RESEARCH PROGRAM IN FINANCE AT THE
WALTER A. HAAS SCHOOL OF BUSINESS,
UNIVERSITY OF CALIFORNIA, BERKELEY

The Research Program in Finance in the Walter A. Haas School of Business at the University of California has as its purpose the conduct and encouragement of research in finance, investments, banking, securities markets, and financial institutions. The present reprint and working paper series were established in 1971 in conjunction with a grant from the Dean Witter Foundation.

INSTITUTE OF BUSINESS AND ECONOMIC RESEARCH
Carl Shapiro, Director

The Institute of Business and Economic Research is an organized research unit at the University of California, Berkeley, whose mission is to promote research by faculty and graduate students in the fields of business and economics. The Institute carries out its mission by organizing programs and activities that enrich the research environment, administering extramural research awards, publishing working papers, and making direct grants for research.
A Spatial Model of Housing Returns and Neighborhood Substitutability

William N. Goetzmann
Yale School of Management

and

Matthew Spiegel
University of California, Berkeley
Haas School of Business

September 1995

Finance Working Paper #253

The authors thank participants in the University of British Columbia, U.C. Berkeley Real Estate workshops for helpful comments, and Arturo Noriega, and Xongshin Zhang for research assistance. We also thank the Yale Finance Institute, and the University of California for research support.
Abstract

This paper presents a new spatial model for analyzing return indices for infrequently traded assets, and applies it to housing data. Within many asset classes, particularly real estate, one expects there to exist a spatial correlation in deviations from the index due to omitted explanatory variables in the econometric model. This error structure can be useful in estimating location-specific indices, whether that location is defined in terms of geography or exposure to common economic or social factors. The econometric design presented in this paper allows the use of distance, broadly defined, to accurately estimate housing return series at the level of individual zip code neighborhoods in the San Francisco Bay area. While a paucity of transactions data would normally make this impossible, the use of spatial and factor correlations provides sufficient information to estimate zip code level returns.

We use these indices to examine the degree to which housing market participants in one major metropolitan statistical area view neighborhoods as substitutes. Using distance defined in terms of geographical proximity, median household income, average educational attainment and racial composition, we find that median household income is the salient variable explaining covariance of neighborhood housing returns. Racial composition and educational attainment, while significant are much less influential and geographical proximity is nearly meaningless. Our methodology has applications to a range of infrequently traded assets, including bonds, commercial real estate and collectibles. The approach may be viewed as an extension of "non-parametric" spatial correlation models. In the non-parametric approach a distance function and decay rate are exogenously specified. In a spatial model one estimates the distance metric and uses statistical rules to obtain the resulting decay rates. The results of our analysis of housing substitutability in the San Francisco Bay area have implications for estimates of the covariance of housing returns within metropolitan areas. In particular, low covariances imply gains to diversification for lenders, equity-holders and tax authorities.
Single family homes are one of the most widely held assets in the United States. Moreover, for many families that own their own house real estate represents the largest single component of their personal portfolio. For these two reasons alone it is important to understand the factors that drive housing returns. A number of studies have shown that the capital appreciation of housing is correlated to fluctuations in local economic variables such as unemployment, as well as to national financial factors such as interest rates. While such widespread variables certainly impact the average regional return to owning a home, local variables -- variables tied to location -- are likely to be important as well. What are the factors that make certain homes good substitutes within a metropolitan area? Despite the vast quantity of housing transaction data, this question is difficult to address empirically. Unlike stocks or bonds, houses sell infrequently, and even houses next door to each other are imperfect substitutes. Thus, any analysis of the dynamics of housing prices typically requires a relatively high level of aggregation to develop meaningful indices.

In this paper, we address the question of housing substitutability on a neighborhood level by examining the socio-economic and locational factors that influence the correlation of neighborhood-level housing returns. We also deal with the econometric problem of creating "micro-indices" of housing at the zip-code level though the development of a spacial model of error-covariances. This spatial model projects individual housing return deviations from the metropolitan index on to the space of factors derived from the 1980 U.S. Census, including income, education and ethnic variables. This projection allows the efficient estimate of quarterly return indices even for zip codes with few transactions.

Our analysis, performed on housing in the San Francisco Bay area over the period 1980 through 1994, yields some interesting results about what factors make certain
neighborhoods substitutable from the perspective of housing market participants. We find that physical distance is a relatively poor indicator of whether two neighborhood housing indices will move together. On the other hand, certain socio-economic variables are good indicators. In particular, the median household income and educational attainment of the neighborhood captures much of the covariance of housing returns. The inclusion of ethnic composition variables adds little, if anything, to the prediction of covariance in returns. Thus, the local price dynamics seem to indicate that (1) physical proximity counts for little, if the income or educational level of the neighborhoods are different, and (2) race does not appear to be the salient factor in housing choice.

A major challenge to the analysis of neighborhood level data is the infrequency of transactions with which to construct an index of housing returns. When each asset in a given market trades every period, calculating a price index is straightforward. The observed price of each asset at the end of the period is simply weighted to provide an index value. For infrequently traded assets the problem becomes more complex. During any one period the observer may not have a transaction price for some or even most of the assets in the market. To create an index, the econometrician must somehow infer the missing prices. This paper presents a technique designed to solve this problem that we call distance weighted repeat sales (DWRS).

Within many asset classes, particularly real estate, a spatial correlation in deviations from the index exist, due to omitted explanatory variables in the econometric model used to construct the index. This error structure can be useful in estimating location-specific indices, whether that location is defined in terms of geography or exposure to common economic or social factors. The distance weighted repeat sales procedure presented in
this paper allows the use of distance, broadly defined, to estimate housing return series at the level of individual zip code neighborhoods. We apply it to a data set of repeat-sales of homes in the San Francisco Bay area. While paucity of transactions data would normally make detailed index construction of this nature impossible, the use of spatial and factor correlations provide sufficient information to estimate zip code level returns.

The DWRS methodology also has applications to a range of infrequently traded assets, including bonds, commercial real estate and collectibles, or any other asset whose heterogeneity can be described within an econometrically meaningful characteristic space. The approach may be viewed as an extension of "non-parametric" spatial correlation models. Non-parametric models exogenously specify a distance metric and a decay rate. In contrast, the spatial model estimates parameters that determine spatial proximity, and then employs statistical updating rules for the decay rate. The results of our analysis of housing substitutability in the San Francisco Bay area have implications for estimates of the covariance of housing within metropolitan areas. In particular, low covariances imply gains to diversification for lenders, equity-holders and tax authorities.

---

While financial models have not taken advantage of the spatial correlations among assets other fields have. Cressie (1993) discusses the problem from the viewpoint of geostatistics. In geology, the problem concerns estimating the amount of ore at various locations based upon a sample drawn at discrete locations. By weighting data so that local observations have the largest impact estimates can be derived even for locations from which samples have not been drawn. The model presented here differs from the analysis in Cressie in several respects since geologists can draw controlled samples, and do not need to form indices. The reader is referred to Cressie for specific details.
1 Background

Two techniques are commonly used to construct housing price indices. The first, weighted repeats sales (WRS), is described by Case and Shiller (1987) and Case and Quigley (1991) among others. This methodology uses matched purchase and sale price-date pairs for homes within some area. The total capital appreciation return from each repeated sale is the dependent variable in a weighted least squares regression that "explains" these returns by the time periods over which the asset was held.

The second method estimates the value of a "representative" house in the market each period via a set of priced characteristics. This uses individual home prices rather than matched sale pairs, and in addition makes use of attribute information such as the number of bedrooms, baths and other amenities. Research on the relative merits of the "hedonic" and repeat sales methods include Halvorsen and Pollakowksi (1981), and Meese and Wallace (1991, and 1995), Case, Pollakowski, and Wachter (1991), and Clapp and Giaccotto (1992).

The primary methodological advantage of the WRS lies in its ability to estimate housing returns without an explicit specification of how characteristics map into prices. Shiller (1991) shows that the matched prices used in the WRS offers a perfect hedonic control, as long as characteristics of the assets do not change between purchase and sale dates. One limitation is that the approach ignores housing's spatial nature. A WRS index assumes that in any one period all homes in the sample appreciate at the same rate times a white noise error term. Since data sets often cover fairly large cities, the appreciation
assumptions must hold over potentially distant and dissimilar neighborhoods. Further problems arise if the appreciation assumptions fail and the concentration of housing sales varies from neighborhood to neighborhood over time.\textsuperscript{2} Hedonic indices cure some of the problems associated with a WRS index by estimating the value of individual housing characteristics. Thus, dissimilar areas can appreciate at different rates. However, hedonic models also make some fairly strong assumptions, in particular, the independence of the error terms across homes. Consider two homes that are located next to each other. Now suppose the city installs a new park at a nearby location. If the data set contains both homes, then the model will underpredict the value of each house since the park is not likely to be part of the hedonic model. Nevertheless, the statistical technique assumes that the pricing errors are independent. In contrast, casual observation suggests that for houses located close together omitted variables from the estimation will lead to correlated errors.

To help alleviate the omitted variables problem Meese and Wallace (1991) employ a nonparametric estimation technique. This technique estimates a separate index at each location by using a weighted average of the local observations. The nonparametric hedonic approach comes closest to the one used here. Our model offers a potential improvement to the nonparametric methodology by providing an explicit technique for optimally reweighting the data. Other differences lie in the derivation of the models, and

\textsuperscript{2}For example suppose neighborhood A appreciates at 10\%, and neighborhood B at 6\%. Now consider a data set containing one home in A that sells in periods 0 and 2, and another home in B that sells in periods 1 and 2. The WRS index will report appreciation rates of 14.1\% in period 1, and 6\% in period 2. In reality if the two neighborhoods have equal housing densities, then the “true” citywide index should appreciate at a steady 8\% per period.
the interpretation of the parameters. The paper provides additional details in section 2.2.

2 Empirical Methodology

2.1 Spatial Modeling of Housing Returns

The standard weighted repeat sales (WRS) technique provides a useful introduction to the distance weighted methodology developed here. A WRS analysis examines a data set that contains paired observations on various homes over time. An observation pair consists of a date \( b \), and a price \( P_b \) when a family buys the house, and a date \( s \), and a price \( P_s \) when the family sells it. Let \( r_t \) equal one plus the return to housing in period \( t \), and \( \varepsilon_t \) an error term. The WRS model then assumes that the price process can be written as

\[
P_s = P_b \prod_{t=b+1}^{s} r_t \varepsilon_t.
\]

Taking logs produces the familiar linear system

\[
\ln(P_s) - \ln(P_b) = \sum_{t=b+1}^{s} \ln(r_t) + \ln(\varepsilon_t).
\]

Typically, the WRS model imposes the assumption that the \( \ln(\varepsilon_t) \) are independently identically distributed normal random variables with mean zero, and variance \( \sigma^2_\varepsilon \). This allows the researcher to use standard generalized least squares techniques to produce
efficient estimates for the log returns.

Goetzmann and Spiegel (1995) argue that the WRS equation (2) should be modified to allow for a return component associated with housing transactions. Their analysis indicates that this component tends to be positive, and they find evidence that it may be due to home improvements that tend to occur around the time of a sale. Whatever the cause, a transaction specific return can be added to (2) by introducing a return \( h \) that occurs whenever a home changes hands, and an error term \( \eta \) on the sale date. This leads to the following variant of (2)

\[
\ln(P_s) - \ln(P_b) = \ln(h) + \ln(\eta) + \sum_{t=b-1}^{s} [\ln(r_t) + \ln(\varepsilon_t)].
\]

(3)

While the WRS model provides an easily estimated model, it completely ignores housing's spatial nature. As the famous real estate cliché says, the three most important elements of real estate value are "location, location, and location." Yet, equations (2), and (3) implicitly assume a uniform return structure across the entire sample area. As even the most casual observer knows, housing returns display very strong spatial patterns. If the high school in town A improves relative to the school in town B, then homes in A will appreciate relative to B. Accounting for spatial relations in the model requires one to abandon some of the assumptions underlying the WRS model.

Deriving return estimates in a spatial setting requires an explicit model describing housing price formation. This paper assumes that the return on a house at location \( i \), in period \( t \), consists of three components. The first component \( r \) represents one plus the per
period expected return on housing. One can generalize this return in a manner similar to the APT by assuming that \( r \) depends upon a set of factors.\(^3\) If these factors vary over time and by location then one can instead write expected housing returns as \( r_i(t) \).

A second element that enters into a home's price is a location-specific return. Let \( r_i(t) \) equal one plus the local return at time \( t \). Suppose that a home at location \( i \) goes up in value by 10% over some period of time. Then intuitively, nearby homes should also go up by approximately 10%. After all, homes in the same neighborhood share many characteristics, and are to one degree or another substitutes for each other. Conversely, there is much less reason to believe that a home across town must have risen by 10%, since the house acts as only a poor substitute for those in \( i \)'s immediate neighborhood. This intuitive argument leads to the following model. Assume that \( \ln(r_i(t)) \) follows a normal distribution, with mean zero, and variance \( \sigma_i^2 \). Then the covariance between \( \ln(r_i(t)) \) and \( \ln(r_i(m)) \) declines monotonically as the distance between \( i \) and \( m \) increases. Note that distance does not have to mean physical distance. Distance can include any number of characteristics such as school quality, town services, and other factors. To allow for general characteristic spaces let \( d_i(i,m) \) represent some measure of the distance between the two locations at time \( t \). The model estimated in this paper makes the particular assumption that \( \text{cov}(\ln(r_i(t)), \ln(r_i(m))) = \sigma_i^2 \exp(-d_i(i,m)) \). In general one can use any functional form such that the covariance declines monotonically in \( d \).

While housing returns in a neighborhood may be highly correlated, obviously there still remain individual differences among homes. To account for changes peculiar to a

\(^3\)As will be seen, the model estimated here employs several factors including a neighborhood's racial composition, median income, and educational attainment.
house, the model introduces a final random variable $\varepsilon_i(t)$. This variable represents one plus a house specific return in period $t$. The model assumes that $\ln(\varepsilon_i(t))$ is normally distributed with mean zero, and variance $\sigma^2_{\varepsilon}$. Because $\varepsilon$ represents events unique to a particular house it seems natural to assume that the $\ln(\varepsilon_i(t))$ are uncorrelated across both time and space.

Based upon the assumptions given above the model in equation (3) can be rewritten as

$$\ln(P_e(t)) - \ln(P_b(t)) = \ln(h) + \ln(\eta) + \sum_{t=b(t)-1}^{s(t)} [\ln(\tau_1(t)) + \ln(\tau_2(t)) + \ln(\varepsilon_i(t))],$$

(4)

where $b(t)$ and $s(t)$ represent the purchase and sale dates for the house at location $i$. While equation (4) provides a description of housing returns it cannot be estimated directly. In order to produce estimates of the locational returns, the empirical procedure must make use of the variance, and covariance relationships. Note that the return structure in equation (4) can depend upon individual housing characteristics. This implies that with the proper data set, any hedonic model can be modified for use in a DWRS estimator.

The DWRS model assumes that the $\ln(\eta)$, $\ln(\tau_i(t))$, and $\ln(\varepsilon_i(t))$ terms in (4) are all normally distributed with zero means and variances of $\sigma^2_{\eta}$, $\sigma^2_{\tau}$ and $\sigma^2_{\varepsilon}$ respectively. Define

$$e(t) = \ln(P_e(t)) - \ln(P_b(t)) - \ln(h) - \sum_{t=b(t)-1}^{s(t)} \ln(\tau_1(t)).$$

(5)

Then treating $\ln(\eta)$, $\ln(\tau_i(t))$, and $\ln(\varepsilon_i(t))$ as error terms one can write down the following
\[
E[(e(t) - e(m))^2] = 2\sigma_n^2 + (n(t) + n(m))(\sigma_r^2 + \sigma_e^2) + 2\alpha(t, m)\sigma_\epsilon^2 e^{-d(t, m)}.
\] (6)

Here, \(n(t) = s(t) - b(t)\) represents the number of periods between sales for the house at location \(t\), and \(\alpha(t, m) = \max\{0, \min[s(t), s(m)] - \max[b(t), b(m)]\}\) which equals the number of periods both homes overlap in the data set. Equation (6) provides a means for estimating the model's variance and distance parameters and distinguishes the DWRS methodology from both WRS and hedonic models. When potential homeowners enter the market, they must choose among imperfect substitutes. However, while public services, and other housing attributes do vary there exists a limit to which housing prices can differ among neighborhoods. The distance function (d) in equation (6) provides an estimate of that limit.

If an attribute has a large coefficient in d, that means individuals place a high value on it, and thus if two homes differ in it the homes are poor substitutes. Conversely, homes can differ widely on an attribute with a small coefficient, and buyers will still consider substituting one house for the other.

One should note that just because an attribute has a large impact on the value of a house does not mean that it necessarily has a major impact on substitutability, and vice versa. For example, imagine that people will pay a great deal more for a home in an area with superior recreational facilities. This does not imply that homes in an area with substandard recreational facilities cannot act as a perfect substitute. If the housing prices in the poor quality district are low enough, then homeowners may decide that given the price difference they can afford to purchase recreational services in the private market.
Under these conditions the returns to the two neighborhoods will remain nearly identical. Location can act in exactly the opposite manner. Homes on the north end of town may sell for the same amount as those on the south end. However, if the distances are large enough, then the neighborhoods will be poor substitutes for each other.

Using equations (4) and (6) one can use various statistical techniques to estimate the model's parameters. This paper employs a two step procedure that first estimates (4) and then uses the residuals to estimate (6).

Once the model has been estimated, the task of producing local area indices still remains. At this point in the estimation process one knows the expected return in each area \( \bar{r}_i(t) \) and the variance and covariance relationship among housing returns. However the actual \( r_i(t) \) terms have yet to be estimated. From the projection theorem one can write

\[
\ln(r_i(t)) = \ln(\bar{r}_i(m)) \exp(-d_i(t,m)) - \delta_i(t,m),
\]

(7)

where \( \delta_i \) equals a normally distributed error term with mean zero. To calculate the variance of \( \delta_i(t,m) \) square both sides of (7) to get,

\[
\sigma^2_i = \sigma^2 \exp(-2d_i(t,m)) + \sigma^2_\delta,
\]

(8)

which rearranges to

\[
\sigma^2_\delta = \sigma^2_i [1 - \exp(-2d_i(t,m))].
\]

(9)

Notice that when two houses occupy the same location \( d_i(t,m) = 0 \), and thus the two homes
have the same return index. As the distance between the homes increases, \( \sigma^2 \) increases which implies that the correlation between the housing return indices declines.

To utilize (9) replace the location parameter in (4) with \( m \). Now multiply and divide the \( \ln(r_i(m)) \) terms by \( \exp(-d_i(t,m)) \). Then substitute out the \( \ln(r_i(m))\exp(-d_i(t,m)) \) by using (7), to get

\[
\ln(P_s(m)) - \ln(P_b(m)) = \ln(h) + \ln(\eta) + \sum_{t=0}^{s} \ln(r_i(m)) + \frac{\ln(r_i(t)) + \delta_i(t,m)}{\exp(-d_i(t,m))} + \ln(\epsilon(t)).
\]  

(10)

By treating the \( \ln(r_i(t)) \) and \( \ln(h) \) terms as known, one can use least squares techniques to produces consistent estimates of the local area returns.

Notice that equation (10) allows one to construct neighborhood indices. Thus, the model can produce accurate indices even if returns vary from one local to the next, or regional sales densities change over time.

2.2 Comparing Nonparametric Estimators with DWRS

In general nonparametric models assume that the researcher has access to observations \( y \) at various locations \( t \), generated via an equation of the form

\[
y(t) = f(t) + \epsilon(t).
\]  

(11)

Here, \( f \) represents an unknown function, and \( \epsilon \) a white noise error term. What makes nonparametric models so appealing is that they allow one to estimate \( f \), without specifying
its functional form. Essentially, one estimates \( f \) at location \( \ell \) by using a weighted average of the available observations closest to \( \ell \), with the more distant observations having less weight. If \( f \) does not vary too much over the region from which the data has been selected, then the estimated value of \( f \) should be reasonable accurate.

The spatial model presented in the previous section bears many similarities to the nonparametric methodology. In particular, both models estimate the value of the unknown function at a location \( \ell \) by reweighting data depending upon its proximity to \( \ell \). While both techniques use a weighted regression analysis to form estimates, the derivation of the distance estimates, and the weighting function differs across the models. With a nonparametric model the researcher must exogenously specify the number of observations used to estimate \( f \) at each location, exogenously specify a hedonic space in which to calculate distance, and a weighting function based upon the specified distances. In contrast the spatial model produces estimates of the appropriate distance metric, and statistical updating rules then determine the appropriate weights.

To highlight the differences consider the housing model from the previous section. A typical nonparametric model may use the "closest" 100 observations to location \( \ell \). However, when the characteristic space contains several variables determining how far away two observations are from each other may not be trivial. For example, suppose the model states the distance between observations as \( \Delta \text{miles}(\ell,m)+\Delta \text{feet}^2(\ell,m) \), which equals the number of miles plus the difference in square feet between homes at locations \( \ell \) and \( m \). This model implicitly assumes that one extra mile of distance has the same impact as one extra square foot. One can easily see the problem by considering the choices a European researcher may employ. Since Europe uses the metric system, the researcher
may prefer a distance metric of $\Delta$kilometers$(t,m)+\Delta$meters$^2(t,m)$. In this case, and in
general, changing the units for measuring distance will also change the empirical
estimates. In contrast, a spatial model of the same problem employs a distance metric of
$\beta_0\Delta$miles$(t,m)+\beta_1\Delta$feet$^2(t,m)$ and then estimates parameters $\beta_0$ and $\beta_1$. Thus, a change in
scale will simply result in different estimates of $\beta_0$ and $\beta_1$, leaving the estimated distances
unaffected.

Once a suitable distance metric has been specified both the nonparametric and
spatial models must define a weighting system. In a nonparametric model the researcher
exogenously specifies a mechanism by which more distant observations are given less
weight in the estimates. For example, one may use $\exp(-d(t,m))$, where $d(t,m)$ equals the
specified distance between observations. Absent additional information, nonparametric
methods do not provide any particular guidance as to the optimal weighting scheme. On
the other hand, a spatial model provides an exact formula. Using the housing model in the
previous section, the covariance of the returns equals $\sigma^2 \exp(-d(t,m))$, where both $\sigma^2$ and
the function $d$ are both estimated. The optimal weighting scheme then follows from the
mathematics governing equation (10).

As the above discussion indicates, any nonparametric estimator can be duplicated
within a spatial model, by exogenously specifying the distance metric, and the weighting
function. What a spatial model offers is the opportunity to estimate the parameters
governing the statistical distance between observations, and the proper weighting scheme
to produce unbiased estimates.
3 San Francisco Bay Area Housing Returns

3.1 Data and Hedonic Controls

The data base contains 131,603 repeat sales in the San Francisco Bay Area covering the period 1980 through 1994. Each observation consists of the zip code in which the sale took place, and the date and price of each sale. While it may be useful, for estimation purposes, to obtain the distances between every home in the data set this is not possible in practice. Instead the statistical model assumes that all homes within a zip code are located at the population centroid for that zip code. These spatial locations are particularly useful since there exist several commercial data bases with this information. In addition to the repeat sales data, the model makes use of Census Bureau information. Fortunately, census data can be obtained on a zip code by zip code bases, and this forms the paper's hedonic data set.

One important justification for using the hedonic measures as the basis for estimating the local indices is that they act as controls for variation in housing types and quality. Selection bias is a major concern in the construction of housing indices. Gatzlaff and Haurin (1993) point out that repeat-sales indices may be upward biased due to observations being conditioned upon sale. Taylor (1992 and 1983) shows that in a market with heterogeneous assets a spurious negative autocorrelation may be induced through quality-selection filtering. Goetzmann (1995) simulates this filtration for various seller reserve rules and shows how the index may be biased. The use of hedonic controls has
the potential for reducing or eliminating certain types of selection bias, if the selection is conditional upon socio-economic or geographical variables. To illustrate this point, suppose that homes in lower income neighborhoods have lower capital appreciation, and turn over infrequently when compared to their higher income counterparts. Low income areas will thus be under-represented in a repeat-sale dataset, and as a consequence the estimated city-wide index will be biased because it is based solely upon homes that sold, i.e. homes in high-income neighborhoods.

In fact, the analysis in this paper finds clear evidence of exactly this kind of selection in the San Francisco Bay repeat-sales data. Table 3, reports the percentage of transactions, by quintiles of socio-economic characteristics in the database. It shows that less than 6% of the observations in the database come from zip codes representing the lowest quintile of median household income, while more than 30% of the sales data comes from the highest quintile. We do not know whether this discrepancy is due to the differential rates of home ownership across median family income levels, or is due to the low turnover rates for homes in lower income neighborhoods. The table also shows that zip codes with the highest proportion of white residents are under-represented, and the zip codes with the lowest proportions of black and Hispanic residences are likewise under-represented. In fact, the table indicates that an index for the San Francisco Bay area that equally weights all repeat-sales observations will mostly capture the behavior of middle-income neighborhoods of average racial composition and educational attainment. Another way to look at the fluctuating composition of a city-wide index that does not employ hedonic controls is to consider the variation in the number of transactions which come from any given neighborhood. Figure 1 shows the fluctuations one might expect. The graph takes
three sample zip codes and plots the variation through time of the fraction of repeat-sales data that it contributes to the estimation procedure. The contribution is normalized to the zip code's average contribution. Thus, a value of two at a specific point in time indicates that, for that particular quarter, the zip code is represented by twice as many observations as is typical for that zip code. The zip codes were chosen as the two largest in terms of sample size (94550 and 94583 with 7,312 and 5,976 observations respectively) and the median zip code in terms of sample size (94401 with 1,044 observations). Notice that proportions vary dramatically though time, as sales are concentrated first in one area and then another. If the rate of sales is conditional upon the capital appreciation, as conjectured in Gatzlaff and Haurin (1991) and Goetzmann (1995), then this variation will almost certainly induce a positive bias in a resulting index. The distance-weighting procedure implicitly controls for this geographic variation in transactions used in the index by creating an index for one specific location. Once the individual location indices are estimated, the researcher is free to choose any weighting scheme, tilted towards any set of hedonic characteristics or sales processes he or she chooses. In the description that follows, we present the methodology for the distance-weighted repeat-sale measure.

3.2 Empirical Model and Estimation Technique

The actual model uses the following functional specification for the expected return per quarter,
\[
\ln(r_t(t)) = a_0 + a_1 \cdot MEDINC + a_2 \cdot EDUC16P + a_3 \cdot BLACK + \\
a_4 \cdot ASIAN + a_5 \cdot HISPANIC + a_6 \cdot OTHER
\]  
(12)

The \(a_i\)'s represent parameters for estimation, and the variables in the above equation represent the following information for the zip code in which a home is located:

**MEDINC:** Median income in thousands of dollars,

**EDUC16P:** Percentage of the population with a college or graduate degree,

**BLACK:** Percentage of the population identifying themselves as black,

**ASIAN:** Percentage of the population identifying themselves as Asian,

**HISPANIC:** Percentage of the population identifying themselves as Hispanic,

**OTHER:** Percentage of the population identifying themselves as neither white, or in any of the above categories.

To obtain the covariance among housing returns we estimate two euclidean distance specifications. The first uses several census variables in addition to physical distance,

\[
de(t,m) = (b_0 \cdot \Delta MILES^2 + b_1 \cdot \Delta MEDINC^2 + b_2 \cdot \Delta EDUC15L^2 + \\
b_3 \cdot \Delta EDUC16P^2 + b_4 \cdot \Delta WHITE^2 + b_5 \cdot \Delta BLACK^2 + \\
b_6 \cdot \Delta ASIAN^2 + b_7 \cdot \Delta HISPANIC^2 + b_8 \cdot \Delta OTHER^2)^{1/2}
\]  
(13)

The variable \(\Delta MILES^2\) equals the squared distance, in miles, between the homes located at \(t\) and at \(m\). The term \(\Delta WHITE^2\) equals the squared difference in the percentage of the population identifying themselves as white. Finally, \(\Delta EDUC15L^2\) equals the squared difference in the percentage of the population with less than four years of college. All of
the other variables have meanings analogous to those in equation (12), except that the Δ symbol represents the difference in the attribute between the zip codes for house \(i\) and \(m\), and the superscript 2 indicates that the analysis uses the squared difference.

To form a base case, we estimate a second specification that uses only the millage variable

\[
dp(i,m) = b_o \cdot \text{abs}(\Delta \text{MILES}). \tag{14}
\]

Estimation of the model's parameters and standard errors occurs in several stages. The estimation process begins with the standard WRS methodology by running (12) with an OLS procedure. As is well known, this produces consistent estimates of the parameters in question. Next, for each observation \(i\), the prediction error from the first stage \((e_i)\) is paired with the prediction error from observation \(i+1\), except for the last observation which is paired with observation 1. These errors are then used in (6) to produce a system of \(N\) equations. The parameters for either equation (13), or (14) are then estimated via nonlinear least squares (NLLS).

Because of the spatial correlation structure, the variance-covariance matrix has an entry at every location, which impacts the calculation of the standard errors. The variance of the \(e_i\)'s can be found in equation (6),

\[
E[e(i)^2] = \sigma_n^2 + n(i)(\sigma_r^2 + \sigma_t^2), \tag{15}
\]

and from the same equation the covariances equal

\(^4\)Prior to estimation the observations are randomly ordered, to prevent spatial or temporal clustering in this stage of the estimation process.
\[ E[e(t)e(m)] = \sigma^2 e^{-d(t,m)}. \] (16)

Let \( V \) equal the variance-covariance matrix formed from (15) and (16). Then the formula for the covariance of the OLS parameter estimates can be written as \((x'x)^{-1}x'Vx(x'x)^{-1}\). In principle one can obtain a more efficient estimate of the parameters in (12) by using GLS for estimation. Given the large sample sizes in this study, and the technical difficulties associated with inverting very large matrices, we are only able estimate the OLS equation, and report the OLS standard errors. Because of the covariances that exist across locations these standard errors are underestimated. Nevertheless, under the null hypothesis that there does not exist any spatial correlation (so that either \( \sigma^2 \) equals zero, or all of the homes are "infinitely" far from each other) they should be approximately correct.\(^5\) As another test the paper also presents standard errors generated from a bootstrap procedure which we describe below.

Given a sample size in excess of 100,000 observations, present day computer resources do not permit the calculation of analytical standard errors for the parameters in equation (6). As an alternative, the analysis uses a bootstrap procedure in which the computer randomly draws repeat-sales observations from the data with replacement, and then re-estimates the model. This procedure is justified under the null hypothesis that all of the observations have independent errors.

\(^5\)Only approximately since the time between sales still induces some heterogeneity in the variance of the error terms.
3.3 **Empirical Results: Factors Affecting Covariance**

The San Francisco Bay area has a diverse population and a diverse set of neighborhoods. Table 1 displays the distribution of the data by zip code. For example, the median column lists the median value among all zip codes in the data set. As one can see while most neighborhoods are predominantly white, quite a few have high minority concentrations. The white population varies from as low as 7.8% to 100% of the residents in particular zip codes. Among the minority populations blacks show the most neighborhood heterogeneity, ranging from 0% to 84% of the population. While Asians and Hispanics represent somewhat smaller population groups, some zip codes have concentrations of over 33%.

Since only 1980 and 1990 census data are available, we obviously measure several of the independent variables with error. To get some idea as to how serious this problem may be Table 2 lists the correlation between the 1980 and 1990 census population figures. Except for the OTHER race the figures along the diagonal are quite high.\(^6\) This implies that neighborhoods tend to be quite stable in demographic terms, thus mitigating the errors in variables problem. Among the cross-correlations, the Asian-black, and Asian-Hispanic results seem quite surprising. In general one expects the racial variables to display negative correlations, since the percentages are constrained to add up to 100. Nevertheless, these two cross correlations are positive. This implies high 1980 levels in

---

\(^6\) The other category tends to be very small and thus may be volatile on a percentage basis. Further, we suspect that from 1980 to 1990 many individuals that may have previously identified themselves as "OTHER" may have moved into one of the larger categories.
the black or Hispanic populations tend to be associated with above average 1990 levels in the Asian population.

Estimates of equation (12) can be found in Table 3. Contrary to the finding in Goetzmann and Spiegel (1995) the ln(h) estimate comes out negative. This implies that on average sales "reduce" the value of a house. A more reasonable interpretation is that once a house has been sold, an immediate resale must go to the next highest bidder, resulting in a negative intercept.

The median income parameter has a negative coefficient, which implies homes in wealthy areas appreciated at a lower rate than homes in poorer neighborhoods. In contrast to the median income variable, the education variable has a positive sign. All else equal, better educated neighborhoods had higher returns than less educated neighborhoods. Roughly, these two results seem to imply that, on average, university neighborhoods provided superior returns.

The omitted racial variable is white. Thus, increasing the fraction of the neighborhood population that is either black or other and reducing the white population reduces returns. In contrast, increasing either the Asian or Hispanic populations, while decreasing the white population increases returns. However, one must keep in mind that the estimated coefficients are very small. To provide some feeling for the magnitudes, returns in a hypothetical 100% black neighborhood (the lowest possible value) were .0091 lower per quarter than returns in a hypothetical 100% Hispanic neighborhood (the highest possible value).

Since the spatial model allows us to estimate what factors affect neighborhood substitutability we now turn to the estimates of equation (6) for further evidence. The
estimates from the three specifications used to estimate the distance between houses can be found in Table 4. A priori we know that the parameters on each variable cannot be negative. Thus, it seems sensible to restrict the estimation process to the nonnegative values. One way to do this is by estimating $\exp(\text{parameter})/100$, which also ensures that confidence intervals will remain within feasible parameter values.\(^7\) Thus, our estimate of $b_0$ (the mileage parameter) in equation (13) equals $\exp(-14.515398)=0.0000000049664$. Similar calculations for equation (14) yields an estimate of 0.000031894168. Notice that even when the characteristic space only includes physical distance, the mileage parameter plays only a minor roll. Consider two homes one located 30 miles north of San Francisco, the other 30 miles south. The model (14) estimates indicate that their local component returns will have a covariance about 99.8% as large as the return covariance for two homes in the same neighborhood. One might have expected this result a priori. Imagine there are three neighborhoods A, B, and C, with B between A and C. Then the returns in A and C will be closely linked through B. If prices in A go up that will force up prices in B, which in turn will force up prices in C. If people moving into the area are sufficiently flexible among adjoining areas, the return correlation among A and C will be very high.

The inability of pure physical distance model (14) to fit the data can be seen in the estimate of $\sigma^2_i$. The combination of the small parameter estimates for mileage, and $\sigma^2_i$ cause the model to treat all homes as if they exist in the same neighborhood, with substantial degrees of house specific risk. In contrast, the full model allows for a much greater degree of correlation in the return to homes at similar locations, and less house

\(^7\)The division by 100 helps to scale the numbers and speed up convergence.
specific risk. The peculiar structure of the San Francisco Bay Area housing market may explain why physical distance plays such a small role. The Bay Area has numerous hills and valleys. As a result, the value of a home often depends upon its altitude above sea level. Homes high up on a hill are worth more. Thus, knowing the distance between two homes may tell a buyer much less than knowing their relative altitudes. Since most other cities do not share this geological feature, physical distance may play a larger role in their estimated housing returns.

While physical distance plays a minor role in the correlation estimates when it appears alone, it becomes completely negligible when other neighborhood characteristics enter the model. Rather, median household income dominates the estimates. From Table 1 the difference between two neighborhoods in the top and bottom quartile equals 7.7 thousand dollars. For model (13), if all the other differences equal zero, this difference in median income will reduce the return covariance between the neighborhoods by 34%. While the race and education variables also play a strong role, they are much less important. Nevertheless changing the racial composition of a neighborhood seems to change the covariance of returns with respect to other neighborhoods. Based upon the estimates from model (13) the return covariance between homes in an all white neighborhood and an otherwise identical all black, Asian and Hispanic will exhibit 20%, 19%, and 17% of the covariance of homes in the same neighborhood. If whites are replaced in the above comparison with any other group the return covariance percentages will fall even further. Estimates of conditional covariances between neighborhoods is useful for a number of reasons. Low covariances among areas separated in terms of
median household income implies that there are potentially large reductions in risk to portfolios that are diversified across neighborhoods. For mortgage lenders, the dynamics of the housing pricing indices within a metropolitan area reflect changing loan-to-value ratios. Our analysis indicates that lending in only high median family income neighborhoods, while potentially attractive in terms of qualifying a borrower, may have adverse portfolio effects. Conversely, a policy of lending across income-diversified neighborhoods has potential for reducing the volatility of the portfolio. The same guidelines hold true for equity investors in housing, as well as for municipalities. Property tax flows are conditioned, over the long-term, upon capital appreciation. Analysis of how rates differ across neighborhoods may provide useful information about the uncertainty of future cash flows, and guidelines for targeted urban development. While the evidence for mortgage lender red-lining is mixed (see Holmes and Horvitz (1994), and Schill and Wachter (1993) for example), it would appear to have been a poor strategy from a risk and return perspective in the San Francisco Bay area over the period of our study. Diversification across neighborhoods with varying income and educational compositions would have reduced portfolio volatility, measured in loan-to-value ratios.

An important caveat in the interpretation of the estimated coefficients is that the variables are not independent. Median household income in the 1990 census is positively correlated with the proportion of white residents, and negatively correlated to the proportion of black residents. This colinearity is likely to affect the magnitude of the coefficient estimates, and indeed may even affect the sign of a less detailed model. For instance, suppose one omits median household income from the specification. Then racial composition will appear to be a fair instrument. The specification may thus produce a
positive relative return for non-white neighborhoods. The same potentially holds true for geographical distance. In cities where a home’s location on a hill plays less of a role the household income will vary geographically. In this case, the omission of median household income may make physical distance appear to be a key variable defining the covariance of errors. While economists are willing to make the assumption of ceteris paribus for purposes of analysis, home buyers, lenders and policy-makers may not have that luxury, and any "comparative statics" applied should be based upon reasonable expectations about covariation in variables.

3.4 Empirical Results: Indices

The indices that result from the distance-weighted repeat-sales procedure demonstrate how widely the dynamics of capital appreciation may vary, even within a single metropolitan area. Figure 2 plots five zip code indices chosen from the sample of 188. Notice that the cumulative returns over fourteen years vary from a low of 25% to a high of 275%. The figure also shows two distinct “groups.” Zip codes 94554, 94586 and 95450 have all had nearly flat returns since 1990, while 94544 and 94621 have shown increases in the 1990’s. Our study is confined to the San Francisco Bay area, and thus differences in the capital appreciation dynamics across zip codes may not be as dramatic for different metropolitan areas. In Goetzmann and Spiegel (1995) we found that the average correlation in housing returns within the San Francisco Bay area to be small relative to three other areas: Dallas, Chicago and Atlanta. Map 1 illustrates the differences in the five-year capital appreciation returns by quartile. It is a patchwork of
different rates of return, with areas of high return interspersed with areas of low return. A closer look at the map shows, however, that there is some aggregation at the local level. Contiguous zip codes typically “lump” together in the same capital appreciation quartile, due to continuities in socio-economic variables within the region. For instance, in the lowest quartile a dollar invested at the end of 1988 returned between $.97 and $1.16 in nominal terms. Typical areas with this low capital appreciation are cities in the Northeast Bay such as Martinez, Pinole and Benizia, as well as towns in the South Bay such as Hayward City and Union City. Zip codes with the high capital appreciation, ranging from 23% to 36% over the past five years are a bit more concentrated in Marin County, Berkeley and the environs west of Palo Alto. The range of rates of return in the Bay area suggests that one city-wide index is likely to do a poor job at estimating the dynamics of housing appreciation within any single neighborhood. This is bad news for practitioners using indices constructed on a city-wide level to calculate loan-to-value ratios. The magnitude of the deviation from a single index can be important. or example, the five year returns for the zip codes in our sample range from -3% to +36%.

Map 2 shows why the estimates in Table 5 indicate that zip codes with different incomes tend to have weakly correlated returns. In this map the darker areas correspond to higher income levels. Now compare the results with Map 1. Notice that, to some degree, the dark areas in Map 1 (high income) correlate with the light areas of Map 2 (low returns). While the relationship is by no means perfect, it does appear stronger than the physical distance-return relationship. The reason for this comes from the dispersal of high income neighborhoods throughout the Bay Area. As a result, high return neighborhoods (for the five year period) are also scattered in a somewhat similar pattern.
4 Conclusion

Previous methods for estimating the returns to illiquid assets have fallen into either the repeat sales or hedonic modeling classes. The spatial model presented here can be interpreted as an extension of both these approaches. In particular, it overcomes some of the selection bias problems associated with the estimation of a city-wide index when socio-economic characteristics are associated with resale frequency. In addition, the distance-weighted repeat-sales procedure offers potential improvements over nonparametric estimators. Nonparametric models assert a distance metric and weighting scheme. In contrast, spatial models permit one to estimate the distance metric and then use statistical updating rules to determine the proper weighting scheme.

An examination of housing returns in the San Francisco Bay area using the distance-weighted repeat-sales model reveals patterns of interest to homeowners, mortgage lenders and civic authorities. First, it appears clear that zip codes in the Bay area had widely differing rates of return over the past fourteen years. This is certainly of interest to lenders wishing to recalculate loan-to-value ratios of housing loan portfolio. It is also of interest to homeowners wishing to gauge how far their neighborhood return might deviate from the return of homes in a typical neighborhood. It is also useful to both lenders and home investors to know that low income and better educated neighborhoods had higher returns than high income or low education neighborhoods, at least in the metropolitan area examined in this study. A cautionary note to “redliners” in mortgage lending is that, while race had some impact on returns it was minor at best.
The motivation for decomposing a regional housing return index into its neighborhood-level constituents is that the price dynamics contain information about intra-regional substitutability. What we found was people seem to be much less concerned with physical distance than with social distance. The most important empirical result of this study is the finding that socio-economic variables strongly influence the covariance of neighborhood housing returns. The return correlation between high and low income neighborhoods is surprisingly small, and no other factor seems to have as much influence. Apparently, the difference between the "haves" and the "have-nots" extends to the pattern of capital appreciation of their homes. To the extent that the home is the major asset in the lower-income home-owner's portfolio, the differential appreciation rate over the 1980 through 1994 period was effectively a trend towards parity. Whether this trend was due to endogenous or exogenous factors is an open question.
Bibliography


Table 1

Distribution Summary of Socio-Economic Characteristics

This table reports five quantiles of socio-economic variables for the zip codes in the San Francisco Bay area used in the paper. The variables are taken from the 1980 census, and census tracts were mapped into the appropriate zip codes. "MEDINC" refers to the Median Household Income of residents in the zip code. "EDUC15L" refers to the fraction of adult residents surveyed who attained an education level of at least three years in college. "EDUC16P" refers to the fraction of adult residents surveyed who attained an education level of at least a four-year college program. "WHITE", "BLACK", "ASIAN", "HISPANIC" and "OTHER" refer to the fraction of residents in the zip code who reported themselves as belonging to the respective ethnic group. Thus, for instance, the median percentage of white residents across all zip-codes in the area in 1980 was 87.30%.

<table>
<thead>
<tr>
<th>Variable</th>
<th>Min</th>
<th>25%</th>
<th>Median</th>
<th>75%</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEDINC</td>
<td>4.32</td>
<td>17.64</td>
<td>21.53</td>
<td>25.35</td>
<td>42.08</td>
</tr>
<tr>
<td>EDUC15L</td>
<td>28.00</td>
<td>69.40</td>
<td>81.30</td>
<td>86.60</td>
<td>96.70</td>
</tr>
<tr>
<td>EDUC16P</td>
<td>3.30</td>
<td>13.40</td>
<td>18.70</td>
<td>30.60</td>
<td>72.00</td>
</tr>
<tr>
<td>WHITE</td>
<td>7.80</td>
<td>76.00</td>
<td>87.30</td>
<td>93.10</td>
<td>100.00</td>
</tr>
<tr>
<td>BLACK</td>
<td>0.00</td>
<td>0.07</td>
<td>2.10</td>
<td>6.00</td>
<td>84.20</td>
</tr>
<tr>
<td>ASIAN</td>
<td>0.00</td>
<td>2.50</td>
<td>4.50</td>
<td>8.20</td>
<td>36.40</td>
</tr>
<tr>
<td>HISPANIC</td>
<td>0.00</td>
<td>1.30</td>
<td>2.90</td>
<td>6.40</td>
<td>37.10</td>
</tr>
<tr>
<td>OTHER</td>
<td>0.00</td>
<td>0.30</td>
<td>0.60</td>
<td>1.00</td>
<td>8.50</td>
</tr>
</tbody>
</table>
### Table 2

**Correlation between 1980 and 1990 Census Variables**

The table reports the correlation between the values of each variable reported in the 1980 and 1990 census surveys. Census tracts were mapped into the appropriate zip codes. "MEDINC" refers to the Median Household Income of residents in the zip code. "EDUC15L" refers to the fraction of adult residents surveyed who attained an education level of at least three years in college. "EDUC16P" refers to the fraction of adult residents surveyed who attained an education level of at least a four-year college program. "WHITE", "BLACK", "ASIAN", "HISPANIC" and "OTHER" refer to the fraction of residents in the zip code who reported themselves as belonging to the respective ethnic group.

<table>
<thead>
<tr>
<th></th>
<th>MEDINC</th>
<th>EDUC15L</th>
<th>EDUC16P</th>
<th>WHITE</th>
<th>BLACK</th>
<th>ASIAN</th>
<th>HISPANIC</th>
<th>OTHER</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEDINC</td>
<td>0.83</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EDUC15L</td>
<td>-0.51</td>
<td>0.68</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EDUC16P</td>
<td>0.51</td>
<td>-0.68</td>
<td>0.89</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WHITE</td>
<td>0.41</td>
<td>0.07</td>
<td>0.32</td>
<td>0.93</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>BLACK</td>
<td>-0.38</td>
<td>-0.08</td>
<td>-0.20</td>
<td>-0.73</td>
<td>0.98</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ASIAN</td>
<td>-0.05</td>
<td>-0.10</td>
<td>0.01</td>
<td>-0.57</td>
<td>0.18</td>
<td>0.85</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HISPANIC</td>
<td>-0.23</td>
<td>0.06</td>
<td>-0.45</td>
<td>-0.55</td>
<td>0.01</td>
<td>0.50</td>
<td>0.86</td>
<td></td>
</tr>
<tr>
<td>OTHER</td>
<td>-0.35</td>
<td>0.20</td>
<td>-0.32</td>
<td>-0.07</td>
<td>-0.03</td>
<td>0.02</td>
<td>0.18</td>
<td>0.21</td>
</tr>
</tbody>
</table>
Table 3
Percentage of All Observations by Characteristic Quintile

This table reports the socio-economic composition of the sample used in the distance-weighted repeat sales estimates for the San Francisco Bay area. For example, the table indicates that 15.28% of the observations used in the estimation procedure come from zip codes in the second income quintile. Census tracts were mapped into the appropriate zip codes. "MEDINC" refers to the Median Household Income of residents in the zip code. "EDUC15L" refers to the fraction of adult residents surveyed who attained an education level of at least three years in college. "EDUC16P" refers to the fraction of adult residents surveyed who attained an education level of at least a four-year college program. "WHITE", "BLACK", "ASIAN", "HISPANIC" and "OTHER" refer to the fraction of residents in the zip code who reported themselves as belonging to the respective ethnic group.

<table>
<thead>
<tr>
<th>Education 4+ Years College</th>
<th>Income</th>
<th>White</th>
<th>Black</th>
<th>Asian</th>
<th>Hispanic</th>
<th>Other</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quintile 1 (low)</td>
<td>0.0588</td>
<td>0.1691</td>
<td>0.2043</td>
<td>0.0912</td>
<td>0.0659</td>
<td>0.1469</td>
</tr>
<tr>
<td>Quintile 2</td>
<td>0.1528</td>
<td>0.2179</td>
<td>0.2012</td>
<td>0.2652</td>
<td>0.2368</td>
<td>0.2093</td>
</tr>
<tr>
<td>Quintile 3</td>
<td>0.1917</td>
<td>0.2800</td>
<td>0.2245</td>
<td>0.2353</td>
<td>0.2550</td>
<td>0.2088</td>
</tr>
<tr>
<td>Quintile 4</td>
<td>0.2694</td>
<td>0.1674</td>
<td>0.2422</td>
<td>0.2137</td>
<td>0.2224</td>
<td>0.2310</td>
</tr>
<tr>
<td>Quintile 5 (high)</td>
<td>0.3274</td>
<td>0.1657</td>
<td>0.1277</td>
<td>0.1946</td>
<td>0.2200</td>
<td>0.2039</td>
</tr>
</tbody>
</table>
Table 4

Parameter Estimates For the OLS Regression Model

This table reports the ordinary least-squares parameter estimates for a model explaining deviations from the area-wide housing index via a set of socio-economic variables taken from the 1980 census data. OLS and Bootstrap standard errors for the coefficients are reported. The model estimated is:

\[ \ln(r_i(t)) = a_0 + a_1\text{MEDINC} + a_2\text{EDUC16P} + a_3\text{BLACK} + a_4\text{ASIAN} + a_5\text{HISPANIC} + a_6\text{OTHER} \]  

(17)

<table>
<thead>
<tr>
<th>Parameter</th>
<th>$\ln(h)$</th>
<th>$\ln(\hat{T})$</th>
<th>MEDINC</th>
<th>EDUC16P</th>
<th>BLACK</th>
<th>ASIAN</th>
<th>HISPANIC</th>
<th>OTHER</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estimate</td>
<td>-0.011550</td>
<td>0.020195</td>
<td>-0.000078</td>
<td>0.000017</td>
<td>-0.000046</td>
<td>0.000024</td>
<td>0.000045</td>
<td>-0.001581</td>
</tr>
<tr>
<td>OLS SE</td>
<td>0.001301</td>
<td>0.000236</td>
<td>0.000007</td>
<td>0.000004</td>
<td>0.000003</td>
<td>0.000006</td>
<td>0.000007</td>
<td>0.000010</td>
</tr>
<tr>
<td>Bootstrap SE</td>
<td>0.001170</td>
<td>0.000276</td>
<td>0.000010</td>
<td>0.000006</td>
<td>0.000004</td>
<td>0.000007</td>
<td>0.000008</td>
<td>0.0000125</td>
</tr>
</tbody>
</table>
Table 5

Parameter Estimates and Bootstrapped Confidence Ranges:
Socio-Economic Variables and Physical Distance

The bootstrap values are coefficient estimates generated under the null hypothesis that the observations are independent from each other. This conforms to a null that the exponentiated values of the distance parameters are zero. This is generated by randomizing over the repeated-sales observations with replacement. The fractile value is the empirical fractile of the actual value, based upon 250 bootstrap iterations. Extreme fractile values indicate that the actual estimated values do not conform well to the bootstrap under the null. Zero values in the .05 quantile column indicate that, under the null hypothesis, a coefficient value of zero cannot be rejected at traditional confidence levels. The coefficients are generated by the model:

\[ \delta(t,m) = \left[ e^{b_1 \cdot \Delta MILES^2} \cdot e^{b_2 \cdot \Delta MEDINC^2} \cdot e^{b_3 \cdot \Delta EDUC15L^2} \cdot e^{b_6 \cdot \Delta EDUC16P^2} \right. \]
\[ \left. e^{b_4 \cdot \Delta BLACK^2} \cdot e^{b_5 \cdot \Delta ASIAN^2} \cdot e^{b_7 \cdot \Delta HISPANIC^2} \cdot e^{b_8 \cdot \Delta OTHER^2} \right) ^{1/2} \]

(18)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Actual</th>
<th>Min</th>
<th>0.05</th>
<th>0.25</th>
<th>0.50</th>
<th>0.75</th>
<th>0.95 Max</th>
<th>Fractile</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \sigma^2 )</td>
<td>0.066521</td>
<td>0.038294</td>
<td>0.047579</td>
<td>0.059060</td>
<td>0.065833</td>
<td>0.073597</td>
<td>0.081530</td>
<td>0.091020</td>
</tr>
<tr>
<td>( \sigma^2_i )</td>
<td>0.014243</td>
<td>0.011815</td>
<td>0.013904</td>
<td>0.015215</td>
<td>0.016097</td>
<td>0.017001</td>
<td>0.018576</td>
<td>0.019880</td>
</tr>
<tr>
<td>( \sigma^2_r )</td>
<td>0.017786</td>
<td>0.012869</td>
<td>0.013973</td>
<td>0.015372</td>
<td>0.016232</td>
<td>0.017253</td>
<td>0.018547</td>
<td>0.019922</td>
</tr>
<tr>
<td>Miles</td>
<td>0.000000</td>
<td>0.000000</td>
<td>0.000000</td>
<td>0.000000</td>
<td>0.000000</td>
<td>0.000000</td>
<td>0.000000</td>
<td>0.000000</td>
</tr>
<tr>
<td>Med. Income</td>
<td>0.002997</td>
<td>0.000000</td>
<td>0.000446</td>
<td>0.001016</td>
<td>0.001403</td>
<td>0.002013</td>
<td>0.003147</td>
<td>0.005932</td>
</tr>
<tr>
<td>Educ. 15</td>
<td>0.000312</td>
<td>0.000016</td>
<td>0.000104</td>
<td>0.000332</td>
<td>0.000464</td>
<td>0.000606</td>
<td>0.000886</td>
<td>0.001418</td>
</tr>
<tr>
<td>Educ. 16</td>
<td>0.000254</td>
<td>0.000085</td>
<td>0.000211</td>
<td>0.000326</td>
<td>0.000439</td>
<td>0.000568</td>
<td>0.000793</td>
<td>0.001433</td>
</tr>
<tr>
<td>White</td>
<td>0.000000</td>
<td>0.000000</td>
<td>0.000000</td>
<td>0.000000</td>
<td>0.000000</td>
<td>0.000000</td>
<td>0.000000</td>
<td>0.000002</td>
</tr>
<tr>
<td>Black</td>
<td>0.000088</td>
<td>0.000000</td>
<td>0.000000</td>
<td>0.000000</td>
<td>0.000017</td>
<td>0.000091</td>
<td>0.000256</td>
<td>0.000551</td>
</tr>
<tr>
<td>Asian</td>
<td>0.000102</td>
<td>0.000000</td>
<td>0.000000</td>
<td>0.000003</td>
<td>0.000068</td>
<td>0.000271</td>
<td>0.000915</td>
<td>0.001801</td>
</tr>
<tr>
<td>Hispanic</td>
<td>0.000138</td>
<td>0.000000</td>
<td>0.000000</td>
<td>0.000000</td>
<td>0.000015</td>
<td>0.000015</td>
<td>0.000283</td>
<td>0.001054</td>
</tr>
<tr>
<td>Other</td>
<td>0.000494</td>
<td>0.000009</td>
<td>0.000255</td>
<td>0.000391</td>
<td>0.000463</td>
<td>0.000547</td>
<td>0.001717</td>
<td>0.083597</td>
</tr>
</tbody>
</table>

Notes: The coefficients reported are the exponentiated values. The bootstrapped quantiles are based upon 250 iterations.
Table 6
Parameter Estimates and Bootstrapped Confidence Ranges:
Physical Distance Model Only

The bootstrap values are coefficient estimates generated under the null hypothesis that the observations are independent from each other. This conforms to a null that the exponentiated values of the distance parameters are zero. This is generated by randomizing over the repeated-sales observations with replacement. The fractile value is the empirical fractile of the actual value, based upon 250 bootstrap iterations. Extreme fractile values indicate that the actual estimated values do not conform well to the bootstrap under the null. Zero values in the .05 quantile column indicate that, under the null hypothesis, a coefficient value of zero cannot be rejected at traditional confidence levels. The coefficients are generated by the model:

$$de(l,m) = \frac{e^{b_0} \Delta MILES^2}{100} \right)^{1/2}$$

(19)

<table>
<thead>
<tr>
<th>Variable</th>
<th>Actual</th>
<th>Min</th>
<th>0.05</th>
<th>0.25</th>
<th>0.50</th>
<th>0.75</th>
<th>0.95 Max</th>
<th>Fractile</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma^2_{\eta}$</td>
<td>0.066378</td>
<td>0.035896</td>
<td>0.048648</td>
<td>0.059942</td>
<td>0.065475</td>
<td>0.071970</td>
<td>0.080638</td>
<td>0.088859</td>
</tr>
<tr>
<td>$\sigma^2_i$</td>
<td>0.008882</td>
<td>0.008290</td>
<td>0.008995</td>
<td>0.009790</td>
<td>0.010295</td>
<td>0.010883</td>
<td>0.012102</td>
<td>0.012613</td>
</tr>
<tr>
<td>$\sigma^2_{e}$</td>
<td>0.023156</td>
<td>0.020146</td>
<td>0.020669</td>
<td>0.021466</td>
<td>0.021976</td>
<td>0.022542</td>
<td>0.023297</td>
<td>0.023731</td>
</tr>
<tr>
<td>Miles</td>
<td>0.000032</td>
<td>0.000000</td>
<td>0.000000</td>
<td>0.000000</td>
<td>0.000000</td>
<td>0.000001</td>
<td>0.000006</td>
<td>0.000017</td>
</tr>
</tbody>
</table>

Notes: The coefficients reported are the exponentiated values. The bootstrapped quantiles are based upon 250 iterations.
Figure 1: This figure takes three sample zip codes and plots the variation through time of the fraction of repeat-sales data that it contributes to the estimation procedure. The contribution is normalized to the zip code's average contribution. Thus, a value of two at a specific point in time indicates that, for that particular quarter, the zip code is represented by twice as many observations as is typical. The zip codes were chosen as the two largest in terms of sample size (94550 and 94583) and the median zip code in terms of sample size (94401).
Sample Zip Code Indices in the SF Area
Capital Appreciation 1980 - 1994

Index Level

0 0.5 1 1.5 2 2.5 3


Year

94544 94621 95450
94709 94586

**Figure 2:** This figure shows the capital appreciation index over the 1980 through 1994 period for five representative zip codes in the San Francisco Bay area, estimated by the distance-weighted repeat-sales method. The index is scaled to show the changing value of $1 invested in housing in December, 1979.