An Autoregressive Approach to House Price Modeling

Chaitra H. Nagaraja, Lawrence D. Brown, Linda H. Zhao

December 19, 2009

Abstract

A statistical model for predicting individual house prices is proposed utilizing only information regarding sale price, time of sale, and location (ZIP code). This model is composed of a fixed time effect and a random ZIP (postal) code effect combined with an autoregressive component. The latter piece is applied only to homes sold repeatedly while the former two components are applied to all of the data. In addition, the autoregressive component incorporates heteroscedasticity in the errors. To evaluate the proposed model, single-family home sales for twenty U.S. metropolitan areas from July 1985 through September 2004 are analyzed. The model is shown to have better predictive abilities than the benchmark S&P/Case-Shiller model, which is a repeat sales model, and a conventional mixed effects model. It is also shown that the time effect in the proposed model can be converted into a house price index. Finally, the special case of Los Angeles, CA is discussed as an example of history repeating itself in regards to the current housing market meltdown.

Keywords: autoregressive, time series, repeat sales.

1 Introduction

Modeling house prices presents a unique set of challenges. Houses are distinctive, each with its own set of hedonic characteristics: number of bedrooms, square footage, location, amenities, and so forth. Moreover the price of a house, or the value of the bundle of characteristics, is observed only when sold. Sales, however, occur infrequently. As a result, during any length of time, out of the entire population of homes, only a small sample are actually sold. From this limited information, our objective is to develop a practical model to predict prices. Our
A common approach for modeling house prices, called repeat sales, utilizes homes that sell multiple times to track market trends. Bailey, Muth, and Nourse (1963) (hereby referred to as BMN) first proposed this method and Case and Shiller (1987, 1989) extended it to incorporate heteroscedastic errors. In both models, the log price difference between two successive sales of a home is used to construct a house price index using linear regression. The previous sale price acts as a surrogate for hedonic information, provided the home remains substantially unchanged between sales. Much research has been done on improving the index estimates produced by this approach. Shiller (1991) and Goetzmann (2002) propose arithmetic average versions of the repeat sales estimator as an alternative to the original geometric average estimator. Gatzlaff and Haurin (1997) suggest a repeat sales model that corrects the correlation between economic conditions and the chance of a sale occurring. A modified form of the repeat sales models is used for the Home Price Index produced by the Office of Federal Housing Enterprise Oversight (OFHEO) [5]. With a few changes, the repeat sales index is used commercially by Standard and Poors (S&P/Case-Shiller Home Price Index). We will be using this index in our analysis as it is the most well-known.

Several criticisms have been made about repeat sales methods. For instance, a house is assumed to have undergone no changes between sales. If it has, the home is omitted from the analysis. Many have commented on the difficulty of detecting such changes without the availability of additional information about the home [16, 10]. Goetzmann, et al (1995) propose an alternate model to correct for the effect of changes to homes around the time the house is sold. Furthermore, once homes which have changed are removed from the data set, an index constructed out of the remaining homes may still not reflect the true effect of time. Case, et al (1991) argue that houses age and this has been shown to have a depreciating effect on a house price. Therefore, the repeat sales indices provide an estimate of time effects confounded with the effect of age [6, p. 289].

In a sample period, out of the entire population of homes, only a small fraction are actually sold. A fraction of these sales are repeat sales homes with no significant changes. Recall, the remaining sales, those of the single sales homes, are excluded from the analysis. If repeat sales indices are used to describe the housing market as a whole, one would hope that the sample of repeat sales homes have similar characteristics to all homes. If not, the indices would be affected by sample selection bias [6, p. 290]. England, et al (1999) found in a study of Swedish home sales and Meese and Wallace (1997) in a study of Oakland and Freemont home sales that repeat sales homes are indeed different from single sale homes. Both studies found that in addition to being older, repeat sales homes are smaller and are more “modest”
Therefore, repeat sales indices only provide information about a very specific type of home and may not apply to the entire housing market. However, published indices do not seem to be interpreted in that manner.

Case and Quigley (1991) proposed a hybrid model that combined repeat sales methodology with hedonic information so all sales could be included; however, this requires housing characteristics that may be difficult to collect. The model proposed in our paper attempts to improve the repeat sales methodology while still maintaining the simplicity and reduced data requirements that the original BMN method had. Our primary goal is prediction, but we believe the resulting index is could be a better general description of housing sales than traditional repeat sales methodology.

We feel that the premise of repeat sales methodology is valuable although the current models in use such as the S&P/Case-Shiller and OFHEO Home Price Index suffer from the issues described above. The proposed model applies the repeat sales idea albeit in a different manner in an attempt to correct some of the criticisms. Log prices are modeled as the sum of a time effect (index), a location effect modeled as a random effect for ZIP code, and an underlying first-order autoregressive time series (AR(1)). This setting offers four advantages. First, the price index is estimated with all sales—single and repeat. In essence, the index is a weighted sum of price information from single and repeat sales. The latter component receives a much higher weight because more useful information is available for those homes. Second, the previous sale price becomes less useful the longer it has been since the last sale. The AR(1) series includes this feature into the model more effectively than the Case-Shiller method. Third, metropolitan areas are diverse and neighborhoods may have disparate trends. We include ZIP (postal) codes to model these differences in location. (ZIP code was readily available in our data; other geographic variables at roughly this scale might have been even more useful had they been available.) Finally, the proposed model is straightforward to interpret even while including the features described above. We believe the model captures trends in the overall housing market better than existing repeat sales methods and is a practical alternative.

To test the predictive capacity, the model is applied to single family home sales from July 1985 through September 2004 for twenty US metropolitan areas. These data are described in Sec. 2. The autoregressive model is proposed and the estimation using maximum likelihood is given in Sec. 3. In Secs. 4 and 5, results are discussed. For comparison, two alternative models are fit: a conventional mixed effects model and the method used in the S&P/Case-Shiller Home Price Index. In Sec. 6 we examine the case of Los Angeles, CA where the proposed model does not perform as well. We end with a general discussion in Sec. 7.

# 2 House Price Data

The data are comprised of single family home sales qualifying for conventional mortgages from the twenty US metropolitan areas listed in Table 1. These sales occurred between July 1985 and September 2004. Not included in these data are those homes with prices that are too high to be considered for a conventional mortgage or those homes that are sold at
Table 1: Metropolitan Areas in Data

| Metropolitan Area    | Ann Arbor, MI | Kansas City, MO | Minneapolis, MN | Raleigh, NC |
|----------------------|---------------|-----------------|-----------------|------------|
|                      | Atlanta, GA   | Lexington, KY   | Orlando, FL     | San Francisco, CA |
|                      | Chicago, IL   | Los Angeles, CA | Philadelphia, PA | Seattle, WA |
|                      | Columbia, SC  | Madison, WI     | Phoenix, AZ     | Sioux Falls, SD |
|                      | Columbus, OH  | Memphis, TN     | Pittsburgh, PA  | Stamford, CT |

Table 2: Summary of Sample Cities

| Metropolitan Area | Sales | Houses |
|-------------------|-------|--------|
| Stamford, CT      | 14,602| 11,128 |
| Ann Arbor, MI     | 68,684| 48,522 |
| Pittsburgh, PA    | 104,544| 73,871 |
| Los Angeles, CA   | 543,071| 395,061 |
| Chicago, IL       | 688,468| 483,581 |

subprime rates. However, subprime loans were not prevalent during the time period covered by our data. The same type of data are used by Fannie Mae and Freddie Mac and are used in constructing the OFHEO Home Price Index. For each observation, the following information is available: address with ZIP code, month and year of sale, and price. To ensure adequate data per time period, we divide the sample period into three month intervals for a total of 77 periods, or quarters. We make an attempt to remove non-arms-length sales by omitting homes that are sold more than once in a single quarter. In this section, we provide a brief overview for five cities: Stamford, CT, Ann Arbor, MI, Pittsburgh, PA, Los Angeles, CA, and Chicago, IL. Complete tables are provided in Appendix A.

Table 2 shows the number of sales and unique houses sold in the sample period. Since houses can sell multiple times (repeat sales), the total number of sales is always greater than the number of houses. Perhaps more illuminating is Table 3 which breaks down houses by the number of times each was sold. As expected, as the number of sales per house increases,

Table 3: Sale Counts for Sample Cities

| Metropolitan Area     | 1 sale | 2 sales | 3 sales | 4+ sales |
|-----------------------|--------|---------|---------|----------|
| Stamford, CT          | 8,200  | 2,502   | 357     | 62       |
| Ann Arbor, MI         | 32,458 | 12,662  | 2,781   | 621      |
| Pittsburgh, PA        | 48,618 | 20,768  | 3,749   | 718      |
| Los Angeles, CA       | 272,258| 100,918 | 18,965  | 2,903    |
| Chicago, IL           | 319,340| 130,234 | 28,369  | 5,603    |
The number of houses drops off rapidly. Note that there are a significant number of homes which sell more than twice. With such a long sample period (nearly twenty years), this is not unusual; however, single sales are the most common even with a long sample period. The first column of Table 3 shows this clearly. This pattern holds for all cities in our data.

For all metropolitan areas in our data, the time of a sale is fuzzy as there is often a lag between the day when the price is agreed upon and the day the sale is recorded (around 20-60 days). Theoretically, the true value of the house would have changed between these two points. Therefore, in the strictest sense, the sale price of the house does not reflect the price at the time when the sale is recorded. Dividing the year into quarters reduces the importance of this lag effect.

3 The Proposed Model

The log house price series is modeled as the sum of an index component, an effect for ZIP code (as an indicator for location), and an AR(1) time series. Let $y_{i,j,z}$ be the log sale price of the $j$th sale of $i$th house in ZIP code $z$. The sale prices of a particular house are treated as a series of sales: $y_{i,1,z}, y_{i,2,z}, \ldots, y_{i,J_i,z}, \ldots$. Note that $y_{i,1,z}$ is defined as the first sale price in the sample period; as a result, both new homes and old homes sold for the first time in the sample period are indicated with this notation. Say there are $1, \ldots, T$ discrete time periods where house sales occur. Let $t(i,j,z)$ denote the time period when the $j$th sale of the $i$th house in zip code $z$ occurs and $\gamma(i,j,z)$ to be $t(i,j,z) - t(i,j-1,z)$, or the gap time between sales. Finally, there are a total of $N = \sum_{z=1}^{Z} \sum_{i=1}^{J_i} J_i$ observations in the data where there
are \( Z \) ZIP codes, \( I_z \) houses in each ZIP code, and \( J_t \) sales for a given house.

The log sale price \( y_{i,j,z} \) can now be described as follows:

\[
y_{i,1,z} = \mu + \beta_{t(i,1,z)} + \tau_z + \varepsilon_{i,1,z} \\
y_{i,j,z} = \mu + \beta_{t(i,j,z)} + \tau_z + \phi^{\gamma(i,j,z)} (y_{i,j-1,z} - \mu - \beta_{t(i,j-1,z)} - \tau_z) + \varepsilon_{i,j,z} \quad j > 1
\]

(1)

where:

1. The parameter \( \beta_{t(i,j,z)} \) is the log price index at time \( t(i,j,z) \). Let \( \beta_1, \ldots, \beta_T \) denote the log price indices which are assumed to be fixed effects.

2. \( \phi \) is the autoregressive coefficient and \( |\phi| < 1 \).

3. \( \tau_z \) is the random effect for ZIP code \( z \). \( \tau_z \overset{iid}{\sim} \mathcal{N}(0, \sigma^2_{\tau}) \) where \( \tau_1, \ldots, \tau_Z \) are the ZIP code random effects which are distributed normally with mean 0 and variance \( \sigma^2_{\tau} \), and \( iid \) denotes independent and identically distributed.

4. We impose the restriction that \( \sum_{t=1}^{T} n_t \beta_t = 0 \) where \( n_t \) is the number of sales at time \( t \). This allows us to interpret \( \mu \) as the overall mean.

5. Finally, let \( \varepsilon_{i,1,z} \sim \mathcal{N} \left( 0, \frac{\sigma^2_{\varepsilon}}{1-\phi^2} \right) \), \( \varepsilon_{i,j,z} \sim \mathcal{N} \left( 0, \frac{\sigma^2_{\varepsilon}(1-\phi^{2\gamma(i,j,z)})}{1-\phi^2} \right) \), and assume that all \( \varepsilon_{i,j,z} \) are independent.

Note that there is only one process for the series \( y_{i,1,z}, y_{i,2,z}, \ldots \). The error variance for the first sale, \( \sigma^2_{\varepsilon}/(1 - \phi^2) \) is a marginal variance. For subsequent sales, as we have information about previous sales, it is appropriate to use the conditional variance (conditional on the previous sale), \( \sigma^2_{\varepsilon} (1 - \phi^{2\gamma(i,j,z)})/(1 - \phi^2) \), instead.

The underlying series for each house is given by \( u_{i,j,z} = y_{i,j,z} - \mu - \beta_{t(i,j,z)} - \tau_z \). We can rewrite this series as: \( u_{i,j,z} = \phi^{\gamma(i,j,z)} u_{i,j-1,z} + \varepsilon_{i,j,z} \) where \( \varepsilon_{i,j,z} \) is as given above. This autoregressive series is stationary, given a starting observation \( u_{i,1,z} \), because \( E[u_{i,j,z}] = 0 \), a constant, where \( E[\cdot] \) is the expectation function and the covariance between two points depends only on the gap time. Specifically, \( \text{Cov}(u_{i,j,z}, u_{i,j'}z) = \sigma^2_{\varepsilon} \phi^{\gamma(i,j',z)-\gamma(i,j,z)}/(1 - \phi^2) \) if \( j' < j \). Thus the covariance between a pair of sales depends only on the gap time between sales. Consequently, the time of sale is uninformative for the underlying series, only the gap time is required. Furthermore, the series \( u_{i,j,z} \) where \( i \) and \( z \) are fixed and \( j \geq 1 \) is a Markov process as a result.

The autoregressive component adds an important feature to the model. Intuitively, the longer the gap time between sales, the less useful the previous sale price should become when predicting the next sale price. For the model described in (1), as gap time increases, the variance of the error term increases. This indicates that the information contained in the previous sale price is less useful than if the gap time had been shorter. Moreover, as the gap time increases, the autoregressive coefficient decreases by construction \( (\phi^{\gamma(i,j,z)}) \) meaning that sales prices of a home with long gap times are less correlated with each other. See Remark 3.1 at the end of this section for additional discussion on the form of \( \phi \).
To fit the model, we write the autoregressive model in (1) in matrix form:

\[ y = X\beta + Z\tau + \varepsilon^*. \]  

(2)

where \( y \) is the vector of log prices, \( X \) and \( Z \) are the design matrices for the fixed effects \( \beta = [\mu_1 \beta_2 \cdots \beta_{T-1}]' \) and random effects \( \tau \) respectively. Then, log price can be modeled as a mixed effects model with autocorrelated errors, \( \varepsilon^* \), and with covariance matrix \( V \).

We apply a transformation matrix \( T \) to the model (2) to simplify the computations; essentially, this matrix applies the autoregressive component of the model to both sides of (2). It is an \( N \times N \) matrix and is defined as follows. Let \( t_{(i,j,z),(i',j',z')} \) be the cell corresponding to the \((i,j,z)\)th row and \((i',j',z')\)th column. Then,

\[
t_{(i,j,z),(i',j',z')} = \begin{cases} 
1 & \text{if } i = i', j = j', z = z' \\
-\phi \gamma(i,j) & \text{if } i = i', j = j' + 1, z = z' \\
0 & \text{otherwise.}
\end{cases} \]  

(3)

As a result, \( T\varepsilon^* \sim \mathcal{N}(0, \frac{\sigma^2}{1-\phi^2} \text{diag}(r)) \) where \( \text{diag}(r) \) is a diagonal matrix of dimension \( N \) with the diagonal elements \( r \) being given by:

\[
r_{i,j,z} = \begin{cases} 
1 & \text{when } j = 1 \\
1 - \phi^{2\gamma(i,j)} & \text{when } j > 1.
\end{cases} \]  

(4)

Thus, using the notation from (1), let \( \varepsilon = T\varepsilon^* \). Finally, we restrict \( \sum_{t=1}^{T} n_t \beta_t = 0 \) where \( n_t \) is the number of sales at time \( t \). Therefore, \( \beta_T = -\frac{1}{nr} \sum_{t=1}^{T-1} n_t \beta_t \).

The likelihood function for the transformed model is:

\[
L(\theta; y) = (2\pi)^{-N/2}|V|^{-1/2} \exp \left\{ -\frac{1}{2} (T(y - X\beta))'V^{-1}(T(y - X\beta)) \right\}
\]  

(5)

where \( \theta = \{\beta, \sigma^2_\varepsilon, \sigma^2_\tau, \phi\} \) is the vector of parameters, \( N \) is the total number of observations, \( V \) is the covariance matrix, and \( T \) is the transformation matrix. We can split \( V \) into a sum of the variance contributions from the time series and the random effects. Specifically,

\[
V = \frac{\sigma^2_\varepsilon}{1-\phi^2} \text{diag}(r) + (TZ)D(TZ)'
\]  

(6)

where \( D = \sigma^2_\tau I_Z \) and \( I_Z \) is an identity matrix with dimension \( Z \times Z \).

We use the coordinate ascent algorithm to compute the maximum likelihood estimates (MLE) of \( \theta \) for the model in (1). This iterative procedure maximizes the likelihood function with respect to each group of parameters while holding all other parameters constant [3, p. 129]. The algorithm terminates when the parameter estimates have converged according to the specified stopping rule. For models in the exponential family, the coordinate ascent algorithm can be proven to converge to the MLE. The proposed model, however, is a member of the differentiable exponential family; therefore the proof does not directly apply [4].
Nonetheless, we find empirically that the likelihood function is well behaved so the MLE appears to be reached for this case as well. The algorithm is outlined below. The equations for updating the parameters and random effects estimates are given in Appendix B.

**AR Model Fitting Algorithm**

1. Set a tolerance level $\epsilon$ (possibly different for each parameter) and a maximum number of iterations $K$.
2. Initialize the parameters: $\theta^0 = \{\beta^0, \sigma^2_{\varepsilon}, \sigma^2_\tau, \phi^0\}$.
3. For iteration $k$,
   (a) For $t \in \{1, \ldots, T\}$, calculate $\beta^k$ using (16) in Appendix B. That is, $\beta^k = f(\sigma^2_{\varepsilon}^{k-1}, \sigma^2_\tau^{k-1}, \phi^{k-1})$.
   (b) Compute $\sigma^2_{\varepsilon}^k$ by computing the zero of (17) using $\{\beta^k, \sigma^2_{\varepsilon}^{k-1}, \sigma^2_\tau^{k-1}, \phi^{k-1}\}$.
   (c) Compute $\sigma^2_\tau^k$ by calculating the zero of (18) using $\{\beta^k, \sigma^2_{\varepsilon}^k, \sigma^2_\tau^{k-1}, \phi^{k-1}\}$.
   (d) Find the zero of (19) to compute $\phi^k$ using $\{\beta^k, \sigma^2_{\varepsilon}^k, \sigma^2_\tau^k\}$.
   (e) If $|\theta_i^{k-1} - \theta_i^k| > \epsilon$ for any $\theta_i \in \theta$ and the number of iterations is less than $K$, repeat Step 3 after replacing $\theta^{k-1}$ with $\theta^k$. Otherwise, stop (call this iteration $K'$ where $K' \leq K$).
4. Solve for $\beta_T$ by computing: $\hat{\beta}_T = -\frac{1}{n_T} \sum_{i=1}^{T-1} n_i \hat{\beta}_i^{K'}$.
5. Plug in $\{\beta^{K'}, \sigma^2_{\varepsilon}^{K'}, \sigma^2_\tau^{K'}, \phi^{K'}\}$ to compute the estimated values for $\tau$ using (20).

To predict a log price, we substitute the estimated parameters and random effects into (1):

\[
\hat{y}_{i,j,z} = \hat{\mu} + \hat{\beta}_{t(i,j,z)} + \hat{\tau}_z + \hat{\phi}^{\gamma(i,j,z)} (y_{i,j-1,z} - \hat{\mu} - \hat{\beta}_{t(i,j-1,z)} - \hat{\tau}_z).
\]

(7)

We then convert $\hat{y}_{i,j,z}$ to the price scale (denoted as $\hat{Y}_{i,j,z}$) using:

\[
\hat{Y}_{i,j,z} (\sigma^2) = \exp \left\{ \hat{y}_{i,j,z} + \frac{\sigma^2}{2} \right\}.
\]

(8)

where $\sigma^2$ denotes the variance of $y_{i,j,z}$. The additional term $\frac{\sigma^2}{2}$ approximates the difference between $E[\exp\{X\}]$ and $\exp\{E[X]\}$ where $E[\cdot]$ is the expectation function. We must adjust the latter expression to approximate the conditional mean of the response. We improve the efficiency of our estimates by using this adjustment [20, p. 3025]. In (8), $\sigma^2$ is estimated
from the mean squared residuals (MSR), where MSR = \( \frac{1}{N} \sum_{i=1}^{N} \) where \( N \) is the total number of observations. Therefore, the log price estimates, \( \hat{y}_{ij,z} \) are converted to the price scale by:

\[
\hat{Y}_{ij,z} = \exp \left\{ \hat{y}_{ij,z} + \frac{\text{MSR}}{2} \right\}.
\] (9)

Goetzmann (1992) suggests a similar transformation for the index values computed using a traditional repeat sales method and Calhoun (1992) prescribes using this adjustment when using an index value to predict a particular house price. This adjustment is not needed when estimating the index. The standard error of the index is sufficiently small that the efficiency adjustment has a negligible impact on the estimated index. Therefore, \( \exp \{ \hat{\beta}_t \} \) is used to convert the index to the price scale. Finally, we rescale the vector of indices so that the first quarter has an index value of 1.

Remark 3.1. The form of the autoregressive coefficient, \( \phi^{\gamma(i,j,z)} \), deserves further explanation. For each house indexed by \((i,z)\) let \( t_1(i,z) = t(i,1,z) \) denote the time of the initial sale. Condition on the (unobserved) values of the parameters \( \{ \mu, \beta_t, \sigma^2_\varepsilon, \sigma^2_\tau \} \) and on the values of the random ZIP code effects, \( \{ \tau_z \} \). Let \( \{ u_{i,z,t} : t = t_1(i, z), t_1(i, z) + 1, \ldots \} \) be a (latent) AR (1) process, as described below. The values of \( u_{i,z,t} \) are to be interpreted as the potential sale price adjusted by \( \{ \mu, \beta_t, \sigma^2_\varepsilon, \sigma^2_\tau \} \) of the house indexed by \((i,z)\) if the house were to be sold at time \( t \).

To be more precise, \( u_{i,z,t} \) is a conventional, centered and stationary AR(1) process, defined by:

\[
u_{i,z,t} \begin{cases} 
\varepsilon_{i,1,z} & \text{if } t = t_1(i, z), \\
\phi u_{i,z,t-1} + \varepsilon_{i,1,z} & \text{if } t > t_1(i, z)
\end{cases}
\] (10)

where if \( t = t(i,j,z) \) then \( \varepsilon_{i,z,t(i,j,z)} = \varepsilon_{i,j,z} \) and otherwise \( \varepsilon_{i,z,t} \overset{iid}{\sim} \mathcal{N} \left( 0, \frac{\sigma^2_\varepsilon}{1-\phi^2} \right) \). Then the observed log sale prices are given by \( \{ y_{i,j,z} \} \) where \( u_{i,z,t(i,j,z)} = y_{i,j,z} - (\mu + \beta_{t(i,j,z)} + \tau_z) \).

For housing data like ours, the value of the autoregressive parameter, \( \phi \), for this latent process will be near the largest possible value, \( \phi = 1 \). Thus, if the latent process were actually an observed process from which one wanted to estimate \( \phi \), then estimation of \( \phi \) could be a delicate matter. However, sales generally occur with fairly large gap times and so the values of \( \phi^{\gamma(i,j,z)} \) occurring in the data will generally not be close to 1. Because of this, conventional estimation procedures perform satisfactorily for estimation of \( \phi \). We comment on this further in Sec. 4 in connection with the results in Table 4.

4 Estimation Results

To fit and validate the autoregressive (AR) model, we divide the observations for each city into training and test sets. The test set contains all final sales for homes that sell three or more times. Among homes that sell twice, the second sale is added to the test set with probability 1/2. As a result the test set for each city contains around 15% of the sales. The
remaining sales (including single sales) comprise the training set. Table 7, in Appendix A, lists the training and test set sizes for each city. In this section, we fit the model on the training set and examine the estimated parameters. The test set will be used in Sec. 5 to validate the AR model against other models.

In Table 4 the estimates for the overall mean $\mu$, the autoregressive parameter $\phi$, the variance of the error term $\sigma^2_{\varepsilon}$, and the variance of the random effects $\sigma^2_{\tau}$ are provided for each metropolitan area. The more expensive cities have the highest values of $\mu$: Los Angeles, CA, San Francisco, CA and Stamford, CT. Also, in Fig. 2, the indices for a sample of the twenty cities are provided; there are clearly different trends across cities.

The estimates for the AR model parameter $\phi$ are close to one. This is expected as the adjusted log sale prices, $u_{i,j,z}$, for sale pairs with short gap times are expected to be closer than those with longer gap times. It may be tempting to assume that since $\phi$ is so close to one, the prices form a random walk instead of an AR(1) time series. However, this is clearly not the case. Recall that $\phi$ enters the model not by itself but as $\phi^{\gamma(i,j,z)}$ where $\gamma(i,j,z)$ is the gap time. These gap times are high enough (the mean gap time is around 22 quarters) that the correlation coefficient $\phi^{\gamma(i,j,z)}$ is considerably lower than 1. For example, in the case of Ann Arbor, MI, $\hat{\phi}^{22} = 0.993247^{22} \approx 0.8615$ which is clearly less than 1. Therefore, our autoregressive analysis does not involve the types of sensitivity often entailed as a consequence of near unit roots in autoregressive models.

Recall that we model the adjusted log prices, $u_{i,j,z} = y_{i,j,z} - \beta_t(i,j,z) - \tau_z$ as a latent AR(1) time series. Thus, for each gap time, $\gamma(i,j,z) = h$, there is a different correlation between the sale pairs, namely $\phi^h$. To check that the data support the theory, we compare the correlation between pairs of quarter-adjusted log prices at each gap length.

First, we compute the estimated adjusted log prices $\hat{u}_{i,j,z} = y_{i,j,z} - \hat{\beta}_t(i,j,z) - \hat{\tau}_z$. Next, for each gap time $h$, we find all the sale pairs $(\hat{u}_{i,j-1,z}, \hat{u}_{i,j,z})$ with that particular gap length.
| Metropolitan Area     | $\hat{\mu}$ | $\hat{\phi}$ | $\hat{\sigma}_1^2$ | $\hat{\sigma}_2^2$ |
|-----------------------|-------------|--------------|--------------------|-------------------|
| Ann Arbor, MI         | 11.6643     | 0.993247     | 0.001567           | 0.110454          |
| Atlanta, GA           | 11.6882     | 0.992874     | 0.001651           | 0.070104          |
| Chicago, IL           | 11.8226     | 0.992000     | 0.001502           | 0.110683          |
| Columbia, SC          | 11.3843     | 0.997526     | 0.000883           | 0.028062          |
| Columbus, OH          | 11.5159     | 0.994807     | 0.001264           | 0.090329          |
| Kansas City, MO       | 11.4884     | 0.993734     | 0.001462           | 0.121954          |
| Lexington, KY         | 11.6224     | 0.996236     | 0.000968           | 0.048227          |
| Los Angeles, CA       | 12.1367     | 0.981888     | 0.002174           | 0.111708          |
| Madison, WI           | 11.7001     | 0.994318     | 0.001120           | 0.023295          |
| Memphis, TN           | 11.6572     | 0.994594     | 0.001120           | 0.101298          |
| Minneapolis, MN       | 11.8327     | 0.992008     | 0.001515           | 0.050961          |
| Orlando, FL           | 11.6055     | 0.993561     | 0.001676           | 0.046727          |
| Philadelphia, PA      | 11.7022     | 0.992349     | 0.001543           | 0.106971          |
| Phoenix, AZ           | 11.3408     | 0.992059     | 0.002546           | 0.103488          |
| Pittsburgh, PA        | 11.7447     | 0.993828     | 0.001413           | 0.047029          |
| Raleigh, NC           | 12.4236     | 0.985644     | 0.001788           | 0.056201          |
| San Francisco, CA     | 11.9998     | 0.989923     | 0.001658           | 0.039459          |
| Seattle, WA           | 11.6025     | 0.995262     | 0.001120           | 0.032719          |
| Sioux Falls, SD       | 12.5345     | 0.987938     | 0.002294           | 0.093230          |
The sample correlation between those sale pairs provides us with an estimate of $\phi$ for gap length $h$. If we repeat this for each possible gap length, we should obtain a steady decrease in the correlation as gap time increases. In particular, the points should follow the curve $\phi^h$ if the model is specified correctly.

In Fig. 3, we plot the correlation of the adjusted log prices by gap time for Columbus, OH. Note that the computed correlations for each gap time were computed with varying number of sale pairs. Those computed with fewer than twenty sale pairs are given in blue on the plot. We also overlay the predicted relationship between $\phi$ and gap time. The inverse relationship between gap time and correlation seems to hold well. We obtain similar results for most cities. One notable exception is Los Angeles, CA which we discuss in detail in Sec. 6.

5 Model Validation

To show that the proposed AR model provides good predictions, we fit the model separately to each of the twenty cities and apply the fitted models to each test set. For comparison purposes, the benchmark S&P/Case-Shiller model is applied to the data along with a mixed effects model. The latter model is a simple, but reasonable, alternative to the AR model. Both models are described below. The root mean squared error $\left( RMSE = \sqrt{\frac{1}{n} \sum_{k=1}^{N} (Y_k - \hat{Y}_k)^2} \right)$, where $Y$ is the sale price and $n$ is the size of the test set, is used to evaluate predictive per-
formance for each city in Sec. 5.3. The price indices and residuals obtained from the model are also analyzed. We see that the AR model provides the best predictions. We show the results from Columbus, OH again as a typical example in this section.

5.1 Mixed Effects Model

A mixed effects model provides another plausible approach for modeling these data. This model treats the time effect \( \beta_t \) as a fixed effect, and the effects of house \( \alpha_i \) and ZIP code \( \tau_z \) are modeled as random effects. There is no time series component to this model. We describe the model as follows:

\[
y_{i,j,z} = \mu + \alpha_i + \tau_z + \beta_{t(i,j,z)} + \varepsilon_{i,j,z}
\]

where \( \alpha_i \sim iid \mathcal{N}(0, \sigma^2_\alpha) \), \( \tau_z \sim iid \mathcal{N}(0, \sigma^2_\tau) \), and \( \varepsilon_{i,j,z} \sim iid \mathcal{N}(0, \sigma^2_\varepsilon) \) for houses \( i \) from 1, \ldots, \( I_z \), sales \( j \) from 1, \ldots, \( J_i \), and ZIP codes \( z \) from 1, \ldots, \( Z \). As before, \( \mu \) is a fixed parameter and \( \beta_{i,j,z} \) is the fixed effect for time. The estimates for the parameters \( \theta = \{ \mu, \beta, \sigma^2_\varepsilon, \sigma^2_\tau \} \) are computed using maximum likelihood estimation.

Finally, estimates for the random effects \( \alpha \) and \( \tau \) are calculated by iteratively calculating the following:

\[
\hat{\alpha} = \left( \frac{\sigma^2_\varepsilon}{\sigma^2_\alpha} I + W'W \right)^{-1} W' \left( y - X \hat{\beta} - Z \hat{\tau} \right),
\]
\[
\hat{\tau} = \left( \frac{\sigma^2_\varepsilon}{\sigma^2_\tau} I + Z'Z \right)^{-1} Z' \left( y - X \hat{\beta} - W \hat{\alpha} \right).
\]

where \( X \) and \( W \) are the design matrices for the fixed and random effects respectively and \( y \) is the response vector. These expressions are derived using Henderson’s method of computing BLUP estimators (1975).

To predict the log prices, we substitute in the estimated values:

\[
\hat{y}_{i,j,z} = \hat{\mu} + \hat{\beta}_{t(i,j,z)} + \hat{\alpha}_i + \hat{\tau}_z
\]

We use transformation (8) to convert these predictions back to the price scale. Finally we construct a price index similar to the autoregressive case. As in Fig. 2, the values of \( \exp\{ \hat{\beta}_t \} \) are rescaled so that the price index at the first quarter is 1.

5.2 S&P/Case-Shiller Model

The original Case-Shiller model (1987) is a repeat-sales model which expands upon the BMN setting by accounting for heteroscedasticity in the data due to the gap times between sales. Borrowing some of their notation, the framework for the model is:

\[
y_{i,t} = \beta_t + H_{i,t} + u_{i,t}
\]
where \( y_{i,t} \) is the log price of the sale of the \( i \)th house at time \( t \), \( \beta_t \) is the log index at time \( t \), and \( u_{i,t} \overset{iid}{\sim} \mathcal{N}(0, \sigma_u^2) \). The middle term, \( H_{i,t} \), is a Gaussian random walk which incorporates the previous log sale price of the house \[9, p. 126\]. Location information, such as ZIP codes, are not included in this model. Like the BMN setup, the Case-Shiller setting is a model for differences in prices. Thus, the following model is fit:

\[
y_{i,t'} - y_{i,t} = \beta_{t'} - \beta_t + \sum_{k=t+1}^{t'} v_{i,k} + u_{i,t'} - u_{i,t} \tag{14}
\]

where \( t' > t \). The random walk steps \( v_{i,k} \overset{iid}{\sim} \mathcal{N}(0, \sigma_v^2) \). Weighted least squares is used to fit the model to account for both sources of variation.

The S&P/Case-Shiller procedure follows in a similar vein but is fit on the price scale instead of the log price scale. This change requires introducing some additional steps to the fitting process. One of these steps is using instrumental variables. The procedure followed is described briefly next; however, full details are available in S&P/Case-Shiller methodology report \[23, p. 22\]. The design matrix \( X \), an instrumental variables (IV) matrix \( Z \), and response vector \( w \) are now defined as follows \[23, p. 22\]:

\[
X_{s,t} = \begin{cases} 
-Y_s & \text{if first sale at } t, \ t > 1, \\
Y_s & \text{if second sale at } t, \\
0 & \text{otherwise}.
\end{cases}
\]

\[
Z_{s,t} = \begin{cases} 
-1 & \text{if first sale at } t, \ t > 1, \\
1 & \text{if second sale at } t, \\
0 & \text{otherwise}.
\end{cases}
\]

\[
w_s = \begin{cases} 
Y_s & \text{first sale at time } 1, \\
0 & \text{else}
\end{cases}
\]

where \( Y_s \) is the sale price (not log price) at time \( s \). We should still think of each observation as part of a sale pair.

The goal is to fit the model \( w = Xb + \varepsilon \) where \( b = (b_1, \ldots, b_T) \) is the vector of the reciprocal price indices. That is, \( B_t = 1/b_t \) is the price index at time \( t \). A three step process is implemented to fit this model. First, \( b \) is estimated using regression with instrumental variables. Second, the residuals from this regression are used to compute weights for each observation. Finally, \( b \) is estimated again, this time incorporating the weights. This process is described in full below \[23\].

1. Estimate \( b \). Run a regression using instrumental variables: \( \hat{b} = (Z'X)^{-1}Z'w \).

2. Calculate the weights for each observation using the squared residuals from the first step. These weights are dependent on the gap time between sales. We denote these as \( \hat{\varepsilon}_i \). This residual is an estimate of: \( u_{i,t'} - u_{i,t} + \sum_{k=1}^{t'-t} v_{i,k} \). The expectation of
\( \varepsilon_i \) is \( E[u_{i,t'} - u_{i,t} + \sum_{k=1}^{t'-t} v_{i,k}] = 0 \) and the variance is \( Var[u_{i,t'} - u_{i,t} + \sum_{k=1}^{t'-t} v_{i,k}] = 2\sigma_u^2 + (t'-t)\sigma_v^2 \) as the errors are independent of each other. To compute the weights for each observation, the squared residuals from the first step are regressed against the gap time. That is,

\[
\hat{\varepsilon}_i^2 = \alpha_0 + \frac{\alpha_1}{2\sigma_u^2} (t'-t) + \eta_i
\]

where \( E[\eta_i] = 0 \). The reciprocal of the square root of the fitted values from the above regression are the weights. We denote this weight matrix by \( \Omega^{-1} \).

3. The final step is to re-estimate \( b \) while incorporating the weights, \( \Omega \):

\[
b = (Z'\Omega^{-1}X)^{-1}Z'\Omega^{-1}w.
\]

The indices are then the reciprocals of each element in \( b \) for \( t > 1 \) and, by construction, \( B_1 = 1 \).

Finally, to estimate the prices in the test set, we simply do the following:

\[
\hat{Y}_{i,j} = \frac{\hat{B}_{t(i,j)-1}}{B_{t(i,j)}} Y_{i,j-1}
\]

where \( Y_{i,j} \) is the price of the \( j \)th sale of the \( i \)th house and \( B_t \) is the price index at time \( t \). We do not apply the Goetzmann correction when estimating prices because the correction is appropriate only on the log price scale which then must be converted to the price scale \([14]\). The S&P/Case-Shiller method is fitted on the price scale so no transformation is required.

### 5.3 Comparing Predictions

We fit all three models on the training sets for each city and predict prices for those homes in the corresponding test set. The RMSE for the test set observations is calculated in dollars for each model to compare performance across models. Note that while the S&P/Case-Shiller produces predictions directly on the price scale, the autoregressive and mixed effects models must be converted back to the price scale using (9). These results are listed in Table 5. The model with the lowest RMSE value is given in italicized font. It is clear that the AR model performs better than the benchmark S&P/Case-Shiller model for all of the cities, reducing the RMSE by up to 21% depending on the city. The improved performance of the AR model over the S&P/Case-Shiller model is robust against alternative loss functions as well.

Note the missing value for Kansas City, MO for the S&P/Case-Shiller model. Some of the observation weights calculated in the second step of the procedure were negative stopping the estimation process. This is another drawback to some of the existing repeat sales procedures. Calhoun (1996) suggests replacing the sale specific error \( u_{i,t} \) (as given in (14)) with a house specific error \( u_i \); however, this fundamentally changes the structure of the error term and the fitting process and is not used in the S&P/Case-Shiller methodology. Therefore, we do not apply it to our data. Three values are also missing for the mixed effect model results. For these three cities, the failure of convergence can be attributed to a combination of the very
### Table 5: Test Set RMSE for Local Models (in dollars)

| Metropolitan Area       | AR (Local) | Mixed Effects (Local) | S&P/C-S |
|-------------------------|------------|-----------------------|---------|
| Ann Arbor, MI           | 41,401     | 46,519                | 52,718  |
| Atlanta, GA             | 30,914     | 34,912                | 35,482  |
| Chicago, IL             | 36,004     | —                     | 42,865  |
| Columbia, SC            | 35,881     | 38,375                | 42,301  |
| Columbus, OH            | 27,353     | 30,163                | 30,208  |
| Kansas City, MO         | 24,179     | 25,851                | —       |
| Lexington, KY           | 21,132     | 21,555                | 21,731  |
| Los Angeles, CA         | 37,438     | —                     | 41,951  |
| Madison, WI             | 28,035     | 30,297                | 30,640  |
| Memphis, TN             | 24,588     | 25,502                | 25,267  |
| Minneapolis, MN         | 31,900     | 34,065                | 34,787  |
| Orlando, FL             | 28,449     | 30,438                | 30,158  |
| Philadelphia, PA        | 33246      | —                     | 35,350  |
| Phoenix, AZ             | 28,347     | 29,286                | 29,350  |
| Pittsburgh, PA          | 26,406     | 28,630                | 30,135  |
| Raleigh, NC             | 25,839     | 27,493                | 26,775  |
| San Francisco, CA       | 49,927     | 48,217                | 50,249  |
| Seattle, WA             | 38,469     | 41,950                | 43,486  |
| Sioux Falls, SD         | 20,160     | 21,171                | 21,577  |
| Stamford, CT            | 57,722     | 58,616                | 68,132  |
large size of the data set and more importantly with the fact that the data do not conform well with the implemented mixed effects model.

Next, we examine a few diagnostic plots to determine whether the model assumptions are satisfied. We start by examining the variance of the residuals. As the gap time increases, we expect a higher error variance as the previous price becomes less useful over time. The proposed autoregressive model and the S&P/Case-Shiller model fit this differently using a latent AR(1) time series and a random walk respectively. The mixed effects model, however, assumes a constant variance regardless of gap time. Fig. 4 examines which of these models provides better modeling of the observed variances. For each model, we plot the variance of the predictions by gap time. The expected variance using the respective models and the parameters estimated from the data is overlayed. Observe the differences in scales of the residuals. The autoregressive and mixed effects models are fit on the log price scale whereas the S&P/Case-Shiller model is fit on the price scale.

There are two features to note here. The first is that heteroscedasticity is clearly present. The variance of the residuals does in fact increase with gap time. The second feature is that while none of the methods perfectly model the heteroscedastic error, the autoregressive model is undoubtedly the best. This pattern holds across all of the cities in the data set.

For both the AR and mixed effects models, the random effects for ZIP codes are assumed to be normally distributed. As a diagnostic procedure we examined the residual ZIP code effects for Columbus, OH for normality for both models. The results are shown in Figure 5. Columbus, OH has a total of 103 ZIP codes, or random effects. We find the normality assumption appears to be satisfactorily satisfied for the mixed effects model but less well satisfied for the autoregressive model. Note, however, that each residual on the plot is estimated using very different sample sizes and the residuals have varying, non-zero covariances with each other. These facts interfere with the routine interpretation of the plots in the figure. In particular, the outlier points (in both plots) correspond to ZIP codes having ten or fewer sales. Across all metropolitan areas, the normality assumption seems to be well satisfied in some cases and not so well in others, but with no clear pattern we could discern as to type of analysis or size or geographic region of the area.

In Fig. 6, we plot four indices constructed from the AR model, the mixed effects model, the S&P/Case-Shiller model and the mean price index for Columbus, OH. The mean index is simply the average price at each quarter rescaled so that the first index value is 1. From the plot, we see that the autoregressive index generally stays between the S&P/Case-Shiller index and the mean index. The mean index treats all sales as single sales. That is, information about repeat sales is not included; in fact, no information about house prices is shared across quarters. The S&P/Case-Shiller index, on the other hand, only includes repeat sales houses. The autoregressive model, since it includes both single sales and repeat sales, is a mixture of the two perspectives. Essentially, the index constructed from the proposed model is a measure of the average house price giving more weight to those homes which have sold more than once.
Figure 4: Comparing Variance Estimates: Columbus, OH
Figure 5: Normality of Zip Code Effects: Columbus, OH

![Normal Q-Q Plot (AR Model)](image1)

![Normal Q-Q Plot (Mixed Effects Model)](image2)

Figure 6: House Price Indices for Columbus, OH

![Index Chart for Columbus, OH](image3)
6 The Case of Los Angeles, CA

While, the autoregressive model has a lower RMSE than the S&P/Case-Shiller model for Los Angeles, CA, the model does not seem to fit the data well. If we examine Fig. 7, a plot of the correlation against gap time, we immediately see two significant issues when what is expected (line) is compared with what the data indicate (dots). First, the value of $\phi$ is not as close to one as expected. Second, the rate of decay, $\phi^y(i,j,z)$, also does not follow the expected pattern. For the remainder of this section, we focus on Los Angeles, CA home sales and discuss these two issues.

We expect $\phi$ to be close to one. For Los Angeles, CA, this does not seem to be the case. In fact, according to the data, for short gap times, the correlation between sale pairs seems to be much lower than one. To explain this feature, we look at sale pairs with gap times between 1 and 5 quarters more closely. In Fig. 8, we create a histogram of the quarters where the second sales occurred when the gap time was short. We pair this histogram with a plot of the price index for Los Angeles, CA. Most of these sales occurred during the late 1980s and early 1990s. This corresponds to the same period when lenders were offering people mortgages where the monthly payment was greater than 33% of their monthly income [22]. The threshold of 33% is set to help ensure that people will be able to afford their mortgage. Those people with mortgages that exceed this percentage have a higher probability of defaulting on their payments. A number of banks including the Bank of California and Wells Fargo were highly exposed to these risky investments especially in the wake of the housing downturn during the early 1990s [2]. If a short gap time is an indication that a foreclosure took place, this would explain why these sale pairs are not highly correlated.
If we look at the period between January 1990 and December 1996 on Fig. 8, the housing index was decreasing. However, if we examine the RMSE of test set sales in this period only, we find that the autoregressive model still performs better than the S&P/Case-Shiller method. The RMSE values are $32,039 and $41,841 respectively. Therefore, the autoregressive model seems to perform better in periods of decline as well as in times of increase.

The second issue with Fig. 7 is that the AR(1) process does not decay at the same rate as the model predicts. In 1978 California voters, as a protest against rising property taxes, passed Proposition 13 which limited how fast property tax assessments could increase per year. Galles and Sexton (1998) argue that Proposition 13 encouraged people to retain homes especially if they have owned their home for a long time [12, p. 124]. It is possible that this feature of Fig. 7 is a long term effect of Proposition 13. On the other hand, it could be that California home owners tend to renovate their homes more frequently than others reducing the decay in prices over time. However, we have no way of verifying either of these explanations given our data.

7 Discussion

Two key tasks in analyzing house prices are predicting sale prices of individual homes and constructing price indices which measure general housing trends. The statistical model we have developed focuses mainly on the former task. Using extensive data from twenty metropolitan areas we have compared our predictive method to two other methods, including one involving the S&P/Case-Shiller Home Price Index. We find that on average the predictions using our method are more accurate in all but one of the metropolitan areas (See Table 5).

Data such as ours often do not contain reliable hedonic information on individual homes, if at all. Some have also incorporated ad hoc adjustments to take account of the gap time between the repeat sales of a home. In contrast our model involves an underlying AR(1) time series that automatically adjusts for the time gap between sales. It also uses the home’s ZIP code as an additional indicator of the its hedonic value. This indicator has some predictive value, although its value is quite weak by comparison with the price in a previous sale, if one has been recorded. Consequently, the estimator that corresponds to our statistical model can be viewed as a weighted average of estimates from single sale and repeat sale homes, with the repeat sales prices having a dramatically higher weight. As noted, the time series feature of the model guarantees that this weight for repeat sales prices slowly decreases in a natural fashion as the gap time between sales increases.

Our results do not provide definitive evidence as to the value of our index by comparison with other currently available indices as a general economic indicator. Indeed, such a determination should involve a study of the general economic uses of such indicators as well as an examination of their formulaic construction and their use for prediction of individual sale prices. We have not undertaken such a study, and so can offer only a few comments about the possible comparative values of our index.

As we have discussed, we feel it may be an advantage that our index involves all home
Figure 8: Examining the Housing Downturn

Los Angeles, CA

Index

Quarter

Bad Loans

Gap Times Between 1 and 5 Quarters

Frequency

Quarter
sales in the data (subject to the naturally occurring weighting described above), rather than only repeat sales. Repeat sales homes are only a small, selected fraction of all home sales. Studies have shown that repeat sales homes may have different characteristics than single sale homes. In particular, they are evidently older on average, and this could be expected to have an effect on their sale price. Since our measure brings all home sales into consideration, albeit in a gently weighted manner, and since it provides improved prediction on average, it may provide a preferable index.

Another asset of our model is that it remains easy to interpret at both the micro and macro levels, in spite of including several features inherent in the data. Future work seems desirable to understand anomalous cases such as those we have discussed in the Los Angeles, CA area. This work might improve this model to fit better for situations like those that held there. For example, it could involve the inclusion of economic indicators which may affect house prices such as interest rates and tax rates and measures of general economic status such as the unemployment rate.

A Data Summary
| City                | No. Sales | No. Houses | No. Houses Per Sale Count | No. ZIP Codes |
|---------------------|-----------|------------|---------------------------|---------------|
| Ann Arbor, MI       | 68,684    | 48,522     | 32,458 12,662 2,781 621   | 57            |
| Atlanta, GA         | 376,082   | 260,703    | 166,646 76,046 15,163 2836 | 184           |
| Chicago, IL         | 688,468   | 483,581    | 319,340 130,234 28,369 5,603 | 317           |
| Columbia, SC        | 7,034     | 4,321      | 2,303 1,470 431 117       | 12            |
| Columbus, OH        | 162,716   | 109,388    | 67,926 31,739 7,892 1,831 | 103           |
| Kansas City, MO     | 123,441   | 90,504     | 62,489 23,706 3,773 534   | 179           |
| Lexington, KY       | 38,534    | 26,630     | 16,891 7,901 1,555 282    | 31            |
| Los Angeles, CA     | 543,071   | 395,061    | 272,258 100,918 18,965 2,903 | 280           |
| Madison, WI         | 50,589    | 35,635     | 23,685 9,439 2,086 425    | 40            |
| Memphis, TN         | 55,370    | 37,352     | 23,033 11,319 2,412 587    | 64            |
| Minneapolis, MN     | 330,162   | 240,270    | 166,811 59,468 11,856 2,127 | 214           |
| Orlando, FL         | 104,853   | 72,976     | 45,966 22,759 3,706 543    | 96            |
| Philadelphia, PA    | 402,935   | 280,272    | 179,107 82,681 15,878 2,606 | 330           |
| Phoenix, AZ         | 180,745   | 129,993    | 87,249 35,910 5,855 968    | 130           |
| Pittsburgh, PA      | 104,544   | 73,871     | 48,618 20,768 3,749 718    | 257           |
| Raleigh, NC         | 100,180   | 68,306     | 42,545 20,632 4,306 818    | 82            |
| San Francisco, CA   | 73,598    | 59,416     | 46,959 10,895 1,413 149    | 70            |
| Seattle, WA         | 253,227   | 182,770    | 124,672 47,406 9,198 1,494 | 110           |
| Sioux Falls, SD     | 12,439    | 8,974      | 6,117 2,353 419 85         | 30            |
| Stamford, CT        | 14,602    | 11,128     | 8,200 2,502 357 62         | 23            |
| City                  | Autoregressive Model | S&P/Case-Shiller® Model |
|-----------------------|----------------------|------------------------|
|                       | Training  | Test   | No. Houses | Training Pairs | No. Houses | Training  | Test   | No. Houses | Training Pairs | No. Houses |
| Ann Arbor, MI         | 58,953    | 9,731  | 48,522     | 10,431         | 9,735      |
| Atlanta, GA           | 319,925   | 56,127 | 260,703    | 59,222         | 55,911     |
| Chicago, IL           | 589,289   | 99,179 | 483,581    | 105,708        | 99,069     |
| Columbia, SC          | 5,747     | 1,287  | 4,321      | 1,426          | 1,279      |
| Columbus, OH          | 136,989   | 25,727 | 109,262    | 27,601         | 25,458     |
| Kansas City, MO       | 107,209   | 16,232 | 90,977     | 16,705         | 16,092     |
| Lexington, KY         | 32,705    | 5,829  | 26,630     | 6,075          | 5,748      |
| Los Angeles, CA       | 470,721   | 72,350 | 395,061    | 75,600         | 72,338     |
| Madison, WI           | 43,349    | 7,240  | 35,635     | 7,714          | 7,221      |
| Memphis, TN           | 46,724    | 8,646  | 37,352     | 9,372          | 8,673      |
| Minneapolis, MN       | 286,476   | 43,686 | 242,790    | 46,206         | 43,764     |
| Orlando, FL           | 89,123    | 15,730 | 73,393     | 16,147         | 15,531     |
| Philadelphia, PA      | 343,354   | 59,581 | 283,773    | 63,082         | 60,068     |
| Phoenix, AZ           | 155,823   | 24,922 | 129,901    | 25,830         | 24,656     |
| Pittsburgh, PA        | 89,762    | 14,782 | 75,370     | 15,891         | 14,956     |
| Raleigh, NC           | 84,678    | 15,502 | 69,176     | 16,372         | 15,388     |
| San Francisco, CA     | 66,527    | 7,071  | 59,416     | 7,111          | 6,948      |
| Seattle, WA           | 218,741   | 34,486 | 184,255    | 35,971         | 34,304     |
| Sioux Falls, SD       | 10,755    | 1,684  | 8,971      | 1,781          | 1,677      |
| Stamford, CT          | 12,902    | 1,700  | 11,128     | 1,774          | 1,654      |
B Updating Equations

In this section, we provide the updating equations for estimating the parameters \( \theta = \{ \beta, \sigma^2_\varepsilon, \sigma^2_\tau, \phi \} \) in the autoregressive model (see Sec 3). Observe that the covariance matrix \( V \) is an \( N \times N \) matrix where \( N \) is the total sample size. Given the size of our data, it is simpler computationally to exploit the block diagonal structure of \( V \). Each block, denoted by \( V_{z,z} \), corresponds to observations in zip code \( z \). Computations are done on the zip code level and the updating equations provided below reflect that. For instance, \( y_z \) and \( T_z \) are the rows of the log price vector and transformation matrix for observations in zip code \( z \) respectively.

To start, an explicit expression for \( \beta \) can be computed:

\[
\hat{\beta} = \left( \sum_{z=1}^{Z} (T_zX_z)'V_{z,z}^{-1}T_zX_z \right)^{-1} \sum_{z=1}^{Z} (T_zX_z)'V_{z,z}^{-1}T_zy_z. \tag{16}
\]

For the remaining parameters, estimates must be computed numerically. As all of these are one-dimensional parameters, methods such as the Newton-Raphson algorithm are highly suitable. We first define \( w_z = y_z - X_z\beta \) for clarity. To update \( \sigma^2_\varepsilon \), compute the zero of:

\[
0 = - \sum_{z=1}^{Z} tr\left( V_{z,z}^{-1} diag(r_z) \right) + \sum_{z=1}^{Z} (T_zw_z)'V_{z,z}^{-1}diag(r_z)V_{z,z}^{-1}(T_zw_z) \tag{17}
\]

where \( tr(\cdot) \) is the trace of a matrix and \( diag(\cdot) \) is as defined in (4). Similarly, to update \( \sigma^2_\tau \), find the zero of:

\[
0 = \sum_{z=1}^{Z} tr\left( V_{z,z}^{-1}(T_z1_{n_z})(T_z1_{n_z})' \right) \nonumber + \sum_{z=1}^{Z} (T_zw_z)'V_{z,z}^{-1}(T_z1_{n_z})(T_z1_{n_z})'V_{z,z}^{-1}(T_zw_z). \tag{18}
\]

Finally, to update the autoregressive parameter \( \phi \), we must calculate the zero of the
function below. Note that \( n_z \) denotes the number of observations in zip code \( z \).

\[
0 = -\sum_{z=1}^{Z} tr \left\{ V_{z,z}^{-1} \left( \sigma_\tau^2 \left( \frac{\partial (T_z \mathbf{1}_{n_z})}{\partial \phi} \right)(T_z \mathbf{1}_{n_z})' + \sigma_\tau^2 (T_z \mathbf{1}_{n_z}) \left( \frac{\partial (T_z \mathbf{1}_{n_z})}{\partial \phi} \right)' \right) + \frac{2\phi\sigma_\tau^2}{(1 - \phi^2)^2} diag(r_z) + \frac{\sigma_\phi^2}{1 - \phi^2} \partial diag(r_z) \right\}
\]

\[
-\sum_{z=1}^{Z} \left( \frac{\partial T_z}{\partial \phi} \mathbf{w}_z \right)' V_{z,z}^{-1} (T_z \mathbf{w}_z) - \sum_{z=1}^{Z} (T_z \mathbf{w}_z)' V_{z,z}^{-1} \left( \frac{\partial T_z}{\partial \phi} \mathbf{w}_z \right)
\]

\[
+ \sum_{z=1}^{Z} \left[ (T_z \mathbf{w}_z)' V_{z,z}^{-1} \left[ \sigma_\tau^2 \left( \frac{\partial (T_z \mathbf{1}_{n_z})}{\partial \phi} \right)(T_z \mathbf{1}_{n_z})' + \sigma_\tau^2 (T_z \mathbf{1}_{n_z}) \left( \frac{\partial (T_z \mathbf{1}_{n_z})}{\partial \phi} \right)' \right] + \frac{2\phi\sigma_\tau^2}{(1 - \phi^2)^2} diag(r_z) + \frac{\sigma_\phi^2}{1 - \phi^2} \partial diag(r_z) \right] V_{z,z}^{-1} (T_z \mathbf{w}_z) \right].
\] (19)

After the estimates converge, the final step is to estimate the random effects. We use Henderson’s (1975) procedure to derive the Best Linear Unbiased Predictors (BLUP) for each zip code. Henderson’s method assumes that the parameters in the covariance matrix, \( V \) are known; however, we use the estimated values.

\[
\hat{\tau}_z = \left[ \frac{2\sigma_z^2}{\sigma_\tau^2} + \left( 1 - \phi^2 \right) \left( \frac{\partial (T_z \mathbf{1}_z)}{\partial \phi} \right)' \mathbf{1}_z \right] \left( \mathbf{1}_z \right) \right]^{-1} \times
\]

\[
\left( \frac{\partial (T_z \mathbf{1}_z)}{\partial \phi} \mathbf{1}_z \right)' \mathbf{1}_z \hat{\tau}_z \mathbf{1}_z \right).
\] (20)

where \( \mathbf{1}_z \) is the inverse of the estimated diagonal matrix \( diag(r) \).

References

[1] Bailey, M.J., Muth, R.F., Nourse, H.O. (1963). A regression method for real estate price index construction. Journal of the American Statistical Association. 58 933-942.

[2] Bates, J. (Dec. 30, 1989). Survey cites four california banks with possibly risky realty loans. Los Angeles Times. 1.

[3] Bickel, P.J. , Doksum, K. A. (2001). Mathematical Statistics–Basic Ideas and Selected Topics, Vol. I, 2nd ed. Prentice Hall, New Jersey.

[4] Brown, L.D. (1986). Fundamentals of Statistical Exponential Families–with Applications in Statistical Decision Theory. Institute of Mathematical Statistics, Hayward, California. Vol. 9.

[5] Calhoun, C. (1996). OFHEO house price indices: HPI technical description. http://www.ofheo.gov
[6] Case, B., Pollakowski, H.O., Wachter, S.M. (1991). On choosing among house price index methodologies. *AREUEA Journal*. 19 286-307.

[7] Case, B., Quigley, J.M (1991). The dynamics of real estate prices. *The Review of Economics and Statistics*. 73 50-58.

[8] Case, K.E., Shiller, R.J. (1987). Prices of single-family homes since 1970: new indexes for four cities. *New England Economic Review*. Sept./Oct. 45-56.

[9] Case, K.E., Shiller, R.J. (1989). The efficiency of the market for single family homes. *The American Economic Review*. 79 125-137.

[10] Englund, P., Quigley, J.M., Redfearn, C.L. (1999). The choice of methodology for computing housing price indexes: comparisons of temporal aggregation and sample definition. *Journal of Real Estate Finance and Economics* 19 91-112.

[11] Fisher, J., Gatzlaff, D., Geltner, D., Haurin, D. (2003). Controlling for the impact of variable liquidity in commercial real estate price indices. *Real Estate Economics*. 31 269-303.

[12] Galles, G.M., Sexton, R.L. (1998). A tale of two tax jurisdictions: the surprising effects of California's Proposition 13 and Massachusetts' Proposition 2 1/2. *American Journal of Economics and Sociology*. 57 123-133.

[13] Gatzlaff, D.H., Haurin, D.R. (1997). Sample selection bias and repeat-sales index estimates. *Journal of Real Estate Finance and Economics*. 14 33-50.

[14] Goetzmann, W.N. (1992). The accuracy of real estate indices: repeat sales estimators. *Journal of Real Estate Finance and Economics*. 5 5-53.

[15] Goetzmann, W.N., Peng, L. (2002). The bias of the RSR estimator and the accuracy of some alternatives. *Real Estate Economics*. 30 13-39.

[16] Goetzmann, W.N., Spiegel, M. (1995). Non-temporal components of residential real estate appreciation. *The Review of Economics and Statistics*. 77 199-206.

[17] Henderson, C.R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*. 31 423-447.

[18] Meese, R.A., Wallace, N.E. (1997). The construction of residential housing price indices: a comparison of repeat-sales, hedonic-regression, and hybrid approaches. *Journal of Real Estate Finance and Economics*. 14 51-73.

[19] Sargan, J.D. (1964). Wages and prices in the United Kingdom: a study in econometric methodology. *Econometric Analysis for National Economic Planning*, eds. Hart, P.E., Mills, G., Whitaker, J.K. Butterworths, London. Reprinted in *Econometrics and Quantitative Analysis*, eds. Wallis, K.F., Hendry, D.F. 275-314. Basil Blackwell, Oxford.
[20] Shen, H., Brown, L.D., Zhi, H. (2006). Efficient estimation of log-normal means with application to pharmacokinetic data. *Statistics in Medicine.* 25 3023-3038.

[21] Shiller, R. (1991). Arithmetic repeat sales price estimators. *Journal of Housing Economics.* 1 110-126.

[22] Sing, B., Furlong, T. (Mar. 29, 1989). Defaults feared if payments keep ballooning fast-rising interest rates causing concern for adjustable mortages. *Los Angeles Times.* 5.

[23] S&P/Case-Shiller® Home Price Indices–Index Methodology (Nov. 2009). http://www.standardandpoors.com