Machine Learning and the Spatial Structure of House Prices and Housing Returns*

Andrew Caplin, Sumit Chopra, John Leahy, Yann LeCun, and Trivikrmaman Thampy†

December 14, 2008

Abstract

Economists do not have reliable measures of current house values, let alone housing returns. This ignorance underlies the illiquidity of mortgage-backed securities, which in turn feeds back to deepen the sub-prime crisis. Using a massive new data tape of housing transactions in L.A., we demonstrate systematic patterns in the error associated with using the ubiquitous repeat sales methodology to understand house values. In all periods, the resulting indices under-predict sales prices of less expensive homes, and over-predict prices of more expensive homes. The recent period has produced errors that are not only unprecedentedly large in absolute value, but highly systematic: after a few years in which the indices under-predicted prices, they now significantly over-predict them. We introduce new machine learning techniques from computer science to correct for prediction errors that have geographic origins. The results are striking. Accounting for geography significantly reduces the extent of the prediction error, removes many of the systematic patterns, and results in far less deterioration in model performance in the recent period.

Key Words: House prices, machine learning, default, sub-prime crisis

*We thank Leslie Greengard, Damien Weldon, and Ed Wike for their help. We thank New York University for its foresight in providing the grant that supported this inter-disciplinary research.

†New York University: Department of Economics; Department of Computer Science; Department of Economics; Department of Computer Science; and Department of Economics. Corresponding author Andrew Caplin: andrew.caplin@nyu.edu

1 Introduction

Despite the essential role they have played in the recent economic turmoil, we remain desperately ignorant of the determinants of housing returns. This ignorance begins at a very primitive level: we do not even have reliable measures even of current house values, let alone changes in value. Our current ignorance underlies the illiquidity of mortgage-backed securities, which in turn feeds back to deepen the current sub-prime crisis.

This paper presents first fruits of a research program in which economists and computer scientists at NYU are looking jointly to push forward our understanding of house prices. Such improved understanding is vital for those interested in assessing the level of house price risk that the financial sector is currently bearing in terms of likely default on outstanding mortgages. It is equally vital in assessing the financial risk is involved in promising use of future taxes to support guarantees being offered by Fannie Mae and Freddie Mac. Improved modeling of house prices will enable us to uncover the economic fundamentals impacting their spatio-temporal evolution, as well as the extent to which individual house price risk can be hedged by index-based securities.

As detailed in section 2, the most important house price measures in current use are based on the repeat sales methodology, including the ubiquitous Case-Shiller (henceforth CS) index. Using a massive new data tape of housing transactions in L.A. comprising over **1.5 million** transactions spread over some **25 years**, we explore the relevance of this index to valuation of homes that have not yet been sold, but may be so in a short period. Contrary to the assumption of the underlying statistical model of house prices, we find systematic patterns in the error term:

- *Mean reversion*: The index systematically under-predicts future returns on cheaper homes and over-predicts future returns on expensive homes.
- *Patterns over turnover time*: Performance deteriorates monotonically as the time between sales increases, and returns on homes that transact after long periods of time are systematically over-predicted.
- *Patterns by date*: In the period Jul:Dec 2007, a period that saw a significant price decline in LA, we find that the CS model has a systematic over-prediction of prices, having under-predicted in the previous years. Moreover the errors in the CS model increased dramatically after 2007.

In the current work, we show that these systematic patterns in house prices are greatly reduced when we take explicit account of geography. While the hypothesis that many patterns in returns have geographic origins is of long standing, practical confirmation has been held back by the massive computational requirements of models that look to uncover spatio-temporal patterns in housing returns. It is in regard of working with such massive data tapes that our introduction of new machine learning techniques from computer science is essential.

The results we derive with our measurement methodology are striking. First, we find that accounting for geography eliminates all patterns in the error term with turnover time. Second, our spatio-temporal model shows no bias in predicting prices of homes with low initial prices (although bias remains for high priced homes). There is also less deterioration of the model performance in the recent period, and far less systematic bias by date. The stark differences between the two models clearly imply that accounting for geography has first order effects on prediction accuracy and bias.

Following the literature review of section 2, we outline in section 3 the data tape of housing transactions in Los Angeles County that we construct for purposes of model estimation. In section 4 we detail the predictive accuracy of the Case-Shiller repeat sales index in this data tape. In section 5 we summarize the prior literature on the effects of location on house prices. In section 6 we outline our own spatio-temporal model of house prices, while section 7 presents our findings on the performance of this empirical framework. The results open the door to new research in the many areas of economics in which house prices play a role, including a wide range of research programs that focus not only on the determinants of house prices, but also mortgage default, and systematic patterns of local house prices.

2 Literature Review

2.1 Hedonic Pricing

The hedonic price equation says that the price of a house is a function of its attributes, each of which has an implicit market price. However, there is no theory directing the researcher to the theoretically correct functional form for the hedonic equation. Dubin [1998] was the first to seriously investigate the choice of the functional form. He used the linear Box-Cox transform of

the hedonic price equation, which is a special case of the following quadratic Box-Cox transform:

$$P^{(\theta)} = \alpha_0 + \sum_{i=1}^{k} \alpha_i Z_i^{(\lambda_i)} + \frac{1}{2} \sum_{i=1}^{k} \sum_{j=1}^{k} \gamma_{ij} Z_i^{(\lambda_i)} Z_j^{(\lambda_j)},$$

where P is the price, Z_i are the attributes, and $P^{(\theta)}$ and $Z_i^{(\lambda_i)}$ are the Box-Cox transforms as given below,

$$\begin{aligned} P^{(\theta)} &= \frac{P^{\theta} - 1}{\theta}, & \theta \neq 0 \\ &= ln(P), & \theta = 0 \\ Z_i^{(\lambda_i)} &= \frac{Z_i^{\lambda_i} - 1}{\lambda_i}, & \lambda_i \neq 0 \\ &= ln(Z_i). & \lambda_i = 0 \end{aligned}$$

The linear Box-Cox transform obviously does not involve the quadratic part. Goodman assumes that $\lambda_i = 1 \; \forall i$. Maximum likelihood estimates reject both the linear ($\theta = 1$) and the semi-log ($\theta = 0$) functional forms.

Halvorsen and Pollakowski [1981] use the more general quadratic transform, while assuming that $\lambda_i = \lambda_j = \lambda$. They use ML to jointly estimate all parameters of the model. They reject all popularly used functional forms from linear, semi log, log linear ($\theta = 0$, $\lambda = 0$, $\gamma_{ij} = 0$), trans-log ($\theta = 0$, $\lambda = 0$), and several quadratic forms. The ML estimates of θ and λ are found to be 0.06 and 0.28 respectively which suggest non-linearities in the pricing equation.

In addition to problems of functional form with given regressors, it is impossible to gather data on all relevant characteristics of a home and its neighborhood. This data might either not be recorded/available or just too costly to obtain. Hence omitted variables are likely to be a major problem in any linear parametric pricing function. Data quality also varies by location. Finally, depending on how the index is computed, hedonic models either assume constant marginal prices over time or run pair-wise regressions for consecutive time periods to compute an index. The latter approach selectively ignores large chunks of information. Given these fundamental problems with using hedonics, most indices in use today are based on the repeat sales methodology.

2.2 Repeat Sales Indices

Bailey et al. [1963] (henceforth BMN) proposed that the heterogeneity problem could be solved by looking only at repeat sales. If houses do not change attributes over time, and if there is no selection bias in selecting only households that have been sold at least twice, then this approach constructs a house price index with desirable properties. The basic model here is as follows:

$$\frac{P_{it'}}{P_{it}} = \frac{B_{t'}}{B_t} U_{itt'},$$

where P_{it} is the price at which house i sells at time t, B_t is the price index at time t and $U_{itt'}$ is an error term. Taking logs we obtain,

$$p_{it'} - p_{it} = b_{t'} - b_t + u_{itt'},$$

where small case represents logs. Defining x_t as 1 in the period t when the house is resold, and as -1 in the period t when the house is first sold, and as 0 in all other periods, the above equation can be written as,

$$r_{itt'} = \sum_{j=1}^{T} b_j x_j + u_{itt'},$$

which in matrix notation becomes,

$$r = xb + u, \tag{1}$$

where $r_{itt'} = p_{it'} - p_{it}$ and T is the total number of periods for which we have sales data. b can now be estimated using OLS and this gives us the log price index.

More than twenty years after BMN wrote their pioneering paper on construction of house price indices using repeat sales in a regression framework, Case and Shiller [1989] (henceforth CS) addressed some of the limitations of the basic model and proposed a method to correct it. In particular CS argue that the variance of the error term must depend on the time interval between sales. The basic intuition is that changes in attributes, which this model does not explicitly account for, are likely to be more significant the longer the time between sales, resulting in a larger unexplained variance in prices for these households i.e. there is a drift in house prices over time. In the BMN method, homes resold after a longer time have a greater impact on the index (because they have a higher unexplained variance as postulated by CS). Hence Case and Shiller argue

that these observations must be under-weighted in determining the price index. They formulate a Weighted Repeat Sales (WRS) index to do this. The basic methodology is as follows.

Log price of the ith house at time t, p_{it} is given by,

$$p_{it} = c_t + h_{it} + n_{it},$$

where c_t denoted the area wide price index, h_{it} (the drift term) is a Gaussian random walk i.e. $h_{it} - h_{it-1}$ is i.i.d $N(0, \sigma_h^2)$, and n_{it} is an i.i.d sale specific random error with distribution $N(0, \sigma_n^2)$. Writing the price difference equation we have,

$$p_{it'} - p_{it} = c_{t'} - c_t + \Delta h_{itt'} + \Delta n_{itt'},$$

where,

$$Var(\Delta h_{itt'} + \Delta n_{itt'}) = \sigma_h^2(t' - t) + 2\sigma_n^2.$$

CS use a 3 stage least squares method to estimate the price index. First they run OLS in the above price difference equation. Then they regress the residuals from that regression on a constant and the time difference between sales. From the above variance formula we know that the constant in that regression would be $2\sigma_n^2$ and the coefficient on the time difference between sales would be σ_h^2. Then they use the fitted values from this second regression to weigh variables appropriately in the first price difference regression to get an estimate of c_t.

The next significant innovation was due to Goetzmann and Spiegel [1995] (henceforth GS), who showed that it is important to account for non-temporal returns from housing in any price index construction using repeat sales data. GS argue that people make changes to the house right around the time of the sale. Since this would not be controlled for in a standard repeat sales model, the price index so constructed would be upward biased. The GS idea can be understood in the CS framework in terms of a simple example. Let us assume that the log of non-temporal return on housing is a constant η. Then the CS log price difference equation can be written as[1],

$$p_{it'} - p_{it} = \eta + c_{t'} - c_t + \Delta h_{itt'} + \Delta n_{itt'}.$$

If the null $\eta = 0$ is rejected, there must be non-temporal returns to housing and a model that does

[1] This equation can easily be derived from a more fundamental price equation

not take this into account would give a biased estimate of the price index. In particular, if $\eta > 0$, the model without the intercept would result in an upward biased price index and vice-versa.

While they are relatively simple in principle to estimate, there are many questions concerning repeat sales indices. One issue concerns the fact that prices are inferred based only on repeat sales in a period, not all sales. Second, the index of today will be revised tomorrow as new data arrives, and there may be systematic patterns in the revisions. In what follows we will apply the index on a wide enough geographic area with many transactions available for each index computation, so that one can hope that the former problem is not overwhelming. Furthermore the revisions problem is of manageable proportions. Nevertheless, the index will be found seriously wanting.

3 The L.A. Data Tape

3.1 Sources of Data

LoanPerformance, a division of First American CoreLogic, supplied NYU with a deeds tape recording all transactions in LA county from Jan 1984 to Apr 2008 as well as the 2007 tax-roll tape. The deeds tape comes from various counties' registry of deeds. The tax-roll tape is compiled information obtained from the tax assessor's office and contains detailed information on all homes in LA county irrespective of their transaction history. Almost all observations have associated with them a valid APN number which uniquely identifies a home. It is therefore possible to combine the tapes and to add in neighborhood specific attributes for all homes in LA county.

The deeds tape has a total of 14,713,346 recorded transactions between Jan 1984 and Apr 2008. Of these, 13,527,413 have valid APNs. We consider only arms length transactions that are resales or new constructions with either a grant deed or foreclosure document type. This leaves us with 3,630,759 property transactions for which we record:

- Sale price: This is the price at which the property transacted.
- Recording date: This is the date on which the property transaction was recorded.
- Sale date: This is the actual sale date of the property.

The 2007 tax-roll tape contains detailed information about all properties in LA county. We use the tape to identify single family residences, which are stand-alone properties that are inhabited by one family only, leaving 2,187,254 transactions for 971,251 distinct homes. While the tax-roll provides information on many home characteristics, we currently limit attention to land square footage, living square feet, and effective year built: this is the last year in which the property had the same physical characteristics as it has in 2007. The list of data items chosen is deliberately short since the main aim of the paper is to show that explicitly accounting for geography, and geography alone, can lead to better understanding of house prices. In addition, the quality of tax-roll tapes varies from county to county, making findings that derive from detailed house characteristics hard to generalize.

The tax-roll tape also provides mailing addresses for all properties. These mailing addresses are used to geocode the home addresses. Roughly 14% of the data has problems with the mailing address. The remaining 86% can be used for any spatio-temporal study that requires geocoordinates. In addition to the mailing address, the tax-roll tape provides the information on the census tract and the associated zipcode. The Chief Executive Office - Public Affairs of LA county provides a break-up of LA county by neighborhoods mapping zipcodes to neighborhoods. This division was used to break LA county into 171 different neighborhoods. Several neighborhoods have only one zipcode associated with them while others some have a couple dozen associated with them. All incorporated cities in LA county, such as Pasadena, Santa Monica, and Beverly Hills, define their own neighborhoods.

3.2 Data Cleaning

The data cleaning, which is designed to replicate the CS methodology as implemented by Standard and Poor's (S&P) for their commercial index, comprises three stages.

Stage 1: The first step of cleaning operates at the home level.

- Transactions for the same home that happen on same exact date but with different prices are all removed.
- Apparently separate transactions for the same home that happen on the same exact date at the same exact price are unified.

- Repeat transactions for homes that happen too quickly (currently within less than 7 months of each other) are both removed. This is identical to the S&P cleaning methodology. Quick transactions very likely involve structural change and hence cannot be used in a repeat sales methodology. Further, transactions on a home before and after this *dirty repeat transaction* are de-coupled due to potential structural change. In essence, it is now two different homes, ones before the *dirty repeat transaction* and one after. Note that because of the de-coupling, a single home is potentially broken down into any combination of single and repeat transactions.

- After the above cleaning, a nx4 matrix S is constructed which is a pure transaction index matrix. In it's first column, this matrix has APNs which are not unique. In it's second column it has the number of transactions on that APN, which can be 1 or 2. In the third and fourth column it has the row indices of these transactions (fourth column is empty for single transactions) in the original cleaned raw tape. This implies that every sale price is only explicitly tied to its last sale price, if available. This is in line with the two models we will be estimating.

Stage 2: In this stage, cleaning is done on repeat transactions only

- From the above matrix S, the subset of transactions that are repeat transactions are extracted.

- Repeat transactions in which the *effective year built* as recorded in the second transaction lies after the first transaction are removed.

- Sale pairs where either the first or second price is less than $5000 and greater than $100,000,000 are removed, since these transactions have a high likelihood of being mis-recorded and are hence removed.

- Sale pairs that have annualized returns less than -50% or greater than 100% are removed.

- Finally, sale pairs that out(under)-perform the median house price index by more(less) than 25% on an annualized basis are removed.[2] The last two steps are implemented because S&P implement some special down-weighting for transactions that display returns *abnormally* above market returns. Since neither the definition of what constitutes an abnormal

[2]The standard deviation of housing returns less median house price index returns on an annualized basis is 12%.

return nor the down-weighting procedure is publicly available, we experiment with reasonable definitions of abnormal returns and then remove transactions that display abnormal returns.[3]

Stage 3: Re-visiting single sales from the matrix S.

- Remove all single sales that have price greater than the maximum transaction price in the repeat sales tape and all sales that have price less than the minimum transaction price in the repeat sales tape.

After the above cleaning, we are left with 591,239 repeat sale transactions and 367,973 single transactions making a total of 1,550,451 transactions. As shown in table 1, transaction volumes were the highest in the late 1980's (a period of high housing returns in LA), staying low through the early and mid 1990's (a period of low to negative housing returns in LA), then starting to climb again through 2005 which corresponds to the housing boom and then dropping sharply in 2006 and 2007. For the repeat transactions, the median turnover time is 60 months or 5 years and maximum number of transactions take place within 2 to 3 years.

4 Repeat Sales Indices as Predictors of Price

We implement the CS repeat sales index methodology following the specifications of Standard and Poor's (S&P) for their commercial index. Unlike the CS index discussed above which takes a geometric average (uses log prices instead of prices), the S&P index uses actual prices and computes an arithmetic average.

In the S&P methodology, the entire time period is divided into a pre-base-period and a post-base-period. The base period is January 2000. The index is normalized to 100 for the base period and all other index numbers are computed relative to the base period index. In particular, the index is computed in two steps. First, a pre-base-period index is computed by using only the pre-base-period data. In the post-base-period sample, the index is computed in real time using data up to that month only. For example, the index for March 2004 is not updated with the arrival of new

[3]Other more and less conservative definitions were also used for abnormal returns. However, the above definitions return an index closest to the S&P Case Shiller index.

Year	Number of transactions
1984	62,151
1985	70,838
1986	86,789
1987	82,139
1988	86,896
1989	74,761
1990	55,843
1991	51,335
1992	46,751
1993	48,510
1994	55,540
1995	49,987
1996	54,200
1997	57,890
1998	66,624
1999	69,620
2000	66,491
2001	67,144
2002	75,018
2003	75,220
2004	73,349
2005	71,202
2006	57,612
2007	36,735
Jan-Apr 2008	7806
Total	1,550,451

Table 1: Final number of clean transactions by year

information in Apr-2004. Fortunately, revisions with the arrival of new information are very small (on average a little under 1% annually) and display no systematic patterns.

Finally, the S&P index is a moving average index in that every transaction is assumed to take place thrice, in the month in which it was recorded and the two months following that month. There are two reasons cited for the moving average estimation. One, there are potential delays in the flow of sale data from county deed recorders, and two, “to keep sample size large enough to create

meaningful price change averages".

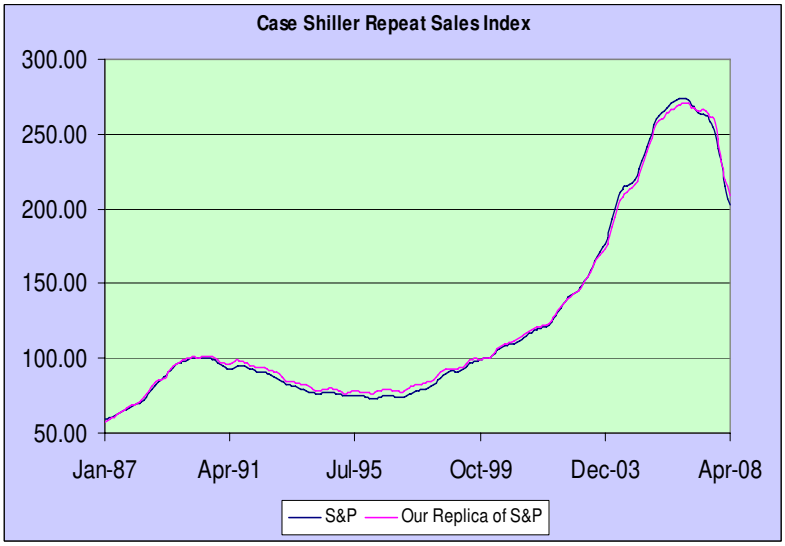

Figure 1: Case Shiller Repeat Sales House Price Index

As a first step, in replication, we estimate the CS repeat sales index using the above version of the S&P methodology for the time period Jan 1987 to Apr 2008 and compare it to the actual published S&P CS repeat sales index.

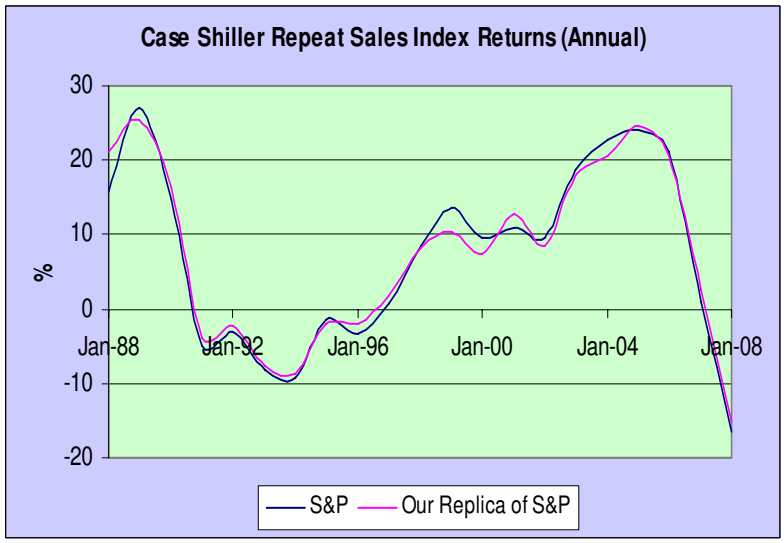

Figure 2: Case Shiller Repeat Sales House Price Index Returns

Figure 1 shows that the indices are very close to each other. In figure 2 we compare annual

returns on the two indices.[4] There are differences in annual returns that the two indices accrue, but they are small and are concentrated mainly in the late eighties and the late nineties.

4.1 Prediction Errors

One would ask of a price index that it not make systematic errors in predicting contemporaneous prices. One would also be interested in understanding any such patterns that did exist and using them as additional data of value in understanding actual contemporaneous house values. To investigate this we develop a methodology for using the index for purposes of one step ahead prediction. The unit of time is months. Prediction is done one period ahead by dynamically dividing the data into training and test sets. The test and training sets are constructed as follows.

- There are 99 monthly test sets running from February 2000 to April 2008.
- A test set of a particular month comprises all homes that transact the second time in that month with their first transaction happening in an earlier period.
- The corresponding training set includes all repeat transactions that took place before the test month.

For each test period t the index values $I_1....I_{t-1}$ are estimated using data on all repeat sales data (training data) till $t-1$. The index is then predicted one period forward by assuming that nominal monthly returns $R_t = I_t/I_{t-1} - 1$ follow a simple AR(1) process as follows, [5]

$$R_t = \rho R_{t-1} + \theta_t,$$

where θ_t is assumed to be i.i.d $N(0, \sigma_\theta^2)$. Case and Shiller [1989] show that the repeat sales index returns exhibit significant positive autocorrelation at the quarterly level. The S&P index returns display significant positive autocorrelation at the monthly level (though a fair amount of this is artificially built in because of the moving average assumption). The value of ρ averaged over the 99 prediction runs is 0.85. It is important to note that the AR(1) model is a simplifying assumption to build in a prediction based methodology for ease of comparison with results from our model in

[4] Monthly and quarterly returns are very volatile.

[5] Using an AR(2) or AR(3) process instead has no effect on our results

the next chapter, and has no effect on the results. None of the results below change by using our exact replica of the S&P index or the full data index.

Having estimated $\hat{\rho}$, the index value in period t is given by $\hat{I}_t = \hat{\rho} R_{t-1} I_{t-1}$. The full series of index values $I_1, ...I_{t-1}, \hat{I}_t$ is then used to predict prices of homes in period t. Prediction error e_{it} for a pair of transactions (i, t, P_{it}) and $(i, t', P_{it'})$ such that $1 \leq t' \leq (t-1)$ is defined as follows:

$$e_{it} = 100(1 - \frac{P_{it'} \frac{\hat{I}_t}{I_{t'}}}{P_{it}}).$$

We measure accuracy of prediction by several different criteria. The first two are the median actual and absolute prediction error, and the others are the percentage of homes for which $|e_{it}| \leq 5\%$, and the percentage of homes for which $|e_{it}| \leq 15\%$.

Time period	**Median err (%)**		**% Homes**		**Homes**
	Absolute	**Actual**	$\|e_{it}\| \leq 5$**%**	$\|e_{it}\| \leq 15$**%**	
Feb:Dec 2000	11.88	-3.96	23.49	59.02	31851
2001	11.39	-3.54	24.54	61.03	35893
2002	10.81	-0.67	25.65	62.99	42405
2003	10.42	1.10	26.60	64.72	43902
2004	10.68	3.13	25.20	63.96	43696
2005	10.63	3.20	25.61	64.40	44618
2006	11.69	1.36	22.92	60.33	37183
2007	13.86	-4.47	19.23	53.35	23913
Jan:Apr 2008	14.74	-4.97	18.15	50.71	5314

Table 2: Error patterns with time for Los Angeles county: 2000-08

We find that for LA county as a whole, the median absolute prediction error is 11.26%, with 24.41% of the homes having $|e_{it}| \leq 5\%$ and 61.67% of the homes having $|e_{it}| \leq 15\%$. The median error for LA county is 0.027%.

Table 2 shows the distribution of errors over time. The prediction accuracy of the CS index stays roughly constant in the period 2000-06, thought it is somewhat lower in the early 2000s. In 2007-08, the prediction accuracy of the CS index takes a significant fall.

The most significant finding is the dramatic change in the median actual error in the 8 year period. The median error goes from -3.96% in 2000 to 3.2% in 2005, a change of over 7%. Subsequently it drops by over 8% from 2005 to 2008.

After accounting for heteroscedasticity, the errors on the CS model should be i.i.d normal. However, we find a lot of structure in the error term. In addition to the patterns over time (as shown in table 2), we find patterns in the error term over initial prices, time between transactions and geography.

4.2 Time Between Sales

Time btwn sales	Median err (%)		% Homes		Homes
	Absolute	Actual	$\lvert e_{it}\rvert \leq 5$%	$\lvert e_{it}\rvert \leq 15$%	
$<$1 year	5.86	2.83	43.52	93.4	12836
1-2 years	7.88	1.58	33.5	77.5	33650
2-3 years	9.22	2.15	29.24	70.89	40283
3-5 years	11.06	2.49	24.12	63.16	56326
5-7 years	13.02	0.69	20.83	55.93	37807
7-10 years	12.78	-2.27	21.66	56.51	39410
10-12 years	12.37	-4.33	22.19	57.62	22743
12-15 years	14.06	-5.08	19.8	52.49	31823
15-20 years	17.12	-4.73	16.31	45.06	29939
$>$ 20 years	21.56	-5.23	12.83	37.34	3958

Table 3: Error patterns with time between sales

Table 3 shows a very strong pattern of errors with time between sales. The repeat sales index seems to do an excellent job predicting prices on quick repeat trades. However, the performance of the repeat sales index falls monotonically with increasing time between sales. For homes that transact a second time after 15-20 years, the median absolute error is 17.12% and only 45% of the test samples are predicted with less than 15% error.

Even more interesting is the result that the median error gets increasingly negative as time

between sales increases, reaching roughly -5% for repeat transactions after 12 or more years, where a negative error implies prices are over-predicted. This implies that turnover time has a powerful correlation with housing returns and that homes with longer turnover times have systematically lower returns. Among the possible reasons for this are: that there are non-temporal returns (GS) because of once-off improvements at the time of sale; that longer turnover times are correlated with poor maintenance; that these homes are fundamentally different and should be included in separate indices.

4.3 Initial Price: Mean Reversion

Table 4 shows that the CS index does a very poor job at predicting prices of homes with either very low or very high initial prices. The error is below 15% for only 2% of the homes with initial price less than $25,000. Though one might view very low priced homes like these with skepticism in the light of possible measurement errors, the index does only marginally better with homes in the initial price range $25,000-50,000. The index performance improves as initial price increases. It peaks for homes in the initial price range of $200,000-400,000, predicting roughly 70% of the homes with less than 15% error and a median absolute error of about 9%. As initial price increases beyond $400,000, prediction accuracy again starts to fall. The error is below 15% for only a little under half of the homes with initial prices between $1 and 2 million, and the median absolute error is 15-16%.

The most striking finding in table 4 is that the index systematically under-predicts prices of homes with very low initial prices and over-predicts prices of homes with very high initial prices. Homes with initial prices in the range $25,000-50,000 have a median error of 55%, which continues to stay high (17%) if one looks at homes with initial price in the range $50,000-75,000. Similarly, homes with initial prices greater than $500,000 are systematically over-predicted with median error rising from about 8% from homes with initial prices in the range $500,000-600,000 to almost 11% for homes with initial prices in the range $1 to 2 million.

One might ask if the above mean reversion results are in fact just an artifact of the time between sales effect with homes in the mid-price range transacting more frequently and homes at the end of the price range spectrum transacting less frequently. Though that would explain the magnitude of the error, it would not explain its sign. Further, we repeat the same exercise with homes that

Initial price ($)	Median err (%)		% Homes		Homes
	Absolute	Actual	$\|e_{it}\| \leq 5$%	$\|e_{it}\| \leq 15$%	
$\leq$25,000	81.43	81.1	0.51	2.24	1958
25-50,000	57.08	55.61	2.41	7.86	3269
50-75,000	25.87	17.13	9.82	29.97	8079
75-100,000	17.85	5.48	15.04	43.3	17817
100-125,000	14.84	2.3	17.68	50.49	26641
125-150,000	12.8	1.54	20.78	57.4	38289
150-200,000	10.18	2.56	25.84	67.25	65096
200-250,000	8.89	1.7	30.07	71.75	40784
250-300,000	8.65	0.26	31.42	72.26	29237
300-350,000	8.87	-1.77	30.94	70.7	21363
350-400,000	9.1	-3.06	30.17	69.61	17248
400-450,000	9.97	-4.7	27.59	66.18	12726
450-500,000	10.64	-6	25.84	64	9718
500-600,000	11.91	-8.11	22.67	59.72	12741
600-750,000	12.99	-9.18	21.14	55.81	10601
750,000-1 million	13.69	-8.98	20.38	53.89	7515
1-2 million	15.30	-10.90	18.26	49.13	5058
$>$ 2 million	19.79	-14.48	15.92	40.07	791

Table 4: Error patterns with initial price in Jan 1984-Jan 2000

had their first transaction in the period Jan 1984 to Dec 1988 in table 5 and also for homes that had their first transaction in the period Jan 1989 to Dec 1993 in table 6. The prediction accuracy and median errors stay roughly the same for homes with initial prices in the lower range. However, the median errors on homes with very high prices are now even more negative. In other words, mean reversion is much stronger than table 4 suggests. Mean reversion is also a slow process. If errors are measured on homes which transacted for the first time in the 1994-98 or 1999-2003 interval, the median errors get much closer to zero and prediction accuracy improves.

Initial price ($)	Median err (%)		% Homes		Homes
	Absolute	Actual	$\|e_{it}\| \leq 5$%	$\|e_{it}\| \leq 15$%	
$\leq$ 25,000	81.5	81.18	0.28	0.86	1393
25-50,000	57.34	55.77	2.53	7.92	1615
50-75,000	25.16	6.33	11.07	32.94	3466
75-100,000	16.14	-5.3	17.63	47.78	7865
100-125,000	13.76	-7.04	19.03	53.58	7638
125-150,000	13.18	-6.92	21.41	55.21	6618
150-175,000	12.82	-7.91	19.75	56.66	4571
175-200,000	13.94	-8.98	19.45	53.43	3279
200-250,000	16.9	-12.5	14.86	44.66	3674
250-300,000	18.68	-14.79	12.26	41.14	2282
300-400,000	21.93	-17.57	11.44	33.67	2159
400-500,000	25.47	-20.27	10.73	30.94	866
500-750,000	25.92	-22.95	10.33	31.81	745
750,000-1 million	32.95	-27.73	11.06	22.55	235
$>$ 1 million	47.31	-42.62	5.88	18.23	170

Table 5: Error patterns with initial price in Jan 1984-Dec 1988

4.4 Geography

Finally, in this section we see if the CS errors have a geographic structure to them. We consider nine major areas in LA county as shown in table 7 that are geographically well spread out over LA county, and that are economically and demographically diverse. All the areas in the table are incorporated cities except South Central which is a designated neighborhood by the Public Affairs Office of the County of Los Angeles.

Homes were identified to belong to one area or another based on their zipcodes. The mapping from zipcodes to cities and neighborhoods in LA county was also obtained from the Public Affairs Office of LA county. Overlapping zipcodes resulted in certain cities being combined to form bigger areas. Torrance and Redonodo Beach which are adjacent cities in the South Bay region were combined. Agoura, Agoura Hills, Calabassas, Oak Park and Hidden Hills all cities in the western

Initial price ($)	Median err (%)		% Homes		Homes
	Absolute	Actual	$\|e_{it}\| \leq 5$%	$\|e_{it}\| \leq 15$%	
$\leq$ 25,000	84.06	84.01	0.24	2.18	412
25-50,000	73.03	72.17	0.14	1.87	694
50-75,000	46.98	43.68	6.41	15.8	810
75-100,000	22.84	10.49	11.23	33.71	1993
100-125,000	16.57	-0.99	16.57	46.23	4393
125-150,000	12.57	-3.04	21.35	57.88	7403
150-175,000	9.83	-2.47	27.72	67.71	8667
175-200,000	8.69	-2.3	30.11	71.33	7306
200-250,000	9.87	-4.49	27.49	67.78	8580
250-300,000	11.27	-5.56	23.82	62.29	5028
300-400,000	13.12	-6.91	20.91	55.33	5149
400-500,000	16.21	-9.62	17.64	47.11	2222
500-750,000	19.28	-13.65	13.58	40.12	2253
750,000-1 million	23.51	-17.15	11.96	32.4	719
$>$ 1 million	30.7	-28.28	8.64	26.55	659

Table 6: Error patterns with initial price in Jan 1989-Dec 1993

part of LA county were combined. Industry, La Puente, Bassett, Valinda, Hacienda Heights, Diamond Bar and Walnut, which are all cities in the eastern part of LA county were combined. Finally, the city of Santa Monica was combined with Castellmare, Pacific Highlands and Pacific Palisades which are all neighborhoods in LA city. This combination was necessary to expand the pool of transactions for constructing a local index.

Table 7 shows that there is a lot of geographic structure to the CS error. The index does relatively well in Industry and it neighboring cities, Long Beach, Van Nuys, and Torrance and Redondo Beach. It does relatively poorly in South Central and Santa Monica.

Finally, table 8 shows that there are significant prediction gains to be had by simply estimating indices at the local level. We employ the same prediction methodology as the one we use for LA county as a whole. Prediction accuracy is improved, and in some cases by a significant amount,

Neighborhood	Median err (%)		% Homes		Homes
	Absolute	Actual	$\|e_{it}\| \leq 5$%	$\|e_{it}\| \leq 15$%	
Los Angeles Full	11.26	0.027	24.41	61.67	308,775
Palmdale	13.99	-0.84	19.68	52.77	16387
South Central	16.04	5.86	16.79	46.99	5807
Pasadena	12.29	2.67	23.01	58.56	5587
Santa Monica etc.	14.79	-0.87	18.41	50.33	2265
Torrance etc.	10.63	-3.49	26.6	63.77	6931
Van Nuys	11.1	4.07	23.99	63.59	5431
Long Beach	10.45	2.66	25.99	65.96	14228
Agoura Hills etc.	12.3	-2.9	22.55	57.23	1658
Industry etc.	9.71	-3.02	28.1	67.98	11791

Table 7: Spatial error patterns

Neighborhood	% Homes with $\|e_{it}\| \leq 15$%	
	LA county index	Local index
Palmdale	52.77	67.17
South Central	46.99	58.84
Pasadena	58.56	58.15
Santa Monica etc.	50.33	52.23
Torrance etc.	63.77	70.73
Van Nuys	63.59	69.71
Long Beach	65.96	68.17
Agoura etc.	57.23	58.02
Industry etc.	67.98	71.64

Table 8: Comparing prediction performance

by just estimating indices at the local level.

5 Location and House Prices: Prior Literature

The results of the last section suggest that accounting for location may be important to understanding the patterns that we have uncovered in house prices. This idea not only motivates our later work, but also rationalizes a large body of prior literature.

5.1 Hedonic Models and Location

There is now a large hedonic price literature on the spatio-temporal correlation of housing returns. Spatio-temporal correlation arises for a number of reasons. Missing variables, especially neighborhood variables like crime, schooling quality etc. will lead to spatial correlation of the error term. In pricing homes, Realtors look at prices of nearby homes that transacted in the recent past. This implies that house prices can be predicted by lags of prices of similar houses. Hence it is not surprising that many recent papers find evidence for spatial autocorrelation of the error term in the traditional pricing function. Can [1990] estimates the following spatial autocorrelation model of prices on a data set of 563 single family homes that transacted in 1980 in the Franklin county section of the Columbus metropolitan area (see also Can [1992]),

$$P = \alpha + \rho WP + \beta X + \epsilon,$$

where P is the vector of house prices, W is a generalized weighting matrix, ρ is the autocorrelation coefficient, X is the vector of housing and neighborhood attributes, and β is the marginal prices of attributes. Examples of weighting adjacent house prices she considers are: for the price equation of house j, $W_{ij} = 1/d_{ij}$ where d_{ij} is the distance between house i and house j, or $W_{ij} = 1/d_{ij}^2$. She also consider a discrete version where $W_{ij} = 1$ if $d_{ij} < 5$ miles and $W_{ij} = 0$ otherwise.

Since P is both the dependent variable and a regressor, the error term cannot be assumed to be independent of the regressors and OLS cannot be used. Can estimates this model using maximum likelihood which, given the spatial autocorrelation, involves the use of non-linear optimization techniques. She finds that the autocorrelation coefficient is positive and significant. She further finds that the error term in the traditional hedonic model (without the WP term) displays significant spatial correlation which disappears after the introduction of the WP term.The model seems to perform best when the weighting matrix used in inversely proportional to the square of

the distance.

Rather than incorporating the correlations directly into the pricing equation, Dubin [1992] incorporates spatial features into the model by allowing for spatial autocorrelation of the error term. In Dubin [1998], the assumption is that all individual errors in the standard hedonic pricing equation, $P = \alpha + \beta X + \epsilon$, are normal, $\epsilon_i \sim N(0, \sigma^2)$, but that these errors are correlated across locations,

$$Corr(\epsilon_i, \epsilon_j) = b_1 exp(-\frac{D_{ij}}{b_2}) \forall i \neq j,$$

where D_{ij} is the geographic distance between i and j, and b_1, b_2 are parameters to be estimated. Note that the variance covariance matrix V is given by $\sigma^2 \phi(b_1, b_2)$ where $\phi_{ii} = 1 \ \forall i$ and $\phi_{ij} = b_1 exp(-\frac{D_{ij}}{b_2})$ for $i \neq j$. The estimation is performed using Maximum Likelihood and first the ML estimates are obtained for β and σ^2 as a function of b_1, b_2. These estimates are then plugged back into the ML function, which is then maximized with respect to b_1, b_2.

Both the Dubin papers use a data set of 1493 transactions in 1978 in Baltimore. How severely the current literature is limited by the availability of the right tools is best illustrated by the fact that Dubin [1992] and Dubin [1998] are forced to break this small data set into smaller ones because of computational issues. A separate model is estimated for each test point, using the nearest 100-150 training points. Finally, since there is now a structure on the error term, one can estimate the error term on the test point by a process called kriging, and the estimate for the error term for the i^{th} test point e_i given by,

$$e_i = k'(i) \phi^{-1} \hat{\epsilon},$$

where $k'(i)$ is a vector whose jth element is given by $b_1 exp(-\frac{D_{ij}}{b_2})$, ϕ is the correlation matrix for the nearest neighbors of the test point i, and $\hat{\epsilon}$ is the corresponding estimated errors. Dubin finds that both b_1 and b_2 are significant. The magnitude of the correlation decreases from from 0.64 at 100 feet to 0.53 at 1000 feet to 0.23 at 1 mile and 0.08 at 2 miles. He also reports a 38% reduction in the sum of squared errors in the test sample as compared to the standard OLS model.

Basu and Thibodeau [1998] model the spatial autocorrelation of the error term using a spherical semivariogram function for a sample of 5000 homes that transacted between 1991Q4 and 1993Q1 in metropolitan Dallas. However, they do not do this for the entire market, but in fact for an exogenously defined set of eight housing sub-markets in Dallas, Texas. The housing submarkets are geographic in nature and are separated by major US highways. They find that in

four of the eight housing sub-markets, there is significant spatial autocorrelation of the error term for properties within 1200 meters of each other. For two of the housing sub-markets, the spatial autocorrelation is present across the whole sub-market, while there is no evidence for spatial autocorrelation in two of the sub-markets.

Pace and Gilley [1997] use the simultaneous autoregressive model (SAR) from the spatial statistics literature to model the spatial autocorrelation of the hedonic residuals. Their data covers roughly 72,000 transactions from Fairfax county Virginia and spans the period from 1966 to 1991. Note that this seemingly large data set is still roughly only 5% of the size of our data set. They estimate the model by maximizing the likelihood function and find that the model's sum of squared errors is 44% less than the corresponding OLS model which includes location dummies. Pace et al. [1998] extend the analysis to make W a function of both space and time and also find a significant increase in prediction accuracy of the model.

The above models clearly suggest that location matters, and that one can improve understanding of local house prices by incorporating information that is richer than the city-wide index alone. However they are both conceptually and computationally limited. In conceptual terms, marginal prices β are assumed to be constant across space and time. This assumption is justifiable as long as one works with a small and homogeneous geographic area and a small time span. Moreover, they involve very limited data sets, and as such their conclusions do not have the generality required to impact day to day analysis of patterns in house prices. Our LA data covers a massive set of transactions that are spread out over 4000 square miles of a very heterogeneous housing market and span a period of 25 years. It is for this reason that we provide a more thoroughgoing re-evaluation of index methodology both conceptually and computationally.

5.2 Repeat sales models and location

Goetzmann and Spiegel [1997] incorporate locational returns in their distance weighted repeated sales model. Houses in the same neighborhood are very likely to exhibit co-movement of returns and similarly houses in different neighborhoods are likely to exhibit very little correlation in returns. Given the location dependence in this model, houses will be indexed both by the house number i and the location number l in the original Case-Shiller model. Defining $r_t(l)$ as the error term arising due to the location, GS assume that $cov(r_t(l), r_t(m)) = \sigma_r^2 \exp(-d_t(l,m))$ where $d_t(l,m)$ is

some measure of the distance (not necessarily geographic) between the two locations l and m.

$$p_{ilt'} - p_{ilt} = \eta + c_{t'} - c_t + \Delta h_{itt'} + \Delta n_{itt'} + \Delta r_{tt'}(l)$$

In principle, all the variance parameters can be identified given their functional dependence on measurable factors, time in case of σ_h^2 and the distance measure $d_t(l, m)$ in case of $cov(r_t(l), r_t(m))$. Once this has been done, GLS can be used to estimate the price indices as before. GS estimate their model on roughly 130,000 repeat sales in San Francisco over the period 1980-94. Despite the relatively small size of their tape, GS run into computational problems because of the size of the variance-covariance matrix and therefore only report OLS standard errors.

5.3 Semi-parametric and Nonparametric Techniques

Semi/non parametric models in the housing literature all fall in the class of nearest neighbor model. Pace [1993] and Anglin and Gencay [1996] use kernel based estimation while Meese and Wallace [1991] and Clapp [2004] use locally weighted regressions. We discuss the models with locally weighted regressions as we employ this nearest neighbor technique in our model as well.

A local regression model is a non-parametric model, and as the name suggests is a regression model fitted locally. For each house trade, one finds the "nearest" trades, and fits a polynomial regression through those points by minimizing a weighted sum of squared errors, and uses the fitted plane to predict the dependent variable at the test point. In conceptual terms, major changes are that one allows for non-linearities in the hedonic equation, and uses a kernel function to identify the extent of the neighborhood of each given transaction and the weights to be given to each of the neighbors.

Clapp [2004] estimates a semi-parametric model to identify variations in land value across Boston MA. He finds that the land value surface $f(W)$ seems to capture well land price variations in Boston (MA) based on a casual understanding of the real estate market. He also finds that the out of sample mean square error is reduced by about 11% when compared to OLS. However, Case et al. [2004] find that the semi-parametric model does worse than the model with spatially autocorrelated errors (Dubin [1998]). Clapp rules out any relationship between attribute prices and location by suggesting an equilibrium allocation where arbitrage would wipe out any such relation-

ships (Rosen [1974]). Though the argument is theoretically sound under fairly elastic supply of housing attributes and negligible moving and transaction costs, this is rarely the case. Further, it is not clear why location alone determines land value. Separating neighborhood value from land value suggests that the price of public facilities and amenities are independent of location. Clapp is also severely constrained by the size of his dataset - a little over 5000 transactions spread out over 10 years - in estimating a fully non-parametric price surface.

Meese and Wallace [1991] estimate a non-parametric hedonic index model. Their data consists of 137,000 transactions from Alameda county and 45,000 transactions from San Francisco county, spanning a nineteen year period from 1970 to 1988. For each quarter in this period, they estimate the attribute prices of the median home that quarter using locally weighted regression. The median home is defined as a home whose attributes take values that are given by the median of the attributes of all other homes that traded that quarter. Meese and Wallace find significant differences in attribute prices both across municipalities and across time, thereby rejecting linear parametric forms with constant attribute prices. However, they simplify the analysis in many key dimensions as a result of data and computational constraints. In particular, due to lack of GPS data, nearest neighbors are determined in attribute space rather than in distance-time space. Further, computational complexities prevent them for estimating the entire price surface and they estimate prices at median attributes only. These are constraints that our procedure allows us to relax.

5.4 Sub-markets/Stability of the Housing Price Function

In terms of the hedonic price equation, sub-markets imply several pricing equations with different prices for the same structural and/or neighborhood attributes. If we were to estimate one common index for all sub-markets, the index would be biased towards sub-markets that have more frequently transacted houses. Hence, identification of sub-markets is crucial. Most papers on sub-markets divide the area under study into neighborhoods based on apriori information like school districts, neighboring census tracts, municipalities etc. Though seemingly crude, these kind of models have been extremely popular. Schnare and Struyk [1976] segregate the Boston housing market by inner city/outer city, income and number of rooms. Goodman [1981] divides the New Haven-Connecticut region into sub-markets based on contiguous local government areas. Meese and Wallace divide their sample into municipalities. Michaels and Smith [1990] form sub-markets

based on evaluations by real estate agents. They claim that according to the National Association of Realtors, roughly 80% of home buyers buy homes through Realtors and hence the realtor's categorization of markets into sub-markets should contain all the relevant information.

Case et al. [2004] form neighborhoods by combining census tracts that have similar attribute prices using a clustering method. As mentioned earlier, the neighborhoods they produce in this manner are almost all geographically contiguous. Depending on the study, authors apply different local models to the neighborhoods they construct varying from the basic OLS (Goodman [1981]), local regression models, (Meese and Wallace [1991]) and models with spatially correlated errors (Thibodeau [2003]) to mention a few. All of them find significant variation in attribute prices across sub-markets/neighborhoods. McLennan and Tu [1996] and Bourasa et al. [1999] also use cluster analysis to uncover sub-markets. Bourasa et al. [1999] do K means clustering on factor scores to form sub-markets. This study is conducted for housing markets in Sydney and Melbourne. In both cases, they find that three factors explain about 85% of the variance in the data. The first factor has high scores for distance to central business district, average number of bedrooms, percentage of people driving car to work, average number of cars, owner occupation rate, distance to coast, persons per unit area, and number of dwellings per unit area. Bourassa et al. interpret this as a locational factor that distinguishes between the inner and outer city. The second and third factors are socioeconomic and have high scores for unemployment, income etc. They find evidence for five sub-markets in both Sydney and Melbourne. Given the nature of these factors, it is hard to intuitively interpret the sub-markets.

Thibodeau [2003] forms neighborhoods in Dallas county using two approaches. The first and simpler approach labels every municipality as a neighborhood and estimates a separate linear model for each municipality.[6] In the second approach, Thibodeau combines adjacent census blocks to get approximately 150 homes in each neighborhood. This results in a total of 256 neighborhoods. The only condition imposed is that the combined census blocks belong to the same municipality and the same school district. He evaluates the predictive power of a model by first fitting the model on the training sample and then evaluating its predictive power on the test sample.[7] Predictive power is measured by considering the fraction of homes in the prediction sample whose prices were predicted with less than 10% error and those with less than 20% error. The simple linear model predicts the prices of 36% of the homes with less than 10% error, and 62% of the

[6] The data has 41,985 properties sold in the 1998-99 period, spread over 21 municipalities and 15 school districts

[7] This is necessary to avoid over-fitting.

homes with less than 20% error. The municipality neighborhood model significantly out-performs the linear model, predicting 51% of the homes with less than 10% error and 76% of the homes with less than 20% error. Further, the census block neighborhood model significantly out-performs the municipality neighborhood model, predicting 62% of the homes with less than 10% error and 84% of the homes with less than 20% error.

6 Our Model

Rather than break the housing market into neighborhoods and estimate a separate model on each one of them, we use a local regression model to find nearest neighbors for each home in the test sample. We then use suitable spatio-temporal kernel functions to smooth the unobservable price surface, which is estimated by minimizing the prediction error in a least squares sense. A detailed description of the model follows.

There are a total of $i = 1....M$ transactions spanning $t = 1...T$ months. c_t is the natural log of LA level house price index at time t. Each transaction i is identified by a set $D_i = \{t_i, l_i, p_i, z_i, a_i, b_i\}$:

- t_i: month of transaction
- p_i: natural log of price of transaction
- z_i: natural log of the deviation from c_{t_i} (intrinsic price)
- l_i: GPS co-ordinates of home underlying transaction i
- a_i: natural log of the living area of the home underlying transaction i
- b_i: indicator variable that equals 0 if there is no past transaction and k if there is a past transaction k.

The optimization problem is:

$$Min_{\{c_t\}_{t=1}^{T},\{z_i^s\}_{i=1}^{M}}(MSSE + VC + SC), \tag{2}$$

where

- $MSSE$ is the model[8] sum of squared errors and equals

$$\sum_{i=1}^{M}[p_i - (c_{t_i} + z_i^s)]^2. \tag{3}$$

  z_i^s is a smoothed deviation from the index for transaction i.

- VC is the variability constraint and equals

$$\delta \sum_{i=1}^{M}(z_i^s)^2, \tag{4}$$

  where δ is a parameter that will be exogenously fixed using a procedure called validation, which is described in section 3.5. The VC term drives $\overline{z}$ to 0 and minimizes variability of z_i^s around the mean. This is done to ensure that any common component of house price appreciation is absorbed by the index and not the smoothed intrinsic price surface.

- SC is the smoothing constraint and is given by

$$\alpha \sum_{i=1}^{M}(z_i^s - \sum_{j \in N_i} w_j^i z_j^s)^2 \qquad if \quad b_i = 0 \tag{5}$$

$$\alpha \sum_{i=1}^{M}(z_i^s - [(1 - \lambda(t_i, t_k)) \sum_{j \in N_i} w_j^i z_j^s + \lambda(t_i, t_k) z_k^s])^2 \qquad if \quad b_i = k, \tag{6}$$

  where α is a parameter that will be fixed exogenously using validation. N_i is the set of spatio-temporal neighbors of transaction i. w_j^i are coefficients independent of both the index and the intrinsic price surface and will be determined by a smoothing process that we will shortly explain. $\lambda(t_i, t_k)$ is a weighting function which is decreasing in $t_i - t_k$. When there is no past sale ($b_i = 0$), the smoothing constraint says that the intrinsic price of any transaction must equal a weighted sum of the intrinsic price of its spatio-temporal neighbors. When there is a past sale k the weights are distributed between the past intrinsic price z_k^s and the intrinsic price of the spatio-temporal neighbors depending on the time between sales.

We will now explain how the smoothing constraint is derived. The nearness of a transaction j

[8]The pricing equation is given by $p_i = c_{t_i} + z_i + \epsilon_i$, where ϵ_i is i.i.d $\sim N(0, \sigma_\epsilon^2)$.

to a transaction i is determined by a spatio-temporal distance function $K(i,j)$ which is given by,

$$K(i,j) = Max(0, 1 - \frac{d(l_i, l_j)}{\overline{d}} - \frac{r(i,j)}{\overline{r}}), \tag{7}$$

where

- $d(l_i, l_j)$ is the geographic distance in miles between i and j,
- $r(i,j) = |t_i - t_j|$ is the time distance between i and j determined.
- $\overline{d}$ and $\overline{r}$ are parameters fixed exogenously by validation.

In other words, we assume that the iso-distance (spatio-temporal) contours are straight lines. This is in the spirit of the paper keeping all assumptions at a fairly simplified level to show first order effects of accounting for geography.

A transaction $j \in N_i$ where N_i is the set of nearest neighbors of i if $K(i,j) > 0$. If there are more than m^* neighbors that satisfy this criterion, only the closest m^* neighbors are designated as nearest neighbors. m^* is a parameter fixed by validation.

The intrinsic price z_i^s, or deviations from the index for any transaction i, will be a hedonic function of home as well as neighborhood attributes. For simplicity, and to show first order effects of geography we assume that,

$$z_i^s = q_i + h_i a_i \quad if \quad b_i = 0 \tag{8}$$

where,

- q_i: log of intrinsic price per unit area
- h_i: accounts for possible non-linearities in the effect of increase in area on intrinsic price

(q_i, h_i) is the solution to the following weighted local regression model.

$$(q_i, h_i) = \arg\min_{(q,h)} \sum_{j \in N_i} (z_j^s - q - h a_j)^2 v_j^i, \tag{9}$$

where $v_j^i = \frac{K(i,j)}{\sum_{j \in N_i} K(i,j)}$ is the normalized spatio-temporal distance function.

Note that z_j^s are unknown. Hence q_i and h_i are obtained as functions of $\{z_j^s\}_{j \in N_i}$. We express the solutions in terms of the following variables:

- $\overline{a} = \sum_{j \in N_i} v_j^i a_j$
- $\overline{a^2} = \sum_{j \in N_i} v_j^i a_j^2$
- $\sigma_a^2 = \overline{a^2} - \overline{a}^2$

The solution to the above weighted local regression model is given by

$$q_i = \sum_{j \in N_i} v_j^i \frac{\overline{a^2} - a_j \overline{a}}{\sigma_a^2} z_j^s \tag{10}$$

$$h_i = \sum_{j \in N_i} v_j^i \frac{a_j - \overline{a}}{\sigma_a^2} z_j^s \tag{11}$$

Therefore,

$$z_i^s = q_i + h_i a_i = \sum_{j \in N_i} w_j^i z_j^s \tag{12}$$

where,

$$w_j^i = v_j^i \frac{\overline{a^2} + a_j(a_i - \overline{a}) - \overline{a} a_i}{\sigma_a^2} \tag{13}$$

When, for transaction i a past sale $b_i = k$ exists, we assume that

$$z_i^s = (1 - \lambda(t_i, t_k)) \sum_{j \in N_i} w_j^i z_j^s + \lambda(t_i, t_k) z_k^s \tag{14}$$

where,

$$\lambda(t_i, t_k) = Max(0, 1 - \frac{t_i - t_k}{\bar{t}}), \tag{15}$$

and $\bar{t}$ is a parameter fixed by validation.

Equations (12) and (14) are the smoothing constraints. They imply that controlling for area, transactions that are spatio-temporally close cannot have intrinsic prices that are too different. In case house i has a past transaction i.e. $b_i = k \neq 0$, then both the neighboring transactions and the past transaction k provide information about z_i^s. The closer the past transaction in time, the more information content it has relative to other spatio-temporally close transactions. In the limiting case, with the past transaction happening instantaneously before, all useful information is in the past transaction alone.

Note that the smoothing constraints can be written in matrix form as $z^s = Hz^s$ where $z^s = [z_1^s ... z_M^s]'$ and H is a $M \times M$ matrix of coefficients which can be computed from the weights w_j^i and $\lambda(t_i, t_k)$. Substituting $z^s = Hz^s$ in (2), we can re-write the optimization problem in matrix form as

$$Min_{c,z^s}[p - Rc - Hz^s]'[p - Rc - Hz^s] + \delta(z^s)'z^s + \alpha(z^s - Hz^s)'(z^s - Hz^s)$$

where $p = [p_1 ... p_M]'$, $c = [c_1 ... c_T]'$, R is a $M \times T$ matrix with elements $R_{it} = 1$ if $t = t_i$ and 0 otherwise. It is this quadratic optimization problem that we solve computationally using the conjugate gradient algorithm.

7 Prediction Methodology and Validation

The model is tested on eighteen different test periods which were constructed on the basis of month of transaction. These were Mar:Aug 2000, Jan:Jun 2004 and Jul:Dec 2007[9]. The samples are deliberately chosen to represent three very different periods in LA house price history. The first set of test periods Mar:Aug 2000 saw average growth in house prices, the second set of test periods Jan:Jun 2004 were right in the middle of the housing boom and the last set of test periods Jul:Dec 2007 saw a significant decline in house prices.

The training sample corresponding to each test period on which the model was estimated included all transaction data from Jan 1984 to one month before the test period. For example, for the test sample containing transactions in Mar 2004, the training data included all transactions in the period Jan 1984 to Feb 2004.

[9]It was not possible to test on the entire period 2000-08 as in chapter 2 because the hard disk space required to store all the relevant neighborhood matrices exceeds 250 GBs for just these 18 months

For each training period $1...t$, the vector of parameters c and z^s are solved for using the conjugate gradient method. c_{t+1} or the index for the test period was forecasted exactly like it was in chapter 2 by assuming that nominal returns follow an AR(1) process.

The value of z_i^s in the test set was determined using the smoothing constraints (12) and (14) where the nearest neighbor set N_i is constrained to be a subset of the training set. The final prediction error e_i for the i^{th} test sample is given by

$$e_i = 100(1 - exp(c_{t_i} + z_i^s - p_i))$$

Our model uses two variables that the CS repeat sales model does not - living area and GPS co-ordinates which are obtained from the home address. Lack of clean addresses in some cases implies that some of the transaction data that was used in chapter 2 can no longer be used. Similarly, lack of data on living area resulted in some loss of data. Since we now have a different data set, the CS prediction exercise in chapter 2 is repeated for the smaller data set.

All model parameters $\overline{d}$, $\overline{r}$, $\overline{t}$, m^*, α, and δ are determined using a process called validation. The model is estimated on the data from Jan 1984 to Jan 2000 and tested on Feb 2000. Model parameters are chosen to maximize the number of transactions in Feb 2000 for which $e_i \leq 15\%$.

The validated parameter values are as follows:

- $\overline{d} = 0.5$ (miles)
- $\overline{r} = 36$ (months)
- $\overline{t} = 120$ (months)
- $m^* = 150$
- $\delta = 0.01$
- $\alpha = 1$

Our results are not very sensitive to small variations of any of the parameters around the validated values.

7.1 Results

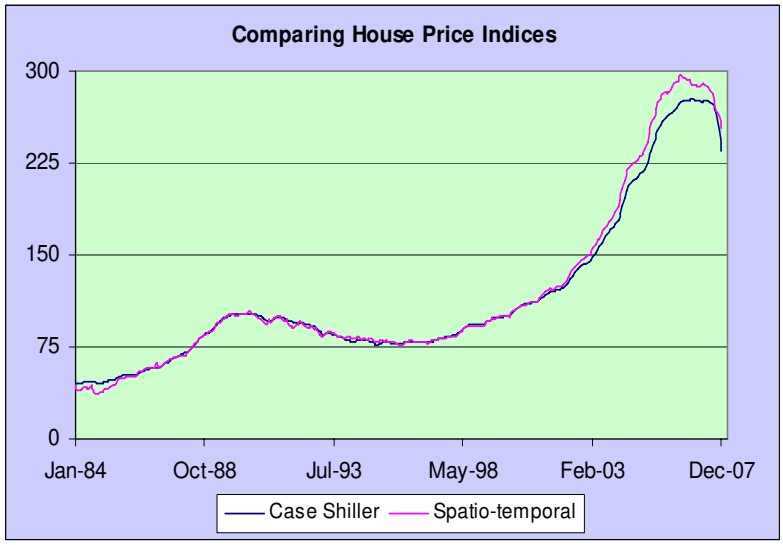

Figure 3: Comparing House Price Indices

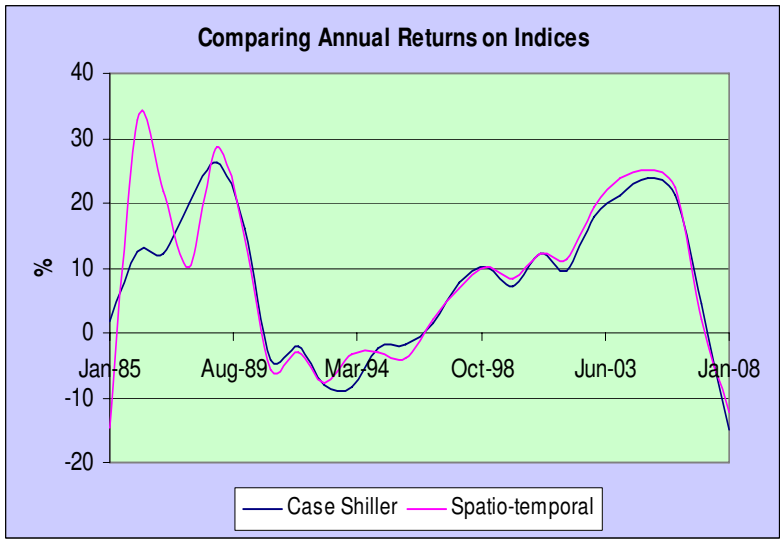

Figure 4: Comparing Annual House Price Index Returns

Figures 3 and 4 compare the CS index with our spatio-temporal index c_t. The index has been normalized so that the index value in Jan 2000 is 100 as in the CS index. Barring the first five years 1984-89, there is very little difference between the two indices. Recall however that the prediction error using the CS index in chapter two displayed significant patterns with turnover

time, initial price and geography. Our hypothesis is that both turnover time and price themselves have geographical patterns and hence controlling for geography should substantially reduce, if not eliminate, these patterns. Further, with an explicit accounting for geography we also expect to reduce prediction errors.

7.1.1 Overall Index Performance

Time period	Median err (%)		% Homes		Homes
	Absolute	Actual	$\|e_{it}\| \leq 5$%	$\|e_{it}\| \leq 15$%	
Mar:Aug 2000	11.29	-3.50	24.79	60.91	15222
Jan:Jun 2004	10.31	3.61	25.95	65.56	16638
Jul:Dec 2007	14.90	-8.06	17.73	50.24	7355

Table 9: Error patterns with time: Case-Shiller Index

Tables 9 and 10 show the overall patterns of the error term for LA county under the two models. The spatio-temporal model comprehensively out-performs the CS index model in prediction accuracy. The difference is particularly stark for the Jul:Dec 2007 period where not only is the CS index drastically out-performed, but it also seems to over-predict on average by about 8% (median error), whereas the spatio-temporal model displays little bias in this period. Figure 4 shows that this difference is unlikely to be because of the index. This suggests that there are selection biases in what kind of homes came up for sale in this period.

Time period	Median err (%)		% Homes		Homes
	Absolute	Actual	$\|e_{it}\| \leq 5$%	$\|e_{it}\| \leq 15$%	
Mar:Aug 2000	9.43	-2.47	29.04	69.30	15222
Jan:Jun 2004	8.00	-3.17	32.95	76.42	16638
Jul:Dec 2007	9.24	+0.5	29.03	71.05	7355

Table 10: Error patterns with time: Spatio-temporal Index

7.1.2 Turnover time

Tables 11 and 12 show the dramatic differences in patterns of the error term with turnover time in the two models. Table 11 demonstrates the previously discussed patterns in the CS prediction error. Prediction accuracy is high at low turnover times, and falls monotonically as turnover time increases. Also, the median error becomes increasingly negative as turnover time increases implying that homes that transact after a long time are over-predicted.

Time btwn sales	**Median err (%)**		**% Homes**		**Homes**
	Absolute	**Actual**	$\|e_{it}\| \leq 5$**%**	$\|e_{it}\| \leq 15$**%**	
$<$ 1 year	6.36	2.61	41.78	90.18	1589
1-2 years	8.65	0.6	31.04	72.12	4133
2-3 years	9.22	2.41	28.68	71.08	4665
3-5 years	11.18	1.13	23.23	62.65	673
5-7 years	12.67	-1.2	21.53	57.7	4835
7-10 years	11.5	-2.54	23.98	61.26	5370
10-12 years	12.87	-5.51	21.16	55.96	3638
12-15 years	14.74	-5.46	19.66	50.49	4220
15-20 years	18.04	-5.31	16.04	43.24	3515
$>$ 20 years	24.13	-10.51	9.78	30.52	511

Table 11: Error patterns with time between sales: Case-Shiller Index

Table 12 shows no patterns in the error term with turnover time for the spatio-temporal model. The median error stays in the -1.5% to -2.5% range implying some over-prediction in all ranges of turnover time. Prediction accuracy too stays roughly constant with median absolute error at around 9% and a little over 70% of the samples being predicted with less than 15% absolute error.

The spatio-temporal model is significantly out-performed by the CS index model at turnover times less than a year. For transactions that happen between 1 and 3 years, there is little difference between the two models. As turnover time increases, the spatio-temporal model performs increasingly better and the CS index model performance rapidly deteriorates. Overall, the effect of turnover time is completely removed when geography is taken into account. This suggests that turnover time displays strong geographic patterns, with clusters of households that trade infre-

Time btwn sales	Median err (%)		% Homes		Homes
	Absolute	Actual	$\|e_{it}\| \leq 5$%	$\|e_{it}\| \leq 15$%	
$<$ 1 year	7.79	-1.48	32.47	75.39	1589
1-2 years	9.21	-1.98	29.71	72.17	4133
2-3 years	8.71	-3.15	30.71	73.48	4665
3-5 years	8.8	-2.81	30.21	71.93	6739
5-7 years	8.78	-1.94	29.12	72.9	4835
7-10 years	8.53	-2.65	31.43	72.77	5370
10-12 years	8.51	-2.57	30.89	73.44	3638
12-15 years	8.78	-1.82	31.2	71.63	4220
15-20 years	8.4	-1.76	32.51	72.2	3515
$>$ 20 years	9.53	+0.72	28.37	71.81	511

Table 12: Error patterns with time between sales: Spatio-temporal Index

quently due e.g. to catering to empty nesters and retirees, and other clusters that trade frequently e.g. as starter homes.

7.1.3 Mean Reversion

The most stark finding in this section (tables 13 and 14) is that once geography is accounted for the error term for homes with low initial prices no longer mean reverts. Since the indices are the same, this implies that geography proxies for mean reversion. There is dramatic improvement in prediction accuracy in the spatio-temporal model over the CS index model at low initial prices with absolutely no under-prediction. At higher initial prices, both models perform badly and over-predict.

7.1.4 Geography

When broken down into neighborhoods, the spatio-temporal model again significantly out-performs the CS model, except in Pasadena. Another interesting fact to jump out of these tables is that for Pasadena, Van Nuys and Long Beach, the CS model under-predicts by about 2% whereas the

Initial price ($)	Median err (%)		% Homes		Homes
	Absolute	Actual	$\|e_{it}\| \leq 5$%	$\|e_{it}\| \leq 15$%	
$\leq$ 50,000	69.25	67.51	0.92	3.88	540
50-75,000	26.17	15.72	12.28	30.64	855
75-100,000	17.26	-0.77	15.88	44.47	2008
100-125,000	14.7	-0.2	18.12	50.95	3144
125-150,000	12.64	-1.32	21.24	57.48	4702
150-200,000	10.35	0.83	25.24	65.58	8533
200-250,000	9.26	1.45	28.91	71.11	5415
250-300,000	9.26	0.48	29.73	69.81	3810
300-350,000	9.71	-1.2	27.38	67.02	2669
350-400,000	9.85	-2.84	28.65	65.89	2111
400-450,000	10.96	-3.83	25.78	64	1567
450-500,000	11.7	-6.3	23.61	59.83	1245
500-600,000	12.54	-8.5	22.83	57.52	1747
600-750,000	13.45	-8.87	19.36	54.57	1508
750,000-1 million	13.66	-7.66	21.31	53.23	990
$>$ 1 million	13.41	-6.53	21.44	54.53	816

Table 13: Error patterns with initial price: Case-Shiller Index

spatio-temporal model over-predicts by about 5-6%.

8 Conclusion

We have introduced new techniques from the machine learning literature and on this basis have developed a new methodology for estimating house prices. Our methodology is designed to provide richer insights into spatio-temporal variations in returns than do current repeat sales based methods. The results provide clear evidence that our techniques indeed provide such insights. Specifically, they allow us to spot systematic patterns in the supposedly idiosyncratic prediction errors associated with repeat sales indices. These positive findings open the door to new re-

Initial price ($)	Median err (%)		% Homes		Homes
	Absolute	Actual	$\|e_{it}\| \leq 5$%	$\|e_{it}\| \leq 15$%	
$\leq$ 50,0000	10.46	0.91	27.4	62.59	540
50-75,000	10.71	1.22	25.49	62.1	855
75-100,000	9.28	0.07	30.37	69.47	2008
100-125,000	8.83	-0.43	30.18	73.31	3144
125-150,000	8.11	-0.95	32.28	75.84	4702
150-200,000	7.74	-2.39	34.22	77.97	8533
200-250,000	8.00	-3.37	33.22	77.52	5415
250-300,000	8.51	-3.24	30.73	73.62	3810
300-350,000	9.13	-3.48	29.18	72.16	2669
350-400,000	9.15	-2.56	29.79	70.39	2111
400-450,000	9.7	-2.64	28.2	68.41	1567
450-500,000	9.63	-1.59	27.38	68.99	1245
500-600,000	10.32	-1.09	27.41	66.34	1747
600-750,000	10.51	-2.46	25.19	65.64	1508
750,000-1 million	12.75	-5.00	23.23	57.07	990
$\geq$ 1 million	16.59	-13.94	17.03	44.73	816

Table 14: Error patterns with initial price: Spatio-temporal Index

Neighborhood	Median err (%)		% Homes		Homes
	Absolute	Actual	$\|e_{it}\| \leq 5$%	$\|e_{it}\| \leq 15$%	
Palmdale	13.96	-5.35	19.78	52.6	1673
Pasadena	11.12	2.87	24.42	61.29	819
Torrance etc.	10.16	-1.79	24.78	64.01	831
Van Nuys	13.72	2.19	18.61	54.3	779
Long Beach	10.95	2.26	25.14	64.82	1845
Industry etc.	10.53	-4.35	25.79	63.55	1663

Table 15: Spatial error patterns: Case-Shiller Index

Neighborhood	Median err (%)		% Homes		Homes
	Absolute	Actual	$\|e_{it}\| \leq 5$%	$\|e_{it}\| \leq 15$%	
Palmdale	8.4	-0.46	31.2	75.31	1673
Pasadena	12.43	-6.03	21.85	57.63	819
Torrance	8.48	-1.44	30.92	74.84	831
Van Nuys	9.38	-5.03	26.44	72.65	779
Long Beach	9.19	-4.3	29.21	74.25	1845
Industry etc.	6.91	+0.27	38.18	81.9	1663

Table 16: Spatial error patterns: Spatio-temporal Index

search in the many areas of economics in which house prices play a role, including a wide range of research programs that focus not only on the determinants of house prices, but also mortgage default, and systematic patterns of local house prices.

References

Anglin, Paul M. and Gencay, Ramazan (1996) "Semiparametric Estimation of a Hedonic Price Function," *Journal of Applied Econometrics*, 11(6).

Bailey, Martin J., Muth, Richard F., and Nourse, Hugh O. (1963) "A Regression Model for Real Estate Price Index Construction," *Journal of the American Statistical Association*, 58(304), 933–942.

Basu, Sabyasachi and Thibodeau, Thomas G. (1998) "Analysis of Spatial Autocorrelation in Home Prices," *Journal of Real Estate Finance and Economics*, 16(1), 61–85.

Bourasa, Steven C., Hamelink, Foorte, and Hoseli, Martin (1999) "Defining Housing Sub-markets," *8*, 8, 160–183.

Can, Ayse (1990) "The Measurement of Neighborhood Dynamics in Urban House Prices," *Economic Geography*, 66(3).

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

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

Case, Karl E. and Shiller, Robert J. (1989) “The Efficiency of the Market for Single Family Homes,” *American Economic Review*, 79(1), 125–137.

Clapp, John M. (2004) “A Semiparametric Method for Estimating Local House Price Indices,” *Real Estate Economics*, 32(1), 127–160.

Dubin, Robin A. (1992) “Spatial Autocorrelation and Neighborhood Quality,” *Regional Science and Urban Economics*, 22, 432–452.

Dubin, Robin A. (1998) “Predicting House Prices Using Multiple Listings Data,” *Journal of Real Estate Finance and Economics*, 17(1).

Goetzmann, William N. and Spiegel, Matthew (1995) “Non-Temporal Components of Residential Real Estate Price Appreciation,” *The Review of Economics and Statistics*, 77(1), 199–206.

Goetzmann, William N. and Spiegel, Matthew (1997) “A Spatial Model of Housing Returns and Neighborhood Substitutability,” *Journal of Real Estate Finance and Economics*, 14, 11–31.

Goodman, Alan C. (1981) “Housing Sub-markets within Urban Areas: Definitions and Evidence,” *Journal of Regional Science*, 21, 175–185.

Halvorsen, Roberts and Pollakowski, Henry (1981) “Choice of Functional Form for Hedonic Price Equations,” *Journal of Urban Economics*, 10, 37–49.

McLennan, D. and Tu, Y. (1996) “Economic Perspectives on the Structure of Local Housing Systems,” *Housing Studies*, 11(3), 387–406.

Meese, Richard and Wallace, Nancy (1991) “Nonparametric Estimation of Dynamic Hedonic Price Models and the Construction of Residential Housing Price Indices,” *AREUEA Journal*, 19, 308–332.

Michaels, Gregory R. and Smith, Keryy V. (1990) “Market Segmentation and Valuing Amnesties with Hedonic Models: The Case of Hazardous Waste Sites,” *Journal of Urbam Economics*, 28, 232–242.

Pace, R. Kelley (1993) “Non-parametric Methods with Applications to Hedonic Models,” *Journal of Real Estate Finance and Economics*, 7, 185–204.

Pace, R. Kelley, Barry, Ronald, Clapp, John M., and Rodriguez, Mauricio (1998) “Spatio-Temporal Autoregressive Models of Neighborhood Effects,” *Journal of Real Estate Finance and Economics*, 17(1), 15–33.

Pace, R. Kelley and Gilley, Otis (1997) “Using the Spatial Configuration of the Data to improve Estimation,” *Journal of Real Estate Finance and Economics*, 14(3), 333–340.

Rosen, Sherwin (1974) “Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition,” *The Journal of Political Economy*, 82(1), 34–55.

Schnare, A.B. and Struyk, R.J. (1976) “Segmentation in Urban Housing Markets,” *Journal of Urban Economics*, 3(2), 146–166.

Thibodeau, Thomas G. (2003) “Marking Single-Family Property Values to Market,” *Real Estate Economics*, 31(1), 1–22.