, I compare the approach of modeling only the intercept with the suggested SVC method using root-mean-square cross-validation errors. Table 2 reports the cross-validation errors from the two approaches along with those from standard linear regression. As expected, allowing all parameters to vary non-parametrically in space still outperforms the alternative approaches.

Incorporating Postal Zip Codes into SVC

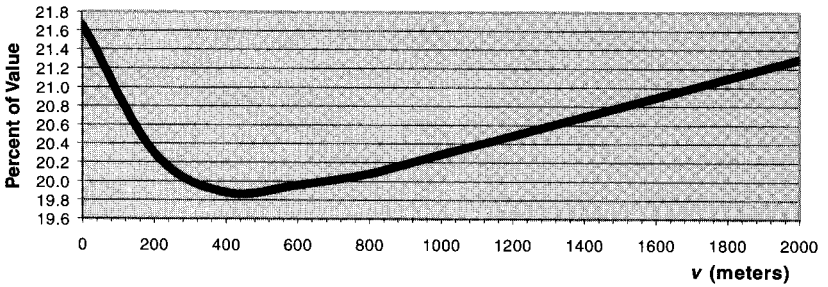
The vast improvement that postal zip codes brought to standard linear regression suggests that they do indeed contain information about neighborhoods and property values. This leads me to consider incorporating postal zip codes into SVC. Since zip codes are another measure of space, it is imperative to incorporate them together with the (x, y) coordinates within the non-parametric estimation. Including zip codes within the parametric components of the SVC method through indicator variables would severely undermine the consistency of the SVC method. The consistency of most non-parametric methods rests on the search radius declining to zero as the number of observations increases. Including indicator variables into the parametric components will result in singular design matrices for sufficiently large numbers of observations. Thus, all neighbors will be in the same zip code and there will be no variation in the indicator variables.

Considering the above discussion, the only way to include zip codes in the estimation without violating the consistency properties of the proposed method is to distort space according to zip codes. Instead of using the usual Euclidean space, I exaggerate the distance between observations if they fall in different zip codes. More precisely, I add a constant v to the distance between two observations if they are in different zip codes. This makes all observations within a zip code appear closer together relative to observations in distinct zip codes. More formally, the distance between two observations (x, y) and (x', y') becomes

$$\text{Distance}(x, y, x', y') = \begin{cases} \sqrt{(x - x')^2 + (y - y')^2} & \text{if both observations are in the same zip code} \\ \sqrt{(x - x')^2 + (y - y')^2} + v & \text{otherwise} \end{cases} \tag{19}$$

This introduces the issue of selecting the appropriate constant v. As with the case of choosing the optimal neighborhood size k, I use cross-validation to choose the optimal v. Computationally, this translates into minimizing the root-mean-square cross-validation error over both k and v. To this end I repeat the estimation of the SVC model, described above and in the appendix, for a wide range of values of v. In particular, I estimate the mean cross-validation error for values of v between 100 and 2,000 meters. For each v, an optimal neighborhood size k is selected. Figure 19 depicts the cross-validation residuals as a function of the parameter v. For each v the optimal k has been selected. The optimal v is 450 meters, with optimal k of

Figure 19 ■ Cross-validation residuals and zip codes. The plot presents the root-mean-square cross-validation residuals for various values of the parameter v. The residuals are minimized for $v = 450$ meters. In other words, the best model adds 450 meters to the distance between two homes that fall in different postal zip codes.



100, which is slightly larger than the optimal neighborhood size from the standard SVC method described above.

As evident from Figure 19, the average cross-validation residual from the SVC method incorporating postal zip codes is less than 20%, which represents by far the best fit among the other comparable approaches discussed above. Furthermore, postal zip codes do indeed contain information relevant to property values, and the SVC method is fully capable of utilizing this information.

Conclusion

This paper examined the issue of real estate valuation utilizing transaction data from Los Angeles County. As evident by the surprisingly high variability of home values, both locally and for the entire area, traditional approaches to modeling real estate markets are unlikely to produce reasonable estimates. This was confirmed by the extensive cross-validation prediction comparisons reported above.

The proposed approach to modeling real estate markets substantially outperforms a number of comparable alternative approaches in terms of cross-validation residuals. It also allows us to make a number of observations regarding the segmentation of the market.

The proposed approach can be used to address a number of other issues in real estate. For instance, for income-producing properties I could estimate the correlation between every property and the rest of the market. This will

determine to what extent a portfolio of properties can be diversified or hedged within the same local market, and to what extent diversification in other markets is required. The proposed technique can give insight into the evolution of urban regions and the speed at which upward or downward movements spread.

In addition to standard Euclidean space, I also incorporate zip codes into what remains a non-parametric estimation. This improvement retains the optimality properties of the method while utilizing the information contained in postal zip codes.

Finally, the proposed approach can be applied to estimation problems well outside real estate valuation. It is particularly useful when a reasonable specification of the functional form of the relationship of interest is available only with respect to some, but not all, of the included variables. In such cases efficiency can be gained by incorporating parametric components into non-parametric estimations.

Special thanks to Stephen Cauley, Larry Kimbell and Edward Leamer from the Anderson School of Management and to Donald Ylvisaker from the Department of Statistics at UCLA for the numerous stimulating and insightful discussions related to this paper. The Anderson Real Estate Center at UCLA provided the data, computer hardware and software, and other resources without which this project would not have been possible. I would also like to thank the three anonymous referees for their helpful comments.

References

Anselin, L. 1989. Some Robust Approaches to Testing and Estimation in Spatial Econometrics. *Regional Science and Urban Economics* 20: 141–163.

Basu, S. and T. Thibodeau. 1997. Analysis of Spatial Correlation in House Prices. Presented at American Real Estate and Urban Economics Meetings. New Orleans.

Can, A. 1990. The Measurement of Neighborhood Dynamics in Urban Housing Prices. *Economic Geography* 66: 254–272.

———. 1992. Specification and Estimation of Hedonic Housing Price Models. *Regional Science and Urban Economics* 22: 453–474.

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

Case, A. 1991. Spatial Patterns in Household Demand. *Econometrica* 59: 953–965.

Case, B., H. Pollakowksi and S. Wachter. 1991. On Choosing among House Price Index Methodologies. *Journal of the American Real Estate & Urban Economics Association* 19(3): 286–307.

Case, K. and R. Shiller. 1987. Prices of Single Family Homes Since 1970: New Indexes for Four Cities. *New England Economic Review* 45–56.

———. 1990. Forecasting Prices and Excess Returns in the Housing Market. *American Real Estate and Urban Economics Journal* 18(3): 253–273.

Cassetti, E. 1972. Generating Models by the Expansion Method: Applications to Geographical Research. *Geographical Analysis* 4: 89–91.

———. 1986. The Dual Expansion Method: An Application to Evaluating the Effects of Population Growth on Development. *IEEE Transactions on Systems, Man and Cybernetics* 16: 29–39.

Cauley, S. 1997. Los Angeles Home Values: Analysis at the Zip Code Level. *UCLA-Anderson Forecast.* First Quarter.

Cleveland, W. and S. Devlin. 1988. Locally Weighted Regression: An Approach to Regression Analysis by Local Fitting. *Journal of the American Statistical Association* 83(403):87–114.

Devroye, L. 1978. The Uniform Convergence of Nearest Neighbor Regression Function Estimators and Their Application to Optimization. *IEEE Transactions on Information Theory* 24: 142–151.

Dubin, R. 1988. Estimation of Regression Coefficients in the Presence of Spatially Autocorrelated Error Terms. *Review of Economics and Statistics* 70: 466–474.

———. 1992. Spatial Autocorrelation and Neighborhood Quality. *Regional Science and Urban Economics* 22: 433–452.

Epanechnikov, V. 1969. Nonparametric Estimates of a Multivariate Probability Density. *Theory of Probability and Its Applications* 14:153–158.

Goetzmann, W. and M. Spiegel. 1997. A Spatial Model of Housing Returns and Neighborhood Substitutability. *Journal of Real Estate Finance and Economics* 14(1–2): 11–31.

Gyofty, L. 1981. The Rate of Convergence of *k*-NN Regression Estimation and Classification. *IEEE Transactions on Information Theory* 27: 500–509.

Hardle, W. 1997. *Applied Nonparametric Regression.* Econometric Society Monographs No. 19. Cambridge University Press.

Hastie, T. and R. Tibshirani. 1993. Varying-Coefficient Models. *Journal of the Royal Statistical Society, Series B—Methodological* 55(4): 757–796.

Kain, J. and J. Quigley. 1970. Measuring the Value of Housing Quality. *Journal of the American Statistical Association* 65: 532–548.

Lai, S. 1977. Large Sample Properties of *k*-Nearest Neighbor Procedures. Ph.D. Dissertation, Department of Mathematics, University of California, Los Angeles.

Lai, T. Y. and K. Wang. 1996. Comparing the Accuracy of the Minimum-Variance Grid Method to Multiple Regression in Appraised Value Estimates. *Real Estate Economics* 24(4): 531–549.

Loftsgaarden, D. and C. Quesenberry. 1965. A Nonparametric Estimate of a Multivariate Density Function. *Annals of Mathematical Statistics* 36: 1049–1051.

Mack, Y. 1981. Local Properties of *k*-NN Regression Estimates. *Journal of Algebraic Discrete Methods* 2: 311–323.

McMillen, D. 1996. One Hundred Fifty Years of Land Values in Chicago: A Nonparametric Approach. *Journal of Urban Economics* 40: 100–124.

Meese, R. and N. Wallace. 1991. Non-parametric Estimation of Dynamic Hedonic Price Models and the Construction of Residential Housing Price Indices. *Journal of the American Real Estate and Urban Economics Association* 19(3): 308–330.

Pace, R. and O. Gilley. 1997. Using the Spatial Configuration of the Data to Improve Estimation. *Journal of Real Estate Finance and Economics* 15(3).

Rosen, S. 1974. Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition. *Journal of Political Economy* 82: 34–55.

Stone, C. 1977. Consistent Nonparametric Regression. *The Annals of Statistics* 5(4): 595–645.

Appendix
Estimation Procedure

Load vectors X, Y and Z containing the coordinates of each observation and the transaction value, and a matrix S containing a column of ones, the size of the house, the number of bedrooms and the number of bathrooms.

Repeat for a wide range of k
    FOR $i = 1$ to N
        Exclude observation i.
        Calculate the weights W.
        Compute the parameters using weighted least squares:

$$\begin{pmatrix} p_{(x_i,y_i)} \\ q_{(x_i,y_i)} \end{pmatrix} = (S'\ W\ S)^{-1}\ (S'\ W\ Z).$$

        Compute the cross-validation error at location (x_i, y_i):

$$cverr = cverr + (z_i - p_{(x_i,y_i)} - q_{(x_i,y_i)}\ s)^2.$$

    Move to the next i.
Move to the next k.

Find k for which the cross-validation error *cverr* is the smallest.
Create a fine grid to estimate the values of the functions p and q.
For each point on the grid
    Calculate the weights W using the optimal k.
    Compute the parameters p and q using weighted least squares.
Move to the next point on the grid.
All computations were implemented using Matlab and took less than a day on a fast PC.