STATISTICAL METHODS FOR MODELING HOUSE PRICES AND INDICES

Chaitra Haikady Nagaraja

A DISSERTATION

in

Statistics

For the Graduate Group in Managerial Science and Applied Economics

Presented to the Faculties of the University of Pennsylvania in Partial

Fulfillment of the Requirements for the Degree of Doctor of Philosophy

2008

Lawrence D. Brown

Supervisor of Dissertation

Graduate Group Chairperson

Dedicated to my grandmother, Mrs. Mandakini Madhyasta, who never had the chance to study mathematics herself.

ACKNOWLEDGEMENTS

Prof. Lawrence D. Brown for being the best mentor one could ever wish for.

Kartikeya S. Ghia for being my best friend.

My parents Dr. H.N. Nagaraja and Jyothi Nagaraja and sister Smitha H. Nagaraja for believing in me even when I did not.

Prof. Dylan Small, Prof. Susan Wachter, Prof. Edward I. George, and Prof. J. Michael Steele for their advice and support throughout my time at Wharton.

ABSTRACT

STATISTICAL METHODS FOR MODELING HOUSE PRICES AND INDICES

Chaitra Haikady Nagaraja

Lawrence D. Brown

Repeat sales techniques are a common approach for modeling house prices. This methodology presumes the previous sale price acts as a proxy for hedonic variables, such as size and number of bedrooms. Capturing the spirit of the repeat sales setup, the proposed model includes the previous price as a predictor of current price. However, the model also includes an adjustment so that the more time which has elapsed between sales, the less useful the previous price becomes. To incorporate this property into the model framework, a two-part, nonlinear model is proposed which consists of a general price index and an autoregressive component (AR). The latter element can be thought of as the result of a latent AR(1) process for each house which is observed only in time periods when sales occur. In the fitting process, all sales contribute to estimating the time effect but only repeat sales factor in the autoregressive coefficient estimate. The resulting index, constructed from the time effects, is therefore more representative of the housing market compared to existing repeat sales models which ignore single sales. Moreover, the proposed model outperforms benchmark models including the S&P/Case-Shiller® model in terms of predictive power when applied to single-family home sales from July 1985 through September 2004 for twenty U.S. metropolitan areas. Finally, an extension to this model is proposed to incorporate local effects. Here, zip code is introduced to the model as a random effect. Predictive performance is further improved with this addition.

Contents

1 Introduction **1**

2 Data **5**

3 The Autoregressive Model **9**

3.1 Complete Data Case 9

3.2 Observed Data Case 11

3.3 Model Fitting 16

3.3.1 Maximum Likelihood Estimation 16

3.3.2 Convergence of the Coordinate Ascent Algorithm 20

3.4 Final Details 22

3.4.1 Asymptotic Variance of the Parameter Estimates 22

3.4.2 Converting Back to the Price Scale 23

3.4.3 Index Construction 23

4 Existing Methods **25**

4.1 Hedonic Models 25

4.2 Repeat Sales Models 26

4.2.1 The BMN Model 27

4.2.2 The Case-Shiller Method 28

4.2.3 The S&P/Case-Shiller® Method 30

4.2.4 Discussion of Repeat Sales Methods 33

4.3 Hybrid Models 38

4.4 Spatial Models 40

4.5 Commercial Indices 42

5 Results for the Global Models **48**

5.1 Parameter Estimates 49

5.2 The Asymptotic Variance Assumption 51

5.3 The AR(1) Assumption 54

5.4 Model Validation 54

5.4.1 Competing Models 55

5.4.2 Prediction 59

5.4.3 The Regression Assumptions 60

5.4.4 Comparing Indices 68

5.5 A Closer Look at Los Angeles, CA 69

5.6 Comparing Repeat Sales Indices 71

6 The Local Autoregressive Model 75

6.1 Time Series Models With Covariates 76

6.2 Introducing the Random Effects 80

6.3 Model Fitting 82

6.3.1 Convergence of the Fitting Algorithm 86

6.4 Additional Computations 88

7 Results for Local Models **89**

7.1 Parameter Estimates for the Local Model 89

7.2 Validation of Local Models 92

7.2.1 Mixed Effects Model 92

7.2.2 Results 93

7.2.3 Residual Analysis 95

8 Future Work **100**

A Data Summary **102**

B Derivations for the Global Autoregressive Model **107**

B.1 MLE Derivations 108

B.2 Computing the Observed Information matrix 111

C Computational Methods **117**

C.1 Residual Regression 117

C.2 Conjugate Gradient Method 119

D Additional Plots for Global Model **121**

E Derivations for the Local Autoregressive Model **130**

E.1 Updating Formulas 131

E.2 Deriving the BLUP 137

E.2.1 The Local Autoregressive Model 137

E.2.2 The Local Mixed Effects Model 140

F Additional Plots for Local Model **142**

List of Tables

2.1 Metropolitan Areas in Data 5

2.2 Summary of Sample Cities 6

2.3 Sale Counts for Sample Cities 6

4.1 Removing Single Sales 33

5.1 Estimated Values of ϕ and σ_ε^2 50

5.2 Comparing Simulated and Expected Results 52

5.3 Test Set RMSE for Global Models (in dollars) 60

5.4 Test Set RMSE for Global Models (in dollars) 74

6.1 Sample Data 81

7.1 Estimated Values of Parameters for Local Model 90

7.2 Test Set RMSE for Local Models (in dollars) 94

A.1 House and Sale Counts 104

A.2 Zip Code and Census Tract Counts 105

A.3 Training and Test Set Sizes 106

List of Figures

2.1 Sales Counts Over Time 7

2.2 Examining Seasonality 8

2.3 Median Prices for Sample Cities 8

3.1 The Log Likelihood Function at Each Iteration 21

3.2 Maximizing ϕ 22

4.1 Comparing US Commercial Indices 46

5.1 Indices from the Autoregressive Model for All Metropolitan Areas 50

5.2 Simulation Results 53

5.3 ϕ vs Gap Time (Columbus, OH) 55

5.4 Residual Plots for Columbus, OH (log scale) 63

5.5 Normality of Residuals for Autoregressive Model (Columbus, OH) 64

5.6 Normality of Residuals for S&P/Case-Shiller® (Columbus, OH) 65

5.7 Normality of Residuals for Fixed and Mixed Effects Models (Columbus, OH) 66

5.8 Variance of Residuals for Competing Models (Columbus, OH) 67

5.9 Comparing Indices (Columbus, OH) 69

5.10 Actual vs. Computed S&P/Case-Shiller® Index 70

5.11 Problems with Assumptions 70

5.12 Examining the Housing Downturn 72

5.13 Prices Versus Log Prices 74

6.1 Indices for All Metropolitan Areas 76

6.2 Plot of the Local Log Likelihood Function 87

6.3 Examining the Convergence of ϕ 87

7.1 ϕ vs Gap Time for Local Model (Columbus, OH) 91

7.2 Normality of Random Effects: Minneapolis, MN 92

7.3 House Price Indices for Columbus, OH 95

7.4 Local Model Residual Plots for Columbus, OH (log scale) 97

7.5 Normality of Residuals for Local Autoregressive Model (Columbus, OH) . . 98

7.6 Normality of Residuals for Local Mixed Effects Model (Columbus,OH) . . . 99

7.7 Variance of Residuals for Competing Local Models (Columbus, OH) 99

D.1 AR(1) Assumption Check: Ann Arbor, MI-Kansas City, MO 122

D.2 AR(1) Assumption Check: Lexington, KY-Orlando, FL 123

D.3 AR(1) Assumption Check: Philadelphia, PA-Seattle, WA 124

D.4 AR(1) Assumption Check: Sioux Falls, SD-Stamford, CT 125

D.5 Global Indices: Ann Arbor, MI-Kansas City, MO 126

D.6 Global Indices: Lexington, KY-Orlando, FL 127

D.7 Global Indices: Philadelphia, PA-Seattle, WA 128

D.8 Global Indices: Sioux Falls, SD-Stamford, CT 129

F.1 Local AR(1) Assumption Check: Ann Arbor, MI-Kansas City, MO 143

F.2 Local AR(1) Assumption Check: Lexington, KY-Orlando, FL 144

F.3 Local AR(1) Assumption Check: Philadelphia, PA-Seattle, WA 145

F.4 Local AR(1) Assumption Check: Sioux Falls, SD-Stamford, CT 146

F.5 Local Indices: Ann Arbor, MI-Kansas City, MO 147

F.6 Local Indices: Lexington, KY-Orlando, FL 148

F.7 Local Indices: Philadelphia, PA-Seattle, WA 149

F.8 Local Indices: Sioux Falls, SD-Stamford, CT 150

Chapter 1

Introduction

Analyzing housing markets is a challenging task; each house is unique involving a collection of many attributes such as size, location, and amenities. The value of these characteristics is difficult to determine, especially in the presence of a lack of data availability. Moreover, not every house is sold at each time period which adds further complexity. As a result, the visible market is only a small subset of the entire population of homes. From this sample, models are developed to both predict individual house prices and construct a price index. For the latter, the goal is to aggregate information across all home sales to provide an idea of how prices change over time. These changes can often be applied to help assess the expected change in a home's price and for other economic purposes.

A common approach to modeling house prices is to use homes that sell multiple times to track overall market trends. The repeat sales method was first proposed by Bailey, Muth, and Nourse (BMN) to construct a house price index using price differences between two successive sales of a home. In this setting, the previous sale price is assumed to be a surrogate for house characteristics provided that the home is unchanged between sales.

Existing repeat sales methods utilize only repeat sales homes; all single sales are ignored. As a result, repeat sales indices are often criticized as being unrepresentative of the housing market.

We utilize the idea that repeat sales homes contain additional information about the market in the autoregressive model introduced in Chapter 3. We compute a separate model for each metropolitan area in our data. Other units of geography are feasible and we later consider this issue.

Specifically, the log price $y_{i,j}$ of the jth sale of the ith house is modeled as:

$$\begin{aligned} y_{i,1} &= \beta_{t(i,1)} + \varepsilon_{i,1} && \text{for initial sales } (j = 1) \\ y_{i,j} &= \beta_{t(i,j)} + \phi^{\gamma(i,j)}(y_{i,j-1} - \beta_{t(i,j-1)}) + \varepsilon_{i,j} && \text{for subsequent sales } (j > 1) \end{aligned} \tag{1.1}$$

where $t(i,j)$ is the time of the sale given in quarter units and $\gamma(i,j)$ is the gap time between sales. We define $y_{i,j} - \beta_{t(i,j)}$ to be the quarter-adjusted log sale price. The random variation for the initial sale has distribution, $\varepsilon_{i,1} \overset{iid}{\sim} \mathcal{N}\left(0, \frac{\sigma_\varepsilon^2}{1-\phi^2}\right)$ and for subsequent sales, $\varepsilon_{i,j} \overset{iid}{\sim} \mathcal{N}\left(0, \frac{\sigma_\varepsilon^2(1-\phi^{2\gamma(i,j)})}{1-\phi^2}\right)$ where $\mathcal{N}\left(\mu, \sigma^2\right)$ denotes a normal distribution with mean μ and variance σ^2 and "iid" stands for independent and identically distributed. In a sense to be later explained, this structure produces a stationary quarter-adjusted log price series.

A key feature of this model is that it handles houses differently based on the gap time between sales. We expect the sale price of a home to be less informative if the sale occurred a long time ago. This belief has a two-fold impact on the model. First, the model forces prices for pairs of sales with long gap times to be less correlated. This is because the autoregressive coefficient applied to a pair of sales depends on the gap time $\left(\phi^{\gamma(i,j)}\right)$. Second, the variance of $\varepsilon_{i,j}$ increases with gap time. Therefore, information contained in the previous sale price becomes less valuable as time passes.

Unlike existing repeat sales methods, all sales are used to construct the index and thus, we believe, captures trends in the overall housing market better. As we will show, the log price index can be thought of as a weighted average of single and repeat sales where the latter are assigned higher weights.

The autoregressive model is fitted using maximum likelihood techniques for single-family home sales from July 1985 through September 2004 in twenty US metropolitan areas. For comparison, three alternative models are fitted as well: a fixed effects model, a mixed effects model, and a model based on the S&P/Case-Shiller® method. When comparing predictive performance on test samples of individual sales, the autoregressive model performs best for seventeen out of twenty cities.

In Chapter 6, we investigate models that incorporate spatial information. To this end, we divide each metropolitan area into regions by zip code. The zip codes are modeled as random effects and are added to the global autoregressive model. The resulting model can be described as follows. Let $y_{i,j,z}$ be the jth log price of the ith house in zip code z. As before, let $\beta_1, \ldots, \beta_T$ denote the fixed, log price indices. The extra term is the zip code, τ_z where $\tau_z \overset{iid}{\sim} \mathcal{N}\left(0, \sigma_\tau^2\right)$. Introducing the random effect term requires the addition of μ, which can be interpreted as the overall mean log price if we require $\sum_{t=1}^{T} n_t\beta_t = 0$ where n_t is the number of sales at time t. Then, the model is:

$$\begin{aligned} y_{i,1,z} &= \mu + \beta_{t(i,1,z)} + \tau_z + \varepsilon_{i,1,z} && j = 1 \\ y_{i,j,z} &= \mu + \beta_{t(i,j,z)} + \tau_z + \phi^{\gamma(i,j,z)}\left(y_{i,j-1,z} - \mu - \beta_{t(i,j-1,z)} - \tau_z\right) + \varepsilon_{i,j,z} && j > 1 \end{aligned} \tag{1.2}$$

where $\varepsilon_{i,1,z} \overset{iid}{\sim} \mathcal{N}\left(0, \frac{\sigma_\varepsilon^2}{1-\phi^2}\right)$, $\varepsilon_{i,j,z} \overset{iid}{\sim} \mathcal{N}\left(0, \frac{\sigma_\varepsilon^2\left(1-\phi^{2\gamma(i,j,z)}\right)}{1-\phi^2}\right)$ where $j > 1$.

We use maximum likelihood estimation again to fit the local model to each of the twenty cities. As a benchmark model, we fit a mixed effects model which also models zip code as a random effect but omits the autoregressive factor. The local autoregressive model has even better predictions than both the benchmark mixed effects model and the global models examined earlier for most cities.

We now provide a brief overview of the subsequent chapters. In Chapter 2, we describe our data and present some summary statistics for the twenty US metropolitan areas. The global autoregressive model is introduced in Chapter 3. We also review basic results for autoregressive time series and outline the coordinate ascent algorithm used to fit the model. A review of existing housing literature is provided in Chapter 4. Four types of models are discussed: hedonic, repeat sales, hybrid, and spatial models; however, we focus on repeat sales models. We end the chapter with a summary of commercial indices in the US and the UK. Results for the global autoregressive model are analyzed in Chapter 5. The case of Los Angeles, CA, for which the model performs relatively poorly, is examined. Local models are investigated in Chapter 6 and results are provided in Chapter 7. Finally, future work is outlined Chapter 8.

Chapter 2

Data

The data are comprised of single family home sales from the twenty US metropolitan areas listed in Table 2.1. These sales occurred between July 1985 and September 2004 and were for homes which qualified for conventional mortgages. For each observation, the following information is available: address, month and year of sale, price, ZIP code, ZIP+4, and census tract. To ensure adequate data per time period, the sample period is divided into three month intervals for a total of 77 periods, or quarters. In this chapter, we provide a brief overview for five cities: Stamford, CT, Ann Arbor, MI, Pittsburgh, PA, Los Angeles, CA, and Chicago, IL. Complete tables are given in Appendix A.

Table 2.1: Metropolitan Areas in Data

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 2.2: 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

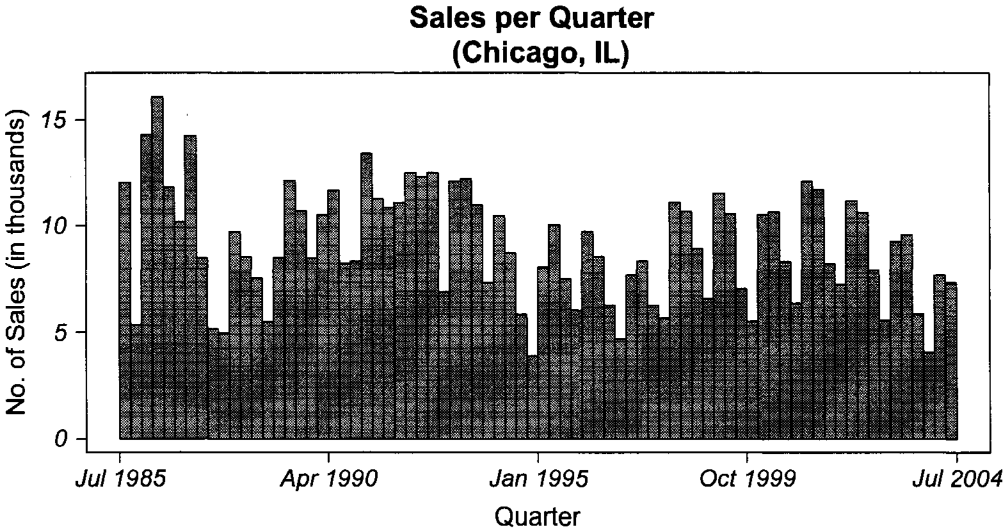
Table 2.3: Sale Counts for Sample Cities

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

Table 2.2 provides the number of sales and unique houses sold in the sample period. As houses sell multiple times (repeat sales), the total number of sales is always greater than the number of houses. Perhaps more illuminating is Table 2.3 which breaks down houses by number of sales. As expected, when the number of sales increases, the number of houses drops off rapidly. Note that there are a significant number of homes which sell more than twice. Since the sample period is long (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 2.3 shows this clearly. Except for Columbus, OH, this pattern holds for all cities.

To obtain stable price index estimates, there must be a substantial number of sales per quarter. In Fig. 2.1 graphs sales (in thousands) per quarter for Chicago, IL. Each bar on the graph represents one quarter. A feature of this plot, apart from the large number of sales each quarter, is seasonal patterns in the data. For instance, more homes are sold during warmer months especially when school is not in session. This is clearer in Fig. 2.2 which plots the mean and standard deviation of sales per month for Chicago, IL.

Figure 2.1: Sales Counts Over Time



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.

Finally, we examine the sale prices in the sample period. Fig. 2.3 is a plot of the median price (in thousands of dollars) for each quarter. In general, prices tend to increase over time; however there is a general decrease in prices in Los Angeles, CA and Stamford, CT during the late 1980's and early 1990's. There is also a difference in the level of prices among cities: Stamford, CT is the most expensive whereas Pittsburgh, PA is the least. In Chapter 6, we add local information, in the form of zip codes, to the autoregressive model. Thus, zip code and census tract breakdowns for each city can be found in Table A.2.

Figure 2.2: Examining Seasonality

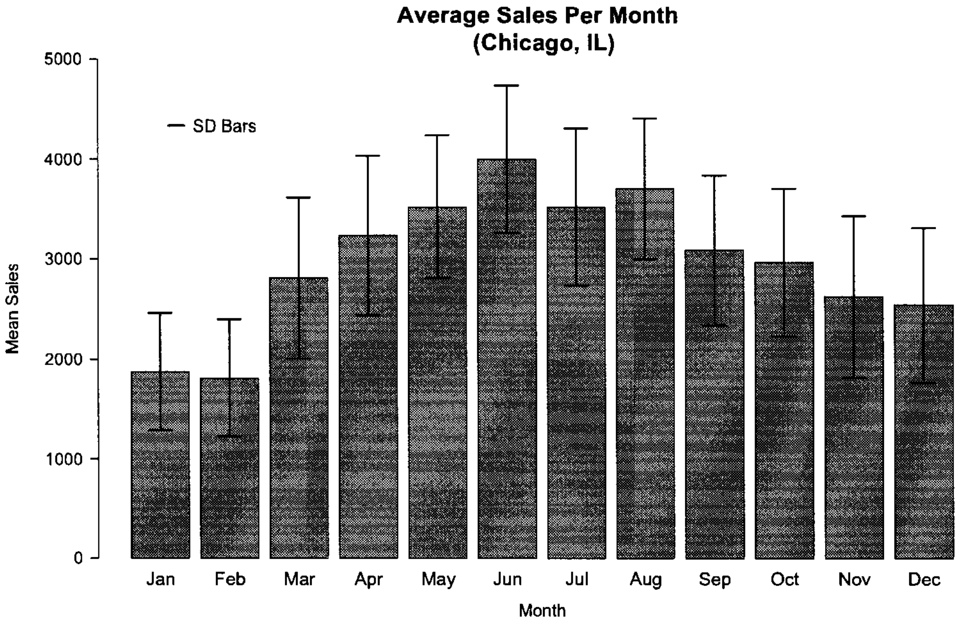
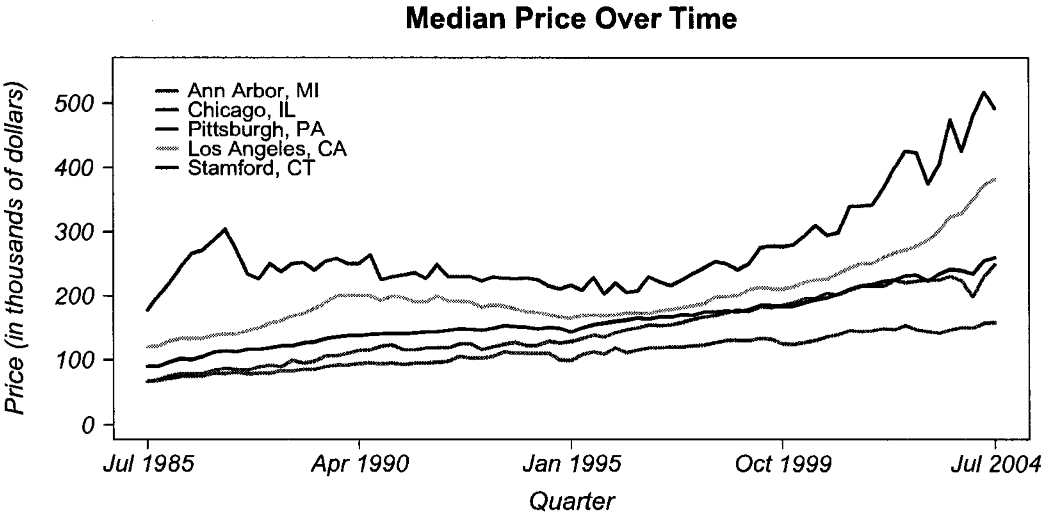


Figure 2.3: Median Prices for Sample Cities



Chapter 3

The Autoregressive Model

Repeat sales analysis assumes that the previous sale price of a house contains all relevant information for modeling; thus, any hedonic information would be redundant. The autoregressive model we propose in this chapter utilizes the essence behind repeat sales while adding two modifications: (a) we account for gap time between sales by incorporating an autoregressive component directly into the model and (b) we use all of the available data. Repeat sales homes are used to compute the autoregressive component but all sales are used in calculating the price index component. These two features combine to create a more powerful house price model.

3.1 Complete Data Case

Before we introduce the autoregressive model, we start with a simpler setup. Suppose we have a series $w_1, w_2, w_3, \ldots$ which follows a stationary autoregressive process of order 1 (denoted as AR(1)). This series can be described as follows:

$$\begin{aligned} w_1 &= \frac{1}{\sqrt{1-\phi^2}}\varepsilon_1 \\ w_j &= \phi w_{i-1} + \varepsilon_j \end{aligned} \tag{3.1}$$

where ϕ is the autoregressive parameter ($|\phi| < 1$). In addition, $\varepsilon_j \overset{iid}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2)$ are independent and identically distributed (iid) random variations where $\mathcal{N}\left(\mu, \sigma^2\right)$ denotes the normal distribution with mean μ and variance σ^2.

Under this scenario, there is a starting point for the series. Therefore, the series extends only forward in time, not infinitely in both directions. To accommodate this feature while preserving stationarity, we have multiplied the error term by $\frac{1}{\sqrt{(1-\phi^2)}}$. Weak stationarity (hereby referred to as stationarity) is defined as [42, p. 24]:

Definition 1. The series $w_1, w_2, w_3, \ldots$ is *weakly stationary* as long as the following conditions are satisfied:

(1) $E[w_t] = \mu$ for all t where $E[\cdot]$ is the expectation function and $\mu < \infty$.

(2) The covariance between two observations w_t and w_{t+h}, denoted as $Cov(t, t+h)$, is a function only of the gap time h where $h \geq 0$.

Next, we define the autocovariance and autocorrelation functions for a stationary series:

Definition 2. For a stationary series $w_1, w_2, w_3, \ldots$, the *autocovariance* between two observations w_t and w_{t+h}, denoted by $Cov(t, t+h)$, is defined as

$$\begin{aligned} Cov(t, t+h) &= Cov(h) \\ &= E[(w_t - \mu)(w_{t+h} - \mu)] \end{aligned}$$

where $E[\cdot]$ is the expectation function, μ is the expected value of the series, and h is the gap length.

Definition 3. Given a stationary series $w_1, w_2, w_3, \ldots$, the *autocorrelation* between two observations w_t and w_{t+h}, denoted by $\rho(h)$, is given by:

$$\rho(h) = \frac{Cov(h)}{Cov(0)}$$

where $Cov(\cdot)$ is the autocovariance function and h is the gap length.

Using these definitions, we now describe the model in (3.1) completely in Proposition 1; a proof is provided in the next section.

Proposition 1. *Define a time series process as follows:*

$$\begin{aligned} w_1 &= \frac{\varepsilon_1}{\sqrt{1-\phi^2}} \\ w_t &= \phi w_{t-1} + \varepsilon_t \end{aligned}$$

where $\varepsilon_t \overset{iid}{\sim} \mathcal{N}\left(0, \sigma_\varepsilon^2\right)$*, and* $|\phi| < 1$*. Then this series is stationary with correlation function* $\rho(h) = \phi^h$*,* $\forall\, h \in \{0\} \cup \mathbb{Z}^+$ *where* h *is the gap time.*

3.2 Observed Data Case

The setting described in Sec. 3.1 is applicable if a house is sold at every period; however, this is unrealistic. Instead, we presume there is an underlying price series which is observed only when the house is sold. In addition, we assume the sale price is a correct measure of a house's value. That is, there is no measurement error.

We model this scenario as follows. Let $y_{i,j}$ be the log price of the jth sale of the ith house. The parameter $\beta_{t(i,j)}$ is the effect of time $t(i,j)$ where $t(i,j)$ denotes the time period when the jth sale of the ith house occurs. Assume there are $1,\ldots,T$ discrete time periods where house sales occur. We define $\gamma(i,j)$ to be $t(i,j)-t(i,j-1)$; that is, the gap time between sales. The model, then, is:

$$\begin{aligned} y_{i,1}-\beta_{t(i,1)} &= \varepsilon_{i,1} & j&=1 \\ y_{i,j}-\beta_{t(i,j)} &= \phi^{\gamma(i,j)}\left(y_{i,j-1}-\beta_{t(i,j-1)}\right)+\varepsilon_{i,j} & j&>1 \end{aligned} \tag{3.2}$$

where ϕ is the autoregressive parameter and $|\phi|<1$. If we define adjusted log prices as $w_{i,j}=y_{i,j}-\beta_{t(i,j)}$, we can interpret ϕ as the "correlation" between consecutive adjusted log prices. Finally, the random variations, $\varepsilon_{i,j}$, have the following distributions: $\varepsilon_{i,1} \overset{iid}{\sim} \mathcal{N}\left(0,\frac{\sigma_\varepsilon^2}{1-\phi^2}\right)$ for the first sale and $\varepsilon_{i,j} \overset{iid}{\sim} \mathcal{N}\left(0,\frac{\sigma_\varepsilon^2(1-\phi^{2\gamma(j)})}{1-\phi^2}\right)$ for $j>1$ (or subsequent sales). Assume all $\varepsilon_{i,j}$ are independent as well. Note that "first sale" includes both sales of new homes and the first sale of older homes *in the sample period.*

We prove shortly that the series described in (3.2) is stationary. First, we reexpress the definition of stationarity for this type of series.

Definition 4. Let $w_1,w_2,\ldots,w_j$ be an intermittently observed time series process where $t(j)$ is the time of the jth observation. Let $\gamma(j)=t(j)-t(j-1)$ be the gap time between two consecutive observations. This series is *stationary* if the following conditions are satisfied:

(1) $E[w_j]=\mu$ for all j where $E[\cdot]$ is the expectation function and $\mu<\infty$.

(2) The covariance between two observations w_j and w_{j+k}, denoted as $Cov(t(j),t(j+k))$, is a function only of the gap time $h=\sum_{i=j+1}^{j+k}\gamma(i)$.

In Proposition 2 we prove that the adjusted log price series $w_{i,1},w_{i,2},\ldots$ is stationary for the model in (3.2) using Definition 4. Without loss of generality, we show stationarity for a single house.

Proposition 2. *Define a time series process as follows:*

$$\begin{aligned} w_1 &= \varepsilon_1 \\ w_j &= \phi^{\gamma(j)} w_{j-1} + \varepsilon_j \end{aligned}$$

where $\varepsilon_1 \overset{iid}{\sim} \mathcal{N}\left(0, \frac{\sigma_\varepsilon^2}{1-\phi^2}\right)$, $\varepsilon_j \overset{iid}{\sim} \mathcal{N}\left(0, \frac{\sigma_\varepsilon^2(1-\phi^{2\gamma(j)})}{1-\phi^2}\right)$ *where* $j \neq 1$ *and all* $\varepsilon_{i,j}$ *are independent. Let* $t(j)$ *be the time period of the* j*th observation and* $\gamma(j) = t(j) - t(j-1)$ *if* $j > 1$*. Finally, assume* $|\phi| < 1$ *and* $\phi \neq 0$*. This series is stationary with correlation function* $\rho(h) = \phi^h$, $\forall\, h \in \{0\} \cup \mathbb{Z}^+$ *where* h *is the gap time.*

Proof. We can write this process as follows:

$$\begin{aligned} w_1 &= \varepsilon_1 \\ w_2 &= \phi^{\gamma(2)}\varepsilon_1 + \varepsilon_2 \\ w_3 &= \phi^{\gamma(3)+\gamma(2)}\varepsilon_1 + \phi^{\gamma(3)}\varepsilon_2 + \varepsilon_3 \\ \vdots &= \\ w_j &= \left(\sum_{k=1}^{j-1} \phi^{\sum_{i=k+1}^{j}\gamma(i)}\varepsilon_k\right) + \varepsilon_j \end{aligned}$$

Using the above expression for w_j, it is clear that $E[w_j] = 0\ \forall\, j \in \mathbb{Z}^+$. Next, we derive the covariance function, $Cov(j,l)$. Without loss of generality assume $j \leq l$. The gap time h, then, is $h = t(l) - t(j) = \gamma(j+1) + \cdots + \gamma(l)$.

$$
\begin{aligned}
Cov(j,l) &= E\left[\left(\sum_{k=1}^{j-1}\phi^{\sum_{i=k+1}^{j}\gamma(i)}\varepsilon_k+\varepsilon_j\right)\left(\sum_{k=1}^{l-1}\phi^{\sum_{i=k+1}^{l}\gamma(i)}\varepsilon_k+\varepsilon_l\right)\right]\\
&= \sum_{k=1}^{j-1}\phi^{\sum_{i=k+1}^{j}\gamma(i)+\sum_{i=k+1}^{l}\gamma(i)}E[\varepsilon_k^2]+\phi^{\sum_{i=j+1}^{l}\gamma(i)}E[\varepsilon_j^2]\\
&= \frac{\sigma_\varepsilon^2}{1-\phi^2}\left(\phi^{\sum_{i=2}^{j}\gamma(i)+\sum_{i=2}^{l}\gamma(i)}+\sum_{k=2}^{j-1}\phi^{\sum_{i=k+1}^{j}\gamma(i)+\sum_{i=k+1}^{l}\gamma(i)}\left(1-\phi^{2\gamma(k)}\right)\right.\\
&\qquad \left.+\,\phi^{\sum_{i=j+1}^{l}\gamma(i)}\left(1-\phi^{2\gamma(j)}\right)\right)\\
&= \frac{\sigma_\varepsilon^2}{1-\phi^2}\left(\phi^{\sum_{i=2}^{j}\gamma(i)+\sum_{i=2}^{j}\gamma(i)+\sum_{i=j+1}^{l}\gamma(i)}\right.\\
&\qquad +\sum_{k=2}^{j-1}\phi^{\sum_{i=k+1}^{j}\gamma(i)+\sum_{i=k+1}^{j}\gamma(i)+\gamma_{i=j+1}^{l}\gamma(i)}\left(1-\phi^{2\gamma(k)}\right)\\
&\qquad \left.+\phi^{\sum_{i=j+1}^{l}\gamma(i)}\left(1-\phi^{2\gamma(j)}\right)\right)\\
&= \frac{\sigma_\varepsilon^2}{1-\phi^2}\left(\phi^{2\sum_{i=2}^{j}\gamma(i)}\phi^h+\sum_{k=2}^{j-1}\phi^{2\sum_{i=k+1}^{j}\gamma(i)}\phi^h\left(1-\phi^{2\gamma(k)}\right)\right.\\
&\qquad \left.+\,\phi^h\left(1-\phi^{2\gamma(j)}\right)\right)\\
&= \frac{\sigma_\varepsilon^2\phi^h}{1-\phi^2}\left(\phi^{2\sum_{i=2}^{j}\gamma(i)}+\sum_{k=2}^{j-1}\phi^{2\sum_{i=k+1}^{j}\gamma(i)}\left(1-\phi^{2\gamma(k)}\right)+\left(1-\phi^{2\gamma(j)}\right)\right)\\
&= \frac{\sigma_\varepsilon^2\phi^h}{1-\phi^2}\left(\phi^{2\sum_{i=2}^{j}\gamma(i)}+\sum_{k=2}^{j-1}\phi^{2\sum_{i=k+1}^{j}\gamma(i)}-\sum_{k=2}^{j-1}\phi^{2\sum_{i=k}^{j}\gamma(i)}+\left(1-\phi^{2\gamma(j)}\right)\right)\\
&= \frac{\sigma_\varepsilon^2\phi^h}{1-\phi^2}\left(\phi^{2\sum_{i=2}^{j}\gamma(i)}-\phi^{2\sum_{i=2}^{j}\gamma(i)}+\phi^{2\sum_{i=3}^{j}\gamma(i)}-\phi^{2\sum_{i=3}^{j}\gamma(i)}\right.\\
&\qquad \left.+\cdots-\phi^{2(\gamma(j-1)+\gamma(j))}+\phi^{2\gamma(j)}-\phi^{2\gamma(j)}\right)\\
&= \left(\frac{\sigma_\varepsilon^2\phi^h}{1-\phi^2}\right)\cdot 1\\
&= \frac{\sigma_\varepsilon^2\phi^h}{1-\phi^2}.
\end{aligned}
$$

As $Cov(j,l)$ depends only on the gap time h, the series is weakly stationary. Therefore, the correlation function is

$$\rho(h) \quad = \quad \frac{\frac{\sigma_\varepsilon^2 \phi^h}{1-\phi^2}}{\frac{\sigma_\varepsilon^2}{1-\phi^2}} = \phi^h.$$

□

Note that Proposition 1 is a special case of Proposition 2; we need only let $\gamma(j) = 1$ $\forall\, j \in \mathbb{Z}^+$.

If we think of the price series as a latent process which is only observed when a house is sold, the observed time series is Markovian. We define a Markov process next [37, p. 358].

Definition 5. A stochastic process X_t with realizations x_t is *Markovian* if
$\mathbb{P}\left(X_t \le x_t | X_s,\ s < t\right) = \mathbb{P}\left(X_t \le x_t | X_{t-1}\right)$.

Following directly from Proposition 2, we can conclude:

Corollary 3. *For $k < l$, $w_l | w_k \sim \mathcal{N}\left(\phi^{\gamma(l)-\gamma(k)}, \left|\phi^{l-k}\right| \sigma_\varepsilon^2\right)$. Therefore, the observed process $w_1, w_2, \ldots$ described in Proposition 2 is Markovian.*

The autoregressive component adds an important feature to the model. Intuitively, the longer the gap time between sales, the less useful the previous price should be when predicting the next sale price. For the model described in (3.2), 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 $\left(\phi^{\gamma(i,j)}\right)$. For short gap times, however, we expect the adjusted log price of each sale to be similar. This will hold only if the ϕ is very close to 1.

3.3 Model Fitting

In this section, we describe the coordinate ascent algorithm which is used to compute the maximum likelihood estimates (MLE) of the model parameters. This iterative procedure, maximizes the likelihood function with respect to each parameter while holding all other parameters constant [4, p. 129]. The algorithm stops when the parameter estimates have converged. We discuss convergence further in Sec. 3.3.2.

3.3.1 Maximum Likelihood Estimation

The likelihood function can be expressed as the product of conditional densities for each observation. To start, let $\boldsymbol{\theta} = \{\boldsymbol{\beta},\ \phi,\ \sigma_\varepsilon^2\}$ denote the set of estimable parameters where $\boldsymbol{\beta} = \{\beta_1, \ldots,\ \beta_T\}$. If $f(\cdot)$ is the density of the first sale and $f(\cdot|\cdot)$ the conditional density of a sale given the previous sale, the likelihood function $L(\boldsymbol{\theta}; \mathbf{y})$ where $\mathbf{y}$ is the vector of log prices is:

$$
\begin{aligned}
L(\boldsymbol{\theta};\ \mathbf{y}) &= \prod_{i=1}^{I} f(y_{i,1}) \prod_{j=2}^{J_i} f(y_{i,j}|y_{i,j-1}) \\
&= \prod_{i=1}^{I} \frac{1}{\sqrt{2\pi\left(\frac{\sigma_\varepsilon^2}{1-\phi^2}\right)}} \exp\left\{ -\frac{1}{2} \frac{\left(y_{i,1} - \beta_{t(i,1)}\right)^2}{\frac{\sigma_\varepsilon^2}{1-\phi^2}} \right\} \times \\
&\quad \prod_{j=2}^{J_i} \frac{1}{\sqrt{2\pi\left(\frac{\sigma_\varepsilon^2\left(1-\phi^{2\gamma(i,j)}\right)}{1-\phi^2}\right)}} \times \\
&\quad \exp\left\{ -\frac{1}{2} \frac{\left(y_{i,j} - \beta_{t(i,j)} - \phi^{\gamma(i,j)}(y_{i,j-1} - \beta_{t(i,j-1)})\right)^2}{\frac{\sigma_\varepsilon^2(1-\phi^{2\gamma(i,j)})}{1-\phi^2}} \right\}
\end{aligned}
$$

To simplify computations, we reparamaterize and let:

$$\tau^2 = \frac{\sigma_\varepsilon^2}{1-\phi^2}. \tag{3.3}$$

Incorporating this change, the log likelihood function, $l(\boldsymbol{\theta}; \mathbf{y})$ is:

$$\begin{aligned} l(\boldsymbol{\theta};\, \mathbf{y}) &= -\frac{N}{2}\log(2\pi\tau^2) - \frac{1}{2\tau^2}\sum_{i=1}^{I}\left(y_{i,1} - \beta_{t(i,1)}\right)^2 - \frac{1}{2}\sum_{i=1}^{I}\sum_{j=2}^{J_i}\log\left(1 - \phi^{2\gamma(i,j)}\right) \\ &\quad -\frac{1}{2\tau^2}\sum_{i=1}^{I}\sum_{j=2}^{J_i}\frac{\left(y_{i,j} - \beta_{t(i,j)} - \phi^{\gamma(i,j)}\left(y_{i,j-1} - \beta_{t(i,j-1)}\right)\right)^2}{1-\phi^{2\gamma(i,j)}} \end{aligned} \tag{3.4}$$

where $N = \sum_{i=1}^{I} J_i$ is the total number of observations.

The Coordinate Ascent Algorithm

Next, we outline the coordinate ascent algorithm tailored to the proposed model. On the practical side, the model was fit using R. Even for large metropolitan areas, such as Chicago, IL, the algorithm worked quickly never exceeding fifteen minutes of computing time. The algorithm requires starting estimates for the parameters to be specified. These values can be anything and the algorithm should eventually converge to the MLEs. However, to save computing time, formulas for useful starting values have been provided at the end of this section. The updating functions are provided after the algorithm.

AR Model Fitting Algorithm

1. Set a tolerance level ϵ (possibly different for each parameter) and a maximum number of iterations K.

2. Initialize the parameters: $\theta^0 = \left\{\beta_1^0, \ldots, \beta_T^0, \phi^0, \tau^{2,0}\right\}$. Let $\{\beta_t\}$ denote the set of β parameters.

3. For iteration k,

    (a) For $t \in \{1, \ldots, T\}$, update β_t using (3.6), to follow. That is, $\beta_t^k = f\left(\left\{\beta^k_{s:s<t}\right\}, \left\{\beta^{k-1}_{s:s>t}\right\}, \tau^{2\ k-1}, \phi^{k-1}\right)$. After each β_t update, recompute: $w_{i,j} = y_{i,j} - \beta_t^k$ where $t(i,j) = t$.

    (b) Update τ^2 using (3.7). That is, $\tau^{2,k} = f\left(\left\{\beta_t^k\right\}, \phi^{k-1}\right)$.

    (c) Find the zero of (3.8) to update ϕ^k using the estimates $\left\{\left\{\beta_t^k\right\}, \tau^{2,k}\right\}$.

    (d) If $\left|\boldsymbol{\theta}^{k-1} - \boldsymbol{\theta}^k\right| > \boldsymbol{\epsilon}$ for any $\theta_i \in \boldsymbol{\theta}$ and $k < K$, repeat Step 3 after replacing $\boldsymbol{\theta}^{k-1}$ with $\boldsymbol{\theta}^k$. Otherwise, stop and denote k' denote the final iteration.

4. Solve for σ_ε^2 using the relation (3.3) and $\left\{\boldsymbol{\beta}^{k'}, \phi^{k'}, \tau^{2,k'}\right\}$.

To obtain updating functions, we must differentiate the log likelihood function with respect to each parameter, set the derivatives to zero, and solve for that parameter. For $\boldsymbol{\beta}$ and τ^2, it is possible to solve these equations exactly. However, for ϕ, the zero must be computed numerically. As ϕ is a one-dimensional parameter, numerical methods, such as the Newton-Raphson algorithm, are highly suitable. The functions are listed below with the derivations provided in Appendix B.1:

$$
\begin{aligned}
w_{i,j} &= y_{i,j} - \beta_{t(i,j)} \qquad (3.5)\\
\beta_t &= \left(\frac{1}{|i : t(i,1) = t| + \sum_{\substack{i:t(j)=t\\ j>1}} \frac{1}{1-\phi^{2\gamma(i,j)}} + \sum_{\substack{i:t(i,j-1)=t\\ j>1}} \frac{\phi^{2\gamma(i,j)}}{1-\phi^{2\gamma(i,j)}}} \right) \times \\
&\Bigg(\sum_{i:t(i,1)=t} y_{i,1} + \sum_{\substack{i:t(i,j)=t\\ j>1}} \frac{1}{1-\phi^{2\gamma(i,j)}} \Big(y_{i,j} - \phi^{\gamma(i,j)} \left(y_{i,j-1} - \beta_{t(i,j-1)} \right) \Big) \Bigg) \\
&- \sum_{\substack{i:t(i,j-1)=t\\ j>1}} \frac{\phi^{\gamma(i,j)}}{1-\phi^{2\gamma(i,j)}} \Big(y_{i,j} - \beta_{t(i,j)} - \phi^{\gamma(i,j)} y_{i,j-1} \Big) \Bigg) \qquad (3.6)
\end{aligned}
$$

$$
\begin{aligned}
\tau^2 &= \frac{1}{N} \left[\sum_{i=1}^{I} w_{i,1}^2 + \sum_{i=1}^{I} \sum_{j=2}^{J_i} \frac{\left(w_{i,j} - \phi^{\gamma(i,j)} w_{i,j-1} \right)^2}{1-\phi^{2\gamma(i,j)}} \right] \qquad (3.7)\\
0 &= \sum_{i=1}^{I} \sum_{j=2}^{J_i} \frac{\gamma(i,j)\phi^{2\gamma(i,j)-1}}{1-\phi^{2\gamma(i,j)}} \\
&+ \frac{1}{\tau^2} \sum_{i=1}^{I} \sum_{j=2}^{J_i} \frac{\left(w_{i,j} - \phi^{\gamma(i,j)} w_{i,j-1} \right) \left(\gamma(i,j) w_{i,j-1} \phi^{\gamma(i,j)-1} \right)}{1-\phi^{2\gamma(i,j)}} \\
&- \frac{1}{\tau^2} \sum_{i=1}^{I} \sum_{j=2}^{J_i} \left(\frac{w_{i,j} - \phi^{\gamma(i,j)} w_{i,j-1}}{1-\phi^{2\gamma(i,j)}} \right)^2 \gamma(i,j)\phi^{2\gamma(i,j)-1} \qquad (3.8)
\end{aligned}
$$

Recall, a major criticism of repeat sales models is that since only repeat sales are used, a large portion of the data is omitted. In the proposed model, this is not the case. As can be seen in the above description, all of the data is used to estimate the time effect (see (3.6)) and σ_ε^2. Specifically, if we examine (3.6) more closely, we see that a log index value is computed as a weighted sum of single sales and repeat sales. More weight is given to the repeat sales observations. Finally, observe that only repeat sales observations are used to estimate the value of ϕ.

Starting Estimates

A simple way to obtain initial estimates for each of the parameters is described next starting with $\{\beta_t\}$. For every time period, let $1 < t < T$. Then,

$$\beta_t^0 = \frac{1}{|i : t(j) = t, 1 < j < J_i|} \sum_{i:t(i,j)=t} y_{i,j}$$

where $|\cdot|$ is the cardinality of a set. For the remaining two parameters, extra calculations are required. First, we compute $\mathbf{w}$:

$$w_{i,j} = y_{i,j} - \beta^0_{t(i,j)}.$$

For each gap time h, find all of the pairs of coordinates $(x_{h,1} = w_{i,j-1}, x_{h,2} = w_{i,j})$ such that $\gamma(i,j) = h$. Let H be the maximum gap time. Then, estimate ϕ and τ by:

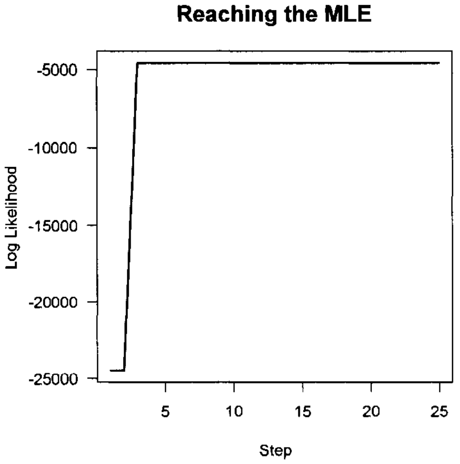
$$\begin{aligned} \phi^0 &= \frac{1}{H}\sum_{h=1}^{H} \left[Cor\left(\mathbf{x}_{h,1}, \mathbf{x}_{h,2}\right)\right]^{\frac{1}{h}} \\ \tau^{2,0} &= \frac{Var(\mathbf{w})}{(1-\phi^{2,0})} \end{aligned}$$

where $Cor(\cdot,\cdot)$ is the sample correlation function and $Var(\cdot)$ is the sample variance function.

3.3.2 Convergence of the Coordinate Ascent Algorithm

For models in the exponential family, the MLE can be proven to exist and be unique. These features help us prove that the coordinate ascent algorithm converges to the MLE [4, p. 130]. The proposed model, however, is a differentiable exponential family [5]. Therefore, the proof does not directly apply; nonetheless, we find empirically that the likelihood function

Figure 3.1: The Log Likelihood Function at Each Iteration

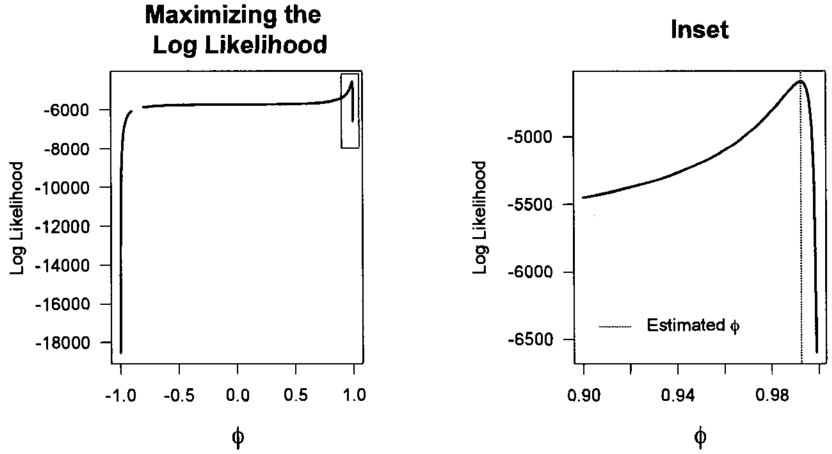


is well behaved so the MLE should be reached for this case as well. We use the likelihood plots from Stamford, CT as a typical example.

In the exponential family case, the log likelihood value will never decrease at each iteration when applying the algorithm. In Fig. 3.1, we plot the log likelihood value after each parameter update to check this for our model. We see that the resulting curve never decreases.

The nonlinear parameter in the proposed model is ϕ. For this parameter, we need to determine whether there is a unique maximum point. We maximize the log likelihood function given fixed values of ϕ; that is, we maximize the log likelihood function with respect to only $\boldsymbol{\beta}$ and τ^2 only. If we plot the maximized log likelihood value for each ϕ, the maximum should be equal to the value of ϕ which the algorithm converges to. We show that is the case in Fig. 3.2. The area of the plot near 1 is magnified in the second plot and the $\hat{\phi}$ which the coordinate ascent algorithm converges to is shown in red. Thus, we are confident that the fitting algorithm converges to the correct MLEs.

Figure 3.2: Maximizing ϕ



3.4 Final Details

3.4.1 Asymptotic Variance of the Parameter Estimates

To compute the asymptotic variance of the parameter estimates, we assume that the MLE estimates of the proposed model are consistent. Consequently, we can use the observed information matrix to obtain estimates for the variances [20, p. 481]. Using Greene's (2003) notation,

$$\left[\hat{\mathbf{I}}\left(\hat{\theta}\right)\right]^{-1} = \left(-\frac{\partial^2 l\left(\hat{\theta};\mathbf{y}\right)}{\partial\hat{\theta}\partial\hat{\theta}'}\right)^{-1}$$

where θ is a parameter. Expressions for the components of the observed information matrix are derived in Appendix B.2.

3.4.2 Converting Back to the Price Scale

Predictions for future observations are made only for repeat sales houses:

$$\hat{y}_{i,j} = \beta_{t(i,j)} + \phi^{\gamma(i,j)} \left(y_{i,j-1} - \beta_{t(i,j-1)}\right) \tag{3.9}$$

As a result, we consider our method to be a repeat-sales method.

To interpret the predictions, we must convert $\hat{y}_{i,j}$ to the price scale (denoted as $\hat{Y}_{i,j}$). Note, however, this is not as simple as exponentiating the fitted values. Instead, to compute the conditional mean of the response, we need to add the following adjustment:

$$\hat{Y}_{i,j}\left(\sigma^2\right) = \exp\left\{\hat{y}_{i,j} + \frac{\sigma^2}{2}\right\} \tag{3.10}$$

where σ^2 denotes the variance of of $y_{i,j}$. The additional term $\frac{\sigma^2}{2}$ approximates the difference between $E[\exp\{x\}]$ and $\exp\{E[x]\}$. We must adjust the latter expression to approximate the former. We improve the efficiency of our estimates by using this adjustment [40, p. 3025]. In formula (3.10), σ^2 is estimated from the mean squared residuals (MSR), where $MSR = \frac{1}{N}\sum_{i=1}^{N}\left(y_{i,j} - \hat{y}_{i,j}\right)^2$ with N specifying the number of observations. Therefore, the log price estimates, $\hat{y}_{i,j}$ are converted to the price scale using:

$$\exp\left\{\hat{y}_{i,j} + \frac{MSR}{2}\right\}. \tag{3.11}$$

3.4.3 Index Construction

The log price indices are given by $\hat{\beta}$; therefore $\exp\{\beta_t\}$ is the index on the price scale. An efficiency adjustment, such as the one in Sec. 3.4.2, is ignored as the standard error for $\hat{\beta}_t$

is so small that it has a negligible impact. Given $1, \ldots, T$ quarters, we let the first quarter be the base year and set the index level for that quarter to 1. Rescaling with respect to the first quarter, the final estimated price index series is:

$$1, \; \exp\left\{\hat{\beta}_2 - \hat{\beta}_1\right\}, \; \exp\left\{\hat{\beta}_3 - \hat{\beta}_1\right\}, \ldots, \; \exp\left\{\hat{\beta}_T - \hat{\beta}_1\right\}. \tag{3.12}$$

Chapter 4

Existing Methods

Numerous models have been proposed for predicting house prices and constructing indices. The most common types fall into four major categories: hedonic, repeat sales, hybrid, and spatial models. We describe each with special emphasis on repeat sales methods. Finally, we end the chapter with a brief overview of the major commercial house price indices.

4.1 Hedonic Models

The simplest type of hedonic model is a regression of house price against hedonic variables: square feet, number of bathrooms, amenities, and so forth. Location is often included as an indicator variable. These models follow the standard multiple regression setup:

$$y_i \quad = \quad \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_p x_{pi} + \varepsilon_i$$

where y is the sale price (or log price), $x_1, \ldots, x_p$ are the p hedonic variables, and $\varepsilon_i \overset{iid}{\sim} \mathcal{N}\left(0, \sigma_\varepsilon^2\right)$. Most often, such models are cross-sectional and changes over time are examined by refitting the model yearly; however, more complex models do exist [47].

Pure hedonic models have been largely abandoned in favor of alternative methods due to various limitations. The availability of relevant variables and model form are two key problems that arise [7]. Other proposed methods, such as repeat sales or spatial models, attempt to circumvent such issues by using previous sale price and geography respectively as surrogates for hedonic variables; however, Meese and Wallace (1997) still advocate the use of hedonic models for constructing local indices.

4.2 Repeat Sales Models

Bailey, Muth, and Nourse (1963), introduced the landmark concept of repeat sales analysis. This method is based on the premise that the previous sale price of a house acts as a proxy for hedonic variables. Essentially, the log price difference between pairs of sales of a house is used to construct a price index. Therefore, only houses which have been sold twice are used to calculate the index; the remaining observations are omitted. However, homes that are known to have undergone significant improvement or degradation are usually also excluded from the analysis. For such homes, the previous sale price is not an appropriate surrogate for hedonic information. The Bailey, Muth, and Nourse (BMN) method was extended by Case and Shiller (1987, 1989) to incorporate heteroscedasticity in the error term. Finally, the Case-Shiller model was converted into a commercial index, the S&P/Case-Shiller® index. This is also a repeat sales index, however, it is computed using prices instead of log prices. We describe all three procedures in detail in the following subsections.

4.2.1 The BMN Model

Let the subscript t index the time of sale of a given house. Let there be $T + 1$ time periods of sales from 0, 1 ,..., T. Using the BMN notation, for a pair of sales of a given house i, prices and indices are related by the following expression:

$$\frac{P_{it'}}{P_{it}} = \frac{B_{t'}}{B_t} U_{itt'} \tag{4.1}$$

where P_{it} a the sale price of the ith house at the tth time period $(t' > t)$. Thus, t is the time of the first sale and t' the time of the second. B_t denotes the general house price index at time t. Finally, $U_{itt'}$ is the error term and has a log normal distribution log $U_{itt'} \overset{iid}{\sim} \mathcal{N}\left(0, \sigma_u^2\right)$ [1, p. 934]. The model is fitted on the logarithmic scale:

$$p_{it'} - p_{it} = b_{t'} - b_t + u_{itt'} \tag{4.2}$$

where p, b, and u are the logarithmic versions of the terms in (4.1). Basically, the expected difference in log prices for two sales of a house is surmised to equal the difference in the corresponding log indices.

Linear regression is used to fit the model. For the ith house, the complete regression equation is:

$$p_{it'} - p_{it} = \sum_{j=1}^{T} b_j x_{ij} + u_{itt'} \tag{4.3}$$

where

$$x_{ij} = \begin{cases} -1 & \text{if } j = t \text{ and } t > 0 \text{ (first sale),} \\ +1 & \text{if } j = t' \text{ (second sale),} \\ 0 & \text{otherwise.} \end{cases}$$

By construction, $b_0 = 0$ and therefore $B_0 = 1$. If there are N sale pairs, the resulting design matrix, $\mathbf{X}$, is an $N \times T$ matrix. $\mathbf{y}$ is the vector of log price differences for sale pairs. The log index is computed using least squares: $\hat{\mathbf{b}} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$; exponentiating the results gives us the desired index on the price scale.

4.2.2 The Case-Shiller Method

Case and Shiller (1987, 1989) expand the BMN setup by further assuming that the error terms are heteroscedastic. They reason that the length of time between sales should increase the variance of the log price differences between sale pairs. To compute the house price index while accounting for the heteroscedasticity, they follow the BMN procedure but add a small twist: in constructing estimates, the observations are weighted depending on the gap time. Sale pairs with larger gap times are given lower weights. The corresponding model is:

$$p_{it} = b_t + H_{it} + u_{it} \tag{4.4}$$

where p_{it} is the log price of the sale of the ith house at time t, b_t is the log index at time t, and $u_{it} \overset{iid}{\sim} \mathcal{N}\left(0, \sigma_u^2\right)$. The middle term, H_{it}, is a Gaussian random walk which contains the previous log sale price of the house [10, p. 126].

A random walk $z_1, z_2, \ldots$ can be written as a sum of random variations added to the initial term. That is,

$$\begin{aligned}
&z_1 \\
z_2 &= z_1 + v_2 \\
z_3 &= z_1 + v_2 + v_3 \\
&\vdots \\
z_t &= z_1 + \sum_{j=2}^{t} v_j .
\end{aligned}$$

Case and Shiller assume that the random walk in their model is Gaussian, meaning that $v_t \overset{iid}{\sim} \mathcal{N}\left(0, \sigma_v^2\right)$. Like the BMN model, the Case-Shiller setting is a model for *differences* in prices. Thus, we want to fit the following model:

$$p_{it'} - p_{it} = b_{t'} - b_t + \sum_{j=t+1}^{t'} v_{ij} + u_{it'} - u_{it} \qquad (4.5)$$

where $t' > t$. They use weighted least squares to fit the model to account for both sources of variation. The three-step procedure is described below:

1. Fit the BMN model as in (4.2) using log price.

2. Compute the residuals of the setup in (4.2), and denote these as $\hat{\varepsilon}_i$. This residual is an estimate of: $u_{it'} - u_{it} + \sum_{j=1}^{t'-t} v_{ij}$. The expectation of ε_i is $E[u_{it'} - u_{it} + \sum_{j=1}^{t'-t} v_{ij}] = 0$ and the variance is $Var[u_{it'} - u_{it} + \sum_{j=1}^{t'-t} v_{ij}] = 2\sigma_u^2 + (t' - t)\sigma_v^2$ since the errors are independent of each other. The square of the residuals is an unbiased estimate of this variance. To compute the weights for each observation, the squared residuals from Step 1 are regressed against the gap time. That is,

$$\hat{\varepsilon}_i^2 = \underbrace{\beta_0}_{2\sigma_u^2} + \underbrace{\beta_1}_{\sigma_v^2}(t' - t) + \eta_i \qquad (4.6)$$

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 $\mathbf{\Omega}^{-1}$.

3. To obtain the adjusted index, we essentially repeat the BMN procedure as in Step 1; however, instead we run a weighted least squares where $\hat{\mathbf{b}} = \left(\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{X}\right)^{-1}\mathbf{X}'\mathbf{\Omega}^{-1}\mathbf{y}$. Incorporating the weight matrix $\mathbf{\Omega}^{-1}$ down-weights sale pairs with large gap times.

4.2.3 The S&P/Case-Shiller® Method

A variation of the model proposed by Case and Shiller is now used to create house price indices for Standard and Poors called the S&P/Case-Shiller® Home Price Index [9, 10, 44]. Two changes have been made from the procedure outlined in Sec. 4.2.2. First, prices instead of log prices are used to compute the index. Nevertheless, they attempt to preserve the random walk component of the model despite the multiplicative relationship between prices and indices as in (4.1). Second, measurement error in the sale price is introduced into the model. To handle this, instrumental variables are used when fitting the model.

As before, we have $T + 1$ time periods from 0, 1, ..., T. We will ignore the "random walk" part of the model for now; we will see why in a minute. For sale pair i, we can write their model as:

$$\begin{aligned} P_{i0} &= \beta_{t'} P_{it'} + U_{i0t'} && \text{first sale at time } 0, \\ 0 &= \beta_{t'} P_{it'} - \beta_t P_{it} + U_{itt'} && \text{first sale at time } t > 0 \end{aligned} \tag{4.7}$$

where P_{it} is the sale price of house i at time t, β_t is the inverse of the index at time t, and $U_{itt'} \overset{iid}{\sim} \mathcal{N}\left(0, \sigma_U^2\right)$. The index $B_t = \frac{1}{\beta_t}$ and $B_0 = 1$. As before, the price index at time 0 is set to 1 by convention. Even with this transformation from β_t to B_t, we still cannot replicate the original BMN repeat-sales model as given in (4.1) for price because of the additive error term $U_{itt'}$ which was originally multiplicative on the price scale. Moreover, it is unclear on how to incorporate the contribution of the random walk since we are *multiplying* the prices by the β values. Ultimately, there is a disconnect between the written and actual models. Meissner and Satchell (2007) observe this problem as well in their paper comparing this index with the Financial Times House Price Index used in the UK.

With this version, observe that the response vector contains mostly zeros as the vast majority of sales do not occur in the base time period. However, as sales in the base period are the only sales to not be multiplied by an index since $B_0 = 1$ by construction, one must assume that this is why they are the only prices that appear in the response vector. Moreover, it seems misleading to create a model where *future* sales are used to explain a *preceding* sale. For those sale pairs where the first sale is in the base period, this is exactly what occurs.

This model also assumes that prices do not reflect the true value of the house. That is, there is some measurement error. In least-squares regression, the explanatory variables are assumed to be fixed, not variable. Introducing measurement error into the prices violates this assumption resulting in biased coefficient estimates. To accommodate this, instrumental variables (IV) can be used when fitting the model [46, p. 7578].

Irrespective of any correspondence to a model such as (4.7), the S&P/Case-Shiller® procedure follows in essence the three step pattern outlined in Sec. 4.2.2. The design matrix **X**, IV matrix, **Z**, and response vector **y** are now defined as follows [44, p. 22]:

$$X_{ij} = \begin{cases} -P_{ij} & \text{if } j = t \text{ and } t > 0 \text{ (first sale),} \\ P_{ij} & \text{if } j = t' \text{ (second sale),} \\ 0 & \text{otherwise.} \end{cases}$$

$$Z_{ij} = \begin{cases} -1 & \text{if } j = t \text{ and } t > 0 \text{ (first sale),} \\ 1 & \text{if } j = t' \text{ (second sale),} \\ 0 & \text{otherwise.} \end{cases}$$

$$y_i = \begin{cases} P_{ij} & \text{if } j = 0 \text{ (first sale at time 0),} \\ 0 & \text{if } j \neq 0 \text{ (first sale not at time 0)} \end{cases}$$

Note that the matrix **Z** here was the matrix **X** in the Case-Shiller method. A valid instrument **Z** must satisfy the following two conditions [46, p. 7578]:

1. The instrument must be uncorrelated with the regression error term. In this case, **Z** should be uncorrelated with the $U_{itt'}$ values. Since **Z** simply indicates what type of sale has taken place (first/second), it is be uncorrelated with the error term.

2. **Z** does not have a direct effect on **Y**; it only affects the response variable through **X**. This is true for the S&P/Case-Shiller® setting as well.

Therefore, **Z**, as defined above, is a valid instrument. The regression in the first step of the S&P/Case-Shiller® procedure is now, $(\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{y}$ which is simply the least square estimate when using an instrumental variable. The third step now becomes $(\mathbf{Z}'\boldsymbol{\Omega}^{-1}\mathbf{X})^{-1}\mathbf{Z}'\boldsymbol{\Omega}^{-1}\mathbf{y}$. The second step, where the weights are computed, is the same as in the Case-Shiller procedure [44, p. 25].

Table 4.1: Removing Single Sales

City	No. Obs.	No. Repeat Sales	% Repeat Sales
	Case-Shiller (1970-1986)		
Atlanta, GA	221,876	8,945	4.0%
Chicago, IL	397,183	15,530	3.9%
Dallas, TX	211,638	6,669	3.2%
San Francisco/ Oakland, CA	121,909	8,066	6.6%
	Meese-Wallce (1970-1988)		
Freemont, CA	23,408	3,405	14.5%
Oakland, CA	27,606	3,342	12%

4.2.4 Discussion of Repeat Sales Methods

While repeat sales methods look promising, a number of problems have been highlighted. Perhaps the most obvious issue is that single sales are excluded reducing the sample size significantly. Sample sizes of data used in Case and Shiller (1989) and Meese and Wallace (1997) are shown in Table 4.1. The number of observations that are eliminated is staggering. While data spanning a longer period will result in a higher number of repeat sales, the number of newly built houses also increases. Therefore, the proportion of repeat sales among all house sales does not increase as fast as one might expect.

Among repeat sales homes, further cuts are made if the house has significantly improved or deteriorated between sales. This is because "house quality" would not have been controlled in the intervening period. Most likely, the Case-Shiller percentages are lower in Table 4.1 than those for Meece-Wallace because houses that went under significant renovation and were not arms-length transactions (i.e. houses sold cheaply to relatives, etc.) were excluded, eliminating even more data.

A related issue is that in repeat sales models, a home is ignored *until* it is sold for a second time. To see the impact, say that a house is sold first at time t and again at time t'. If the initial data set includes sales until time $t^{\star}$ where $t \leq t^{\star} < t'$, this particular house would neither be part of the data nor would it be used to compute the index value for time t. Now, if the data is updated to include all sales up to time t', the house is included and is used to compute the index for *both* time t *and* time t'. Thus, indices can be revised retroactively. This is problematic if indices are to be used in a commercial setting. S&P/Case-Shiller® use a "chain-weighting" procedure to avoid revising the indices [44, p. 26].

Omitting single sales raises the issue whether repeat sales homes are representative of the entire housing market. In any time span, houses can be categorized as follows: new home sales, repeat sales with no changes in the house, repeat sales homes with changes, and houses not sold [7, p. 290]. Repeat sales methods only use data in the second category. There are two ways in which repeat sales methodology can be evaluated in terms of "representativeness."

The first method is to investiage whether repeat sales homes are fundamentally different from single sale homes. Recall, in an entire sample period, a single sale can refer to a new home or and old home which sold only once in the sample period. There is a hypothesis that a higher proportion of repeat sales are "starter homes." Young families tend to live in these so called "starter homes" and later trade up to larger homes after only a few years [14, p. 271]. Clapp, et al. (1991) test this hypothesis with inconclusive results. Meese and Wallace (1997) test this claim as well by comparing hedonic models with an indicator for repeat sales and with interaction terms between repeat sales and hedonic variables. The covariates are: number of bathrooms, number of bedrooms, the ratio of bedrooms to total number of rooms, square footage, age, and an "index of house quality" which was not well described in the paper. In their analysis, they found there was a significant difference and "...repeat-sales homes that did not change attributes are slightly smaller, and are in worse condition, than the average for single-sale homes...The repeat-sales homes that did have attribute changes...tend to be slightly larger and in worse condition, than the average for single-sale homes [27, 55]." It is unclear whether this is the case because brand new homes have a higher index of house quality. Furthermore, the choice of variables and form of the model can always be questioned in such analyses.

A second way to look at representativeness is whether the value of various housing characteristics change over time. Meese and Wallace (1997) test this as well using a similar setup to the one described above; however, they do not distinguish between repeat and single sales but do allow for interactions between housing characteristics and time. Recall repeat sales methodology assumes that the value of the housing attributes do not vary over time; therefore, sales of homes can be examined just for the time effects. Meese and Wallace find that this assumption does not hold and, in actuality, the value of attributes do change. For practical reasons, it is impossible to include all relevant covariates for lack of data and because new attributes (such as dishwashers and other amenities) come into existence. In addition, these changes can simply be thought of as adding to the time effect and become part of the price index.

Even if changes in attribute values are ignored, there is one aspect that varies with all houses: age. Case, et al (1991) claim that because age changes over time, repeat sale indices are biased. Basically, the time effect is confounded with the age effect. Specifcally, the general upward trend of the effect of time is countered by the negative effect of age. Palmquist (1979) suggests adding in a depreciation factor to the repeat sales procedure to account for this; however, this factor must be independently computed which adds much to the complexity of the model [29, p. 337].

Both the Case-Shiller and S&P/Case-Shiller® indices are computed using generalized least squares (GLS). The usual GLS procedure would be to define the weight matrix $\mathbf{\Omega}$ as a matrix of estimated variances. Such weights are used so that the best linear unbiased estimates (BLUE) of the regression coefficients are obtained. However, in the Case-Shiller models the estimated *standard deviations* are used instead. Consequently, the resulting index estimates are unbiased but do not have the lowest possible variance. This is undesirable especially if the regression estimates are to be used for prediction and prediction intervals are to be constructed.

The BMN model is equivalent to a two-way fixed effects model:

$$p_{it} = \alpha_i + b_t + \varepsilon_{it} \tag{4.8}$$

where α_i is a fixed house effect and b_t is the log index. It is fit, however, for differences in prices, $p_{it'} - p_{it}$ and so the fixed effect for houses drops out and we obtain (4.2). The setting in (4.8) is more appropriate, however, when houses sell more than twice in the data. If there are three sales, for example, then the house appears in the data as two sets of sale pairs: the first and second sales form a pair and so do the second and third sales. Bailey, Muth, and Nourse (1963) address this by pointing out that this situation causes correlations in the residuals if you assume that there are house specific effects. They suggest introducing a "property" effect into the model and fitting (4.8) instead of (4.2) [1, p. 939]. Thus, they advocate modeling log prices instead of price differences. However, this procedure is not implemented. The Case-Shiller methods do not address this common scenario at all.

A second point arises when examining the case of multiple sales. For instance, say a house is sold thrice: one at time 0, a second at time h, and a third at time $h + g$. Recall that the variance of the difference of a pair of sales is given by $2\sigma_u^2 + (t' - t)\sigma_v^2$ where t and t' are the times when the sales occurred. Say, by chance, we do not know about the second sale. Then, the variance of the difference of the first and third sale should be $2\sigma_u^2 + (g+h)\sigma_v^2$. Ideally, the fact that there was a second sale which was missing from the data should not be informative; that is, the variance of the estimates should not change with this knowledge. However, this is not the solution if derived from the regression equations. Rather, knowing there is a second sale at time h is informative. To see why this is true, we start by writing the regression equations for both pairs of sales:

$$\begin{aligned} y_2 - y_1 &= \beta_2 - \beta_1 + \varepsilon_{0,h} && (4.9) \\ y_3 - y_2 &= \beta_3 - \beta_2 + \varepsilon_{h,h+g} && (4.10) \end{aligned}$$

where $\varepsilon_{t,t'}$ includes both the random error and the cumulative random walk error. Adding (4.9) and (4.10), we obtain:

$$\begin{aligned} y_3 - y_2 + y_2 - y_1 &= \beta_3 - \beta_2 + \beta_2 - \beta_1 + \varepsilon_{0,h} - \varepsilon_{h,h+g} \\ y_3 - y_1 &= \beta_3 - \beta_1 + \varepsilon_{0,h} - \varepsilon_{h,h+g} \\ Var\left[y_3 - y_1\right] &= Var\left[u_{0,h}\right] + Var\left[u_{h,h+g}\right] \\ &= 2\sigma_u^2 + h\sigma_v^2 + 2\sigma_u^2 + g\sigma_v^2 \\ &= 4\sigma_u^2 + (g+h)\sigma_v^2. && (4.11) \end{aligned}$$

The variance of the first and third sales, given the knowledge of the second, is *larger* than if we had simply omitted it from the data. This occurs because the Case-Shiller model is not a stationary time series.

Random walks are nonstationary because the autocovariance function does not depend only on the gap time. The Case-Shiller method, however, models the first difference of the random walk: $y_2 - y_1, y_3 - y_2, \cdots$. This is a white noise process and is stationary. However, introducing the error term u_{it} into the model in (4.4) causes the final model given in (4.5) to be nonstationary.

A final issue also noted by Calhoun (1996) and described in more detail in Sec. 4.5, applies to Case-Shiller indices only. In the second stage when weights are computed, there is always a chance that for a particular sale pair, the computed weight may in fact be negative. For such cases, the third step cannot be executed at all. Calhoun (1996) outlines a method to circumvent this issue.

Ultimately, despite numerous concerns regarding repeat-sales methods, such procedures have been wholeheartedly adopted by the corporate sector. A number of agencies, including Standard and Poor's, use the Case-Shiller repeat-sales method to construct house price indices (see Sec. 4.5).

4.3 Hybrid Models

The most common criticism of repeat sales models is that a large portion of sales are ignored. Hybrid models attempt to address this issue by combining hedonic and repeat sales methods. This allows all of the data to be used without forgoing the extra information provided by repeat sales homes. To implement this class of models, however, hedonic variables are necessary.

Case and Quigley (1991) propose one such model. They implement their model using single family home sales in Honolulu, Hawaii from October 1980 through October 1987. There were 418 total sales for 310 houses. Of these sales, 108 were repeat sales of which 47 had *no* changes made to the house. "Ten out of the thirteen coefficients are statistically significant at the 0.01 level, and the simple correlation between the actual sale price and its predicted value is almost 0.9 [8, p. 56]." However, a countless number of significance tests were conducted and no analysis of residuals was done.

Quigley (1991) proposes an alternate hybrid model.

$$\begin{aligned} y_{i,j} &= Q_{i,j} + \alpha_{t(i,j)} + \omega_{i,j} \\ Q_{i,j} &= \beta X_{i,j} + \xi_i + \eta_{i,j} \end{aligned}$$

where $y_{i,j}$ is the log price of the jth sale of the ith house, $\alpha_{t(i,j)}$ is the log price index for time $t(i,j)$, and $\omega_{i,j}$. The unobserved log of a houses's quality level at the jth sale is given by $Q_{i,j}$. The quality is modeled by a regression of covariates, $X_{i,j}$, a fixed effect for house ξ_i, and another error term, $\eta_{i,j}$. Several assumptions are placed on the expectation and covariance of the error terms. Furthermore, it is assumed that house prices follow a random walk. To fit the model, use the repeat sales data to determine the parameters for the variance-covariance matrix. The second step is to use all of the data (repeat sales and single sales) to fit the above model using GLS with the previously computed variance-covariance matrix. Quigley claims that his procedure has two advantages: (1) important information included only in the repeat-sales data is extracted and (2) the information on single sales is used to "...increase the efficiency of estimation of the parameters β and $\alpha_{t(i,j)}$ [33, p. 6]." Quigley finds that by using this two-step approach, standard errors were reduced. Quigley applies his model to data from condominium sales in Los Angeles from January 1980 through December 1991. There were a total of 843 sales in the sample period. In checking the random walk assumption, Quigley finds that as the gap time increases, the variance of the difference in house price increases but at a decreasing rate [33, p. 9].

While hybrid methods improve upon the repeat-sales concept by including all of the sales, the problem of the availability of hedonic variables and selecting the correct model form still remains. Given these two issues, the hybrid class of models is currently not practical.

4.4 Spatial Models

The final class of models attempts to incorporate location into the pricing model beyond its role in hedonic models. Most commonly, the correlation between the error terms of a model is specified using spatial information. This incorporates a feature which hitherto was ignored: homes in the same neighborhood tend to be priced similarly. Gelfand, et al. (1998) observe that neighborhood or subdivision is not only a spatial property, it also acts as a surrogate for hedonic characteristics. In Gelfand, et al. (2004), it is found that "40-80 percent of the variability [in prices] is spatial [16, p. 163]."

In the spatial model proposed in Gelfand, et al (1998) hierarchical Bayes methods are used to predict house price. Their linear model includes: hedonic covariates, time effects, subdivision effects, and an interaction term between time and subdivision. The final component is included to allow spatial relationships to change over time. The best models did not include an interaction between time and subdivision. Furthermore, models with an additive temporal effect have similar predictive performances to ones where the effects of time are allowed to evolve.

As mentioned, models which incorporate hedonic variables suffer from two problems: the availability of relevant variables and the possibility of an incorrect model. Pavlov's (2000) solution is to use location as a proxy variable. The location of a house should include information about the unobserved hedonic variables.

The proposed model is given below:

$$z_i = p(x_i, y_i) + q_1(x_i, y_i)s_i + q_2(x_i, y_i)b_i + q_3(x_i, y_i)d_i + \varepsilon_i$$

where z_i is the log(price) of the sale, s_i is the size of the house, b_i is the number of bathrooms,

d_i is the number of bedrooms, and ε_i is the error. The intercept, $p(x_i, y_i)$, and slopes $q_1(x_i, y_i)$, $q_2(x_i, y_i)$, and $q_3(x_i, y_i)$ are to be estimated. These are smooth functions of the location coordinates (x_i, y_i). Therefore, the unobserved variables influence *both* the slopes and the intercepts of the model through location. Pavlov denotes these as space-varying coefficients (SVC). The errors are assumed to be independent and identically distributed with an unspecified distribution and independent of location. Note that this model is *static*–there is no time dimension.

To fit the model, weighted least squares is used. The weights for each observation are determined for each point (x_i, y_i) using a k-nearest neighbors type algorithm. Two models are examined: with and without zip codes. For the simpler model, the k-nearest neighbors are determined using Euclidean distance. The weights are computed by combining these distances with a parabolic weight function. To incorporate the zip codes into the model, the k-nearest neighbors are chosen using an alternate distance metric. Euclidean distance is still used; however, if a "nearby" house is in another zip code, a constant ν is added to this distance. The goal is to make houses in another zip code seem "farther" away. Cross-validation is used to determine the optimal values for k and ν.

Home sales from Los Angeles' West Side between April 1st and September 30th 1997 were used. There were 3,000 observations in the data set. Pavlov shows visually that both slopes (marginal effects) and intercepts do seem to vary according to location. In comparisons with other methods, the SVC method seems to predict prices better.

4.5 Commercial Indices

Over the last decade, a number of indices have emerged as more people have looked to the housing market for investment opportunities. In addition, with the current market collapse, housing indicators have become increasingly important in the quest for understanding how such markets operate. The issues that arise when developing house price models have been discussed earlier in this chapter. However, a practical concern still remains. Housing data, unlike stock data, cannot be examined in real time. In fact, most indices start reporting with at least a two-month lag between transaction dates and publication dates. This means, initial reports for January would not be released at least until March. Therefore, house price indices are mostly used for examining long term trends. In this section, we describe five US indices and, as a comparison, two UK indices.

US Indices

The Office of Federal Housing Enterprise Oversight (OFHEO) releases a quarterly repeat sales index, the House Price Index (HPI) for each state, census division, and nationwide. The data are provided by the Federal Home Loan Mortgage Corporation (Freddie Mac) and Federal National Mortgage Association (Fannie Mae) and contain homes which qualify for a conventional mortgage. This criterion excludes some high-end homes and homes bought at subprime rates. For the most part, they follow the Case-Shiller method (see Sec. 4.2.2); however, a few adjustments are made. In the second stage of the method when the weight matrix is computed, there is a chance that some weights are negative. Recall, the regression in (4.6) is fitted to calculate the weights. The intercept of this regression could be negative and large enough to offset the second term resulting in negative weight. In such situations, the third step of the Case-Shiller method cannot be computed. To eliminate this issue, instead of writing the model as in (4.4), the white noise component, u_{it} is replaced by u_i. That is, there is only one error term for each house, not for each house and sale combination [6, p. 9]. The resulting model is:

$$p_{it} = b_t + u_i + H_{it}. \tag{4.12}$$

Essentially, u_i is treated as a fixed effect. As the Case-Shiller index looks at *differences* between prices, u_i drops out of the fitted model.

A second modification to the Case-Shiller procedure is in regards to the random walk assumptions. For two sales of a house at time t and t', recall:

$$H_{it'} - H_{it} = \sum_{j=t+1}^{t'} v_{ij}$$

where $v_{i,j}$ are the random walk steps and $E\left[v_{i,j}\right] = 0$. If we compute $Var\left[H_{it'} - H_{it}\right]$,

$$\begin{aligned}
Var\left[H_{it'} - H_{it}\right] &= Var\left[\sum_{j=t+1}^{t'} v_{ij}\right] \\
&= E\left[\left(\sum_{j=t+1}^{t'} v_{ij}\right)^2\right] \\
&= \sum_{j=t+1}^{t'} E\left[v_{ij}^2\right] + \sum_{j=t+1}^{t'} \sum_{\substack{j'=t+1 \\ j \neq j'}}^{t'} E\left[v_{ij}v_{ij'}\right] \\
&= (t'-t)E\left[v_{ij}^2\right] + (t'-t)((t'-t)-1)E\left[v_{ij}v_{ij'}\right] \\
&= (t'-t)\left(E\left[v_{ij}^2\right] - E\left[v_{ij}v_{ij'}\right]\right) + (t'-t)^2 E\left[v_{ij}v_{ij'}\right]
\end{aligned}$$

where $j \neq j'$. In the Case-Shiller model, $E\left[v_{ij}v_{ij'}\right] = 0 \ \forall \ j \neq j'$ leaving $Var\left[H_{it'} - H_{it}\right] = \sigma_v^2(t'-t)$. However, this assumption is not made when computing the OFHEO index. Consequently, at the second stage of the fitting procedure, the squared residuals are instead regressed against the gap and squared gap times [6, p. 10].

A third adjustment the HPI index makes to the Case-Shiller method is in regards to the weighting scheme. The Case-Shiller procedure applies weights to each observations based solely on the gap time and not on initial sale price. Shiller writes, "The [initial sale price] weighting may make a difference to the estimated index if price changes in more valuable houses are different from price changes in less valuable houses [41, p. 110]. To address this, the HPI index is released with an adjustment factor which can be applied when using the index on a particular house [6, p. 11].

Two other indices are released by The National Association of Realtors and Freddie Mac and Fannie Mae. The former provides a number of indices at the regional level including single family homes, condominiums, and co-ops. These monthly indices are simply the median sale price for that month. The Conventional Mortgage Home Price Index is released by Freddie Mac and Fannie Mae quarterly. While they use the same data as OFHEO, these are separate indices. The Case-Shiller method as given in (4.4) is used to compute indices for numerous US cities and regions [45]. These last set of indices are by far the most granular–even very small US cities are included.

The S&P/Case-Shiller® Home Price Index is a monthly and quarterly index for 20 Metropolitan Statistical Areas (MSA) and three combined indices. An MSA is essentially a city along with surrounding areas which make up the metropolitan area, such as Chicago, IL and its suburbs. Note that MSAs can extend beyond state borders. The commercial Case-Shiller methodology is described in detail in Sec. 4.2.3. In the commercial index, indices are computed using a rolling three-month window. That is, a house sale in October is used for computing the index for October, November, and December. This is done by listing the house three times, for October through December, and weighting each "expanded observation" by 1/3 [44, p. 26]. Furthermore, additional weights are added based on the initial sale price of the house and whether or not the house underwent significant changes [44, p. 7].

A new, competing index is the Radar Logic Daily$^{\text{TM}}$Price Index which includes both single family homes and condominiums for 25 MSAs. They use a rolling window to provide indices for the past 7, 14, and 28 days. They claim to use all available data but after further inspection, some counties seem to have very low inclusion rates.

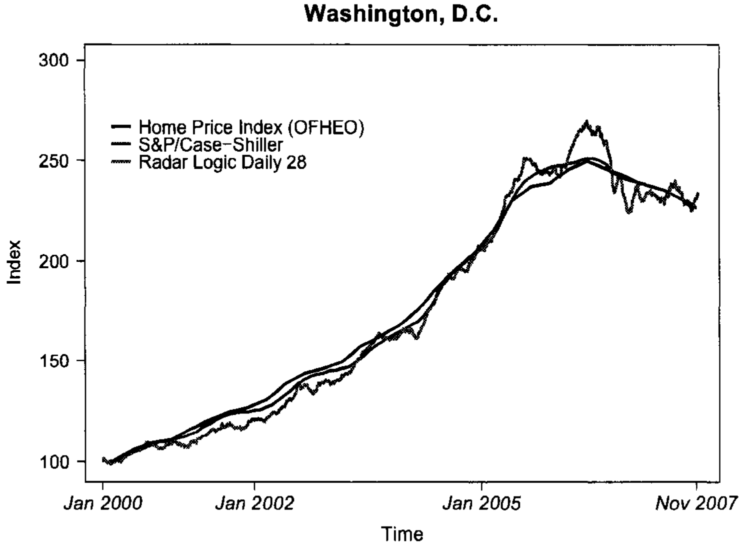
The index is constructed, roughly speaking, using the median price per square foot (ppsf) of the homes sold in the given window. Specifically, let N be the number of houses with a particular ppsf value (or range as in a histogram). Then, for two sets of ppsf and N values, have the following relationship [24, p. 7]:

$$\log N \;=\; \log N' + \beta(\log\ \text{ppsf} - \log\ \text{ppsf}'). \tag{4.13}$$

Essentially, an empirical probability distribution function is fit to the "histogram" of log N against log price per square foot. The log price per square foot values are divided into three categories: low, medium, and high. A piecewise linear function approximates this histogram with a line for each category. They term this the Triple Power Law$^{\text{TM}}$. The median of the resulting distribution is converted back to the price per square foot scale which is the reported index [24, p. 9].

To have an idea of what these indices look like, the OFHEO index, the S&P/Case-Shiller® Home Price Index, and the Radar Logic Daily$^{\text{TM}}$Price Index (28 Day) indices are plotted for Washington, D.C in Fig. 4.1. The indices have been rescaled on the plot to ease comparisons. Observe that the indices track each other quite closely even though the methodologies used vary. In addition the Radar Logic Daily$^{\text{TM}}$Price Index is more volatile since it is a daily index whereas the other two are monthly.

Figure 4.1: Comparing US Commercial Indices



UK Indices

The Land Registry is a UK government agency which collects and provides property data for England and Wales. They also release the monthly Land Registry House Price Index which is computed by Calnea Analytics, Ltd. Essentially, the index is constructed using repeat sales methodology with a seasonal adjustment [25]. Indices are computed both locally and nationwide.

A competing index, the monthly Financial Times House Price Index, is computed by Acadametrics Ltd. employing a wholly different method. Indices are released for regions, counties, and London boroughs. The types of housing included in the analysis are: detached houses, semi-detached houses, terraced houses, and apartments. Using property data, a "mix-adjusted index" is calculated where "mix" simply indicates the use of different types of homes for constructing the index [28, p. 1]. Essentially, a weighted average of the trends among these four categories are calculated accounting for various factors such as seasonality.

This is the only index which does not use individual property prices using averages over small regions instead.

Chapter 5

Results for the Global Models

The autoregressive model proposed in Chapter 3 is as follows. Let $y_{i,j}$ be the log price of the jth sale of the ith house. The parameter $\beta_{t(i,j)}$ is the log price index for quarter, $t(i,j)$, and let $\gamma(i,j)$ be the gap time if it is the second or higher sale. Then,

$$\begin{aligned} y_{i,1} - \beta_{t(i,1)} &= \varepsilon_{i,1} & j = 1 \\ y_{i,j} - \beta_{t(i,j)} &= \phi^{\gamma(i,j)} \left(y_{i,j-1} - \beta_{t(i,j-1)} \right) + \varepsilon_{i,j} & j > 1] \end{aligned}$$

where $\varepsilon_{i,1} \overset{iid}{\sim} \mathcal{N}\left(0, \frac{\sigma_\varepsilon^2}{1-\phi^2}\right)$ and $\varepsilon_{i,j} \overset{iid}{\sim} \mathcal{N}\left(0, \frac{\sigma_\varepsilon^2(1-\phi^{2\gamma(j)})}{1-\phi^2}\right)$ when $j > 1$. The parameters to be fitted are: $\boldsymbol{\beta}$, ϕ, and σ_ε^2.

In this chapter, we will test the applicability of the model to the data using several approaches. Recall that we expect pairs of sales with short gap times to be highly correlated. If true, the estimate for ϕ would be close to one. We confirm this belief in Sec. 5.1. In Sec. 5.2, we investigate through simulation the validity of using the observed information matrix to compute the standard errors of the parameter estimates. Sections 5.3 and 5.4.3, focus on checking model assumptions. We examine the AR(1) assumption and analyze the residuals. Finally, in Section 5.4, we compare the predictive performance of the proposed model with three other methods: a fixed effects model, a mixed effects model, and the S&P/Case-Shiller® method. In Sec. 5.5 we discuss Los Angeles, CA, a city where our proposed model performs relatively poorly. Finally, in Sec. 5.6, we investigate why the autoregressive model performs better than existing repeat sales models.

5.1 Parameter Estimates

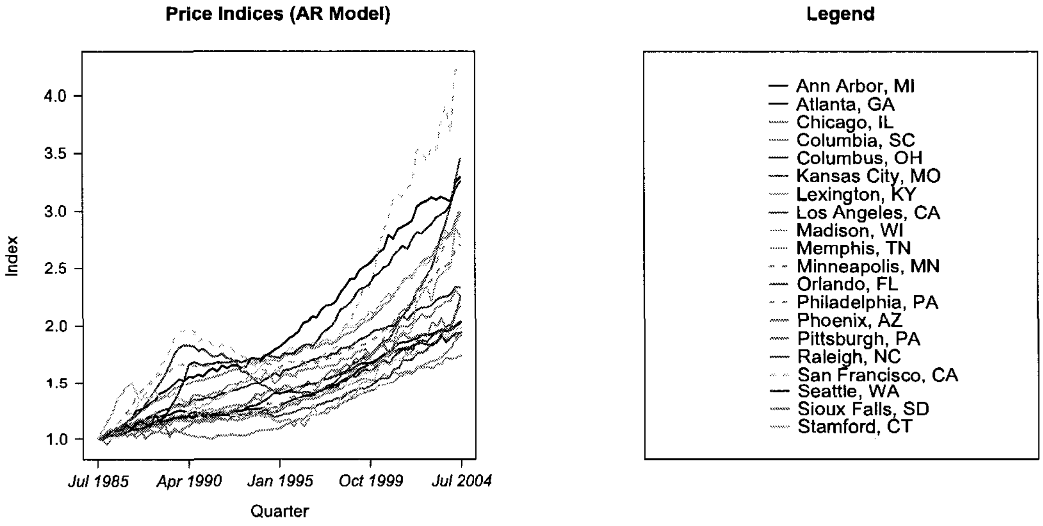
Table 5.1 lists the estimated values for the autoregressive coefficient (ϕ), a component of the error variance $\left(\sigma_{\varepsilon}^{2}\right)$ and their asymptotic variance estimates. The magnitude of the variance estimates for the $\boldsymbol{\beta}$ are a bit larger to those for ϕ and σ_{ε}^{2} and are computed using the observed information matrix (see Sec. 3.4.1). There are two features to note here. First, the variance estimates are quite small; this is because the data sets are large. The second feature is that $\hat{\phi}$ for each area is extremely close to one. Naturally, after subtracting the log index (β_t), the adjusted log prices for homes with a short gap time are expected to be closer than those with a longer gap time.

In Fig. 5.1, the price indices for all cities are plotted together. There is considerable variation among cities in the sample period. The general trend is upwards; however, for several areas such as San Francisco, CA, Los Angeles, CA, and Seattle, WA there is a downward trend during the mid 1990s. In Sec. 5.5, we will be examining the California metropolitan areas more closely. On the relative scale, San Francisco, CA home prices rose the most whereas those of Memphis, TN grew the least.

Table 5.1: Estimated Values of ϕ and σ^2_ε

Metropolitan Area	$\hat{\phi}$ (s.e.)	$\hat{\sigma}^2_\varepsilon$ (s.e)
Ann Arbor, MI	0.995698 (6.8689e-5)	0.001459 (2.1255e-5)
Atlanta, GA	0.995202 (3.2987e-5)	0.001548 (9.6582e-6)
Chicago, IL	0.995785 (2.1285e-5)	0.001333 (6.1202e-6)
Columbia, SC	0.997746 (1.0025e-4)	0.000832 (3.2164e-5)
Columbus, OH	0.996291 (4.0251e-5)	0.001165 (1.0296e-5)
Kansas City, MO	0.996293 (4.7568e-5)	0.001350 (1.5929e-5)
Lexington, KY	0.997196 (5.7846e-5)	0.000897 (1.6753e-5)
Los Angeles, CA	0.991267 (5.7383e-5)	0.002011 (1.2168e-5)
Madison, WI	0.995800 (7.8466e-5)	0.001011 (1.7210e-5)
Memphis, TN	0.996709 (5.7500e-5)	0.000952 (1.4875e-5)
Minneapolis, MN	0.994855 (4.0850e-5)	0.001324 (9.6241e-6)
Orlando, FL	0.995203 (6.2230e-5)	0.001639 (1.9384e-5)
Philadelphia, PA	0.996269 (1.5454e-5)	0.001515 (4.9199e-6)
Phoenix, AZ	0.995439 (4.8433e-5)	0.001456 (1.4099e-5)
Pittsburgh, PA	0.994818 (6.7854e-5)	0.002387 (2.8578e-5)
Raleigh, NC	0.995582 (5.7093e-5)	0.001285 (1.4993e-5)
San Francisco, CA	0.990293 (1.9460e-4)	0.001829 (3.4776e-5)
Seattle, WA	0.992424 (6.9650e-5)	0.001610 (1.3586e-5)
Sioux Falls, SD	0.996416 (1.3851e-4)	0.001019 (3.5962e-5)
Stamford, CT	0.992518 (3.0468e-4)	0.002169 (8.2365e-5)

Figure 5.1: Indices from the Autoregressive Model for All Metropolitan Areas



5.2 The Asymptotic Variance Assumption

In Sec. 3.4.1, we describe using the observed information matrix to compute the asymptotic variances. We determine whether this is appropriate using simulated data. The procedure used to simulate the data is outlined below.

Data Simulation Algorithm

1. Set the parameters $\{\beta_1 \ldots, \beta_T, \sigma_\varepsilon^2, \phi\}$ and fix the maximum number of sales M.
2. For each of I houses,
   (a) Select the number of sales J_i from a discrete uniform distribution: $J_i \sim U(1, M)$.
   (b) Select, without replacement, J_i values from $1, \ldots, T$. These are the quarters when sales occur: $t(i,1), \ldots, t(i,J_i)$ where $t(i,j)$ denotes the time of the jth sale of house i.
   (c) Compute the gap time $\gamma(i,j)$ between each sale of house i only if $J_i > 1$. Recall, gap time is: $\gamma(i,j) = t(i,j) - t(i,j-1)$.
   (d) Simulate the random variations for sales $1, \ldots, J_i$:
      i. Simulate: $\varepsilon_{i,1} \sim \mathcal{N}\left(0, \frac{\sigma_\varepsilon^2}{1-\phi^2}\right)$.
      ii. Simulate: $\varepsilon_{i,j} \sim \mathcal{N}\left(0, \frac{\sigma_\varepsilon^2(1-\phi^{2\gamma(i,j)})}{1-\phi^2}\right)$ when $j > 1$.
   (e) Finally, construct the series $y_{i,j}$:
      i. For $j = 1$: $y_{i,1} = \beta_{t(i,j)} + \varepsilon_{i,1}$.
      ii. If $J_i > 1$, for $j > 1$: $y_{i,j} = \beta_{t(i,j)} + \phi^{\gamma(i,j)}\left(y_{i,j-1} - \beta_{t(i,j-1)}\right) + \varepsilon_{i,j}$.

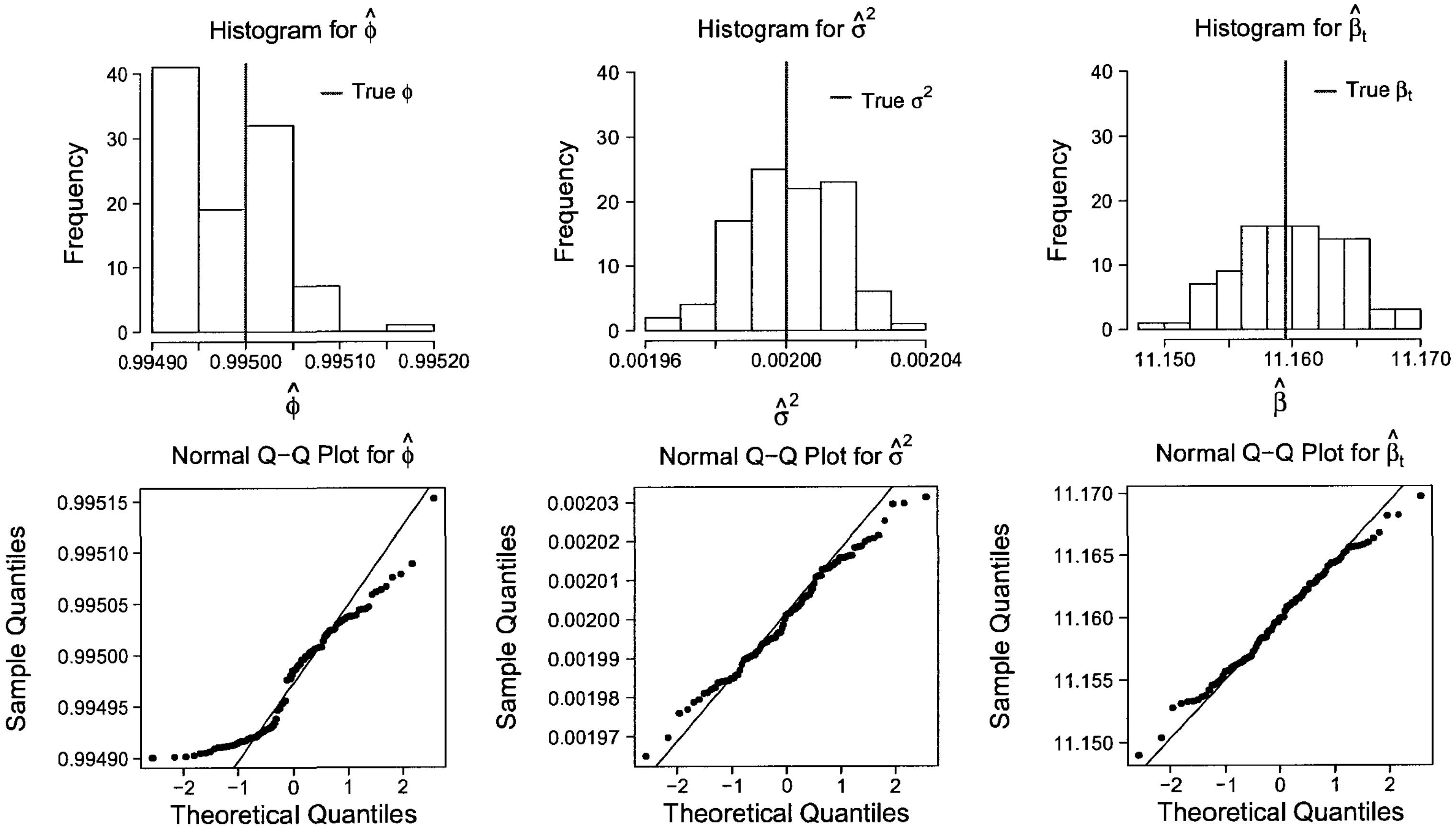
Table 5.2: Comparing Simulated and Expected Results

Parameter	True Value	Mean of Estimates	SD of Estimate	Mean SE from $\left[\hat{\mathbf{I}}(\hat{\boldsymbol{\theta}})\right]^{-1}$
ϕ	0.995	0.994978	5.642e-5	4.494e-5
σ_ε^2	0.002	0.002000	1.398e-5	1.1987e-5
β_t	11.15942	11.15992	4.244e-3	3.634e-3

100 data sets are simulated for the experiment each with 100,000 observations and an average of 40,000 homes. The maximum number of sales M for a single house was set at four. There were 70 quarters of sales where β ranged from 10 to 20, $\phi = 0.995$ and $\sigma_\varepsilon^2 = 0.002$. Table 5.2 provides the results for ϕ, σ_ε^2 and a selected (typical) β_t. We see the that the average parameter estimate is extremely close to the true value; this is also evident in Fig. 5.2 where the estimates from the 100 simulations are plotted. We conclude that the MLE estimates are virtually unbiased for this setting. The second set of columns in Table 5.2 show the standard error estimates. We compare the standard deviation of the parameters across the simulations with the average standard error estimate computed from the observed information matrix. These sets of values are also quite close.

The final check is whether the parameter estimates are normally distributed. This seems like a reasonable assumption for σ_ε^2 and β_t based on the normal quantile plots in Fig. 5.2. For ϕ, however, this does not seem to be the case. In some sense, this is not surprising given that ϕ is so close to 1; for values of ϕ that are less extreme, the normality assumption may still apply.

Figure 5.2: Simulation Results



5.3 The AR(1) Assumption

The model assumes that the quarter-adjusted log prices, $w_{i,j} = y_{i,j} - \beta_{t(i,j)}$, follow a latent AR(1) time series. Thus, for each gap time, $\gamma(i,j) = h$, there is a different correlation between the sale pairs, namely ϕ^h. To check that the data supports the theory, we compare the correlation between pairs of quarter-adjusted log prices at each gap length.

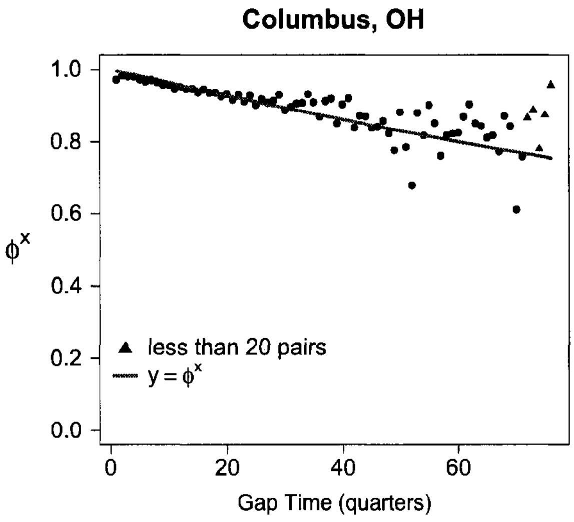
First, we compute the estimated adjusted prices $\hat{w}_{i,j} = y_{i,j} - \hat{\beta}_{t(i,j)}$. Next, for each gap time h, we find all the sale pairs $(\hat{w}_{i,j-1}, \hat{w}_{i,j})$ with that particular gap length. The sample correlation between those sale pairs provides us with an estimate for 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 ϕ^h if the model is specified correctly.

A sample plot for Columbus, OH is shown in Fig. 5.3. To compare, the estimated $\hat{\phi}^h$ curve is plotted as well. Gap lengths with less than 20 sale pairs are denoted with the triangle symbol. The relationship between ϕ and gap time seems to hold moderately well for this city. Plots for all of the metropolitan areas can be found in Appendix D (Figs. D.1- D.4). Not all metropolitan areas seem to have the desired relationship, however. Of particular note is Los Angeles, CA which will be discussed in further detail at the end of this chapter.

5.4 Model Validation

In this section, we investigate the predictive accuracy of our model as compared to others. For this purpose, the observations for each city are divided into training and test sets. The test set contains all final sales for homes that sell three or more times. Among homes that

Figure 5.3: ϕ vs Gap Time (Columbus, OH)



sell twice, the second sale is added to the test set with probably 1/2. As a result, the test set for each city contains around 15% of the observations. The remaining sales (including single sales) comprise the training set. The actual training and test set sizes for each city are given in Table A.3.

5.4.1 Competing Models

We will now describe three alternate models: a fixed effects model, a mixed effects model, and the established S&P/Case-Shiller® model. The first two models are considered simple, benchmark models to evaluate performance. The third is a commercial method used by Standard and Poor's.

Fixed Effects Model

This is a two-way fixed effects model. The expected log price $E[y_{i,j}]$ is modeled as the sum of a house effect (α_i), a quarter effect (β_t), and overall mean μ. Let i be the subscript for house $(1, \ldots, I)$, j be the sale index $(1, \ldots, J_i)$ and t denote the quarter $(1, \ldots, T)$. The form of the model is:

$$y_{i,j} \quad = \quad \mu + \alpha_i + \beta_{t(i,j)} + \varepsilon_{i,j} \tag{5.1}$$

where $\varepsilon_{i,j} \overset{iid}{\sim} \mathcal{N}\left(0, \sigma_\varepsilon^2\right)$. Given the large data size, ordinary least squares methods are not feasible and the residual regression method is used instead for fitting. Details are given in Appendix C.1. After estimating the parameters, an index can be constructed as follows:

$$1, \ \exp\left\{\hat{\beta}_2 - \hat{\beta}_1\right\}, \ \exp\left\{\hat{\beta}_3 - \hat{\beta}_1\right\}, \ldots, \ \exp\left\{\hat{\beta}_T - \hat{\beta}_1\right\}. \tag{5.2}$$

An adjustment to improve efficiency applied when converting the log indices to the price scale has a negligible effect; therefore, it is dropped from the calculations (see Sec. 3.4.3 for more details). Finally, the estimates of house prices can be converted to the price scale by exponentiating the fitted value plus an adjustment for efficiency. Details can be found in Sec. 3.4.

Mixed Effects Model

The second model is a modified version of the two-way fixed effects model. The house effects (α_i) are now modeled as random while the quarter effects (β_t) and mean μ remain as fixed parameters. The resulting model is:

$$y_{i,j} \quad = \quad \mu + \alpha_i + \beta_{t(i,j)} + \varepsilon_{i,j} \tag{5.3}$$

where $\alpha_i \overset{iid}{\sim} \mathcal{N}\left(0, \sigma_\alpha^2\right)$ and $\varepsilon_{i,j} \overset{iid}{\sim} \mathcal{N}\left(0, \sigma_\varepsilon^2\right)$ for houses i from $1, \ldots, I$ and sales j from $1, \ldots, J_i$. This is a standard two-way mixed effects model.

The parameters to be estimated are: $\boldsymbol{\beta}$, σ_ε^2, and σ_α^2. To obtain predictions, we compute estimates for the random effects, $\boldsymbol{\alpha}$, using the Best Linear Unbiased Predictor (BLUP), which is a plug-in estimator. The formula assumes that the variance components, σ_ε^2 and σ_α^2, are known; however, we will use the estimated values. Let $\mathbf{X}$ and $\mathbf{W}$ be the design matrices for the fixed and random effects respectively and $\mathbf{y}$ the response vector. Using Robinson's notation (1991) we can write the variance of random effects and error as follows:

$$Var\begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\varepsilon} \end{bmatrix} = \begin{bmatrix} \frac{\sigma_\alpha^2}{\sigma_\varepsilon^2}\mathbf{I}_I & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_N \end{bmatrix} \sigma_\varepsilon^2 \tag{5.4}$$

where $\mathbf{I}$ is the identity matrix and I and N are the number of houses and observations respectively. To obtain estimates of $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$, iterate between the following equations:

$$\hat{\boldsymbol{\beta}} = \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\left(\mathbf{y} - \mathbf{W}\hat{\boldsymbol{\alpha}}\right) \tag{5.5}$$

$$\hat{\boldsymbol{\alpha}} = \left(\mathbf{W}'\mathbf{W} + \frac{\sigma_\varepsilon^2}{\sigma_\alpha^2}\mathbf{I}_I\right)^{-1}\left(\mathbf{W}'\mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{Z}'\mathbf{y}\right). \tag{5.6}$$

For a full derivation, see Appendix E.2.2.

We have included this model as it is a natural extension of the model in (5.1). Moreover, random effects models exploit shrinkage techniques. When simultaneously estimating many means, it is often advantageous to shrink estimates in order to obtain better estimates. To fit this model, the `R` package `lme4` was used. Adding the random effects component adds significantly to the computing time. The index is constructed according to (5.2) and the method of converting the fitted values into the price scale can be found in Sec. 3.4.

S&P/Case-Shiller®

This model was outlined in Sec. 4.2.3. For all traditional repeat sales methods, model fitting requires the manipulation of large, sparse matrices. To ease computation, the conjugate gradient algorithm is used to solve linear systems quickly. Appendix C.2 describes this algorithm. Unlike the commercial version, we will compute a quarterly index. Moreover, the commercial procedure weights observations by their original sale price along with the weights computed in the second stage of the procedure. We omit this additional weighting scheme in our computations. Finally, we do not screen homes for signs of significant renovation or degradation as we have no means of doing so.

An advantageous feature of repeat sales indices is that it automatically creates an index where the fist time period is the base period. The price index for the first period, thus, is always one. To change the base period, divide each price index by the index of the new base period.

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

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

where $Y_{i,j}$ is the price of the jth sale of the ith house and B_t is the price index at time t. Note that this model is fitted on the *price* scale and not the log price scale. Finally, training set sizes for this model are given in Table A.3.

5.4.2 Prediction

To compare predictive performance, the root mean squared error ($RMSE$) is computed for predictions on the test set. The RMSE provides us with an estimate of the precision of our predictions; smaller values imply a higher value of precision. This quantity is computed using the equation below:

$$RMSE = \sqrt{\frac{1}{N}\sum_{k=1}^{N}\left(Y_k - \hat{Y}_k\right)^2} \tag{5.7}$$

where N is the number of observations in the test set, Y_k is the observed price, and $\hat{Y}_k$ is the predicted price. We hope that the autoregressive model will perform the best since we are taking into account the time series aspect of the data. Furthermore, we expect the random effects model to perform well, and the remaining two models to perform similarly. The random effects model should be performing better than the fixed effects model simply because we are allowing the house-specific estimates to shrink towards the overall mean. Shrinking parameter estimates is generally beneficial when there are several means to estimate; for these data, we have thousands! As long as the variability among the homes is greater than the variability among sales of a specific home, this hypothesis should hold. Neither the S&P/Case-Shiller® nor the fixed effects model allows for shrinking.

In Table 5.3, the RMSE values for the test set are provided for each of the four models. The method with the lowest RMSE value is highlighted in bold font for each city. Note that the S&P/Case-Shiller® RMSE is missing for Kansas City, MO because during the second stage of analysis, some of the computed weights were negative which prevented the procedure from continuing (RMSE was 8,842,578,554 for the second step due to an astoundingly large number of outliers).

For seventeen out of the twenty areas, the autoregressive model provides the lowest RMSE

Table 5.3: Test Set RMSE for Global Models (in dollars)

Metropolitan Area	AR	S&P/C-S	Mixed Effects	Fixed Effects
Ann Arbor, MI	**44,362**	52,718	48,332	54,827
Atlanta, GA	**33,977**	35,482	36,205	37,245
Chicago, IL	**39,201**	42,865	42,090	43,405
Columbia, SC	**36,376**	42,301	38,545	42,978
Columbus, OH	**27,651**	29,863	29,840	31,553
Kansas City, MO	**24,963**	—	26,114	28,369
Lexington, KY	**21,501**	21,731	21,699	21,879
Los Angeles, CA	41,006	41,951	**40,489**	41,484
Madison, WI	**28,687**	30,640	30,573	32,361
Memphis, TN	**25,069**	25,267	25,589	25,838
Minneapolis, MN	**33,233**	34,787	34,535	36,847
Orlando, FL	**29,317**	30,158	30,727	31,525
Philadelphia, PA	**34,811**	35,692	35,410	35,822
Phoenix, AZ	30,231	**29,350**	30,268	29,695
Pittsburgh, PA	**26,507**	30,135	28,772	31,473
Raleigh, NC	**26,563**	26,775	27,632	28,141
San Francisco, CA	50,777	50,249	**49,238**	50,429
Seattle, WA	**42,329**	43,486	43,513	44,864
Sioux Falls, SD	**20,190**	21,577	21,231	22,503
Stamford, CT	**61,805**	68,132	62,079	66,399

values. The random effects model performs better for Los Angeles, CA and San Francisco, CA whereas the S&P/Case-Shiller® performs best only for Phoenix, AZ. In the case of Phoenix, AZ both the autoregressive model and the random effects model perform less well.

5.4.3 The Regression Assumptions

There are number of assumptions we have placed on the error term especially to ensure that the model is stationary for the autoregressive model. In this section, we will examine a variety of residual plots for all four methods to see how well the assumptions are satisfied. All of the plots in this section are for Columbus, OH. To start, we plot a residual by predicted plot along with a residual by gap time plot in Fig. 5.4. In these set of plots, 5% of the observations in the training set have been randomly selected for plotting. Although the S&P/Case-Shiller® model is computed on the price scale, residuals on the log scale are given for comparison purposes for this figure only.

In Fig. 5.4, we see that the fixed and mixed effects models have the narrowest band of residuals. This is because both of these models have a house effect component and we are looking at the training set residuals. As we saw in the previous section, despite the promising look of this plot, neither of these models performed better than the autoregressive model overall in a predictive sense.

In Figs. 5.5-5.7, the residuals are checked for normality. Both a normal quantile plot and a histogram of the residuals are shown. For the two repeat sales models, several gap times are chosen and the residuals from those gap times are plotted as the variance of the residual changes with gap time for both models. None of the four models, unfortunately, seem to satisfy the normality assumption well.

The final set of plots in Fig. 5.8 are plots of the variance of the residuals for each gap length. The curve added to the plot is the expected variance based upon the model fitted. For the autoregressive model, the expected variance is a function of the gap length: $\frac{\sigma_\varepsilon^2(1-\phi^{2\gamma(i,j)})}{1-\phi^2}$. For the mixed effects and fixed effects models, the residuals are expected to be homoscedastic. This constant variance is estimated by computing the mean squared error of the predictions. The S&P/Case-Shiller® method assumes that residuals have variance $2\sigma_\varepsilon^2 + \gamma(i,j)\sigma_v^2$ where σ_v^2 is the variance attributed to the Gaussian random walk. The two variance components are estimated when running the second stage of the method. For Columbus, OH, $\sigma_\varepsilon^2 = -248,918$ and $\sigma_v^2 = 9,179,081$. Note that σ_ε^2 was actually estimated to be negative. The method does not prevent such a situation from happening.

The variance of the residuals clearly increases with gap time so the fixed effects and mixed effects models fail to capture this feature of the data. The S&P/Case-Shiller® also does not capture the trend in the variance although it does expect one to exist. The autoregressive model is best among the four models at describing the variance of the residuals.

Figure 5.4: Residual Plots for Columbus, OH (log scale)

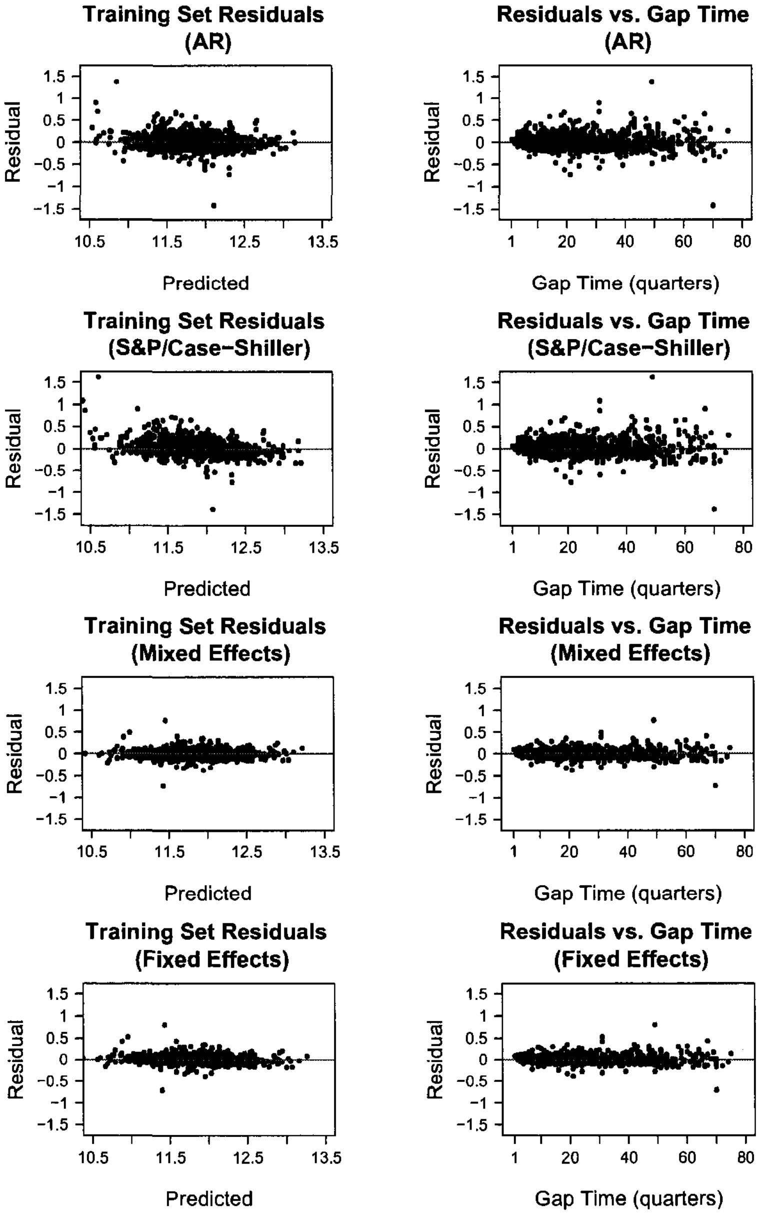


Figure 5.5: Normality of Residuals for Autoregressive Model (Columbus, OH)

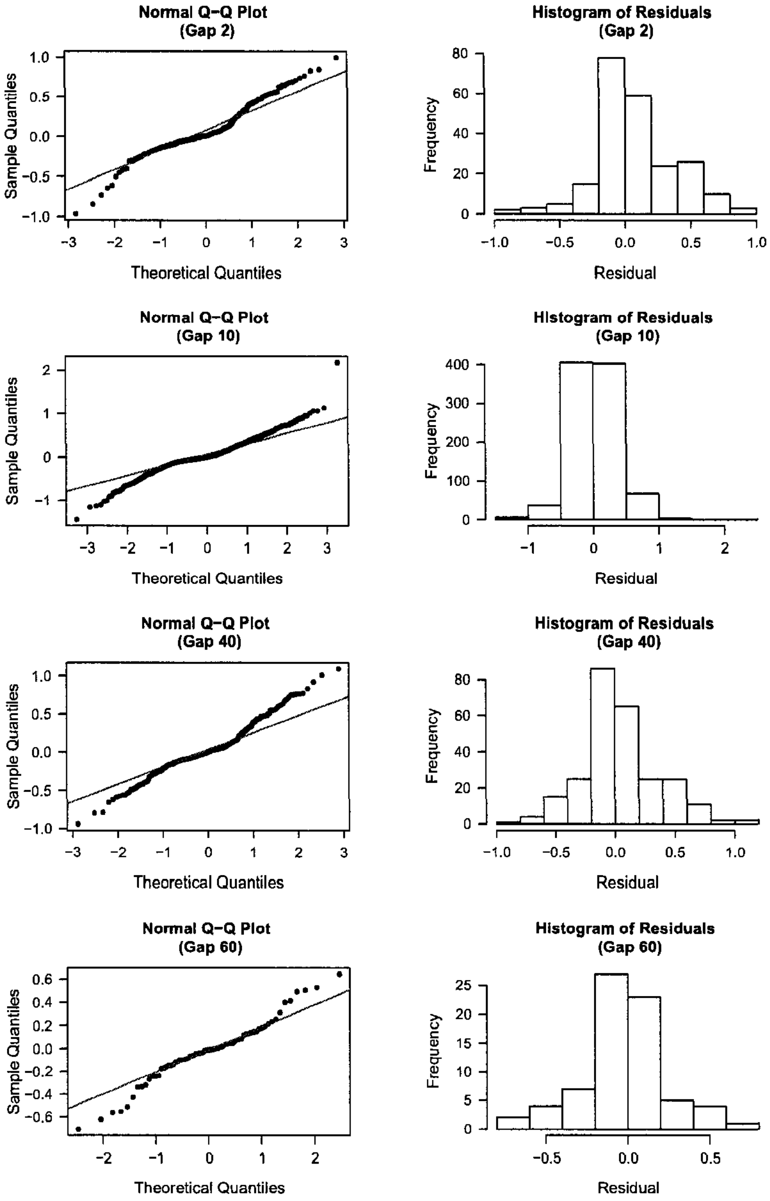


Figure 5.6: Normality of Residuals for S&P/Case-Shiller® (Columbus, OH)

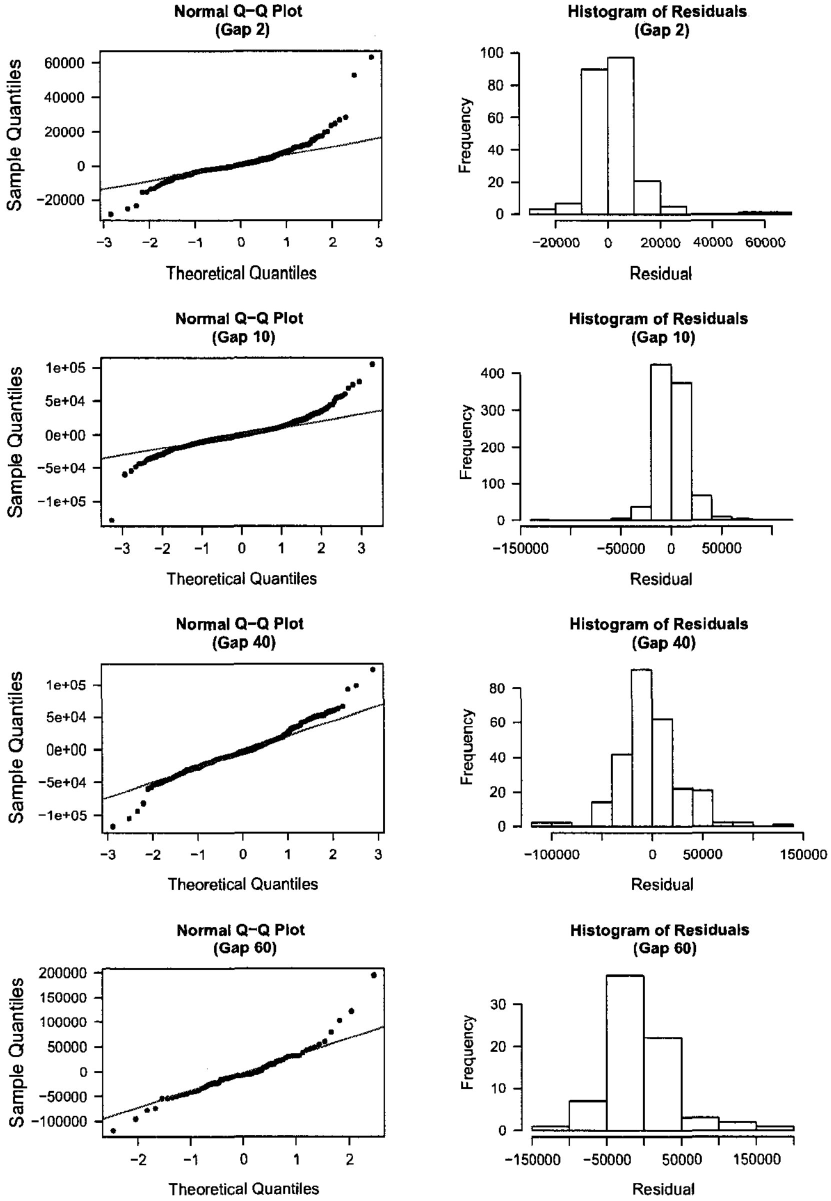


Figure 5.7: Normality of Residuals for Fixed and Mixed Effects Models (Columbus, OH)

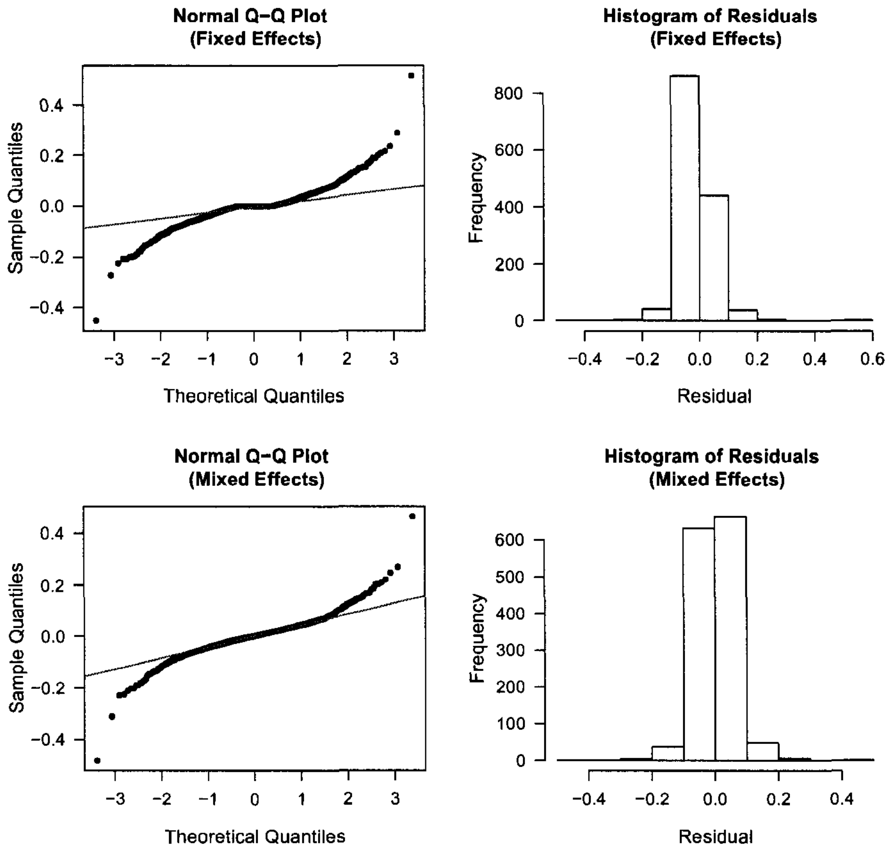
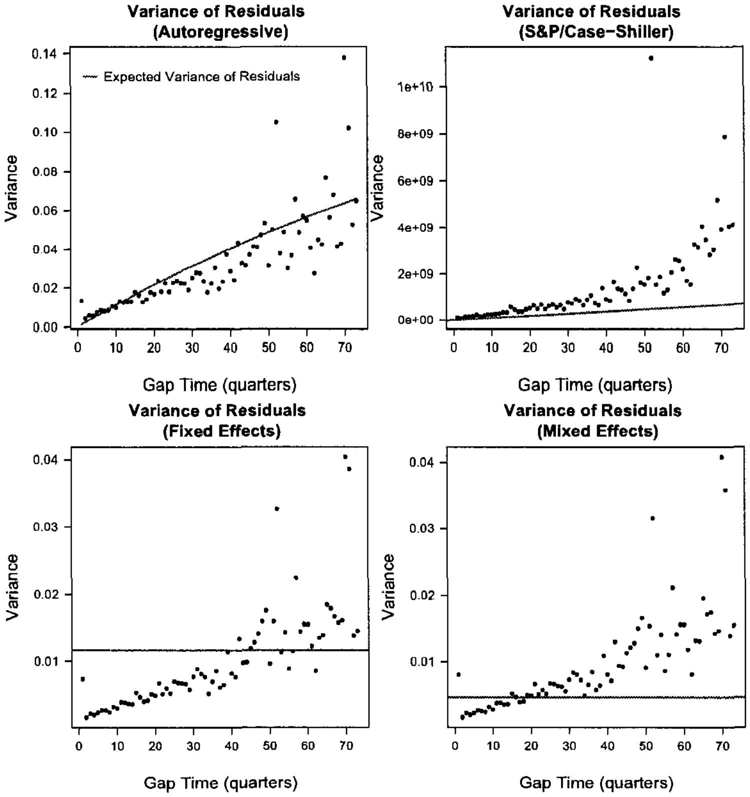


Figure 5.8: Variance of Residuals for Competing Models (Columbus, OH)



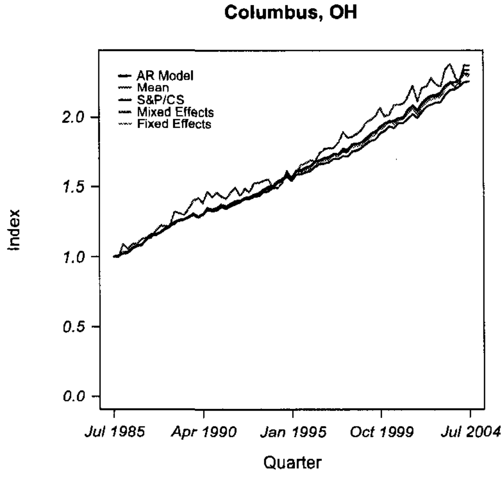
5.4.4 Comparing Indices

Each model can be used to construct a price index. To this collection, we add one more: the mean index. This is computed simply by taking the average price for each quarter and rescaling so that the base value is one (as given in Sec. 3.4). In Fig. 5.9, we plot all five indices for Columbus, OH. A complete set of indices is available in Appendix D in Figs. D.5- D.8. Observe that for smaller metropolitan areas, such as Sioux Falls, SD, where there are fewer observations per quarter, the indices are more volatile.

As we saw in Fig. 4.1, the indices plotted in Fig. 5.9 track each other even though the level of the indices are not the same. That is, they follow the same trends. In general, our autoregressive model index is between the mean index and the computed S&P/Case-Shiller® index. The mean index treats every sale as a single sale. This also explains why the mean index is the most volatile as information is not shared across houses and time periods. On the contrary, the S&P/Case-Shiller® index only uses repeat sales observations ignoring all single sales. The autoregressive model, however, is essentially a weighted average of both single sales and repeat sales. On the premise that we know more about repeat sales homes than those homes that have only sold once, repeat sales homes are weighted more heavily.

The commercial S&P/Case-Shiller® is computed for a number of MSAs across the US. Among the cities in our data, Atlanta, GA, Chicago, IL, Los Angeles, CA, Minneapolis, MN, San Francisco, CA, and Seattle, WA are also computed commercially. To compare the S&P/Case-Shiller® calculated using the available data and the published index, see Fig. 5.10. This plot shows the two indices for Chicago, IL. Note that the base year has been changed to January 2000. The two indices are quite similar although there are some variations most likely accounted by differences in the available data. Recall that the house sales used in this analysis are only those approved for conventional mortgages. However, given that there is not a staggering difference between the two Case-Shiller indices, we

Figure 5.9: Comparing Indices (Columbus, OH)



believe that the conclusions obtained from our analysis are applicable to the overall housing market.

5.5 A Closer Look at Los Angeles, CA

As seen in Table 5.3, the mixed effects model has a lower test RMSE than the autoregressive model for Los Angeles, CA. Why does the proposed model perform poorly for this city? If we examine Fig. 5.11, a plot of the correlation against gap time, we immediately see two significant issues. Recall that we expect ϕ 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.

To explain this feature, we look at sale pairs with gap times between 1 and 5 quarters more closely. In Fig. 5.12, 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

Figure 5.10: Actual vs. Computed S&P/Case-Shiller® Index

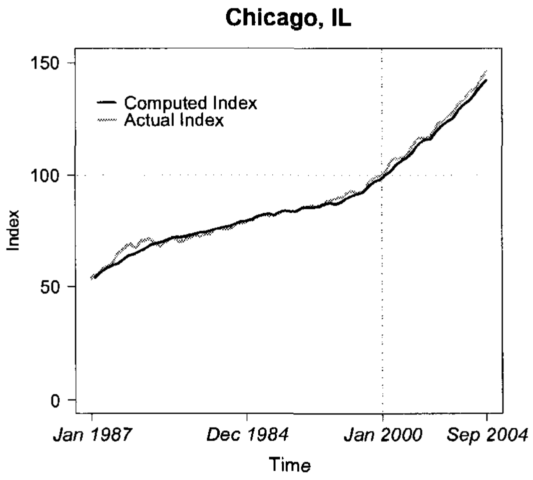
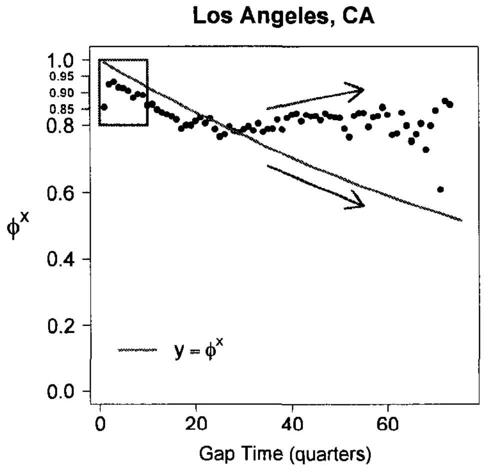


Figure 5.11: Problems with Assumptions



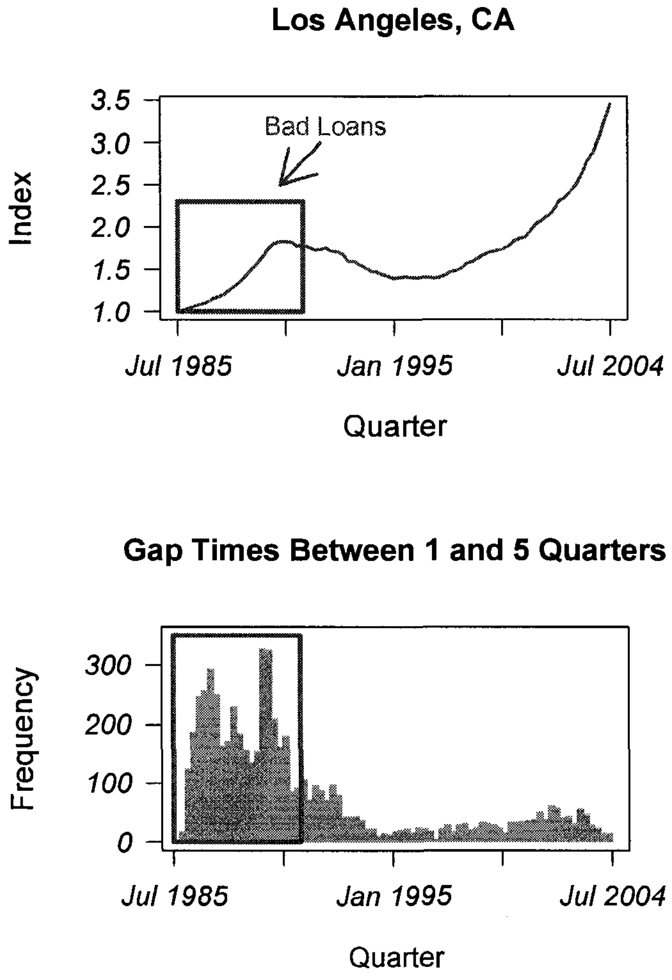
for Los Angeles, CA. Most of these sales occurred during the late 1980s and early 1990s. This corresponds to the same period when lenders were offering people mortgages where the monthly payment was *greater* than 33% of their monthly income [43]. The threshold of 33% is set to help ensure that people will be able to afford their mortgage. Those people with mortgages that exceede this percentage have a higher probability of defaulting on their payments. A number of banks including 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.

The second problem with Fig. 5.11 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 encourages people to stay in their homes and not move especially if they have owned their home for a long time [15, p. 124]. It is possible that this feature of Fig. 5.11 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. Nevertheless, these two issues lead us to conclude that the proposed autoregressive model is not a good description for home sales in Los Angeles, CA.

5.6 Comparing Repeat Sales Indices

In Sec. 5.4, we showed that the autoregressive model has the best overall predictive performance compared to the S&P/Case-Shiller® model. In this section, we investigate why

Figure 5.12: Examining the Housing Downturn



this may be the case. In particular, we want to show that the proposed autoregressive model is truly superior to existing repeat sales methodology and not just because of the methods used to fit the model. Specifically, we want to exclude the following four issues from explaining the difference in results:

1. Using price instead of log price. For skewed variables such as income, models are generally fitted on the log scale. House prices follow a similar asymmetrical distribution which can be seen in Fig. 5.13. The top plot is a histogram of prices for Stamford, CT for all home sales in the second quarter of 1997; the second is the same but for log prices. It is clear that the logarithmic transformation creates a more symmetric distribution.
2. Modeling price differences instead of price.
3. Weighting the observations with estimated standard deviations instead of estimated variances. Recall that the Case-Shiller methods have regression coefficient estimates which are not BLUE because of the weights chosen for each observation.
4. Modeling the sale series as a random walk.

A simple way of exploring these questions is to compare the autoregressive (AR), BMN, Case-Shiller (C-S), and S&P/Case-Shiller® (S&P/C-S) models in terms of predictive performance. In Table 5.4, we provide the test RMSE results for each method. The lowest RMSE value for each city is given in bold font. There are two features to note here. First, the three existing repeat sales methods have RMSE values which are quite similar to each other. Hence, the "improvements" made to the BMN model by the two Case-Shiller methods seem to result in only minor changes to the RMSEs. Second, and more importantly, the autoregressive model performs better than the other three models for eighteen out of the twenty cities. We find, therefore, that the autoregressive process itself is the key to improving repeat sales methodology.

Figure 5.13: Prices Versus Log Prices

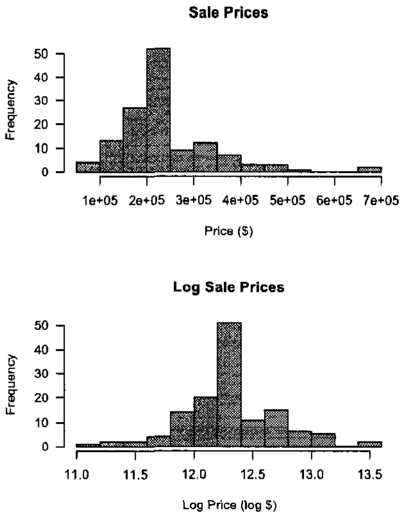


Table 5.4: Test Set RMSE for Global Models (in dollars)

Metropolitan Area	AR	BMN	C-S	S&P/C-S
Ann Arbor, MI	**44,362**	53,709	53,914	52,718
Atlanta, GA	**33,977**	35,456	35,494	35,482
Chicago, IL	**39,201**	42,923	42,960	42,865
Columbia, SC	**36,376**	42,207	42,263	42,301
Columbus, OH	**27,651**	30,176	30,196	29,863
Kansas City, MO	**24,963**	27,682	27,724	—
Lexington, KY	**21,501**	21,748	21,740	21,731
Los Angeles, CA	**41,006**	41,918	41,949	41,951
Madison, WI	**28,687**	30,979	30,942	30,640
Memphis, TN	**25,069**	25,311	25,306	25,267
Minneapolis, MN	**33,233**	35,402	35,538	34,787
Orlando, FL	**29,317**	30,187	30,215	30,158
Philadelphia, PA	**34,811**	35,567	35,637	35,692
Phoenix, AZ	30,231	**29,295**	29,334	29,350
Pittsburgh, PA	**26,507**	30,732	30,812	30,135
Raleigh, NC	**26,563**	26,873	26,856	26,775
San Francisco, CA	50,777	50,513	50,573	**50,249**
Seattle, WA	**42,329**	43,533	43,606	43,486
Sioux Falls, SD	**20,190**	21,527	21,576	21,577
Stamford, CT	**61,805**	67,661	67,668	68,132

Chapter 6

The Local Autoregressive Model

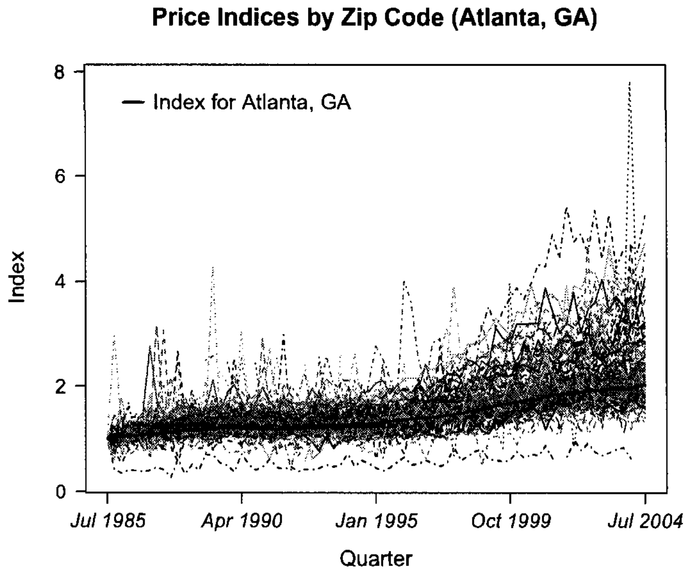
The autoregressive model proposed in Chapter 3 applies a single model for the entire metropolitan area. For larger cities such as Chicago, IL and Atlanta, GA this assumption may not hold; rather, there may be spatial effects. In this chapter, we attempt to capture such effects by introducing zip codes into the autoregressive model. As an initial test, we run separate autoregressive models for the zip codes in Atlanta, GA. The resulting indices are plotted in Fig. 6.1. There are clear differences among the indices. The next step, then, is to build a model which incorporates this variable.

We add zip code as a random effect into the existing model. The resulting model is:

$$\begin{aligned} y_{i,1,z} &= \mu + \beta_{t(i,1,z)} + \tau_z + \varepsilon_{i,1,z} & j &= 1 \\ y_{i,j,z} &= \mu + \beta_{t(i,j,z)} + \tau_z + \phi^{\gamma(i,j,z)} \left(y_{i,j-1,z} - \mu - \beta_{t(i,j-1,z)} - \tau_z \right) + \varepsilon_{i,j,z} & j &> 1 \end{aligned} \tag{6.1}$$

where $y_{i,j,z}$ is the jth log price of the ith house in zip code z. There are a total of $N = \sum_{z=1}^{Z} \sum_{i=1}^{I_Z} J_i$ observations in this model where there are Z zip codes, I_z houses in each zip code, and J_i sales for a given house. Let $\tau_z \overset{iid}{\sim} \mathcal{N}\left(0, \sigma_\tau^2\right)$ where $\tau_1, \ldots, \tau_Z$ are the zip code

Figure 6.1: Indices for All Metropolitan Areas



random effects. Let $\beta_1, \ldots, \beta_T$ denote the log price indices which remain as fixed effects. However, 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 μ as the overall mean. Finally, let $\varepsilon_{i,1,z} \overset{iid}{\sim} \mathcal{N}\left(0, \frac{\sigma_\varepsilon^2}{1-\phi^2}\right)$, $\varepsilon_{i,j,z} \overset{iid}{\sim} \mathcal{N}\left(0, \frac{\sigma_\varepsilon^2\left(1-\phi^{2\gamma(i,j,z)}\right)}{1-\phi^2}\right)$, and assume that all $\varepsilon_{i,j}$ are independent.

6.1 Time Series Models With Covariates

To provide some background, we now describe an algorithm proposed by Sargan (1964) for fitting a regression model with autocorrelated errors. For this setting, the general model to be fitted is:

$$\mathbf{y} \;=\; \mathbf{X}\beta + \varepsilon \tag{6.2}$$

where $\boldsymbol{\varepsilon}$ is a time series process with covariance matrix $\mathbf{V}$. For simplicity, assume that the errors follow an AR(1) process with autoregressive coefficient ϕ. Furthermore, assume we have three observations: $\varepsilon_1, \varepsilon_2, \varepsilon_3$. Then, $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{V})$ where $\mathbf{V}$ is:

$$\mathbf{V} = \frac{\sigma_\varepsilon^2}{1-\phi^2}\begin{bmatrix} 1 & \phi & \phi^2 \\ \phi & 1 & \phi \\ \phi^2 & \phi & 1 \end{bmatrix}$$

Sargan defines a *transformation matrix* $\mathbf{A}$ such that $\mathbf{AVA}' = \sigma_\varepsilon^2 \mathbf{I}_2$ where $\mathbf{I}_2$ is an identity matrix of dimension 2. By applying this transformation to the model, we obtain:

$$\mathbf{Ay} = \mathbf{AX}\boldsymbol{\beta} + \mathbf{A}\boldsymbol{\varepsilon} \tag{6.3}$$

The original regression with autocorrelated errors has now been reduced to an ordinary regression with iid errors. In our example, $\mathbf{A}$, as Sargan defines it, is:

$$\mathbf{A} = \begin{bmatrix} -\phi & 1 & 0 \\ 0 & -\phi & 1 \end{bmatrix}$$

Therefore, $\mathbf{A}\boldsymbol{\varepsilon} \sim \mathcal{N}\left(\mathbf{0}, \sigma_\varepsilon^2 \mathbf{I}_2\right)$. Applying this trick, Sargan outlines a procedure which iterates between estimating the autoregressive parameters and the regression parameters on the transformed data. If the time series process is stationary, the algorithm has been proved to converge to the maximum likelihood estimates as $N \to \infty$ [42, p. 293].

The key to this method is defining the appropriate transformation matrix $\mathbf{A}$. In our scenario we need to make two changes to Sargan's definition. The first is obvious: we must incorporate the random effects into the model. This complicates matters because the covariance matrix $\mathbf{V}$ is now comprised of two components: the random effects variance and the error term variance. Applying $\mathbf{A}$ to $\mathbf{V}$ will affect both components and we will not be able to reduce $\mathbf{V}$ to the identity matrix as before. In addition, the best linear unbiased estimates (BLUP) for the random effects need to be computed which adds a step to the fitting algorithm.

Second, this procedure only provides the maximum likelihood solution when the number of observations in the series is large. Sargan's definition of **A** makes this so. Note that the transformation matrix drops the first observation by construction. This is done to achieve the desirable property of $\mathbf{AVA}' = \sigma_\varepsilon^2 \mathbf{I}_{N-1}$. As the number of time periods increase, the first observation has an increasingly negligible effect on the parameter estimates. Thus, the parameter estimates become closer to the MLEs as the sample size, N, increases. However, since our price series for each house is short, we cannot afford to overlook this point. We describe a modified version of **A** next.

Suppose there are three sales of house i at time 1, 3, and 5. The covariance matrix for this series would be (if there were no random effects):

$$\mathbf{V} = \frac{\sigma_\varepsilon^2}{1-\phi^2} \begin{bmatrix} 1 & \phi^2 & \phi^5 \\ \phi^2 & 1 & \phi^3 \\ \phi^5 & \phi^3 & 1 \end{bmatrix}$$

If we retain the first observation, the transformation matrix would now be:

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 \\ -\phi^2 & 1 & 0 \\ 0 & -\phi^3 & 1 \end{bmatrix}$$

Applying the transformation to the data, we obtain:

$$\begin{aligned}\mathbf{AVA}' &= \frac{\sigma_\varepsilon^2}{1-\phi^2}\begin{bmatrix}1 & 0 & 0\\ 0 & 1-\phi^4 & 0\\ 0 & 0 & 1-\phi^6\end{bmatrix}\\ &= \frac{\sigma_\varepsilon^2}{1-\phi^2}\begin{bmatrix}1 & 0 & 0\\ 0 & 1-\phi^{2\gamma(i,2)} & 0\\ 0 & 0 & 1-\phi^{2\gamma(i,3)}\end{bmatrix}\end{aligned}$$

where $\gamma(i,j)$ is the gap time between the jth and $(j-1)$st sale of house i. Denote $\mathbf{AVA}' = \frac{\sigma_\varepsilon^2}{1-\phi^2} diag(\mathbf{r})$ where $diag(\mathbf{r})$ is a diagonal matrix of dimension N where

$$r_{i,j,z} = \begin{cases} 1 & \text{when } j = 1 \\ 1-\phi^{2\gamma(i,j)} & \text{when } j > 1 \end{cases} \tag{6.4}$$

where $r_{i,j,z}$ denotes the element of $diag(\mathbf{r})$ for the jth sale of the ith house in zip code z.

Thus, we have redefined $\mathbf{A}$ to include the first term of the series. A similar solution has been proposed by Prais and Winston (1954) for the AR(1) setting. Recall, for a stationary AR(1) process with a starting point, the first observation is given by $w_1 = \frac{1}{\sqrt{1-\phi^2}}\varepsilon_1$ where $\varepsilon_1 \overset{iid}{\sim} \mathcal{N}\left(0,\sigma_\varepsilon^2\right)$. Therefore, to accommodate the first observation, and preserve the reduction property: $\mathbf{AVA}' = \sigma_\varepsilon^2\mathbf{I}_N$, they propose multiplying the first observation by $\frac{1}{\sqrt{1-\phi^2}}$. Beach and MacKinnon (1978) point out, however, that this gives the MLE *only* when ϕ is already known. This is because the Prais and Winston procedure uses generalized least squares to fit both components of the model [3, p. 51]. For our purposes, however, we have no need for the reduction property to hold given the introduction of the random effects.

6.2 Introducing the Random Effects

We can now describe the model in (6.1) using matrices:

$$\mathbf{Ay} \;=\; \mathbf{AX}\boldsymbol{\beta} + \mathbf{AZ}\boldsymbol{\tau} + \mathbf{A}\boldsymbol{\varepsilon} \tag{6.5}$$

where $\mathbf{y}$ is log price, $\mathbf{X}$ and $\mathbf{Z}$ are the design matrices for the fixed effects $[\mu \;\; \beta_1 \cdots \beta_{T-1}]'$ and random effects $\boldsymbol{\tau}$ respectively. Let $\boldsymbol{\varepsilon}$ be the random variation, and $\mathbf{A}$ the transformation matrix. Now, $\mathbf{A}\boldsymbol{\varepsilon} \sim \mathcal{N}\left(\mathbf{0}, \frac{\sigma_\varepsilon^2}{1-\phi^2} diag(\mathbf{r})\right)$. From hereon, we let $\boldsymbol{\varepsilon}$ denote $\mathbf{A}\boldsymbol{\varepsilon}$ to simplify the notation. Furthermore, recall that we require $\sum_{t=1}^{T} n_t\beta_t = 0$ where n_t is the number of sales at time t. Therefore, $\beta_T = -\frac{1}{n_T}\sum_{t=1}^{T-1} n_t\beta_t$.

If we define $\boldsymbol{\theta} = \{\boldsymbol{\beta},\, \sigma_\varepsilon^2,\, \sigma_\tau^2,\, \phi\}$, the likelihood function for (6.5) is:

$$L(\boldsymbol{\theta};\, \mathbf{y}) \;=\; (2\pi)^{-N/2}|\mathbf{V}|^{-1/2} \exp\left\{-\frac{1}{2}(\mathbf{A}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}))'\mathbf{V}^{-1}(\mathbf{A}(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}))\right\} \tag{6.6}$$

where N is the total number of observations, $\mathbf{V}$ is the covariance matrix, and $\mathbf{A}$ is the transformation matrix. We can split $\mathbf{V}$ into a sum of the variance contributions from the time series and the random effects. Specifically,

$$\mathbf{V} = \frac{\sigma_\varepsilon^2}{1-\phi^2} diag(\mathbf{r}) + (\mathbf{AZ})\mathbf{D}(\mathbf{AZ})' \tag{6.7}$$

where $\mathbf{D} = \sigma_\tau^2\mathbf{I}_Z$.

The matrix $\mathbf{V}$ is an $N \times N$ matrix which, given the size of our data sets, is incredibly large making it nearly impossible to manipulate. Fortunately, $\mathbf{V}$ is a block-diagonal matrix where each block contains the observations of a given zip code.

As an example, let us use the sample data given in Table 6.1. We have four homes in

Table 6.1: Sample Data

House	Quarter	Zip
1	1	1
1	3	1
2	2	1
2	3	1
3	1	2
3	2	2
3	3	2
4	1	2
4	3	2

this data set, with sales over three time periods. The sales are divided into two zip codes. In this case, $\mathbf{V}$ has the form:

$$\mathbf{V} = \begin{bmatrix} \mathbf{V}_{1,1} & \mathbf{0} \\ \mathbf{0} & \mathbf{V}_{2,2} \end{bmatrix}$$

where $\mathbf{V}_{1,1}$ is:

$$\begin{aligned} \mathbf{V} &= \frac{\sigma_\varepsilon^2}{1-\phi^2} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1-\phi^4 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1-\phi^2 \end{bmatrix} \\ &+\sigma_\tau^2 \begin{bmatrix} 1 & 1-\phi^2 & 1 & 1-\phi \\ 1-\phi^2 & \left(1-\phi^2\right)^2 & 1-\phi^2 & (1-\phi)\left(1-\phi^2\right) \\ 1 & 1-\phi^2 & 1 & 1-\phi \\ 1-\phi & (1-\phi)\left(1-\phi^2\right) & 1-\phi & (1-\phi)^2 \end{bmatrix} \end{aligned}$$

This structure is important because the blocks of a block diagonal matrix can be treated as independent of each other. In particular, for a block diagonal matrix with $\mathbf{Z}$ blocks,

$$|\mathbf{V}| = \prod_{i=z}^{Z} |\mathbf{V}_{z,z}| \quad \text{and} \quad \mathbf{V}^{-1} = \begin{bmatrix} \mathbf{V}_{1,1}^{-1} & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots \\ \mathbf{0} & \cdots & \mathbf{V}_{Z,Z}^{-1} \end{bmatrix} \tag{6.8}$$

where $|\cdot|$ is the determinant of a matrix.

For the zth block (zth zip code), let $\mathbf{y}_z$ be the vector of log prices, $\mathbf{X}_z$ be the design matrix for the fixed, time effects, $\mathbf{A}_z$ be the transformation matrix, and $\mathbf{V}_{z,z}$ be the covariance matrix. If we reduce $\mathbf{V}$ to block form, where $\mathbf{V}_{z,z}$ denotes the block for zip code z, the likelihood function can be rewritten as follows:

$$\begin{aligned} l(\boldsymbol{\theta};\, \mathbf{y}) \;&=\; -\frac{N}{2}\log(2\pi) - \frac{1}{2}\sum_{z=1}^{Z} \log|\mathbf{V}_{z,z}| \\ &\quad -\frac{1}{2}\sum_{z=1}^{Z} (\mathbf{A}_z(\mathbf{y}_z - \mathbf{X}_z\boldsymbol{\beta}))'\mathbf{V}_{z,z}^{-1}(\mathbf{A}_z(\mathbf{y}_z - \mathbf{X}_z\boldsymbol{\beta})) \end{aligned} \tag{6.9}$$

We can reduce (6.7) into block form as well:

$$\mathbf{V}_{z,z} = \frac{\sigma_\varepsilon^2}{1-\phi^2} diag(\mathbf{r}_z) + \sigma_\tau^2 \left(\mathbf{A}_z \mathbf{1}_{n_z}\right)\left(\mathbf{A}_z \mathbf{1}_{n_z}\right)' \tag{6.10}$$

where n_z is the number of observations in zip code z.

6.3 Model Fitting

Just as for the global autoregressive model, we use the coordinate ascent algorithm to estimate the parametersalbeit with more complex equations. The derivations for updating the parameters are given in Appendix E.1. For $\boldsymbol{\beta}$, we can derive an explicit expression; for the remaining three parameters, we have to numerically compute the zero of the partial derivative functions. This procedure is given next:

Local AR Model Fitting Algorithm

1. Set a tolerance level ϵ (possibly different for each parameter) and a maximum number of iterations K.
2. Initialize the parameters: $\theta^0 = \left\{\boldsymbol{\beta}^0, \sigma_\varepsilon^{2,0}, \sigma_\tau^{2,0}, \phi^0\right\}$ (for details see end of section).
3. For iteration k,
   - (a) For $t \in \{1, \ldots, T\}$, calculate $\boldsymbol{\beta}^k$ using (6.11). That is, $\boldsymbol{\beta}^k = f\left(\sigma_\varepsilon^{2,k-1}, \sigma_\tau^{2,k-1}, \phi^{k-1}\right)$.
   - (b) Compute $\sigma_\varepsilon^{2,k}$ by computing the zero of (6.12) using $\left\{\boldsymbol{\beta}^k, \sigma_\tau^{2,k-1}, \phi^{k-1}\right\}$.
   - (c) Compute $\sigma_\tau^{2,k}$ by calculating the zero of (6.13) using $\left\{\boldsymbol{\beta}^k, \sigma_\varepsilon^{2,k}, \phi^{k-1}\right\}$.
   - (d) Find the zero of (6.14) to compute ϕ^k using $\left\{\boldsymbol{\beta}^k, \sigma_\varepsilon^{2,k}, \sigma_\tau^{2,k}\right\}$
   - (e) If $\left|\boldsymbol{\theta}_i^{k-1} - \boldsymbol{\theta}_i^k\right| > \epsilon$ for any $\theta_i \in \boldsymbol{\theta}$ and the number of iterations is less than K, repeat Step 3 after replacing $\boldsymbol{\theta}^{k-1}$ with $\boldsymbol{\theta}^k$. Otherwise, stop (call this iteration K' where $K' \leq K$).
4. Solve for β_T by computing: $\hat{\beta}_T = -\frac{1}{n_T}\sum_{t=1}^{T-1} n_t \hat{\beta}_t^{K'}$.
5. Plug in $\left\{\boldsymbol{\beta}^{K'}, \sigma_\varepsilon^{2,K'}, \sigma_\tau^{2,K'}, \phi^{K'}\right\}$ to compute the BLUPs using (6.15).

Before we provide the updating functions, define $\mathbf{w}_z = \mathbf{y}_z - \mathbf{X}_z\boldsymbol{\beta}$. We start with an explicit function for $\boldsymbol{\beta}$.

$$\hat{\boldsymbol{\beta}} = \left(\sum_{z=1}^{Z} (\mathbf{A}_z\mathbf{X}_z)' \mathbf{V}_{z,z}^{-1} \mathbf{A}_z\mathbf{X}_z\right)^{-1} \sum_{z=1}^{Z} (\mathbf{A}_z\mathbf{X}_z)' \mathbf{V}_{\mathbf{z},\mathbf{z}}{}^{-1} \mathbf{A}_z\mathbf{y}_z. \tag{6.11}$$

To update σ_ε^2, we must compute the zero of:

$$0 = -\sum_{z=1}^{Z} tr\left(\mathbf{V}_{z,z}^{-1} diag(\mathbf{r}_z)\right) + \sum_{z=1}^{Z} (\mathbf{A}_z\mathbf{w}_z)'\mathbf{V}_{z,z}^{-1} diag(\mathbf{r}_z)\mathbf{V}_{z,z}^{-1}(\mathbf{A}_z\mathbf{w}_z) \qquad (6.12)$$

where $tr(\cdot)$ is the trace of a matrix. Similarly, to update σ_τ^2, we need to find the zero of:

$$\begin{aligned} 0 \;=\; & \sum_{z=1}^{Z} tr\left(V_{z,z}^{-1}(\mathbf{A}_z\mathbf{1}_{n_z})(\mathbf{A}_z\mathbf{1}_{n_z})'\right) \\ & + \sum_{z=1}^{Z} -(\mathbf{A}_z\mathbf{w}_z)'\mathbf{V}_{z,z}^{-1}(\mathbf{A}_z\mathbf{1}_{n_z})(\mathbf{A}_z\mathbf{1}_{n_z})'\mathbf{V}_{z,z}^{-1}(\mathbf{A}_z\mathbf{w}_z). \qquad (6.13) \end{aligned}$$

Finally, to update the autoregressive parameter ϕ, we have to calculate the zero of the function:

$$\begin{aligned} 0 \;=\; & -\sum_{z=1}^{Z} tr\left\{\mathbf{V}_{z,z}^{-1}\left(\sigma_\tau^2\left(\frac{\partial(\mathbf{A}_z\mathbf{1}_{n_z})}{\partial\phi}\right)(\mathbf{A}_z\mathbf{1}_{n_z})' + \sigma_\tau^2(\mathbf{A}_z\mathbf{1}_{n_z})\left(\frac{\partial(\mathbf{A}_z\mathbf{1}_{n_z})}{\partial\phi}\right)'\right.\right. \\ & \left.\left. + \frac{2\phi\sigma_\varepsilon^2}{(1-\phi^2)^2} diag(\mathbf{r}_z) + \frac{\sigma_\varepsilon^2}{1-\phi^2}\frac{\partial diag(\mathbf{r}_z)}{\partial\phi}\right)\right\} \\ & -\sum_{z=1}^{Z}\left(\frac{\partial\mathbf{A}_z}{\partial\phi}\mathbf{w}_z\right)'\mathbf{V}_{z,z}^{-1}(\mathbf{A}_z\mathbf{w}_z) - \sum_{z=1}^{Z}(\mathbf{A}_z\mathbf{w}_z)'\mathbf{V}_{z,z}^{-1}\left(\frac{\partial\mathbf{A}_z}{\partial\phi}\mathbf{w}_z\right) \\ & +\sum_{z=1}^{Z}\left[(\mathbf{A}_z\mathbf{w}_z)'\mathbf{V}_{z,z}^{-1}\left[\sigma_\tau^2\left(\frac{\partial(\mathbf{A}_z\mathbf{1}_{n_z})}{\partial\phi}\right)(\mathbf{A}_z\mathbf{1}_{n_z})' + \sigma_\tau^2(\mathbf{A}_z\mathbf{1}_{n_z})\left(\frac{\partial(\mathbf{A}_z\mathbf{1}_{n_z})}{\partial\phi}\right)'\right.\right. \\ & \left.\left. + \frac{2\phi\sigma_\varepsilon^2}{(1-\phi^2)^2} diag(\mathbf{r}_z) + \frac{\sigma_\varepsilon^2}{1-\phi^2}\frac{\partial diag(\mathbf{r}_z)}{\partial\phi}\right]\mathbf{V}_{z,z}^{-1}(\mathbf{A}_z\mathbf{w}_z)\right]. \qquad (6.14) \end{aligned}$$

After the estimates converge, the final step is to estimate the random effects using the (BLUP) formulas:

$$\begin{aligned} \hat{\tau}_z \;=\; & \left[\frac{2\hat{\sigma}_\varepsilon^2}{\hat{\sigma}_\tau^2} + \left(\mathbf{1} - \hat{\phi}^2\right)\left(\hat{\mathbf{A}}_z\mathbf{1}_z\right)' diag^{-1}(\mathbf{r}_z)\left(\hat{\mathbf{A}}_z\mathbf{1}_z\right)\right]^{-1} \times \\ & \left(\left(1 - \hat{\phi}^2\right)\left(\hat{\mathbf{A}}_z\mathbf{1}_z\right)' diag^{-1}(\mathbf{r}_z)\left(\hat{\mathbf{A}}_z\hat{\mathbf{w}}_z\right)\right). \qquad (6.15) \end{aligned}$$

where $diag^{-1}(\hat{\mathbf{r}})$ is the inverse of the estimated diagonal matrix $diag(\mathbf{r})$. A derivation for this formula can be found in Appendix E.2.1.

Sample Initialization for Local Model

While any set of starting values can be used, it is faster to use the following initializations. We start with μ:

$$\mu^0 = \frac{1}{N}\sum_{z=1}^{Z}\sum_{i=1}^{I_z}\sum_{j=1}^{J_i} y_{i,j,z}.$$

Using μ^0, we can compute starting values for the remainder of the time effects:

$$\beta_t^0 = \frac{1}{n_t}\sum_{z=1}^{Z}\sum_{i=1}^{I_z}\sum_{j=1}^{J_i} \mathbb{I}_{t(i,j,z)=t}\left(y_{i,j,z} - \mu^0\right)$$

where $\mathbb{I}_{t(i,j,z)=t}$ is an indicator function denoting whether the observation $y_{i,j,z}$ occurred at time t.

To obtain an initial value for σ_τ^2, we first calculate estimates of the random effects:

$$\tau_z^0 = \frac{1}{n_t}\sum_{i=1}^{I_z}\sum_{j=1}^{J_i}\left(y_{i,j,z} - \mu^0 - \beta^0_{t(i,j,z)}\right)$$

where n_t is the number of observations at time t. Hence,

$$\sigma_\tau^{2,0} = Var\left(\boldsymbol{\tau}^0\right)$$

where $Var(\cdot)$ is the sample variance function.

Finally, let $v_{i,j,z} = y_{i,j,z} - \beta^0_{t(i,j,z)} - \tau_z^0$. For each gap time h, find all pairs $(x_{h,1} = v_{i,j-1,z},$

$x_{h,2} = v_{i,j,z}$) such that $\gamma(i,j) = h$. Let H be the maximum gap time. Then, estimate ϕ and σ_ε^2 by:

$$\begin{aligned}\phi^0 &= \frac{1}{H}\sum_{h=1}^{H}\left[Cor\left(\mathbf{x}_{h,1},\mathbf{x}_{h,2}\right)\right]^{\frac{1}{h}} \\ \sigma_\varepsilon^{2,0} &= Var(\mathbf{v})\end{aligned}$$

where $Cor(\cdot,\cdot)$ is the sample correlation function.

6.3.1 Convergence of the Fitting Algorithm

Given the structure of our model, it is unclear whether the MLEs exist and are unique and if the fitting algorithm will reach these values. We repeat the empirical analysis done in in Sec. 3.3.2 for the local model. The results from Stamford, CT are used as an example here.

In Fig. 6.2, we plot the log likelihood value for at step of the fitting process. As hoped for, the log likelihood value never decreases at any step. This indicates that the fitting algorithm is always heading towards parameter estimates that are more likely given the data.

As before, the parameter which may not converge to the MLE is ϕ. We plot for each ϕ, the log likelihood value if we maximize with respect to the remaining parameters. If the coordinate ascent algorithm converges to the MLE, it should match the maximum point on Fig. 6.2 which it does. Therefore, we are confident that the MLE is reached.

Figure 6.2: Plot of the Local Log Likelihood Function

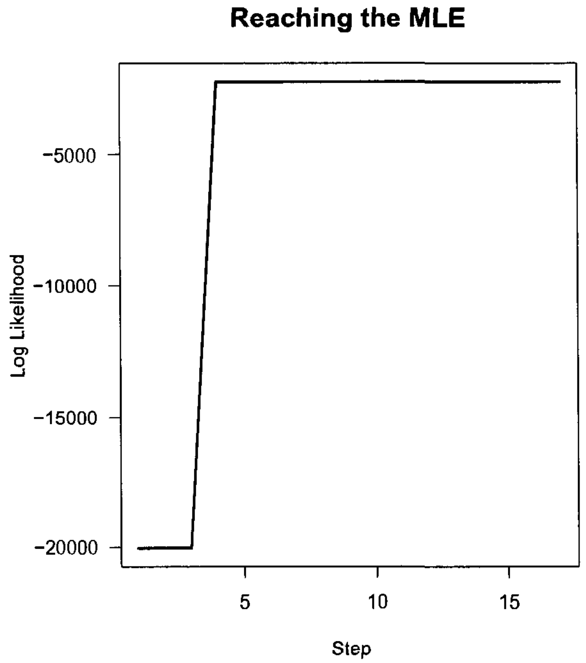
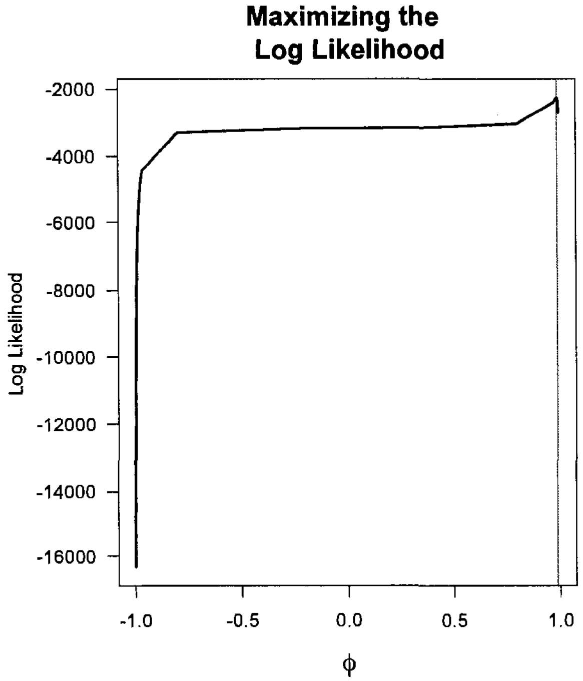


Figure 6.3: Examining the Convergence of ϕ



6.4 Additional Computations

Converting Back to the Price Scale

To predict a log price we simply plug in the estimated parameters into (6.1):

$$\hat{y}_{i,j,z} = \hat{\mu} + \hat{\beta}_{t(i,j,z)} + \hat{\tau}_z + \hat{\phi}^{\gamma(i,j,z)} \left(y_{i,j-1,z} - \hat{\mu} - \hat{\beta}_{t(i,j-1,z)} - \hat{\tau}_z \right) \tag{6.16}$$

To convert this prediction back to the price scale, we exponentiate the prediction above with the adjustment for efficiency. We use the same procedure as outlined in Sec. 3.10.

Constructing an Index

The price index can be constructed as follows:

$$1, \ \exp\left\{\hat{\beta}_2 - \hat{\beta}_1\right\}, \ \exp\left\{\hat{\beta}_3 - \hat{\beta}_1\right\}, \ldots, \ \exp\left\{\hat{\beta}_T - \hat{\beta}_1\right\}. \tag{6.17}$$

We can also incorporate the zip code effect as a multiplier to the index above. That is, we can compute an index for each zip code z relative to the global index:

$$\exp\left\{\hat{\tau}_z\right\}, \ \exp\left\{\hat{\beta}_2 + \hat{\tau}_z - \hat{\beta}_1\right\}, \ \exp\left\{\hat{\beta}_3 + \hat{\tau}_z - \hat{\beta}_1\right\}, \ldots, \ \exp\left\{\hat{\beta}_T + \hat{\tau}_z - \hat{\beta}_1\right\} \tag{6.18}$$

As before, an adjustment for efficiency is omitted since the standard errors of the parameters are too small to have a noticeable effect on the indices.

Chapter 7

Results for Local Models

In this chapter, we analyze the results and check model assumptions for the local autoregressive model introduced in Chapter 6. We start by examining the estimated parameters and end with comparing the local model with an alternate mixed effects model. As hoped, the local autoregressive model seems to describe the data better than the other models previously discussed. A complete set of plots can be found in Appendix F.

7.1 Parameter Estimates for the Local Model

In Table 7.1 the estimates for the overall mean μ, autoregressive parameter ϕ, the variance component of the time series σ_ε^2, and the variance of the random effects σ_τ^2 are provided for each metropolitan area. Similar to the local model, the estimates for ϕ are close to one; however, the estimate for the local model is *lower* than for the global model for all cities. Furthermore, there seems to be a negative correlation between μ, interpreted as the overall

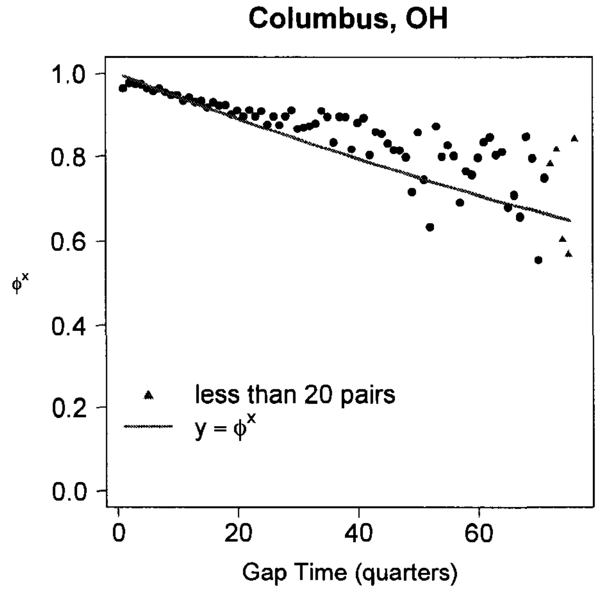
Table 7.1: Estimated Values of Parameters for Local Model

Metropolitan Area	$\hat{\mu}$	$\hat{\phi}$	$\hat{\sigma}^2_\varepsilon$	$\hat{\sigma}^2_\tau$
Ann Arbor, MI	11.6643	0.993247	0.001567	0.11045
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	11.4919	0.994269	0.001258	0.082476
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	1107472	0.992111	0.001588	0.179636
Phoenix, AZ	11.7022	0.992349	0.001543	0.106971
Pittsburgh, PA	11.3408	0.992059	0.002546	0.103488
Raleigh, NC	11.7447	0.993828	0.001413	0.047029
San Francisco, CA	12.4236	0.985644	0.001788	0.056201
Seattle, WA	11.9998	0.989923	0.001658	0.039459
Sioux Falls, SD	11.6025	0.995262	0.001120	0.032719
Stamford, CT	12.5345	0.987938	0.002294	0.093230

mean log price, and ϕ. In particular, for the coastal cities Los Angeles, CA, San Francisco, CA, Seattle, WA and Stamford, CT, $\hat{\phi}$ is lower while $\hat{\mu}$ is higher when compared to the remaining cities.

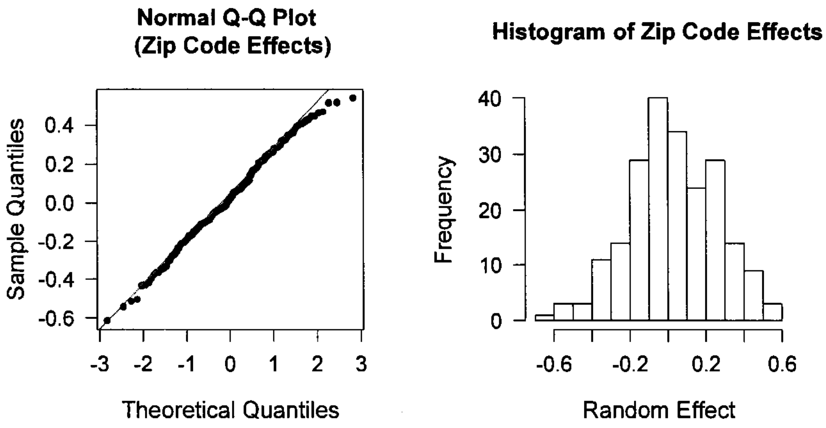
Using the procedure outlined in Sec. 5.3, we evaluate how well the AR(1) assumption applies to the data. The correlation between sale pairs with the same gap time are computed and plotted. The predicted relationship between ϕ and gap time is then overlayed. In Fig. 7.1, an example of such a plot is shown for Columbus, OH. If we compare this plot to the equivalent plot for the global model (Fig. 5.3), we find there is a negligible difference between the two. This the case for nearly all of the cities; these plots can be found in Appendix F.

Figure 7.1: ϕ vs Gap Time for Local Model (Columbus, OH)



Recall, we computed the zip code effects using the BLUP formulas. In the derivation, the random effects are assumed to be normally distributed. In Fig. 7.2, the zip code effects for Minneapolis, MN are examined for normality. Recall that shrinkage estimation becomes increasingly beneficial the more zip codes there are. Minneapolis, MN has a total of 214 zip codes, or random effects, and we find the normality assumption to be well satisfied. Note, however, that the AR(1) assumption also holds well for this city. We find that for those cities where the data are well described by the model, the distribution of the random effects is closer to normal.

Figure 7.2: Normality of Random Effects: Minneapolis, MN



7.2 Validation of Local Models

We apply the local AR model to the test sets for each of the twenty cities. The RMSE for the test set is used to evaluate predictive performance in Sec. 7.2.2. The price indices and residuals obtained from the model are also analyzed. For comparison purposes, a mixed effects model is proposed which is simply an extension of the mixed effects model presented in Sec. 5.4.1. Details are given next.

7.2.1 Mixed Effects Model

In Sec. 5.4.1, we introduced a mixed effects model where houses effects were modeled as random and time effects were fixed. We repeat that model here while adding a second random effect: the zip code. This model is as follows:

$$y_{i,j,z} = \mu + \alpha_i + \tau_z + \beta_{t(i,j,z)} + \varepsilon_{i,j,z} \tag{7.1}$$

where $\alpha_i \overset{iid}{\sim} \mathcal{N}(0, \sigma_\alpha^2)$, $\tau_z \overset{iid}{\sim} \mathcal{N}\left(0, \sigma_\tau^2\right)$, and $\varepsilon_{i,j} \overset{iid}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2)$ 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, μ is a fixed parameter.

The estimates for $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\tau}$ can be found by iteratively using the following formulas:

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\left(\mathbf{y} - \mathbf{W}\hat{\boldsymbol{\alpha}} - \mathbf{Z}\hat{\boldsymbol{\tau}}\right) \\
\hat{\boldsymbol{\alpha}} &= \left(\frac{\sigma_\varepsilon^2}{\sigma_\alpha^2}\mathbf{I}_I + \mathbf{W}'\mathbf{W}\right)^{-1}\mathbf{W}'\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{Z}\hat{\boldsymbol{\tau}}\right) \\
\hat{\boldsymbol{\tau}} &= \left(\frac{\sigma_\varepsilon^2}{\sigma_\tau^2}\mathbf{I}_Z + \mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\left(\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}} - \mathbf{W}\hat{\boldsymbol{\alpha}}\right)
\end{aligned}$$

These expressions are derived in Appendix E.2.2.

To predict the log prices, we plug in the estimates:

$$\hat{y}_{i,j,z} = \hat{\mu} + \hat{\beta}_{t(i,j,z)} + \hat{\alpha}_i + \hat{\tau}_z \qquad (7.2)$$

The procedure in Sec. 3.10 is used to convert these predictions back to the price scale. Finally, to construct a price index, the procedure in Sec. 6.4 is used.

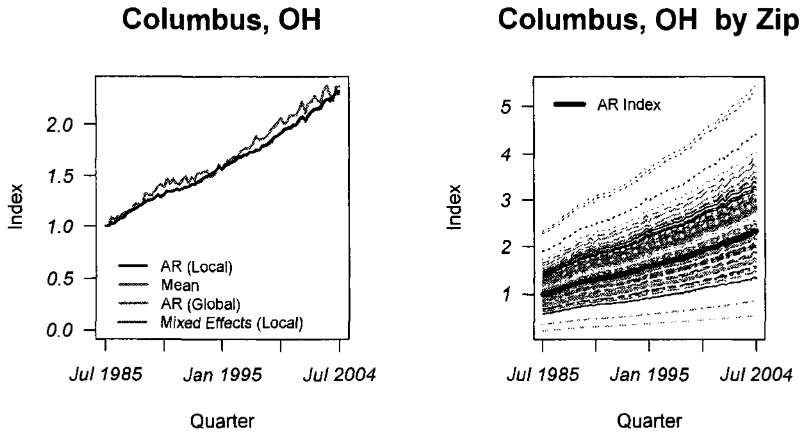
7.2.2 Results

To compare performance across models, the RMSE for the test set observations are calculated. These results are listed in Table 7.2; we add the RMSE results from the global AR model for further comparisons (these values are taken from Table 5.3). The model with the lowest RMSE value is given in bold font. Test RMSEs are not available for Chicago, IL, Los Angeles, CA, and Philadelphia, PA. It is clear that the local AR model not only performs better than the benchmark local mixed effects model, but also provides better predictions than the global AR model for nearly all the cities.

Table 7.2: Test Set RMSE for Local Models (in dollars)

Metropolitan Area	AR (Local)	Mixed Effects (Local)	AR (Global)
Ann Arbor, MI	**41,401**	46,519	44,362
Atlanta, GA	**30,914**	34,912	33,977
Chicago, IL	**36,004**	–	39,201
Columbia, SC	**35,881**	38,375	36,376
Columbus, OH	**26,681**	29,674	27,651
Kansas City, MO	**24,179**	25,851	24,963
Lexington, KY	**21,132**	21,555	21,501
Los Angeles, CA	**37,438**	–	41,006
Madison, WI	**28,035**	30,297	28,687
Memphis, TN	**24,588**	25,502	25,069
Minneapolis, MN	**31,900**	34,065	33,233
Orlando, FL	**28,449**	30,438	29,317
Philadelphia, PA	**33,479**	–	34,811
Phoenix, AZ	**28,247**	29,286	30,231
Pittsburgh, PA	**26,406**	28,630	26,507
Raleigh, NC	**25,839**	27,493	26,563
San Francisco, CA	49,927	**48,217**	50,777
Seattle, WA	**38,469**	41,950	42,329
Sioux Falls, SD	**20,160**	21,171	20,190
Stamford, CT	**57,722**	58,616	61,805

Figure 7.3: House Price Indices for Columbus, OH



In Fig. 7.3, we plot indices constructed from the global and local AR models, the mixed effects model, and the mean price index for Columbus, OH. Ignoring the mean price index, there is not a marked difference among the indices. We find similar results for the cities in Figs. In the second plot in Fig. 7.3, the local price index is plotted with the zip code multiplier (see Sec. 6.4 for details).

7.2.3 Residual Analysis

In this final section, we examine the model assumptions further. For ease of comparison with the results in Sec. 5.4.3, we provide results for Columbus, OH here as well. In Fig. 7.4, we randomly select 5% of the training set and plot the residuals versus the predictions. We also plot the residuals against the gap time for these observations for both local models. As before, the mixed effects model has the narrowest residual plot; however, the mixed effects model does not perform nearly as well as the AR model for the test set.

In Fig. 7.5 and Fig. 7.6 the residuals are checked for normality. The distribution of the residuals depends on the gap time for the local model. As a result, normal quantile plots are provided for a few different gap times. The normality assumption does not seem fully satisfied for all gap times; however, they do indicate an improvement on the global AR model (compare with Fig. 5.5). In contrast, the normality assumption fails to hold for the local mixed effects model altogether.

Both the AR and mixed effects models have specific error structures–the former depends on gap time so the errors are heteroscedastic, for the latter, error variance is assumed to be constant. In Fig. 7.7 the variance of the residuals is plotted against gap time. The estimated relationship between the variance and gap time is plotted as well in red. In both plots, we see a clear association between the two: as the gap time increases, the variance also increases. Unfortunately, while the local AR model does account for the changes in the error variance, the estimated relationship does not match the data well.

Figure 7.4: Local Model Residual Plots for Columbus, OH (log scale)

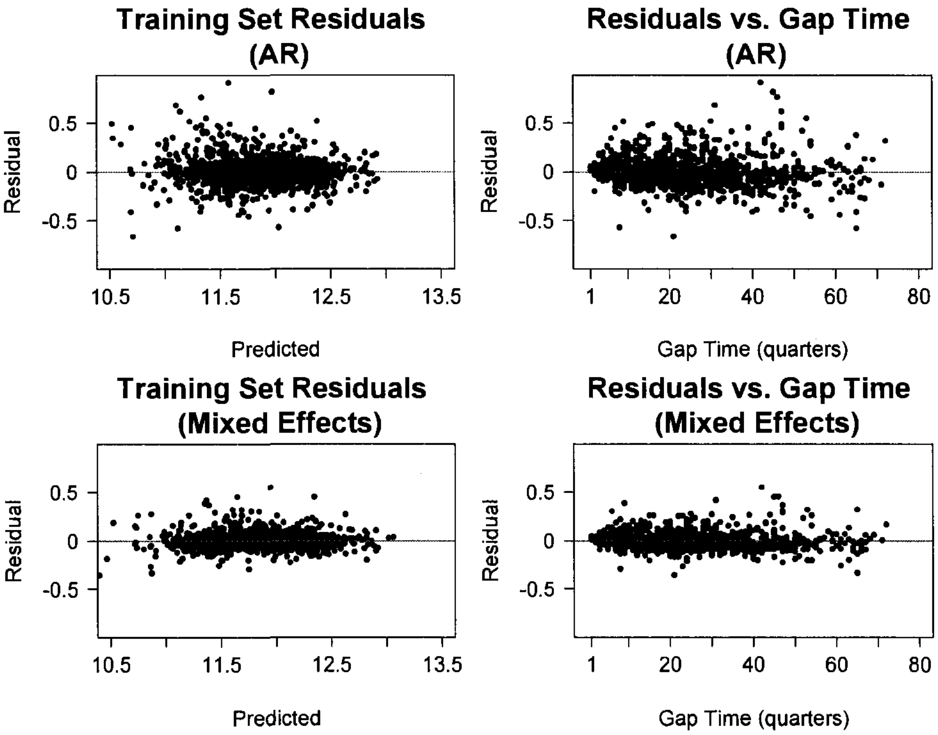


Figure 7.5: Normality of Residuals for Local Autoregressive Model (Columbus, OH)

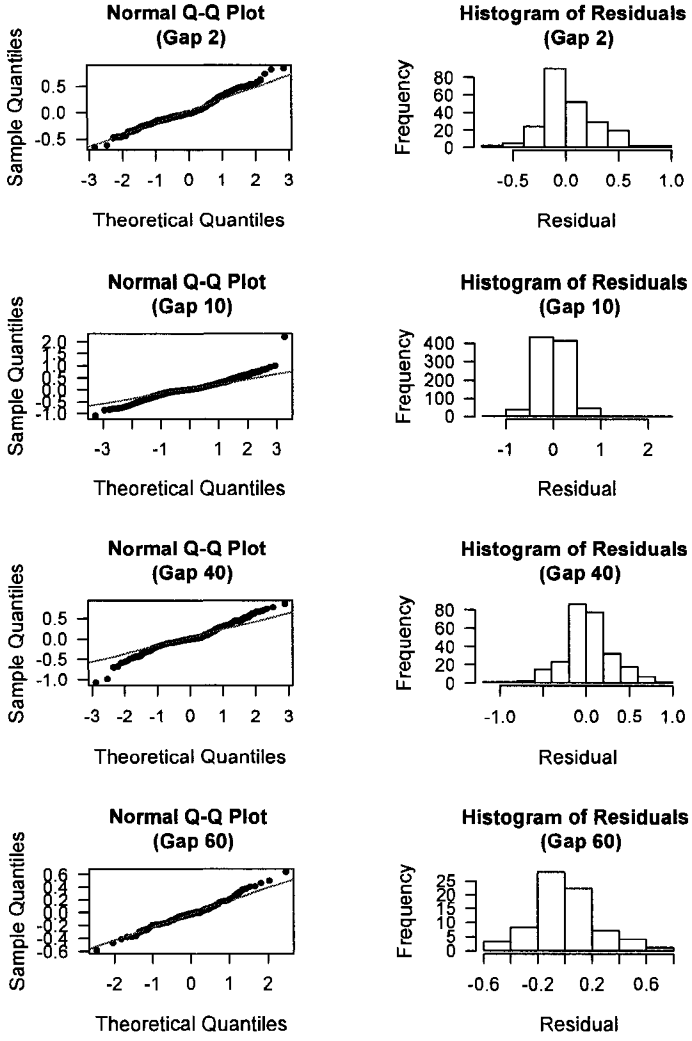


Figure 7.6: Normality of Residuals for Local Mixed Effects Model (Columbus,OH)

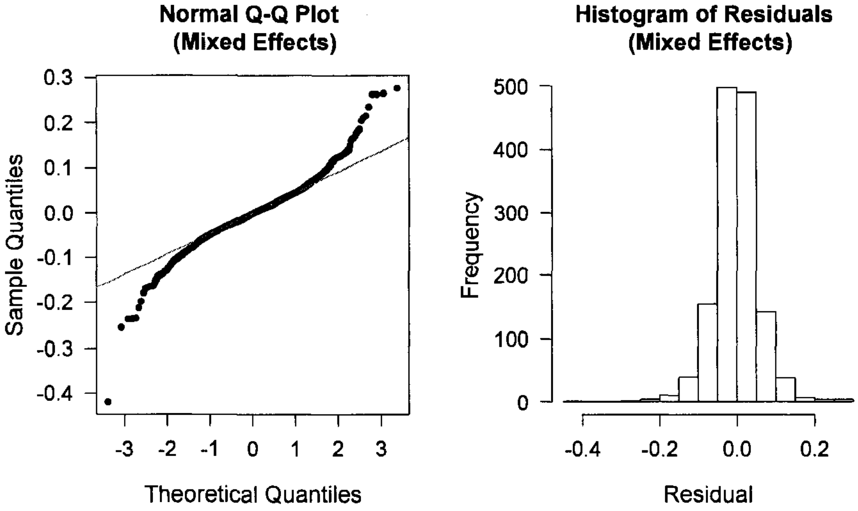
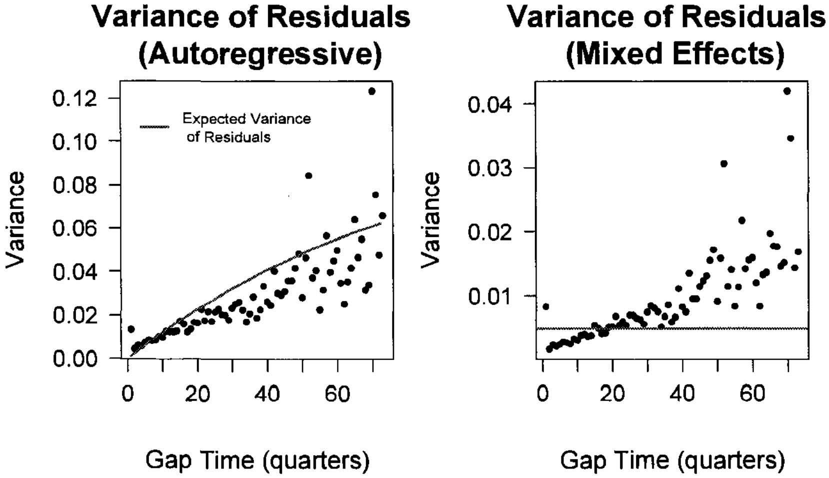


Figure 7.7: Variance of Residuals for Competing Local Models (Columbus, OH)



Chapter 8

Future Work

The two autoregressive models proposed seem to describe the data better than the existing methods. However, it is clear that there is room for improvement. In this chapter, we suggest several directions in which this research can be extended.

All repeat sales models, including the autoregressive model, are based upon the assumption that the previous sale price encompasses all relevant information about a house. This suggests that including hedonic data is redundant. Because the autoregressive method models prices, it is straightforward to add covariates to test this assumption. Say we have p covariates: $x_1, \ldots, x_p$. Then,

$$\begin{aligned} y_{i,1} &= \beta_{t(i,1)} + \textstyle\sum_{k=1}^{p} \alpha_k x_{k,i,1} + \varepsilon_{i,1} \\ y_{i,j} &= \beta_{t(i,j)} + \textstyle\sum_{k=1}^{p} \alpha_k x_{k,i,j} + \phi^{\gamma(i,j)} \left(y_{i,j-1} - \beta_{t(i,j-1)} - \textstyle\sum_{k=1}^{p} \alpha_k x_{k,i,j-1} \right) + \varepsilon_{i,j} \end{aligned} \tag{8.1}$$

with the same distribution for $\varepsilon_{i,j}$ as before. This setup allows one to not only include homes that have changed between sales, but to also obtain a more precise description of single sale homes. The model in (8.1) is a hybrid, similar in style to those proposed by Case and Quigley (see Sec. 4.3). If the parameters $\alpha_1, \ldots, \alpha_p$ were not significant, we would have evidence that hedonic information may indeed be unnecessary.

Throughout this manuscript, we have taken the gap time between sales to be fixed, not random. This simplified the modeling problem. Realistically, however, this is not the case and some effort should be focused on incorporating this feature into the model. Moreover, gap time may be related to price. For instance, people may wait to sell their homes until they feel like they would get a suitably high price. Therefore, it may be inappropriate to treat gap times as given.

We have applied the autoregressive model only to single family home sales. However, many different types of residential properties exist from apartments to condominiums. It would be interesting to see whether these residential properties could also be described by an autoregressive process.

A final assumption we have made throughout is that prices reflect the true value of a house. We could think of a sale price, however, as an *estimate* of value. Factors unique to the particular purchase conditions may add error. That is,

$$\text{price} = \text{value} + \text{error}.$$

If we introduce this new error term to the autoregressive model, we can describe it as a state space model. However, such models often require on a sufficiently long time series to obtain parameter estimates. House price series, unfortunately, are quite short although we do have many such series in a data set. This suggests that traditional state space methods may not be appropriate or even feasible and new methods must be developed.

Appendix A

Data Summary

The tables that follow provide information about all twenty U.S. metropolitan areas. A general description of the data can be found in Chapter 2. Table A.1 shows the number of total sales and unique houses sold from July 1985 to September 2004 along with the breakdown between single and repeat sales.

The name ZIP stands for "Zone Improvement plan" and is a five-digit code given by the United States Postal Service denotes post office facilities around the United States. In 1981 an extra four digits (ZIP+4) were added to the ZIP code to provide more specific information on location [19, p. 126]. Census tracts designate smaller areas than those defined by the ZIP code. The United States Bureau of the Census designed each tract to have homogeneous populations of 2,500 to 8,000 people when first formed and are grouped within counties of a state [26, p. 10-1]. Therefore, unlike the metropolitan areas in the data set, a census tract does not cross state boundaries. The number of ZIP codes and census tracts for each metropolitan area are listed in Table A.2. Note that not all ZIP codes or census tracts have home sales for each quarter of the sample period.

Table A.3 provides the breakdown of the training and test sets for each metropolitan area. The first three columns are relevant for the proposed autoregressive models, and the benchmark fixed effects and mixed effects models. The final two columns refer to the S&P/Case-Shiller® model where only repeat sales are used. Thus the number of sale pairs and the total number of unique houses in this reduced training set are given. Keep in mind, however, that houses with more than two sales appear in multiple sale pairs. As the test set is comprised of final sales, it is the same for all fitted models including the S&P/Case-Shiller® method. The procedure used to split the data into training and test sets is described in Chapter 5.

Table A.1: House and Sale Counts

			No. Houses Per Sale Count			
City	No. Sales	No. Houses	1	2	3	4+
Ann Arbor, MI	68,684	48,522	32,458	12,662	2,781	621
Atlanta, GA	376,082	260,703	166,646	76,046	15,163	2836
Chicago, IL	688,468	483,581	319,340	130,234	28,369	5,603
Columbia, SC	7,034	4,321	2,303	1,470	431	117
Columbus, OH	102,591	49,263	7,801	31,739	7,892	1,827
Kansas City, MO	123,441	90,504	62,489	23,706	3,773	534
Lexington, KY	38,534	26,630	16,891	7,901	1,555	282
Los Angeles, CA	543,071	395,061	272,258	100,918	18,965	2,903
Madison, WI	50,589	35,635	23,685	9,439	2,086	425
Memphis, TN	55,370	37,352	23,033	11,319	2,412	587
Minneapolis, MN	330,162	240,270	166,811	59,468	11,856	2,127
Orlando, FL	104,853	72,976	45,966	22,759	3,706	543
Philadelphia, PA	672,261	295,472	29,948	187,488	53,095	24,730
Phoenix, AZ	180,745	129,993	87,249	35,910	5,855	968
Pittsburgh, PA	104,544	73,871	48,618	20,768	3,749	718
Raleigh, NC	100,180	68,306	42,545	20,632	4,306	818
San Francisco, CA	73,598	59,416	46,959	10,895	1,413	149
Seattle, WA	253,227	182,770	124,672	47,406	9,198	1,494
Sioux Falls, SD	12,439	8,974	6,117	2,353	419	85
Stamford, CT	14,602	11,128	8,200	2,502	357	62

Table A.2: Zip Code and Census Tract Counts

City	No. Zip Codes	No. Census Tracts
Ann Arbor, MI	57	167
Atlanta, GA	184	652
Chicago, IL	317	1,751
Columbia, SC	12	29
Columbus, OH	96	360
Kansas City, MO	179	472
Lexington, KY	31	107
Los Angeles, CA	280	2,013
Madison, WI	40	90
Memphis, TN	64	243
Minneapolis, MN	214	722
Orlando, FL	96	326
Philadelphia, PA	332	1,256
Phoenix, AZ	130	667
Pittsburgh, PA	257	670
Raleigh, NC	82	207
San Francisco, CA	70	360
Seattle, WA	110	524
Sioux Falls, SD	30	29
Stamford, CT	23	84

Table A.3: Training and Test Set Sizes

	Autoregressive Model			S&P/Case-Shiller® Model	
City	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	77,135	25,456	49,263	27,872	25,729
Kansas City, MO	107,209	16,232	90,504	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	75,660	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	240,270	46,206	43,764
Orlando, FL	89,123	15,730	72,976	16,147	15,531
Philadelphia, PA	500,524	171,737	295,472	205,052	171,823
Phoenix, AZ	155,823	24,922	129,993	25,830	24,656
Pittsburgh, PA	89,762	14,782	73,871	15,891	14,956
Raleigh, NC	84,678	15,502	68,306	16,372	15,388
San Francisco, CA	66,527	7,071	59,416	7,111	6,948
Seattle, WA	218,741	34,486	182,770	35,971	34,304
Sioux Falls, SD	10,755	1,684	8,974	1,781	1,677
Stamford, CT	12,902	1,700	11,128	1,774	1,654

Appendix B

Derivations for the Global Autoregressive Model

The global autoregressive model outlined in Chapter 3 is provided below. If $y_{i,j}$ is the log price of the jth sale of the ith house,

$$\begin{aligned} y_{i,1} - \beta_{t(i,1)} &= \varepsilon_{i,1} && j = 1 \\ y_{i,j} - \beta_{t(i,j)} &= \phi^{\gamma(i,j)} \left(y_{i,j-1} - \beta_{t(i,j-1)} \right) + \varepsilon_{i,j} && j > 1 \end{aligned} \tag{B.1}$$

where $\varepsilon_{i,1} \overset{iid}{\sim} \mathcal{N}\left(0, \frac{\sigma_\varepsilon^2}{1-\phi^2}\right)$, $\varepsilon_{i,j} \overset{iid}{\sim} \mathcal{N}\left(0, \frac{\sigma_\varepsilon^2\left(1-\phi^{2\gamma(j)}\right)}{1-\phi^2}\right)$, and all $\varepsilon_{i,j}$ are independent. The parameters in this model are: $\boldsymbol{\theta} = \left\{\boldsymbol{\beta},\ \phi,\ \sigma_\varepsilon^2\right\}$ for a total of 79 parameters since $T = 77$. The MLEs are computed using the coordinate ascent algorithm. The derivations are given in Sec B.1 and the observed information matrix is provided in Sec B.2.

B.1 MLE Derivations

Recall that $\tau = \frac{\sigma_\epsilon^2}{1-\phi^2}$ and for the updating functions for τ and ϕ, we can simplify the log-likelihood function by using the adjusted log prices $w_{i,j} = y_{i,j} - \beta_{t(i,j)}$. The log likelihood function is:

$$\begin{aligned} l(\boldsymbol{\theta};\, \mathbf{y}) &= -\frac{N}{2}\log(2\pi\tau^2) - \frac{1}{2\tau^2}\sum_{i=1}^{I}\left(y_{i,1} - \beta_{t(i,1)}\right)^2 - \frac{1}{2}\sum_{i=1}^{I}\sum_{j=2}^{J_i}\log\left(1 - \phi^{2\gamma(i,j)}\right) \\ & -\frac{1}{2\tau^2}\sum_{i=1}^{I}\sum_{j=2}^{J_i}\frac{\left(y_{i,j} - \beta_{t(i,j)} - \phi^{\gamma(i,j)}\left(y_{i,j-1} - \beta_{t(i,j-1)}\right)\right)^2}{1 - \phi^{2\gamma(i,j)}} \end{aligned}$$

where N is the number of observations. Below are the derivations for updating the parameters.

Update for β_1, …, β_T

Each time effect, β_t, is updated individually:

$$
\begin{aligned}
\frac{\partial l}{\partial \beta_t} &= \frac{1}{\tau^2} \sum_{\substack{i:t(i,j)=t\\ j>1}} \frac{1}{1-\phi^{2\gamma(i,j)}} \Big(y_{i,j} - \beta_t - \phi^{\gamma(i,j)} \left(y_{i,j-1} - \beta_{t(i,j-1)} \right) \Big) \\
&\quad - \frac{1}{\tau^2} \sum_{\substack{i:t(i,j-1)=t\\ j>1}} \frac{\phi^{\gamma(i,j)}}{1-\phi^{2\gamma(i,j)}} \Big(y_{i,j} - \beta_{t(i,j)} - \phi^{\gamma(i,j)} \left(y_{i,j-1} - \beta_t \right) \Big) \\
&\quad + \frac{1}{\tau^2} \sum_{i:t(i,1)=t} (y_{i,1} - \beta_t) \\
0 &= \sum_{i:t(i,1)=t} y_{i,1} - \beta_t |i : t(i,1) = t| - \beta_t \sum_{\substack{i:t(i,j)=t\\ j>1}} \frac{1}{1-\phi^{2\gamma_{i,j}}} \\
&\quad - \beta_t \sum_{\substack{i:t(i,j-1)=t\\ j>1}} \frac{\phi^{2\gamma(i,j)}}{1-\phi^{2\gamma(i,j)}} + \sum_{\substack{i:t(i,j)=t\\ j>1}} \frac{1}{1-\phi^{2\gamma(i,j)}} \Big(y_{i,j} - \phi^{\gamma(i,j)} \left(y_{i,j-1} - \beta_{t(i,j-1)} \right) \Big) \\
&\quad - \sum_{\substack{i:t(i,j-1)=t\\ j>1}} \frac{\phi^{\gamma(i,j)}}{1-\phi^{2\gamma(i,j)}} \Big(y_{i,j} - \beta_{t(i,j)} - \phi^{\gamma(i,j)} y_{i,j-1} \Big) \\
\beta_t &= \left(\frac{1}{|i : t(i,1) = t| + \sum_{\substack{i:t(j)=t\\ j>1}} \frac{1}{1-\phi^{2\gamma(i,j)}} + \sum_{\substack{i:t(i,j-1)=t\\ j>1}} \frac{\phi^{2\gamma(i,j)}}{1-\phi^{2\gamma(i,j)}}} \right) \times \\
&\quad \Bigg(\sum_{i:t(i,1)=t} y_{i,1} + \sum_{\substack{i:t(i,j)=t\\ j>1}} \frac{1}{1-\phi^{2\gamma(i,j)}} \Big(y_{i,j} - \phi^{\gamma(i,j)} \left(y_{i,j-1} - \beta_{t(i,j-1)} \right) \Big) \\
&\quad - \sum_{\substack{i:t(i,j-1)=t\\ j>1}} \frac{\phi^{\gamma(i,j)}}{1-\phi^{2\gamma(i,j)}} \Big(y_{i,j} - \beta_{t(i,j)} - \phi^{\gamma(i,j)} y_{i,j-1} \Big) \Bigg)
\end{aligned}
$$

where $|\cdot|$ is the cardinality of a set.

Update for τ^2

The expression for τ^2 is:

$$\begin{aligned}
\frac{\partial l}{\partial \tau^2} &= -\frac{N}{2\tau^2} + \frac{1}{2(\tau^2)^2}\sum_{i=1}^{I} w_{i,1}^2 + \frac{1}{2(\tau^2)^2}\sum_{i=1}^{I}\sum_{j=2}^{J_i}\frac{\left(w_{i,j} - \phi^{\gamma(i,j)} w_{i,j-1}\right)^2}{1 - \phi^{2\gamma(i,j)}} \\
0 &= -N\tau^2 + \sum_{i=1}^{I} w_{i,1}^2 + \sum_{i=1}^{I}\sum_{j=2}^{J_i}\frac{\left(w_{i,j} - \phi^{\gamma(i,j)} w_{i,j-1}\right)^2}{1 - \phi^{2\gamma(i,j)}} \\
\tau^2 &= \frac{1}{N}\left[\sum_{i=1}^{I} w_{i,1}^2 + \sum_{i=1}^{I}\sum_{j=2}^{J_i}\frac{\left(w_{i,j} - \phi^{\gamma(i,j)} w_{i,j-1}\right)^2}{1 - \phi^{2\gamma(i,j)}}\right].
\end{aligned}$$

Function for ϕ

An explicit expression for ϕ cannot be derived. Instead, the zero of the partial derivative of the log likelihood function with respect to ϕ should be estimated or the log likelihood function should be directly maximized with respect to ϕ.

$$\begin{aligned}
\frac{\partial l}{\partial \phi} &= -\frac{1}{2}\sum_{i=1}^{I}\sum_{j=2}^{J_i}\frac{-2\gamma(i,j)\phi^{2\gamma(i,j)-1}}{1-\phi^{2\gamma(i,j)}} \\
&\quad \frac{2}{2\tau^2}\sum_{i=1}^{I}\sum_{j=2}^{J_i}\frac{\left(w_{i,j}-\phi^{\gamma(i,j)}\right)\left(\gamma(i,j)w_{i,j-1}\phi^{\gamma(i,j)-1}\right)}{1-\phi^{2\gamma(i,j)}} \\
&\quad -\frac{2}{2\tau^2}\sum_{i=1}^{I}\sum_{j=2}^{J_i}\frac{\left(w_{i,j}-\phi^{\gamma(i,j)}w_{i,j-1}\right)^2}{\left(1-\phi^{2\gamma(i,j)}\right)^2}\left(\gamma(i,j)\phi^{2\gamma(i,j)-1}\right) \\
0 &= \sum_{i=1}^{I}\sum_{j=2}^{J_i}\frac{\gamma(i,j)\phi^{2\gamma(i,j)-1}}{1-\phi^{2\gamma(i,j)}} \\
&\quad +\frac{1}{\tau^2}\sum_{i=1}^{I}\sum_{j=2}^{J_i}\frac{\left(w_{i,j}-\phi^{\gamma(i,j)}w_{i,j-1}\right)\left(\gamma(i,j)w_{i,j-1}\phi^{\gamma(i,j)-1}\right)}{1-\phi^{2\gamma(i,j)}} \\
&\quad -\frac{1}{\tau^2}\sum_{i=1}^{I}\sum_{j=2}^{J_i}\left(\frac{w_{i,j}-\phi^{\gamma(i,j)}w_{i,j-1}}{1-\phi^{2\gamma(i,j)}}\right)^2\gamma(i,j)\phi^{2\gamma(i,j)-1}.
\end{aligned}$$

B.2 Computing the Observed Information matrix

The observed information matrix, $\hat{\mathbf{I}}$, is used to estimate the asymptotic variance-covariance matrix of the parameters. This matrix is given by the expression below:

$$\left[\hat{\mathbf{I}}\left(\hat{\theta}\right)\right]^{-1} = \left(-\frac{\partial^2 l\left(\hat{\boldsymbol{\theta}};\mathbf{y}\right)}{\partial\hat{\theta}\partial\hat{\theta}'}\right)^{-1}$$

where θ and θ' are arbitrary parameters. Expressions for the second-order partial derivatives are given next.

Equation for Second Derivative for β_t $\left(\frac{\partial^2 l}{\partial \beta_t^2}\right)$

$$\begin{aligned} -\frac{\partial^2 l}{\partial \beta_t^2} &= \frac{1-\phi^2}{\sigma_\varepsilon^2} \sum_{\substack{i:t(i,j)=t \\ j>1}} \left(\frac{1}{1-\phi^{2\gamma(i,j)}}\right) + \frac{1-\phi^2}{\sigma_\varepsilon^2} \sum_{\substack{i:t(i,j-1)=t \\ j>1}} \left(\frac{\phi^{2\gamma(i,j)}}{1-\phi^{2\gamma(i,j)}}\right) \\ &\quad + \frac{1-\phi^2}{\sigma_\varepsilon^2} |i : t(i,1) = t| \end{aligned}$$

Equation for Partial Derivative of β_t and β_s $\left(\frac{\partial^2 l}{\partial \beta_s \partial \beta_t}\right)$

When $s \neq t$,

$$-\frac{\partial^2 l}{\partial \beta_s \partial \beta_t} = -\frac{1-\phi^2}{\sigma_\varepsilon^2} \sum_{\substack{i:t(i,j)=t, \\ t(i,j-1)=s \\ j>1}} \left(\frac{\phi^{\gamma(i,j)}}{1-\phi^{2\gamma(i,j)}}\right) - \frac{1-\phi^2}{\sigma_\varepsilon^2} \sum_{\substack{i:t(i,j-1)=t, \\ t(i,j)=s \\ j>1}} \left(\frac{\phi^{\gamma(i,j)}}{1-\phi^{2\gamma(i,j)}}\right)$$

Equation for $\frac{\partial^2 l}{\partial\sigma_\varepsilon^2\partial\beta_t}$

$$\begin{aligned}
-\frac{\partial^2 l}{\partial\sigma_\varepsilon^2\partial\beta_t} &= \frac{1-\phi^2}{(\sigma_\varepsilon^2)^2}\sum_{\substack{i:t(i,j)=t\\ j>1}} \frac{1}{1-\phi^{2\gamma(i,j)}}\left(y_{i,j}-\beta_t-\phi^{\gamma(i,j)}w_{i,j-1}\right)\\
&\quad -\frac{1-\phi^2}{(\sigma_\varepsilon^2)^2}\sum_{\substack{i:t(i,j-1)=t\\ j>1}} \frac{\phi^{\gamma(i,j)}}{1-\phi^{2\gamma(i,j)}}\left(w_{i,j}-\phi^{\gamma(i,j)}\left(y_{i,j}-\beta_t\right)\right)\\
&\quad +\frac{1-\phi^2}{(\sigma_\varepsilon^2)^2}\sum_{i:t(i,1)=t}\left(y_{i,1}-\beta_t\right)
\end{aligned}$$

Equation for Second Derivative with Respect to σ_ε^2 $\left(\frac{\partial^2 l}{\partial(\sigma_\varepsilon^2)^2}\right)$

$$\begin{aligned}
-\frac{\partial^2 l}{\partial(\sigma_\varepsilon^2)^2} &= \frac{1-\phi^2}{(\sigma_\varepsilon^2)^3}\sum_{i=1}^{I}\sum_{j=2}^{J_i}\left(\frac{\left(w_{i,j}-\phi^{\gamma(i,j)}w_{i,j-1}\right)^2}{1-\phi^{2\gamma(i,j)}}\right)\\
&\quad -\frac{n}{2(\sigma_\varepsilon^2)^2}+\frac{1-\phi^2}{(\sigma_\varepsilon^2)^3}\sum_{i=1}^{I}w_{i,1}^2
\end{aligned}$$

Partial Derivative with Respect to ϕ and σ_ε^2 $\left(\frac{\partial^2 l}{\partial\phi\partial\sigma_\varepsilon^2}\right)$

$$
\begin{aligned}
-\frac{\partial^2 l}{\partial\phi\partial\sigma_\varepsilon^2} &= \frac{\phi}{(\sigma_\varepsilon^2)^2}\sum_{i=1}^{I}\left(w_{i,1}\right)^2 \\
&-\frac{1}{(\sigma_\varepsilon^2)^2}\sum_{i=1}^{I}\sum_{j=2}^{J_i}\left\{\left(\frac{-\phi+(1-\gamma(i,j))\,\phi^{2\gamma(i,j)+1}+\gamma(i,j)\phi^{2\gamma(i,j)-1}}{(1-\phi^{2\gamma(i,j)})^2}\right)\times\left(w_{i,j}-\phi^{\gamma(i,j)}w_{i,j-1}\right)^2\right\} \\
&+\frac{1}{(\sigma_\varepsilon^2)^2}\sum_{i=1}^{I}\sum_{j=2}^{J_i}\left\{\frac{1-\phi^2}{1-\phi^{2\gamma(i,j)}}\left(w_{i,j}-\phi^{\gamma(i,j)}w_{i,j-1}\right)\times\gamma(i,j)\phi^{\gamma(i,j)-1}w_{i,j-1}\right\}
\end{aligned}
$$

Partial Derivative Respect to ϕ and β_t $\left(\frac{\partial^2 l}{\partial\phi\partial\beta_t}\right)$

$$
\begin{aligned}
-\frac{\partial^2 l}{\partial\phi\partial\beta_t} &= \frac{-2}{\sigma_\varepsilon^2}\sum_{\substack{i:t(i,j)=t\\ j>1}}\left\{\frac{-\phi+(1-\gamma(i,j))\phi^{2\gamma(i,j)+1}+\gamma(i,j)\phi^{2\gamma(i,j)-1}}{(1-\phi^{2\gamma(i,j)})^2}\right\}(y_{i,j}-\beta_t)\\
&+\frac{1}{\sigma_\varepsilon^2}\sum_{\substack{i:t(i,j)=t\\ j>1}}\left\{\frac{\gamma(i,j)\phi^{\gamma(i,j)-1}-(\gamma(i,j)+2)\phi^{\gamma(i,j)+1}+\gamma(i,j)\phi^{3\gamma(i,j)-1}+(2-\gamma(i,j))\phi^{3\gamma(i,j)+1}}{(1-\phi^{2\gamma(i,j)})^2}\right\}w_{i,j-1}\\
&+\frac{1}{\sigma_\varepsilon^2}\sum_{\substack{i:t(i,j-1)=t\\ j>1}}\left\{\frac{\gamma(i,j)\phi^{\gamma(i,j)-1}-(\gamma(i,j)+2)\phi^{\gamma(i,j)+1}+\gamma(i,j)\phi^{3\gamma(i,j)-1}+(2-\gamma(i,j))\phi^{3\gamma(i,j)+1}}{(1-\phi^{2\gamma(i,j)})^2}\right\}w_{i,j}\\
&-\frac{2}{\sigma_\varepsilon^2}\sum_{\substack{i:t(i,j-1)=t\\ j>1}}\left\{\frac{\gamma(i,j)\phi^{2\gamma(i,j)-1}-(\gamma(i,j)+1)\phi^{2\gamma(i,j)+1}+\phi^{4\gamma(i,j)+1}}{(1-\phi^{2\gamma(i,j)})^2}\right\}(y_{i,j-1}-\beta_t)\\
&+\frac{2\phi}{\sigma_\varepsilon^2}\sum_{i:t(i,1)=t}(y_{i,1}-\beta_t)
\end{aligned}
$$

Second Derivative with Respect to ϕ $\left(\frac{\partial^2 l}{\partial \phi^2}\right)$

$$
\begin{aligned}
-\frac{\partial^2 l}{\partial \phi^2} = \; & \frac{N(1+\phi^2)}{(1-\phi^2)^2} - \frac{\phi}{\sigma_\varepsilon^2}\sum_{i=1}^{I} w_{i,1}^2 - \sum_{i=1}^{I}\sum_{j=2}^{J_i} \frac{\gamma(i,j)(2\gamma(i,j)-1)\phi^{2\gamma(i,j)-2} + \gamma(i,j)\phi^{4\gamma(i,j)-2}}{(1-\phi^{2\gamma(i,j)})^2} \\
& + \frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{I}\sum_{j=2}^{J_i}\left\{\frac{-1+(1-\gamma(i,j)(2\gamma(i,j)+1)\phi^{2\gamma(i,j)} + \gamma(i,j)(2\gamma(i,j)-1)\phi^{2\gamma(i,j)-2}}{(1-\phi^{2\gamma(i,j)})^2}\right\}\left(w_{i,j} - \phi^{\gamma(i,j)} w_{i,j-1}\right)^2 \\
& + \frac{4}{\sigma_\varepsilon^2}\sum_{i=1}^{I}\sum_{j=2}^{J_i}\left\{\frac{(1-\phi^{2\gamma(i,j)})\gamma(i,j)\phi^{2\gamma(i,j)-1}(-\phi + (1-\gamma(i,j))\phi^{2\gamma(i,j)+1} + \gamma(i,j)\phi^{2\gamma(i,j)-1})}{(1-\phi^{2\gamma(i,j)})^4}\right\}\left(w_{i,j} - \phi^{\gamma(i,j)} w_{i,j-1}\right)^2 \\
& - \frac{2}{\sigma_\varepsilon^2}\sum_{i=1}^{I}\sum_{j=2}^{J_i}\left\{\frac{-\phi + (1-\gamma(i,j))\phi^{2\gamma(i,j)+1} + \gamma(i,j)\phi^{2\gamma(i,j)-1}}{(1-\phi^{2\gamma(i,j)})^2}\right\}\left(w_{i,j} - \phi^{\gamma(i,j)} w_{i,j-1}\right)\left(w_{i,j-1}\gamma(i,j)\phi^{\gamma(i,j)-1}\right) \\
& - \frac{1}{\sigma_\varepsilon^2}\sum_{i=1}^{I}\sum_{j=2}^{J_i}\left(\frac{1-\phi^2}{1-\phi^{2\gamma(i,j)}}\right)\left(\gamma(i,j) w_{i,j-1}\right)\left((\gamma(i,j)-1) w_{i,j}\phi^{\gamma(i,j)-2} - (2\gamma(i,j)-1) w_{i,j-1}\phi^{2\gamma(i,j)-2}\right) \\
& - \frac{2}{\sigma_\varepsilon^2}\sum_{i=1}^{I}\sum_{j=2}^{J_i}\left\{\frac{-\phi\left(1-\phi^{2\gamma(i,j)}\right) + \gamma(i,j)\left(1-\phi^2\right)\phi^{2\gamma(i,j)-1}}{\left(1-\phi^{2\gamma(i,j)}\right)^2}\right\}\left(\gamma(i,j) w_{i,j-1}\phi^{\gamma(i,j)-1}\right)\left(w_{i,j} - \phi^{\gamma(i,j)} w_{i,j-1}\right)
\end{aligned}
$$

Appendix C

Computational Methods

In this appendix two procedures to reduce matrix inversions for large data sets are described. We start with residual regression which is used to estimate parameters for regressions where the categorical variable has a large number of groups. In Sec. C.2, we outline the conjugate gradient method which can be used to solve systems of the form $\mathbf{Ax} = \mathbf{b}$ where $\mathbf{A}$ is a large, sparse matrix. We use this algorithm when computing the indices for the S&P/Case-Shiller® method.

C.1 Residual Regression

When a regression model has both a large number of observations and groups of a categorical variable, using ordinary least squares may be infeasible due to the computational requirements of manipulating and inverting large matrices. Residual regression can be used in such situations. We use the procedure (notation included) outlined by Chamberlain (1996). His procedure applies the method to a balanced fixed effects models with additional regressors. We extend the method here for unbalanced fixed effects models where there are two fixed effects but only one with a large number of categories. Specifically, we wan to fit the model:

$$y_{i,j,k} = \mu + \alpha_i + \beta_j + \varepsilon_{i,j,k} \tag{C.1}$$

where $\mathbf{y}$ is the response vector, and μ, $\boldsymbol{\alpha}$, and $\boldsymbol{\beta}$ are fixed effects. Finally, $\boldsymbol{\varepsilon} \overset{iid}{\sim} \mathcal{N}\left(0, \sigma_\varepsilon^2 \mathbf{I}_N\right)$ where $\mathbf{I}$ is the identity matrix of dimension N, the number of observations. Assume that the fixed effect $\boldsymbol{\alpha}$ has a large number of categories; for our use, $\boldsymbol{\alpha}$ are the individual house effects. We rewrite the model as follows:

$$\mathbf{y} = \mathbf{A}\boldsymbol{\alpha} + \mathbf{B}\begin{bmatrix} \mu \\ \boldsymbol{\beta} \end{bmatrix} + \boldsymbol{\varepsilon} \tag{C.2}$$

where $\mathbf{A}$ and $\mathbf{B}$ are the design matrices that code the indicator variables for the parameters $\boldsymbol{\alpha}$ and $[\boldsymbol{\beta}\ \ \mu]'$.

The Residual Regression Procedure

1. Compute the following:

$$\begin{aligned}
\tilde{\mathbf{y}} &= \mathbf{y} - \left(\mathbf{A}'\mathbf{A}\right)^{-1}\mathbf{A}'\mathbf{y} \\
\tilde{\mathbf{y}} &= \mathbf{y} - \bar{\mathbf{y}}_i \\
\tilde{\mathbf{B}} &= \mathbf{A} - - \left(\mathbf{A}'\mathbf{A}\right)^{-1}\mathbf{A}'\mathbf{B} \\
\tilde{\mathbf{B}} &= \mathbf{B} - \bar{\mathbf{B}}_i
\end{aligned}$$

where $\bar{y}_i$ and $\hat{\mathbf{B}}_i$ is vector of average value of the response variable and dummy variable vectors by the categories of $\boldsymbol{\alpha}$.

2. Now, to obtain regression estimates for μ and $\boldsymbol{\beta}$,

$$\begin{bmatrix} \hat{\mu} \\ \hat{\boldsymbol{\beta}} \end{bmatrix} = \left(\tilde{\mathbf{B}}'\tilde{\mathbf{B}}\right)^{-1}\tilde{\mathbf{B}}'\tilde{\mathbf{y}} \tag{C.3}$$

3. As a final step, compute $\boldsymbol{\alpha}$:

$$\hat{\boldsymbol{\alpha}} = \bar{\mathbf{y}}_i - \bar{\mathbf{B}}_i \begin{bmatrix} \hat{\mu} \\ \hat{\boldsymbol{\beta}} \end{bmatrix} \tag{C.4}$$

C.2 Conjugate Gradient Method

We need to use special techniques to solve $\mathbf{Ax} = \mathbf{b}$ where $\mathbf{A}$ is a large, sparse matrix. For such matrices, it is important to preserve the zeros in the matrix. The conjugate gradient method is best for this setting as we avoid inverting the matrix entirely [18, p. 105]. This method iteratively solves $Ax = b$ given a starting value $x^{(0)}$ and is outlined next.

The Conjugate Gradient Method

1. Initialize $x^{(0)}$ and set a tolerance level ϵ.

2. Set $k = 0$; $r^{(k)} = b - Ax^{(k)}$; $s^{(k)} = A'r^{(k)}$; $p^{(k)} = s^{(k)}$; and $\gamma^{(k)} = ||s^{(k)}||_2^2$.

3. If $\gamma^{(k)} \leq \epsilon$, set $x = x^{(k)}$ and stop.

4. Set $q^{(k)} = Ap^{(k)}$.

5. Compute $\alpha^{(k)} = \frac{\gamma^{(k)}}{||q^{(k)}||_2^2}$.

6. Update:

   (a) $x^{(k+1)} = x^{(k)} + \alpha^{(k)}p^{(k)}$,

   (b) $r^{(k+1)} = r^{(k)} - \alpha^{(k)}q^{(k)}$,

   (c) $s^{(k+1)} = A'r^{(k+1)}$,

   (d) $\gamma^{(k+1)} = ||s^{(k+1)}||_2^2$, and

   (e) $p^{(k+1)} = s^{(k+1)} + \frac{\gamma^{(k+1)}p^{(k)}}{\gamma^{(k)}}$.

7. Set k to be $k+1$ and return to Step 3.

We can apply this method to the first and third stages of the S&P/Case-Shiller® method. The inputs are:

Stage 1. $A = Z'X$ and $b = Z'Y$

Stage 3. $A = Z'\hat{\Omega}^{-1}X$ and $b = Z'\hat{\Omega}^{-1}Y$.

Appendix D

Additional Plots for Global Model

The plots in this appendix correspond to the results presented in Chapter 5. Figs. D.1- D.4 are plots verifying the AR(1) assumption in the autoregressive model for each metropolitan area. The correlation between adjusted prices of sale pairs with the same gap time are plotted along with the estimated relationship. Correlation values with fewer than twenty sale pairs are marked with the triangle symbol.

The second set of plots, Figs. D.5- D.8 are the computed indices for each model: the autoregressive model, S&P/Case-Shiller® method, fixed effects model, mixed effects model, and a separately constructed mean price index.

Figure D.1: AR(1) Assumption Check: Ann Arbor, MI-Kansas City, MO

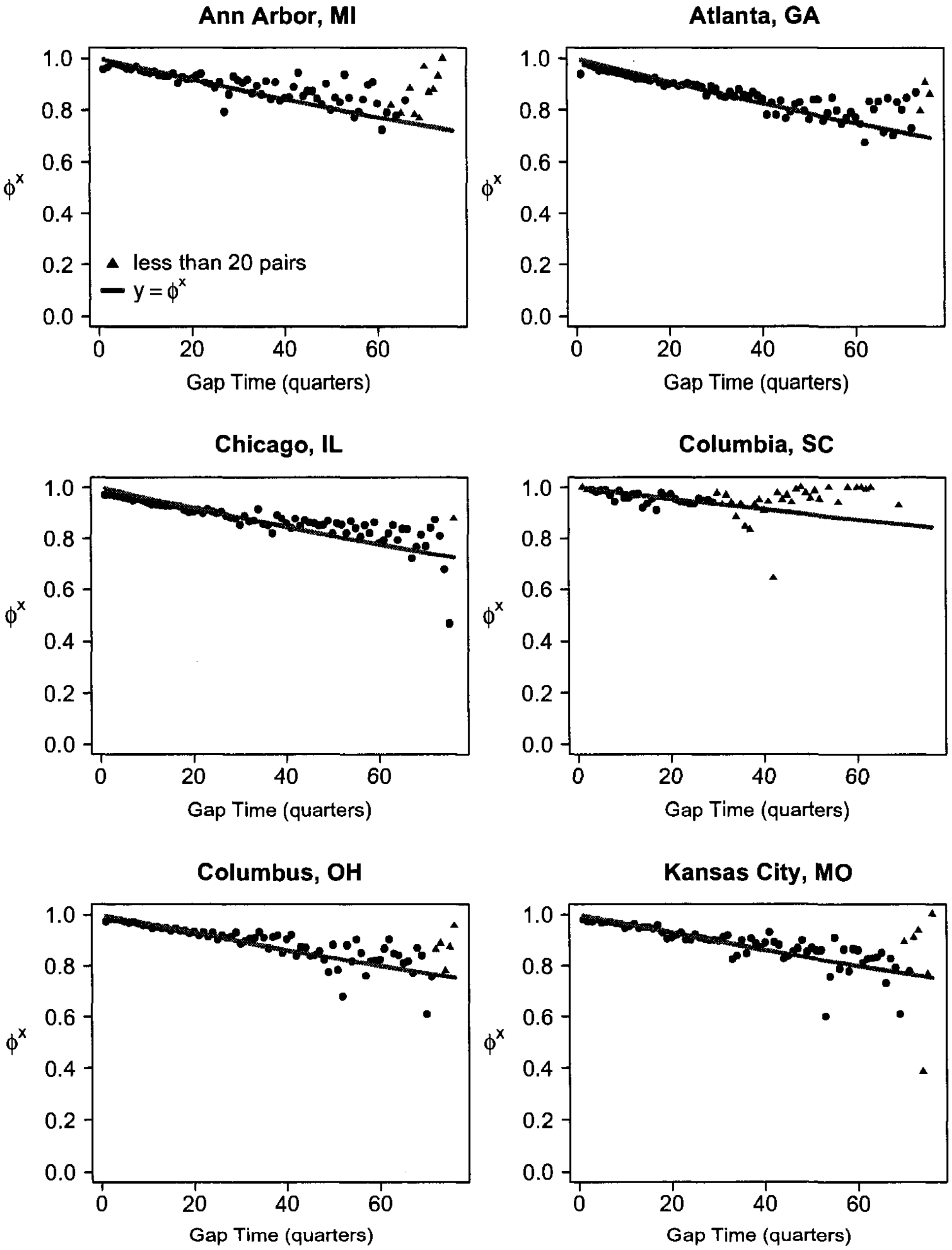


Figure D.2: AR(1) Assumption Check: Lexington, KY-Orlando, FL

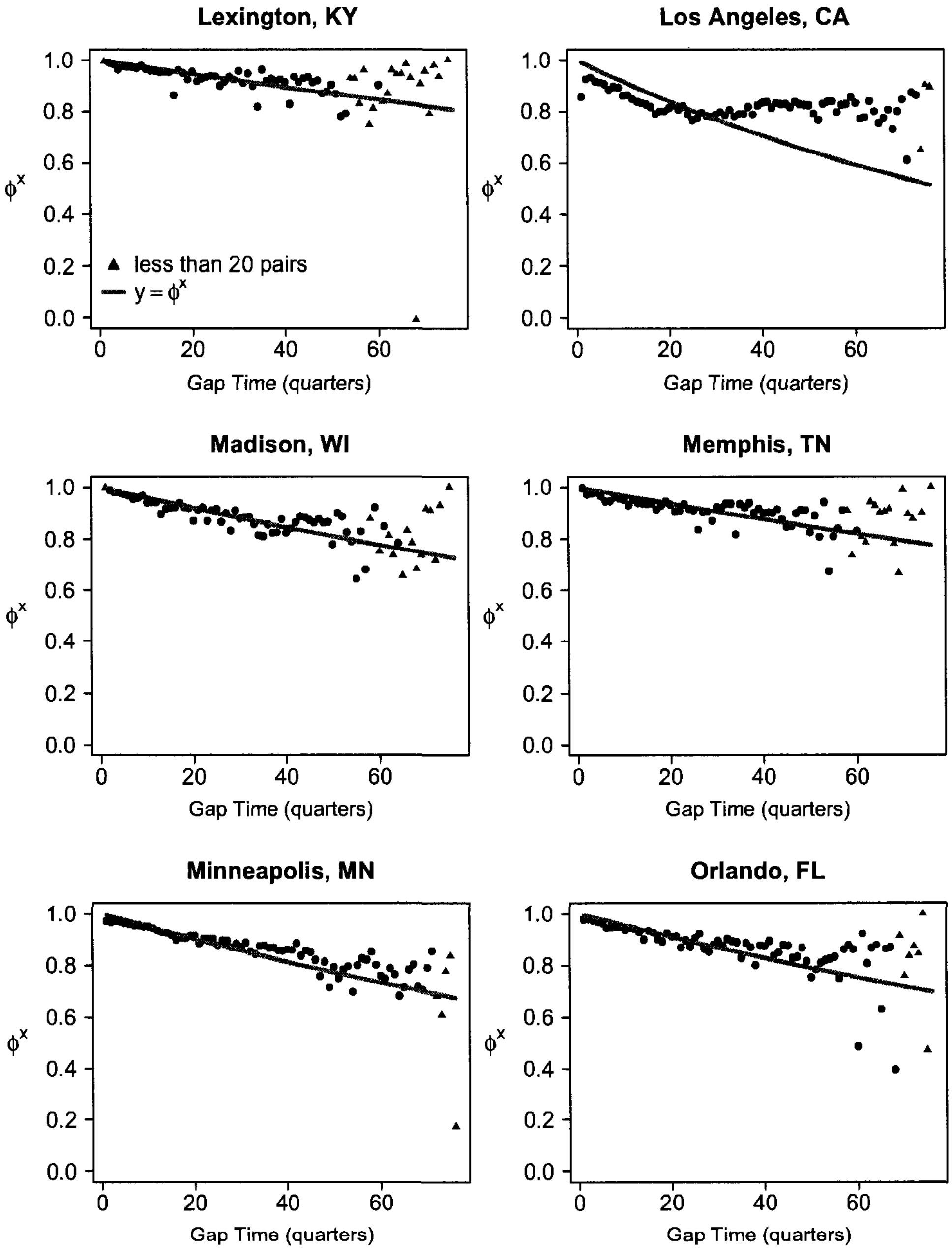


Figure D.3: AR(1) Assumption Check: Philadelphia, PA-Seattle, WA

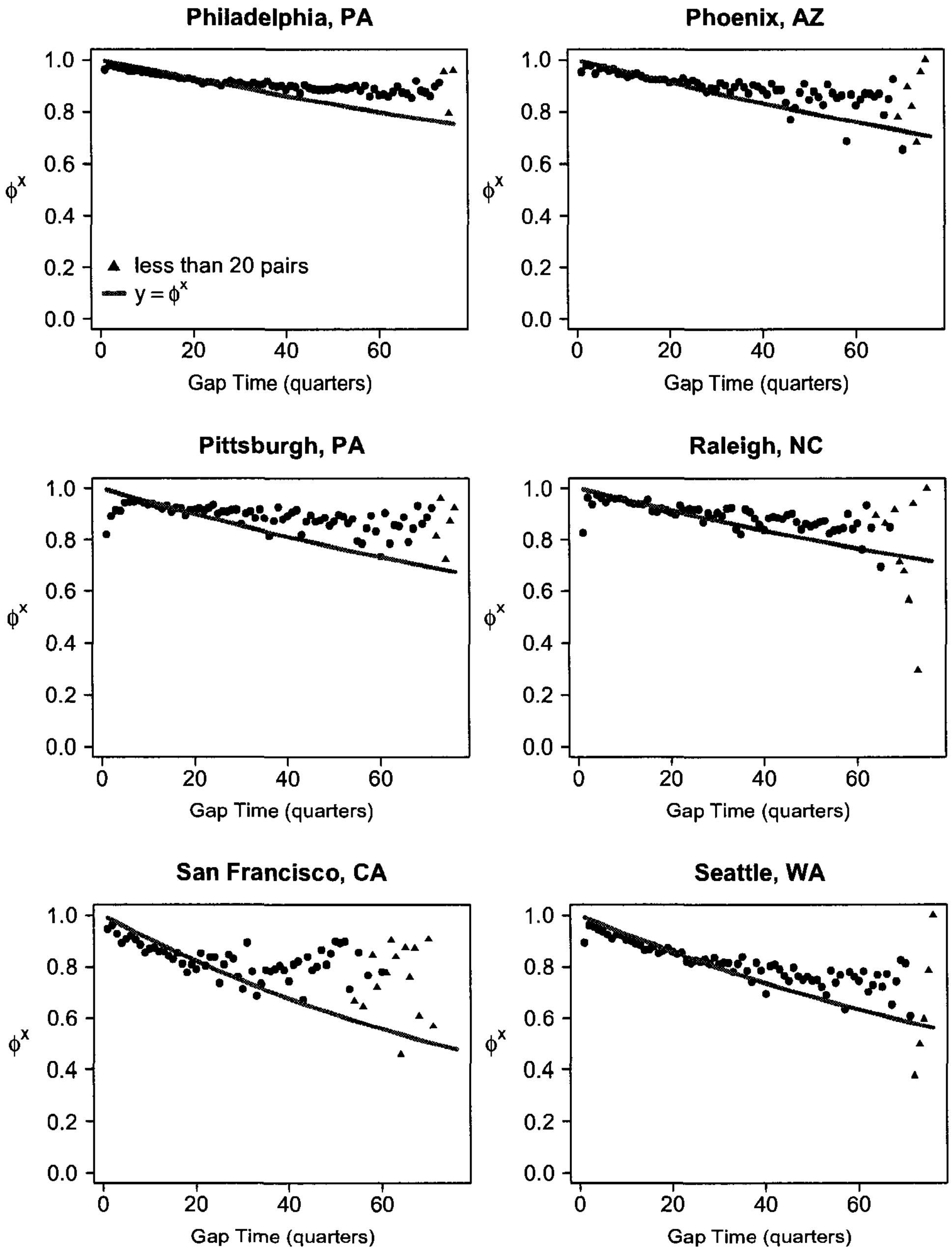


Figure D.4: AR(1) Assumption Check: Sioux Falls, SD-Stamford, CT

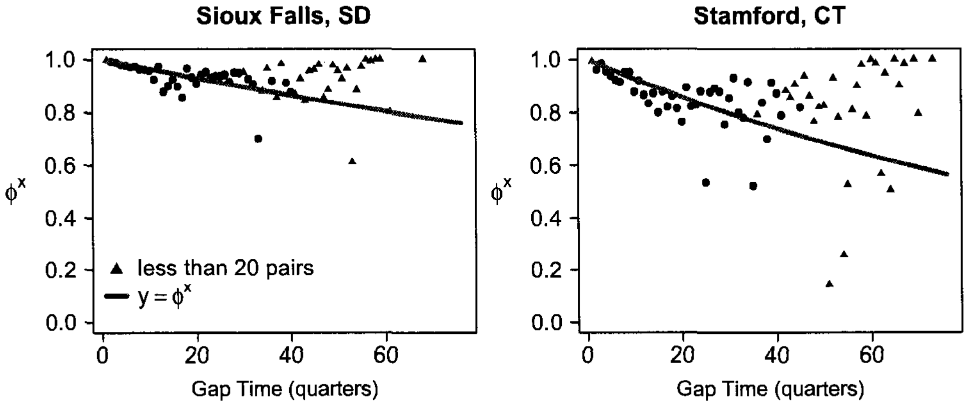
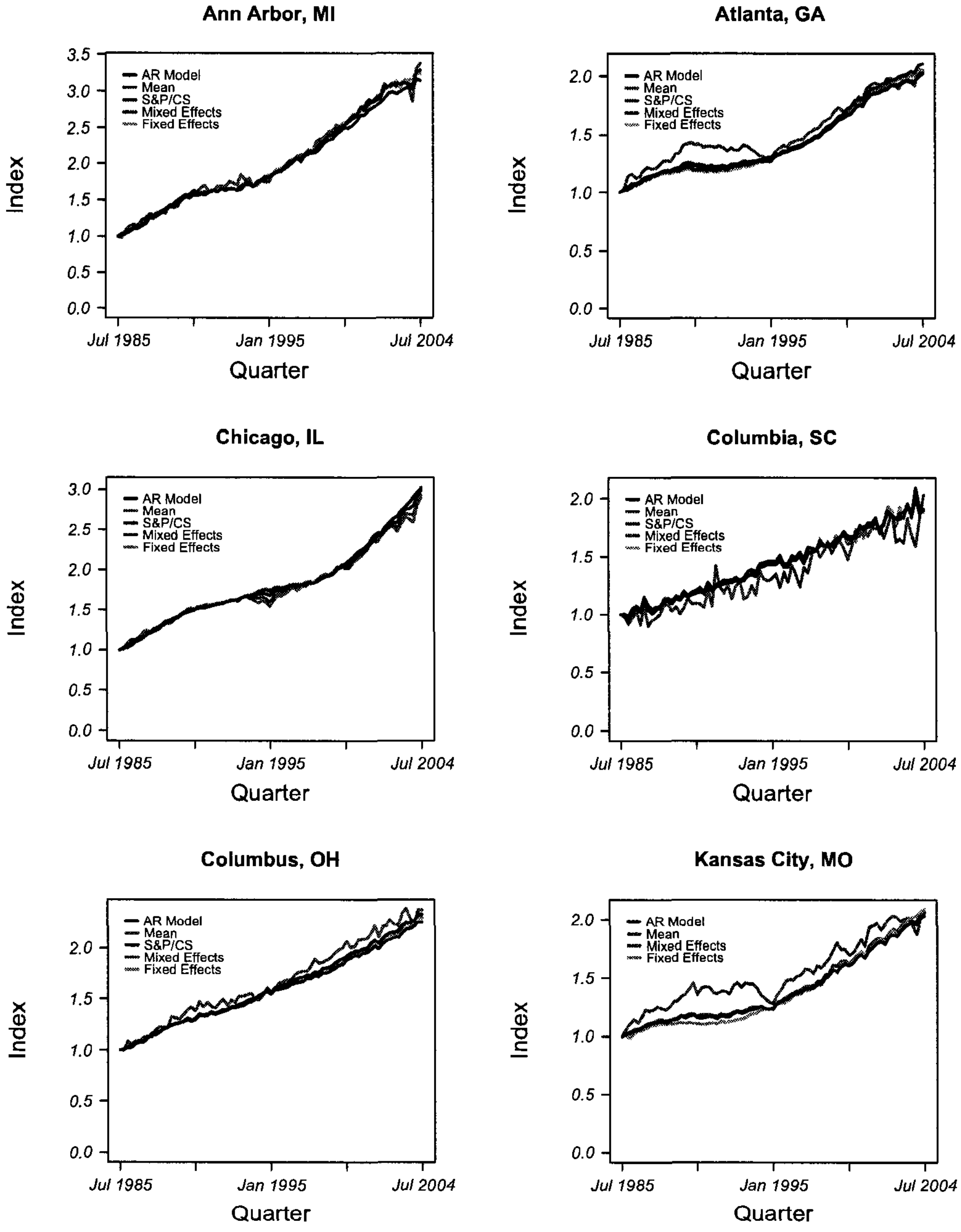


Figure D.5: Global Indices: Ann Arbor, MI-Kansas City, MO


Ann Arbor, MI

Atlanta, GA

Chicago, IL

Columbia, SC

Columbus, OH

Kansas City, MO

Figure D.6: Global Indices: Lexington, KY-Orlando, FL

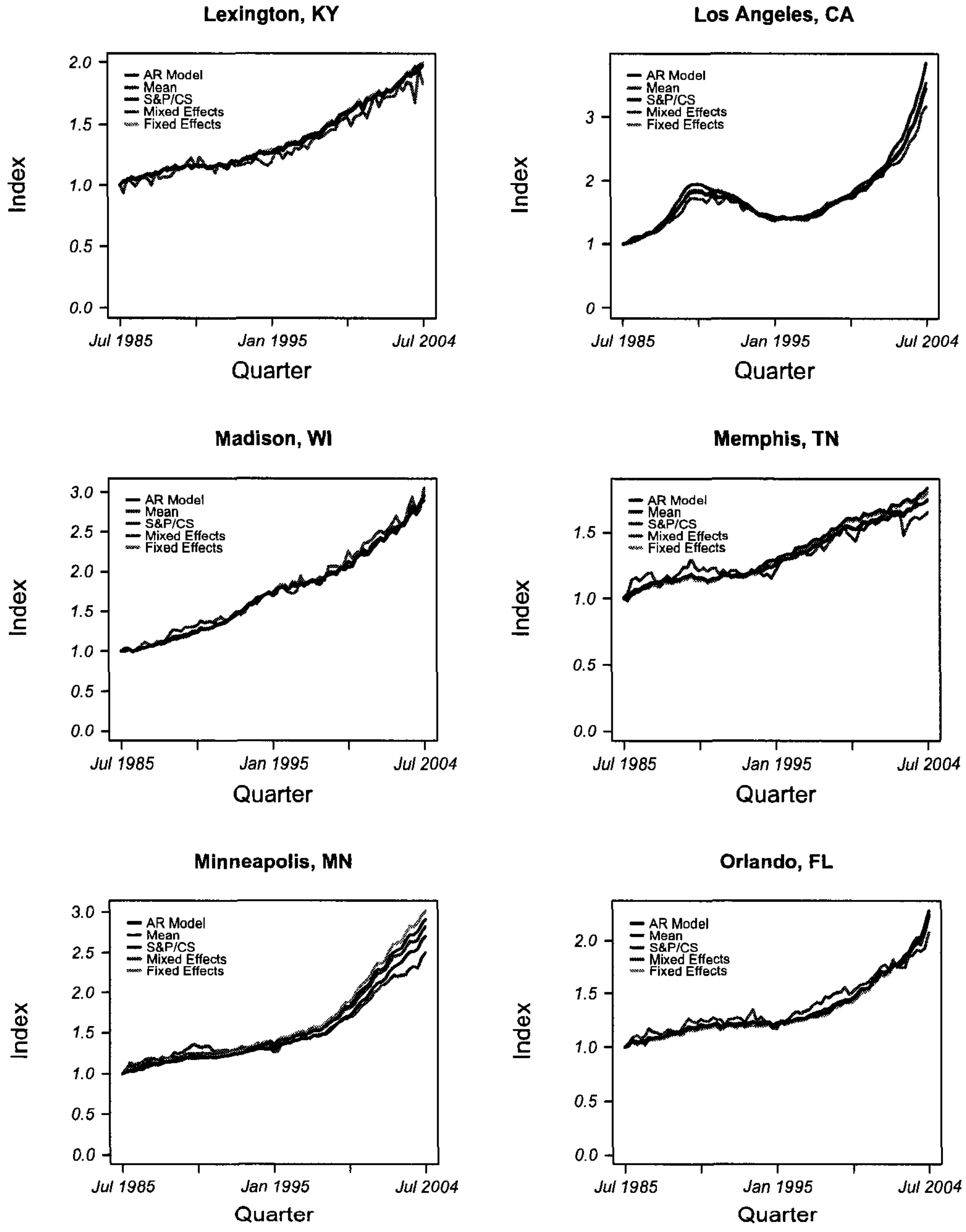


Figure D.7: Global Indices: Philadelphia, PA-Seattle, WA

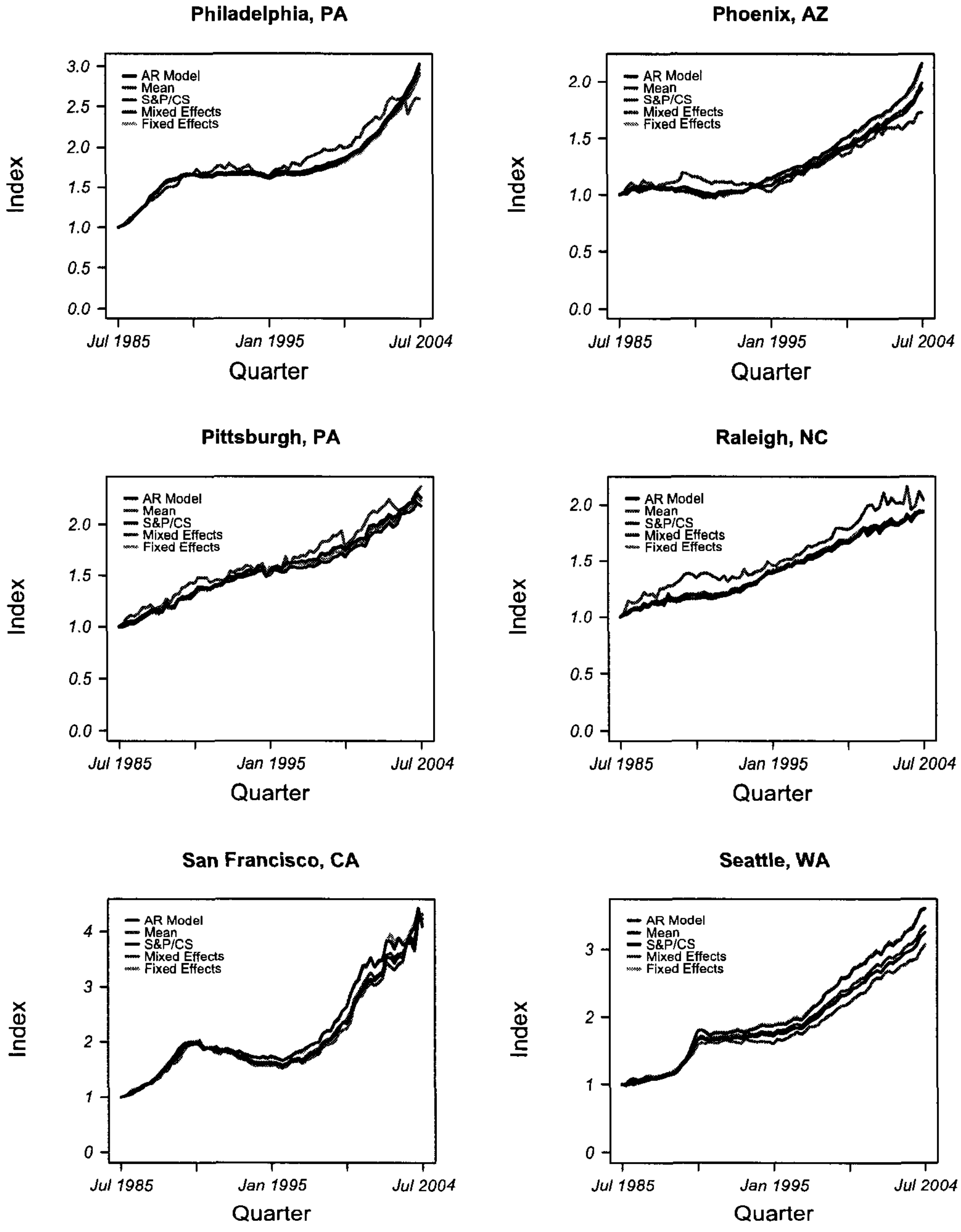
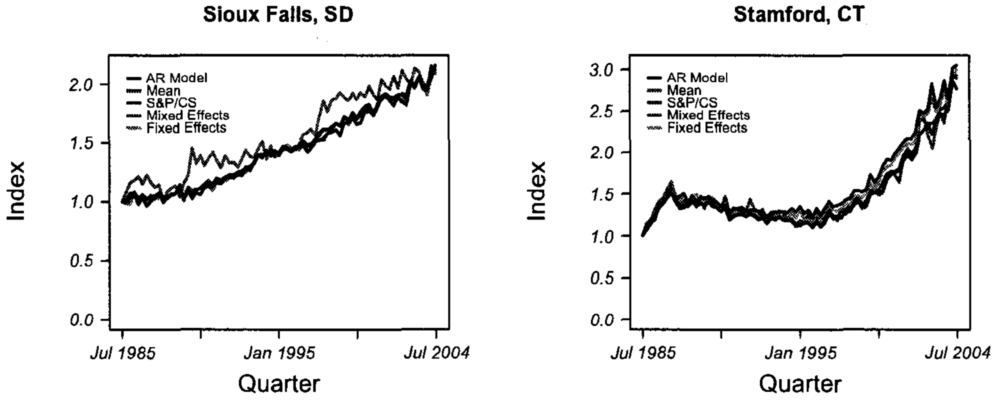


Figure D.8: Global Indices: Sioux Falls, SD-Stamford, CT



Appendix E

Derivations for the Local Autoregressive Model

In this appendix, we derive the updating functions for the local autoregressive model. Here, we add in a random effects term for zip code. The model to be fitted is:

$$\begin{aligned} y_{i,1,z} &= \mu + \beta_{t(i,1)} + \tau_z + \varepsilon_{i,1,z} & j = 1 \\ y_{i,j,z} &= \mu + \beta_{t(i,j)} + \tau_z + \phi^{\gamma(i,j)}\left(y_{i,j-1,z} - \mu - \beta_{t(i,j-1)} - \tau_z\right) + \varepsilon_{i,j,z} & j > 1 \end{aligned} \tag{E.1}$$

where $\tau_z \overset{iid}{\sim} \mathcal{N}\left(0, \sigma_\tau^2\right)$. Moreover, $\varepsilon_{i,1,z} \overset{iid}{\sim} \mathcal{N}\left(0, \frac{\sigma_\varepsilon^2}{1-\phi^2}\right)$, $\varepsilon_{i,j,z} \overset{iid}{\sim} \mathcal{N}\left(0, \frac{\sigma_\varepsilon^2\left(1-\phi^{2\gamma(j)}\right)}{1-\phi^2}\right)$, and all $\varepsilon_{i,j}$ are independent. Finally, $\sum_{t=1}^{T} n_t\beta_t = 0$ where n_t is the number of sales at time t. For clarity, we let $\boldsymbol{\beta} = \{\mu,\ \beta_1, \ldots ;, \beta_T\}$ and denote $\boldsymbol{\theta} = \left(\boldsymbol{\beta},\ \sigma_\varepsilon^2,\ \sigma_\tau^2,\ \phi\right)$.

For this model, the log-likelihood function is:

$$\begin{aligned} l(\boldsymbol{\theta};\, y) &= -\frac{N}{2}\log(2\pi) - \frac{1}{2}\sum_{z=1}^{Z}\log|\mathbf{V}_{z,z}| \\ &\quad -\frac{1}{2}\sum_{z=1}^{Z}(\mathbf{A}_z(\mathbf{y}_z - \mathbf{X}_z\boldsymbol{\beta}))'\mathbf{V}_{z,z}^{-1}(\mathbf{A}_z(\mathbf{y}_z - \mathbf{X}_z\boldsymbol{\beta})) \end{aligned} \tag{E.2}$$

where N is the number of observations and $\mathbf{V}_{z,z} = \sigma_\tau^2(\mathbf{A}_z\mathbf{1}_{n_z})(\mathbf{A}_z\mathbf{1}_{n_z})' + \frac{\sigma_\varepsilon^2}{1-\phi^2}diag(\mathbf{r}_z)$ where $diag(\mathbf{v})$ is a diagonal matrix with elements $\mathbf{u}$. Let n_z be the number of observations in zip code z. Finally, let $\mathbf{w}_z = \mathbf{y}_z - \mathbf{X}_z\boldsymbol{\beta}$.

E.1 Updating Formulas

Given the complexity of the models, we cannot find explicit functions for updating each parameter other than for $\boldsymbol{\beta}$. For the remaining three parameters, we must find the zero of the partial derivative function. We start with a few important identities and formulas used in the following derivations [22]:

$$\begin{aligned} (\mathbf{A}+\mathbf{B})' &= \mathbf{A}'+\mathbf{B}' \\ tr(\mathbf{AB}) &= tr(\mathbf{BA}) \\ \frac{\partial|\mathbf{A}|}{\partial\theta} &= |\mathbf{A}|tr\left(\mathbf{A}^{-1}\frac{\partial\mathbf{A}}{\partial\theta}\right) \\ \frac{\partial\mathbf{x}'\mathbf{A}\mathbf{x}}{\partial\mathbf{x}} &= \mathbf{x}'\left(\mathbf{A}'+\mathbf{A}\right) \\ \frac{\partial\left(\mathbf{A}\mathbf{x}+\mathbf{y}\right)'\mathbf{C}\left(\mathbf{D}\mathbf{x}+\mathbf{w}\right)}{\partial\mathbf{x}} &= \left(\mathbf{D}\mathbf{x}+\mathbf{w}\right)'\mathbf{C}'\mathbf{A} + \left(\mathbf{A}\mathbf{x}+\mathbf{y}\right)'\mathbf{C}\mathbf{D} \\ \frac{\partial\mathbf{XY}}{\partial\theta} &= \left(\frac{\partial\mathbf{X}}{\partial\theta}\right)\mathbf{Y} + \mathbf{X}\left(\frac{\partial\mathbf{Y}}{\partial\theta}\right) \qquad \text{(Product Rule)} \\ \frac{\partial f(\mathbf{A})}{\partial\theta} &= tr\left(\left(\frac{\partial f(\mathbf{A})}{\partial\mathbf{A}}\right)'\frac{\partial\mathbf{A}}{\partial\theta}\right) \qquad \text{(Chain Rule)} \end{aligned}$$

where $\mathbf{A}$ and $\mathbf{B}$ are appropriately sized matrices, θ is a parameter, $f(\cdot)$ is a function, and $tr(\cdot)$ is the trace of a matrix.

Update for $\{\boldsymbol{\beta}\}$

The estimates for $\{\mu, \beta_1, \ldots, \beta_{T-1}\}$ are given by:

$$\begin{aligned}
\frac{\partial l}{\partial \boldsymbol{\beta}} &= -\frac{1}{2}\sum_{z=1}^{Z} -2\left(\mathbf{A}_z\mathbf{X}_z\right)'\mathbf{V}_{z,z}^{-1}\left(\mathbf{A}_z\left(\mathbf{y}_z - \mathbf{X}_z\boldsymbol{\beta}\right)\right) \\
0 &= \sum_{z=1}^{Z}\left(\mathbf{A}_z\mathbf{X}_z\right)'\mathbf{V}_{z,z}^{-1}\mathbf{A}_z\mathbf{y}_z - \sum_{z=1}^{Z}\left(\mathbf{A}_z\mathbf{X}_z\right)'\mathbf{V}_{z,z}^{-1}\mathbf{A}_z\mathbf{X}_z\boldsymbol{\beta} \\
0 &= \sum_{z=1}^{Z}\left(\mathbf{A}_z\mathbf{X}_z\right)'\mathbf{V}_{z,z}^{-1}\mathbf{A}_z\mathbf{y}_z - \left(\sum_{z=1}^{Z}\left(\mathbf{A}_z\mathbf{X}_z\right)'\mathbf{V}_{z,z}^{-1}\mathbf{A}_z\mathbf{X}_z\right)\boldsymbol{\beta} \\
\hat{\boldsymbol{\beta}} &= \left(\sum_{z=1}^{Z}\left(\mathbf{A}_z\mathbf{X}_z\right)'\mathbf{V}_{z,z}^{-1}\mathbf{A}_z\mathbf{X}_z\right)^{-1}\sum_{z=1}^{Z}\left(\mathbf{A}_z\mathbf{X}_z\right)'\mathbf{V}_{z,z}^{-1}\mathbf{A}_z\mathbf{y}_z. \qquad \text{(E.3)}
\end{aligned}$$

After all of the parameters are estimated, $\hat{\beta}_T = -\sum_{t=1}^{T-1}\frac{n_t}{n_T}\hat{\beta}_t$.

Update for σ_τ^2

To find the variance for the random effects, we first note that $\frac{\partial \mathbf{V}_{zz}}{\partial \sigma_\tau^2} = (\mathbf{A}_z \mathbf{1}_{n_z})(\mathbf{A}_z \mathbf{1}_{n_z})'$. This is because $\frac{\partial |A|}{\partial \theta} = tr\left[A^{-1}\frac{\partial A}{\partial \theta}\right]$. Then, if we take the derivative with respect to σ_τ^2,

$$
\begin{aligned}
\frac{\partial l}{\partial \sigma_\tau^2} &= -\frac{1}{2}\sum_{z=1}^{Z} tr\left(V_{z,z}^{-1}(\mathbf{A}_z \mathbf{1}_{n_z})(\mathbf{A}_z \mathbf{1}_{n_z})'\right) \\
&\quad -\frac{1}{2}\sum_{z=1}^{Z} -(\mathbf{A}_z \mathbf{w}_z)'\mathbf{V}_{z,z}^{-1}(\mathbf{A}_z \mathbf{1}_{n_z})(\mathbf{A}_z \mathbf{1}_{n_z})'\mathbf{V}_{z,z}^{-1}(\mathbf{A}_z \mathbf{w}_z) \\
0 &= \sum_{z=1}^{Z} tr\left(V_{z,z}^{-1}(\mathbf{A}_z \mathbf{1}_{n_z})(\mathbf{A}_z \mathbf{1}_{n_z})'\right) \\
&\quad +\sum_{z=1}^{Z} -(\mathbf{A}_z \mathbf{w}_z)'\mathbf{V}_{z,z}^{-1}(\mathbf{A}_z \mathbf{1}_{n_z})(\mathbf{A}_z \mathbf{1}_{n_z})'\mathbf{V}_{z,z}^{-1}(\mathbf{A}_z \mathbf{w}_z). \qquad \text{(E.4)}
\end{aligned}
$$

Update for σ_ε^2

Using the fact that $\frac{\partial \mathbf{V}_{z,z}}{\partial \sigma_\varepsilon^2} = \frac{1}{1-\phi^2} diag(\mathbf{r}_z)$, we can compute the update for the variance parameter in the error term:

$$
\begin{aligned}
\frac{\partial l}{\partial \sigma_\varepsilon^2} &= -\frac{1}{2}\sum_{z=1}^{Z} tr\left(\mathbf{V}_{z,z}^{-1}\frac{1}{1-\phi^2} diag(\mathbf{r}_z)\right) \\
&\quad -\frac{1}{2}\sum_{z=1}^{Z} -\frac{1}{1-\phi^2}(\mathbf{A}_z \mathbf{w}_z)'\mathbf{V}_{z,z}^{-1} diag(\mathbf{r}_z)\mathbf{V}_{z,z}^{-1}(\mathbf{A}_z \mathbf{w}_z) \\
0 &= -\sum_{z=1}^{Z} tr\left(\mathbf{V}_{z,z}^{-1} diag(\mathbf{r}_z)\right) + \sum_{z=1}^{Z}(\mathbf{A}_z \mathbf{w}_z)'\mathbf{V}_{z,z}^{-1} diag(\mathbf{r}_z)\mathbf{V}_{z,z}^{-1}(\mathbf{A}_z \mathbf{w}_z). \qquad \text{(E.5)}
\end{aligned}
$$

Update for ϕ

We will find the derivative of the log likelihood function with respect to ϕ in several parts. To start, we compute $\frac{\partial \mathbf{V}_{z,z}}{\partial \phi}$:

$$\begin{aligned} \frac{\partial \mathbf{V}_{z,z}}{\partial \phi} &= \sigma_\tau^2 \left(\frac{\partial(\mathbf{A}_z \mathbf{1}_{n_z})}{\partial \phi}\right)(\mathbf{A}_z \mathbf{1}_{n_z})' + \sigma_\tau^2 (\mathbf{A}_z \mathbf{1}_{n_z}) \left(\frac{\partial(\mathbf{A}_z \mathbf{1}_{n_z})}{\partial \phi}\right)' \\ &+ \frac{2\phi\sigma_\varepsilon^2}{(1-\phi^2)^2} diag(\mathbf{r}_z) + \frac{\sigma_\varepsilon^2}{1-\phi^2}\frac{\partial diag(\mathbf{r}_z)}{\partial \phi}. \end{aligned} \tag{E.6}$$

Next, we compute $\frac{\partial \log |\mathbf{V}_{z,z}|}{\partial \phi}$:

$$\begin{aligned} \frac{\partial \log |\mathbf{V}_{z,z}|}{\partial \phi} &= tr\left\{\mathbf{V}_{z,z}^{-1}\frac{\partial \mathbf{V}_{z,z}}{\partial \phi}\right\} \\ &= tr\left\{\mathbf{V}_{z,z}^{-1}\left[\sigma_\tau^2 \left(\frac{\partial(\mathbf{A}_z \mathbf{1}_{n_z})}{\partial \phi}\right)(\mathbf{A}_z \mathbf{1}_{n_z})' + \sigma_\tau^2 (\mathbf{A}_z \mathbf{1}_{n_z}) \left(\frac{\partial(\mathbf{A}_z \mathbf{1}_{n_z})}{\partial \phi}\right)'\right.\right. \\ &\left.\left. + \frac{2\phi\sigma_\varepsilon^2}{(1-\phi^2)^2} diag(\mathbf{r}_z) + \frac{\sigma_\varepsilon^2}{1-\phi^2}\frac{\partial diag(\mathbf{r}_z)}{\partial \phi}\right]\right\}. \end{aligned} \tag{E.7}$$

Finally, we compute $\frac{\partial \mathbf{V}_{z,z}^{-1}}{\partial \phi}$:

$$\begin{aligned} \frac{\partial \mathbf{V}_{z,z}^{-1}}{\partial \phi} &= -\mathbf{V}_{z,z}^{-1}\frac{\partial \mathbf{V}_{z,z}}{\partial \phi}\mathbf{V}_{z,z}^{-1} \\ &= -\mathbf{V}_{z,z}^{-1}\left[\sigma_\tau^2 \left(\frac{\partial(\mathbf{A}_z \mathbf{1}_{n_z})}{\partial \phi}\right)(\mathbf{A}_z \mathbf{1}_{n_z})' + \sigma_\tau^2 (\mathbf{A}_z \mathbf{1}_{n_z}) \left(\frac{\partial(\mathbf{A}_z \mathbf{1}_{n_z})}{\partial \phi}\right)'\right. \\ &\left. + \frac{2\phi\sigma_\varepsilon^2}{(1-\phi^2)^2} diag(\mathbf{r}_z) + \frac{\sigma_\varepsilon^2}{1-\phi^2}\frac{\partial diag(\mathbf{r}_z)}{\partial \phi}\right]\mathbf{V}_{z,z}^{-1}. \end{aligned} \tag{E.8}$$

Using (E.8), we can apply the product and chain rules to compute the partial derivative of $f_z(\phi) = (\mathbf{A}_z\mathbf{w}_z)\mathbf{V}_{z,z}^{-1}(\mathbf{A}_z\mathbf{w}_z)'$ with respect to ϕ.

$$\begin{aligned} \frac{\partial f_z}{\partial \phi} &= \left(\frac{\partial \mathbf{A}_z}{\partial \phi}\mathbf{w}_z\right)'\mathbf{V}_{z,z}^{-1}(\mathbf{A}_z\mathbf{w}_z) + (\mathbf{A}_z\mathbf{w}_z)'\mathbf{V}_{z,z}^{-1}\left(\frac{\partial \mathbf{A}_z}{\partial \phi}\mathbf{w}_z\right) \\ &-(\mathbf{A}_z\mathbf{w}_z)'\mathbf{V}_{z,z}^{-1}\left[\sigma_\tau^2 \left(\frac{\partial(\mathbf{A}_z \mathbf{1}_{n_z})}{\partial \phi}\right)(\mathbf{A}_z \mathbf{1}_{n_z})' + \sigma_\tau^2 (\mathbf{A}_z \mathbf{1}_{n_z}) \left(\frac{\partial(\mathbf{A}_z \mathbf{1}_{n_z})}{\partial \phi}\right)'\right. \\ &\left. + \frac{2\phi\sigma_\varepsilon^2}{(1-\phi^2)^2} diag(\mathbf{r}_z) + \frac{\sigma_\varepsilon^2}{1-\phi^2}\frac{\partial diag(\mathbf{r}_z)}{\partial \phi}\right]\mathbf{V}_{z,z}^{-1}(\mathbf{A}_z\mathbf{w}_z). \end{aligned} \tag{E.9}$$

Putting (E.7) and (E.9) together, we obtain the desired partial derivative:

$$\begin{aligned}
\frac{\partial l}{\partial \phi} &= -\frac{1}{2}\sum_{z=1}^{Z} tr\left\{\mathbf{V}_{z,z}^{-1}\left(\sigma_{\tau}^{2}\left(\frac{\partial(\mathbf{A}_z\mathbf{1}_{n_z})}{\partial\phi}\right)(\mathbf{A}_z\mathbf{1}_{n_z})' + \sigma_{\tau}^{2}(\mathbf{A}_z\mathbf{1}_{n_z})\left(\frac{\partial(\mathbf{A}_z\mathbf{1}_{n_z})}{\partial\phi}\right)'\right.\right. \\
&\quad \left.\left. + \frac{2\phi\sigma_{\varepsilon}^{2}}{(1-\phi^2)^2} diag(\mathbf{r}_z) + \frac{\sigma_{\varepsilon}^{2}}{1-\phi^2}\frac{\partial diag(\mathbf{r}_z)}{\partial\phi}\right)\right\} \\
&\quad -\frac{1}{2}\sum_{z=1}^{Z}\left(\frac{\partial \mathbf{A}_z}{\partial\phi}\mathbf{w}_z\right)'\mathbf{V}_{z,z}^{-1}(\mathbf{A}_z\mathbf{w}_z) - \frac{1}{2}\sum_{z=1}^{Z}(\mathbf{A}_z\mathbf{w}_z)'\mathbf{V}_{z,z}^{-1}\left(\frac{\partial \mathbf{A}_z}{\partial\phi}\mathbf{w}_z\right) \\
&\quad +\frac{1}{2}\sum_{z=1}^{Z}\left[(\mathbf{A}_z\mathbf{w}_z)'\mathbf{V}_{z,z}^{-1}\left[\sigma_{\tau}^{2}\left(\frac{\partial(\mathbf{A}_z\mathbf{1}_{n_z})}{\partial\phi}\right)(\mathbf{A}_z\mathbf{1}_{n_z})' + \sigma_{\tau}^{2}(\mathbf{A}_z\mathbf{1}_{n_z})\left(\frac{\partial(\mathbf{A}_z\mathbf{1}_{n_z})}{\partial\phi}\right)'\right.\right. \\
&\quad \left.\left. + \frac{2\phi\sigma_{\varepsilon}^{2}}{(1-\phi^2)^2} diag(\mathbf{r}_z) + \frac{\sigma_{\varepsilon}^{2}}{1-\phi^2}\frac{\partial diag(\mathbf{r}_z)}{\partial\phi}\right]\mathbf{V}_{z,z}^{-1}(\mathbf{A}_z\mathbf{w}_z)\right] \\
0 &= -\sum_{z=1}^{Z} tr\left\{\mathbf{V}_{z,z}^{-1}\left(\sigma_{\tau}^{2}\left(\frac{\partial(\mathbf{A}_z\mathbf{1}_{n_z})}{\partial\phi}\right)(\mathbf{A}_z\mathbf{1}_{n_z})' + \sigma_{\tau}^{2}(\mathbf{A}_z\mathbf{1}_{n_z})\left(\frac{\partial(\mathbf{A}_z\mathbf{1}_{n_z})}{\partial\phi}\right)'\right.\right. \\
&\quad \left.\left. + \frac{2\phi\sigma_{\varepsilon}^{2}}{(1-\phi^2)^2} diag(\mathbf{r}_z) + \frac{\sigma_{\varepsilon}^{2}}{1-\phi^2}\frac{\partial diag(\mathbf{r}_z)}{\partial\phi}\right)\right\} \\
&\quad -\sum_{z=1}^{Z}\left(\frac{\partial \mathbf{A}_z}{\partial\phi}\mathbf{w}_z\right)'\mathbf{V}_{z,z}^{-1}(\mathbf{A}_z\mathbf{w}_z) - \sum_{z=1}^{Z}(\mathbf{A}_z\mathbf{w}_z)'\mathbf{V}_{z,z}^{-1}\left(\frac{\partial \mathbf{A}_z}{\partial\phi}\mathbf{w}_z\right) \\
&\quad +\sum_{z=1}^{Z}\left[(\mathbf{A}_z\mathbf{w}_z)'\mathbf{V}_{z,z}^{-1}\left[\sigma_{\tau}^{2}\left(\frac{\partial(\mathbf{A}_z\mathbf{1}_{n_z})}{\partial\phi}\right)(\mathbf{A}_z\mathbf{1}_{n_z})' + \sigma_{\tau}^{2}(\mathbf{A}_z\mathbf{1}_{n_z})\left(\frac{\partial(\mathbf{A}_z\mathbf{1}_{n_z})}{\partial\phi}\right)'\right.\right. \\
&\quad \left.\left. + \frac{2\phi\sigma_{\varepsilon}^{2}}{(1-\phi^2)^2} diag(\mathbf{r}_z) + \frac{\sigma_{\varepsilon}^{2}}{1-\phi^2}\frac{\partial diag(\mathbf{r}_z)}{\partial\phi}\right]\mathbf{V}_{z,z}^{-1}(\mathbf{A}_z\mathbf{w}_z)\right]. \qquad \text{(E.10)}
\end{aligned}$$

Simplifying Computations for $\mathbf{V}_{z,z}$

Although we have reduced $\mathbf{V}$ into blocks, the dimension of these blocks is often too large for a computer to invert. It is possible to simplify the computations by exploiting the structure of $\mathbf{V}_{z,z}$. To do this, we use the following identities. Let $\mathbf{D}$ be a diagonal matrix and $\mathbf{I}$ the identity matrix. Then, for appropriately sized matrices $\mathbf{U}$, $\mathbf{W}$ and vectors $\mathbf{u}$ and $\mathbf{w}$,

$$\begin{aligned}
(\mathbf{UW})^{-1} &= \mathbf{W}^{-1}\mathbf{U}^{-1} \\
\mathbf{D}+\mathbf{WW}' &= \mathbf{D}^{\frac{1}{2}}\left(\mathbf{I}+\mathbf{D}^{-\frac{1}{2}}\mathbf{WW}'\mathbf{D}^{\frac{1}{2}}\right)\mathbf{D}^{\frac{1}{2}} \\
(\mathbf{I}+\mathbf{UW})^{-1} &= \mathbf{I}-\mathbf{U}\left(\mathbf{I}-\mathbf{WU}\right)^{-1}\mathbf{W} \\
|\mathbf{I}+\mathbf{uw}'| &= 1+\mathbf{u}'\mathbf{w} \\
|\mathbf{UW}| &= |\mathbf{U}||\mathbf{W}| \\
\left|\mathbf{U}^k\right| &= |\mathbf{U}|^k.
\end{aligned}$$

Recall, that $\mathbf{V}_{z,z} = \frac{\sigma_\varepsilon^2}{1-\phi^2}diag(\mathbf{r}_z) + \sigma_\tau^2(\mathbf{A}_z\mathbf{1}_z)(\mathbf{A}_z\mathbf{1}_z)'$. If we let $\mathbf{D}_z = \frac{\sigma_\varepsilon^2}{1-\phi^2}diag(\mathbf{r}_z)$ and $\mathbf{W}_z = \sigma_\tau\mathbf{A}_z\mathbf{1}_z$, we can write:

$$\begin{aligned}
\mathbf{V}_{z,z}^{-1} &= \left(\frac{\sigma_\varepsilon^2}{1-\phi^2}diag(\mathbf{r}_z) + \sigma_\tau^2(\mathbf{A}_z\mathbf{1}_z)(\mathbf{A}_z\mathbf{1}_z)'\right)^{-1} \\
&= \left(\mathbf{D}_z + \left(\sigma_\tau\mathbf{A}_z\mathbf{1}_z\right)\left(\sigma_\tau\mathbf{A}_z\mathbf{1}_z\right)'\right)^{-1} \\
&= \mathbf{D}_z^{-\frac{1}{2}}\left[\mathbf{I}+\left(\mathbf{D}_z^{-\frac{1}{2}}\sigma_\tau\mathbf{A}_z\mathbf{1}_z\right)\left(\mathbf{D}_z^{-\frac{1}{2}}\sigma_\tau\mathbf{A}_z\mathbf{1}_z\right)'\right]^{-1}\mathbf{D}_z^{-\frac{1}{2}} \\
&= \mathbf{D}_z^{-\frac{1}{2}}\left[\mathbf{I}-\left(\mathbf{D}_z^{-\frac{1}{2}}\sigma_\tau\mathbf{A}_z\mathbf{1}_z\right)\left[\mathbf{I}+\left(\mathbf{D}_z^{-\frac{1}{2}}\sigma_\tau\mathbf{A}_z\mathbf{1}_z\right)'\left(\mathbf{D}_z^{-\frac{1}{2}}\sigma_\tau\mathbf{A}_z\mathbf{1}_z\right)\right]^{-1}\times\right. \\
&\quad \left.\left(\mathbf{D}_z^{-\frac{1}{2}}\sigma_\tau\mathbf{A}_z\mathbf{1}_z\right)'\right]\mathbf{D}_z^{-\frac{1}{2}} \\
&= \mathbf{D}_z^{-1} - \sigma_\tau^2\mathbf{D}_z^{-1}\left(\mathbf{A}_z\mathbf{1}_z\right)\underbrace{\left[\mathbf{I}+\sigma_\tau^2\left(\mathbf{1}_z\mathbf{A}_z\right)'\mathbf{D}_z^{-1}\left(\mathbf{1}_z\mathbf{A}_z\right)\right]^{-1}}_{\text{scalar}}\left(\mathbf{1}_z\mathbf{A}_z\right)'\mathbf{D}_z^{-1} \\
&= \mathbf{D}_z^{-1} - \sigma_\tau^2\left[1+\sigma_\tau^2\left(\mathbf{1}_z\mathbf{A}_z\right)'\mathbf{D}_z^{-1}\left(\mathbf{1}_z\mathbf{A}_z\right)\right]^{-1}\mathbf{D}_z^{-1}\mathbf{A}_z\mathbf{1}_z\mathbf{1}_z'\mathbf{A}'\mathbf{D}_z^{-1}. \qquad \text{(E.11)}
\end{aligned}$$

The final expression, (E.11) is much simpler because the inverse of a diagonal matrix is simply the reciprocal of its elements.

We can apply the same rules to compute $\mathbf{V}_{z,z}$.

$$\begin{aligned}
|\mathbf{V}_{z,z}| &= \left|\frac{\sigma_\varepsilon^2}{1-\phi^2} diag(\mathbf{r}_z) + \sigma_\tau^2(\mathbf{A}_z\mathbf{1}_z)(\mathbf{A}_z\mathbf{1}_z)'\right| \\
&= \left|\mathbf{D}_z + (\sigma_\tau\mathbf{A}_z\mathbf{1}_z)(\sigma_\tau\mathbf{A}_z\mathbf{1}_z)'\right| \\
&= \left|\mathbf{D}_z^{\frac{1}{2}}\right|\left|\mathbf{I} + \left(\mathbf{D}_z^{-\frac{1}{2}}\sigma_\tau\mathbf{A}_z\mathbf{1}_z\right)\left(\mathbf{D}_z^{-\frac{1}{2}}\sigma_\tau\mathbf{A}_z\mathbf{1}_z\right)'\right|\left|\mathbf{D}_z^{\frac{1}{2}}\right| \\
&= |\mathbf{D}_z|\left(1 + \sigma_\tau^2(\mathbf{1}_z\mathbf{A}_z)'\mathbf{D}_z^{-1}(\mathbf{1}_z\mathbf{A}_z)\right). \qquad \text{(E.12)}
\end{aligned}$$

This is easier to compute as $|\mathbf{D}_z|$ is just the product of the diagonal elements.

E.2 Deriving the BLUP

E.2.1 The Local Autoregressive Model

After estimating the likelihood parameters, we must compute the random effects. Using Henderson's (1975) notation and procedure, we derive the Best Linear Unbiased Predictors (BLUP) for each zip code. An introduction to BLUPs can be found in Sec. 5.4.1. The model, in matrix form is:

$$\mathbf{Ay} = \mathbf{AX}\boldsymbol{\beta} + \mathbf{AZ}\boldsymbol{\tau} + \mathbf{A}\boldsymbol{\varepsilon} \qquad \text{(E.13)}$$

where $\mathbf{y}$ is the vector of log prices, $\boldsymbol{\beta}$ the vector of time effects, and $\boldsymbol{\tau}$ is the vector of random effects. Let $\mathbf{X}$ and $\mathbf{Z}$ be the design matrices for the fixed (time) and random (zip code) effects respectively. As before, we let $\mathbf{A}$ denote the transformation matrix which applies the observed AR(1) process to the data. The random effects are distributed as $\boldsymbol{\tau} \sim \mathcal{N}\left(\mathbf{0}, \sigma_\tau^2\mathbf{I}_Z\right)$ where $\mathbf{I}_Z$ is the identity matrix of dimension Z, the number of zip codes. Finally, $\mathbf{A}\boldsymbol{\varepsilon} \sim \mathcal{N}\left(\mathbf{0}, \frac{\sigma_\varepsilon^2}{1-\phi^2}diag(\mathbf{r})\right)$. Henderson's method assumes that the parameters in the covariance matrices are known; however, we will be using the estimated values obtained from maximizing the likelihood function.

To obtain the BLUPs, we must maximize the logarithm of the joint density of $\mathbf{y}$ and $\boldsymbol{\tau}$ with respect to the random effects, $\boldsymbol{\tau}$ [36, p. 18]. This density is:

$$\begin{aligned} f(\mathbf{y},\boldsymbol{\tau}) &= \left(2\pi\sigma_\varepsilon^2\right)^{-\frac{N-1}{2}} \begin{vmatrix} \frac{\sigma_\tau^2}{\sigma_\varepsilon^2}\mathbf{I}_Z & \mathbf{0} \\ \mathbf{0} & \frac{1}{1-\phi^2}diag(\mathbf{r}) \end{vmatrix}^{-\frac{1}{2}} \times \\ & \exp\left\{ -\frac{1}{2\sigma_\varepsilon^2} \begin{bmatrix} \boldsymbol{\tau} \\ \mathbf{A}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\boldsymbol{\tau}\right) \end{bmatrix}' \begin{bmatrix} \frac{\sigma_\tau^2}{\sigma_\varepsilon^2}\mathbf{I}_Z & \mathbf{0} \\ \mathbf{0} & \frac{1}{1-\phi^2}diag(\mathbf{r}) \end{bmatrix}^{-1} \times \right. \\ & \left. \begin{bmatrix} \boldsymbol{\tau} \\ \mathbf{A}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\boldsymbol{\tau}\right) \end{bmatrix} \right\} \end{aligned}$$

where N is the number of observations. Taking the log of this density, we obtain:

$$\begin{aligned} \log f(\mathbf{y},\boldsymbol{\tau}) &= -\frac{N-1}{2}\log\left(2\pi\sigma_\varepsilon^2\right) - \frac{1}{2}\log \begin{vmatrix} \frac{\sigma_\tau^2}{\sigma_\varepsilon^2}\mathbf{I}_Z & \mathbf{0} \\ \mathbf{0} & \frac{1}{1-\phi^2}diag(\mathbf{r}) \end{vmatrix} \\ & -\frac{1}{2\sigma_\varepsilon^2} \left(\begin{bmatrix} \boldsymbol{\tau} \\ \mathbf{A}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\boldsymbol{\tau}\right) \end{bmatrix}' \begin{bmatrix} \frac{\sigma_\tau^2}{\sigma_\varepsilon^2}\mathbf{I}_Z & \mathbf{0} \\ \mathbf{0} & \frac{1}{1-\phi^2}diag(\mathbf{r}) \end{bmatrix}^{-1} \times \right. \\ & \left. \begin{bmatrix} \boldsymbol{\tau} \\ \mathbf{A}\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{Z}\boldsymbol{\tau}\right) \end{bmatrix} \right). \end{aligned}$$

If we maximize the log density above with respect to $\boldsymbol{\tau}$, we can drop the first two terms. We plug in the estimated values for $\boldsymbol{\beta}$, ϕ, σ_ε^2, and σ_τ^2. As $diag(\mathbf{r})$ and $\mathbf{A}$ are functions of ϕ, the estimated values of these matrices need to be plugged in as well. Ultimately, we want to minimize the following function.

$$\begin{aligned} g(\boldsymbol{\tau}) &= \begin{bmatrix} \boldsymbol{\tau} \\ \hat{\mathbf{A}}\left(\mathbf{y}-\mathbf{X}\hat{\beta}-\mathbf{Z}\boldsymbol{\tau}\right) \end{bmatrix}' \begin{bmatrix} \frac{\hat{\sigma}^2_\tau}{\hat{\sigma}^2_\varepsilon}\mathbf{I}_Z & \mathbf{0} \\ \mathbf{0} & \frac{1}{1-\hat{\phi}^2} diag(\hat{\mathbf{r}}) \end{bmatrix}^{-1} \begin{bmatrix} \boldsymbol{\tau} \\ \hat{\mathbf{A}}\left(\mathbf{y}-\mathbf{X}\hat{\beta}-\mathbf{Z}\boldsymbol{\tau}\right) \end{bmatrix} \\ &= \frac{\hat{\sigma}^2_\varepsilon}{\hat{\sigma}^2_\tau}\boldsymbol{\tau}'\mathbf{I}_Z\boldsymbol{\tau} \\ &\quad + \left(1-\hat{\phi}^2\right)\left(\hat{\mathbf{A}}\left(\mathbf{y}-\mathbf{X}\hat{\beta}-\mathbf{Z}\boldsymbol{\tau}\right)\right)' diag^{-1}(\hat{\mathbf{r}})\left(\hat{\mathbf{A}}\left(\mathbf{y}-\mathbf{X}\hat{\beta}-\mathbf{Z}\boldsymbol{\tau}\right)\right). \end{aligned}$$

where $diag^{-1}(\hat{\mathbf{r}})$ is the inverse of the diagonal matrix (the reciprocal of the diagonal elements). We take the partial derivative of (E.14) with respect to $\boldsymbol{\tau}$ next:

$$\begin{aligned} \frac{\partial g}{\partial \boldsymbol{\tau}} &= \frac{2\hat{\sigma}^2_\varepsilon}{\hat{\sigma}^2_\tau}\boldsymbol{\tau}'\mathbf{I}_Z - 2\left(1-\hat{\phi}^2\right)\left(\hat{\mathbf{A}}\left(\mathbf{y}-\mathbf{X}\hat{\beta}-\mathbf{Z}\boldsymbol{\tau}\right)\right)' diag^{-1}(\hat{\mathbf{r}})(\hat{\mathbf{A}}\mathbf{Z}) \\ 0 &= \frac{\hat{\sigma}^2_\varepsilon}{\hat{\sigma}^2_\tau}\boldsymbol{\tau}'\mathbf{I}_Z - \left(1-\hat{\phi}^2\right)\left(\left(\hat{\mathbf{A}}\left(\mathbf{y}-\mathbf{X}\hat{\beta}\right)\right)' - \left(\hat{\mathbf{A}}\mathbf{Z}\boldsymbol{\tau}\right)'\right) diag^{-1}(\hat{\mathbf{r}})(\hat{\mathbf{A}}\mathbf{Z}). \end{aligned}$$

Finally, we solve for $\boldsymbol{\tau}$:

$$\left(1-\hat{\phi}^2\right)\left(\hat{\mathbf{A}}\left(\mathbf{y}-\mathbf{X}\hat{\beta}\right)\right)' diag^{-1}(\hat{\mathbf{r}})(\hat{\mathbf{A}}\mathbf{Z}) = \frac{\hat{\sigma}^2_\varepsilon}{\hat{\sigma}^2_\tau}\boldsymbol{\tau}'\mathbf{I}_Z + \left(\hat{\mathbf{A}}\mathbf{Z}\boldsymbol{\tau}\right)' diag^{-1}(\hat{\mathbf{r}})(\hat{\mathbf{A}}\mathbf{Z}).$$

And the BLUP is:

$$\begin{aligned} \hat{\boldsymbol{\tau}} &= \left[\frac{\hat{\sigma}^2_\varepsilon}{\hat{\sigma}^2_\tau}\mathbf{I}_Z + \left(1-\hat{\phi}^2\right)\left(\hat{\mathbf{A}}\mathbf{Z}\right)' diag^{-1}(\hat{\mathbf{r}})\left(\hat{\mathbf{A}}\mathbf{Z}\right)\right]^{-1} \times \\ &\quad \left[\left(1-\hat{\phi}^2\right)(\hat{\mathbf{A}}\mathbf{Z})' diag^{-1}(\hat{\mathbf{r}})\left(\hat{\mathbf{A}}\left(\mathbf{y}-\mathbf{X}\hat{\beta}\right)\right)\right]. \end{aligned} \tag{E.14}$$

In block form, the BLUP is:

$$\begin{aligned} \hat{\tau}_z &= \left[\frac{\hat{\sigma}^2_\varepsilon}{\hat{\sigma}^2_\tau} + \left(1-\hat{\phi}^2\right)\left(\hat{\mathbf{A}}_z\mathbf{1}_z\right)' diag^{-1}(\mathbf{r}_z)\left(\hat{\mathbf{A}}_z\mathbf{1}_z\right)\right]^{-1} \times \\ &\quad \left(\left(1-\hat{\phi}^2\right)\left(\hat{\mathbf{A}}_z\mathbf{1}_z\right)' diag^{-1}(\mathbf{r}_z)\left(\hat{\mathbf{A}}_z\hat{\mathbf{w}}_z\right)\right). \end{aligned} \tag{E.15}$$

E.2.2 The Local Mixed Effects Model

Recall the mixed effects model described in Sec. 7.2.1:

$$\mathbf{y} \;=\; \mathbf{X}\boldsymbol{\beta} + \mathbf{W}\boldsymbol{\alpha} + \mathbf{Z}\boldsymbol{\tau} + \boldsymbol{\varepsilon} \tag{E.16}$$

where

$$\begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\tau} \\ \boldsymbol{\varepsilon} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \frac{\sigma_\alpha^2}{\sigma_\varepsilon^2}\mathbf{I}_I & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \frac{\sigma_\tau^2}{\sigma_\varepsilon^2}\mathbf{I}_Z & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}_N \end{bmatrix} \sigma_\varepsilon^2 \right).$$

Let $\mathbf{X}$, $\mathbf{W}$, and $\mathbf{Z}$ be the design matrices for the time effects, house effects, and zip code effects respectively. We need to estimate the BLUPs for both $\boldsymbol{\alpha}$ and $\boldsymbol{\tau}$. As in the previous section, we use Henderson's method to obtain the formulas.

For the model in (E.16), the log density for $\mathbf{y}$, $\boldsymbol{\alpha}$ and $\boldsymbol{\tau}$ is:

$$\begin{aligned} \log f(\mathbf{y}, \boldsymbol{\tau}, \boldsymbol{\alpha}) \;=\; & -\frac{N-1}{2}\log\left(2\pi\sigma_\varepsilon^2\right) - \frac{1}{2}\log \begin{vmatrix} \frac{\sigma_\alpha^2}{\sigma_\varepsilon^2}\mathbf{I}_I & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \frac{\sigma_\tau^2}{\sigma_\varepsilon^2}\mathbf{I}_Z & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}_N \end{vmatrix}^{-\frac{1}{2}} \\ & -\frac{1}{2\sigma_\varepsilon^2} \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\tau} \\ \mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{W}\boldsymbol{\alpha} - \mathbf{Z}\boldsymbol{\tau} \end{bmatrix}' \begin{bmatrix} \frac{\sigma_\alpha^2}{\sigma_\varepsilon^2}\mathbf{I}_I & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \frac{\sigma_\tau^2}{\sigma_\varepsilon^2}\mathbf{I}_Z & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{I}_N \end{bmatrix} \times \\ & \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\tau} \\ \mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{W}\boldsymbol{\alpha} - \mathbf{Z}\boldsymbol{\tau} \end{bmatrix}. \end{aligned}$$

Taking the partial derivative with respect to $\boldsymbol{\beta}$, $\boldsymbol{\alpha}$, and $\boldsymbol{\tau}$, we have:

$$\begin{aligned}
\frac{\partial \log f}{\partial \boldsymbol{\beta}} &= -2\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{W}\boldsymbol{\alpha}-\mathbf{Z}\boldsymbol{\tau}\right)'\mathbf{X} \\
\frac{\partial \log f}{\partial \boldsymbol{\alpha}} &= \frac{2\sigma_{\varepsilon}^2}{\sigma_{\alpha}^2}\boldsymbol{\alpha}'\mathbf{I}_I - 2\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{W}\boldsymbol{\alpha}-\mathbf{Z}\boldsymbol{\tau}\right)'\mathbf{W} \\
\frac{\partial \log f}{\partial \boldsymbol{\tau}} &= \frac{2\sigma_{\varepsilon}^2}{\sigma_{\tau}^2}\boldsymbol{\tau}'\mathbf{I}_Z - 2\left(\mathbf{y}-\mathbf{X}\boldsymbol{\beta}-\mathbf{W}\boldsymbol{\alpha}-\mathbf{Z}\boldsymbol{\tau}\right)'\mathbf{Z}.
\end{aligned}$$

Finally, we set the partial derivatives to zero and solve for the unknowns. To calculate the BLUPs, iterate through the following expressions until stable estimates are reached:

$$\begin{aligned}
\hat{\boldsymbol{\beta}} &= \left(\mathbf{X}'\mathbf{X}\right)^{-1}\mathbf{X}'\left(\mathbf{y}-\mathbf{W}\hat{\boldsymbol{\alpha}}-\mathbf{Z}\hat{\boldsymbol{\tau}}\right) \\
\hat{\boldsymbol{\alpha}} &= \left(\frac{\sigma_{\varepsilon}^2}{\sigma_{\alpha}^2}\mathbf{I}_I+\mathbf{W}'\mathbf{W}\right)^{-1}\mathbf{W}'\left(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}-\mathbf{Z}\hat{\boldsymbol{\tau}}\right) \\
\hat{\boldsymbol{\tau}} &= \left(\frac{\sigma_{\varepsilon}^2}{\sigma_{\tau}^2}\mathbf{I}_Z+\mathbf{Z}'\mathbf{Z}\right)^{-1}\mathbf{Z}'\left(\mathbf{y}-\mathbf{X}\hat{\boldsymbol{\beta}}-\mathbf{W}\hat{\boldsymbol{\alpha}}\right).
\end{aligned}$$

Appendix F

Additional Plots for Local Model

The complete set of plots for the local autoregressive model are contained in this appendix. Figs. F.1- F.4 are the AR(1) verification plots. Figs. F.5- F.8 are the index plots.

Figure F.1: Local AR(1) Assumption Check: Ann Arbor, MI-Kansas City, MO

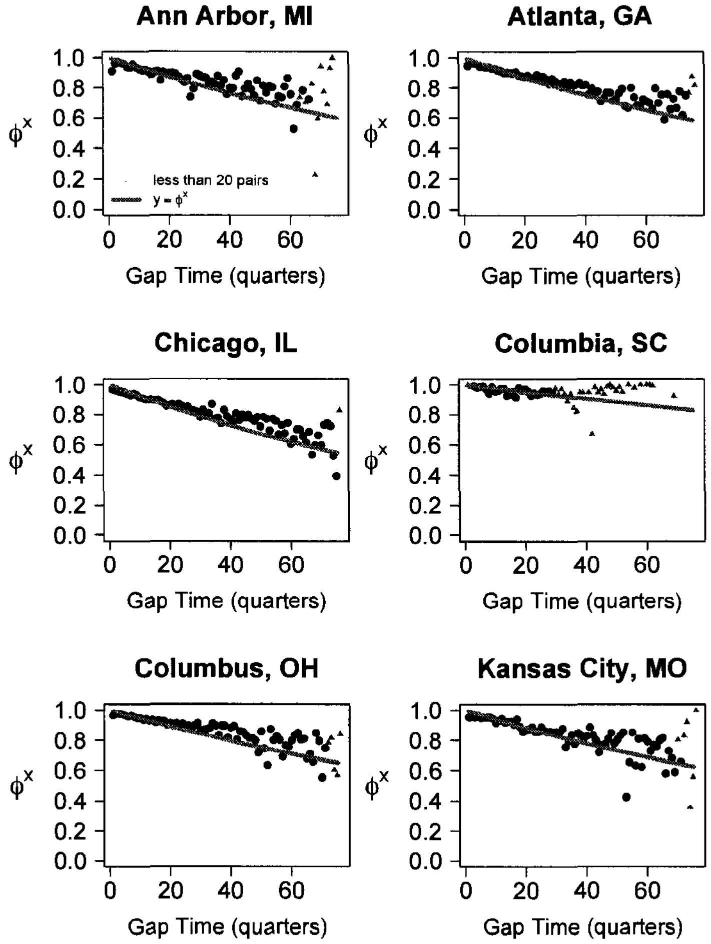


Figure F.2: Local AR(1) Assumption Check: Lexington, KY-Orlando, FL

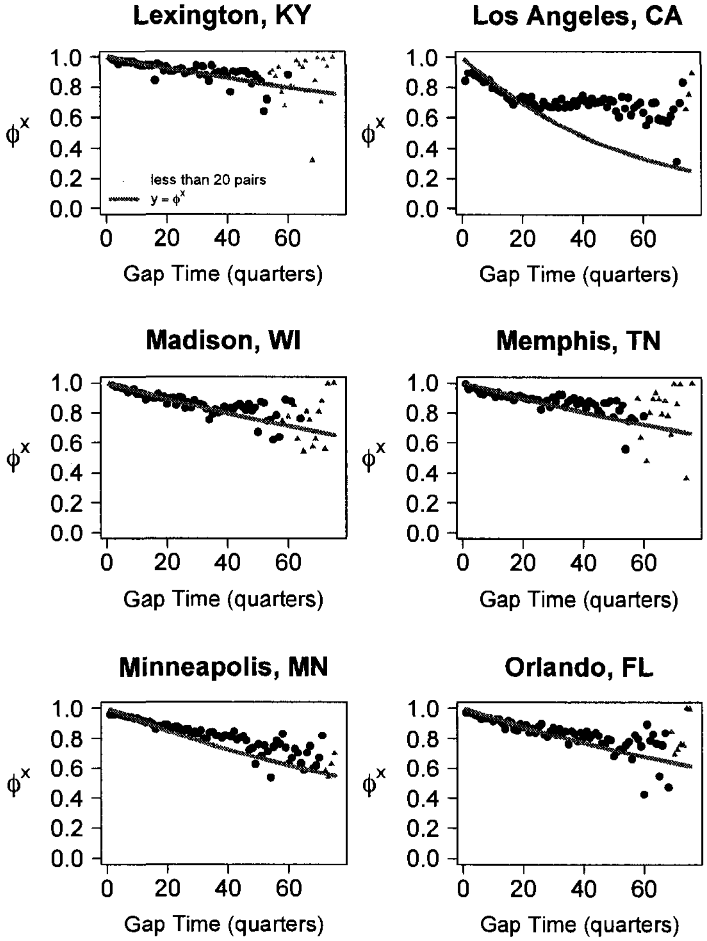


Figure F.3: Local AR(1) Assumption Check: Philadelphia, PA-Seattle, WA

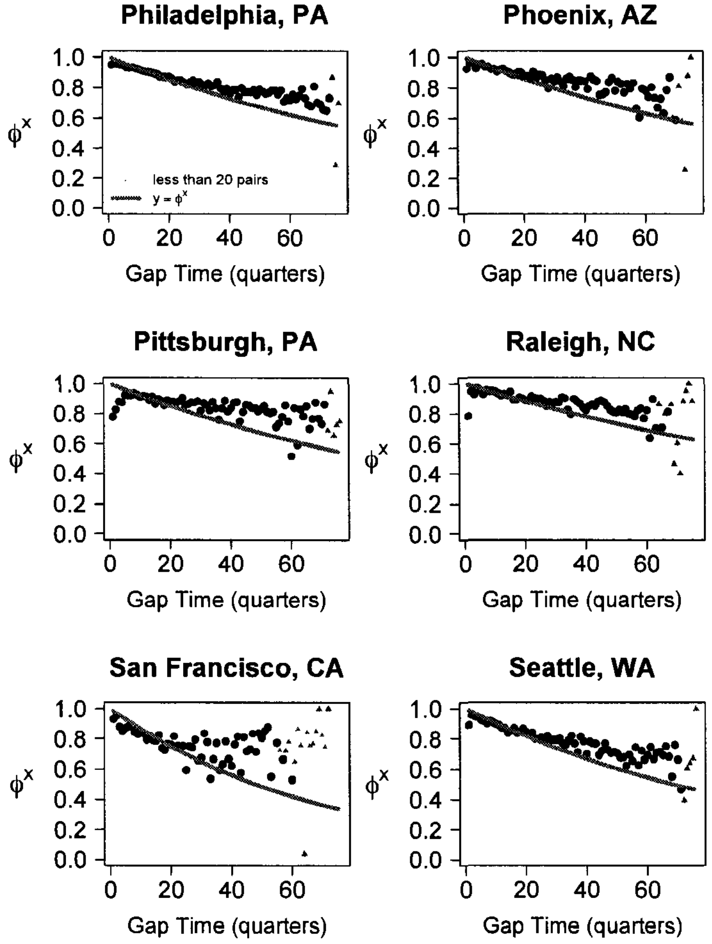


Figure F.4: Local AR(1) Assumption Check: Sioux Falls, SD-Stamford, CT

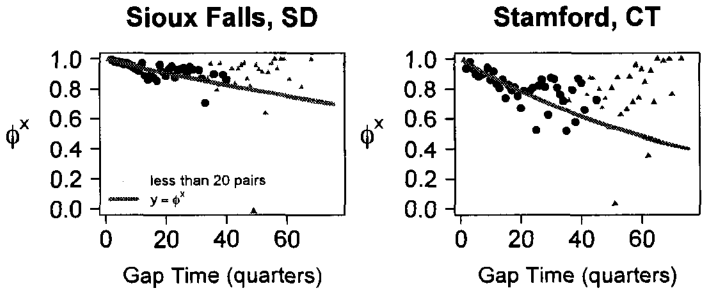


Figure F.5: Local Indices: Ann Arbor, MI-Kansas City, MO

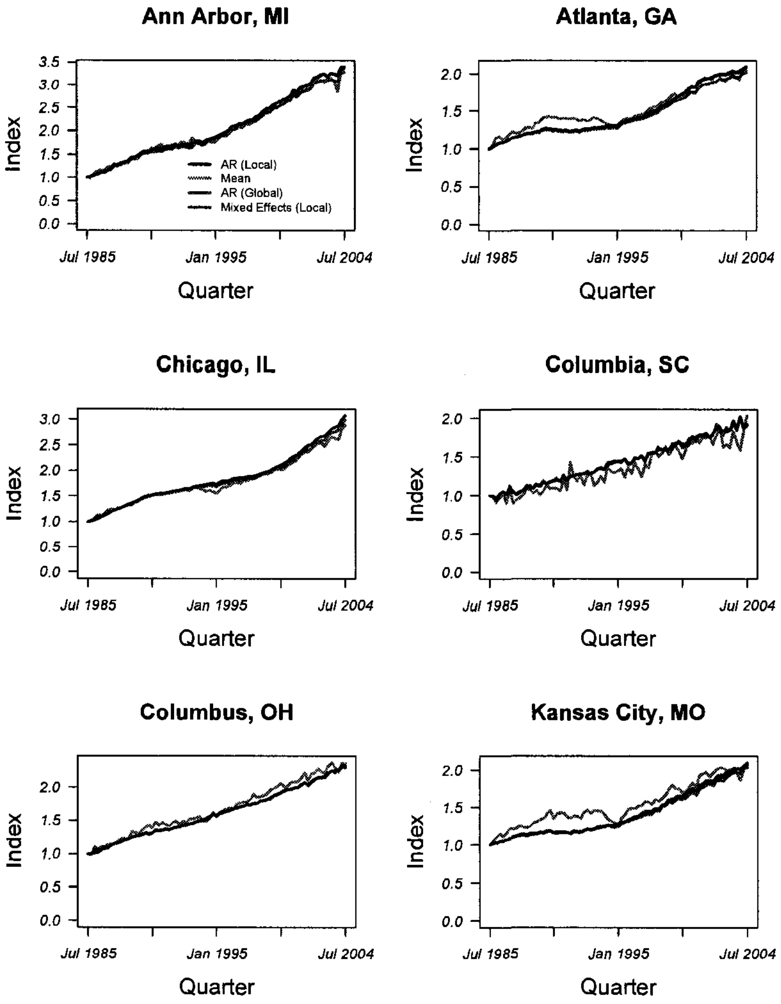


Figure F.6: Local Indices: Lexington, KY-Orlando, FL

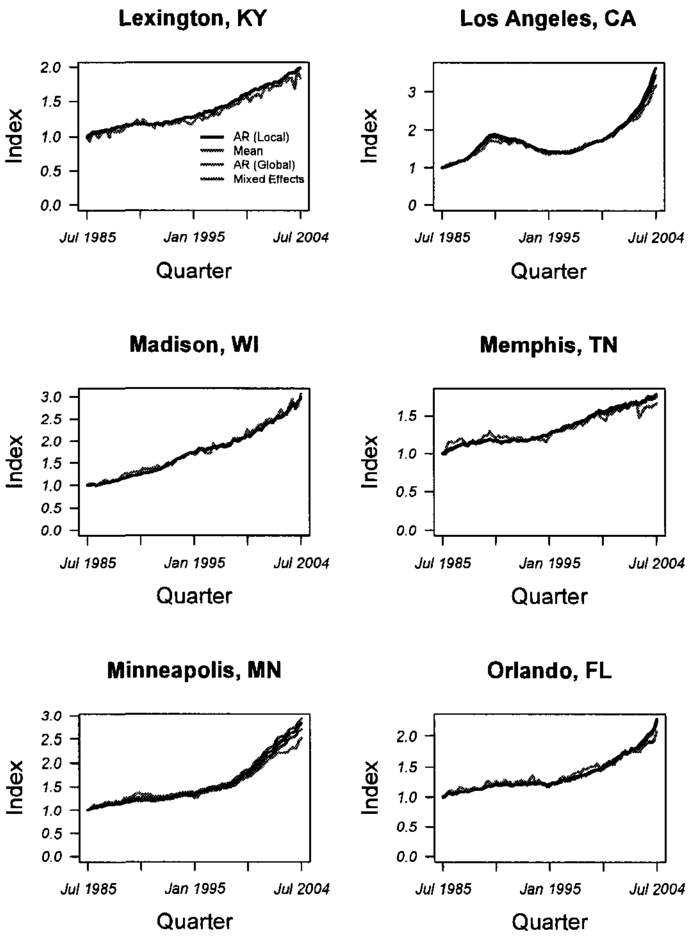


Figure F.7: Local Indices: Philadelphia, PA-Seattle, WA

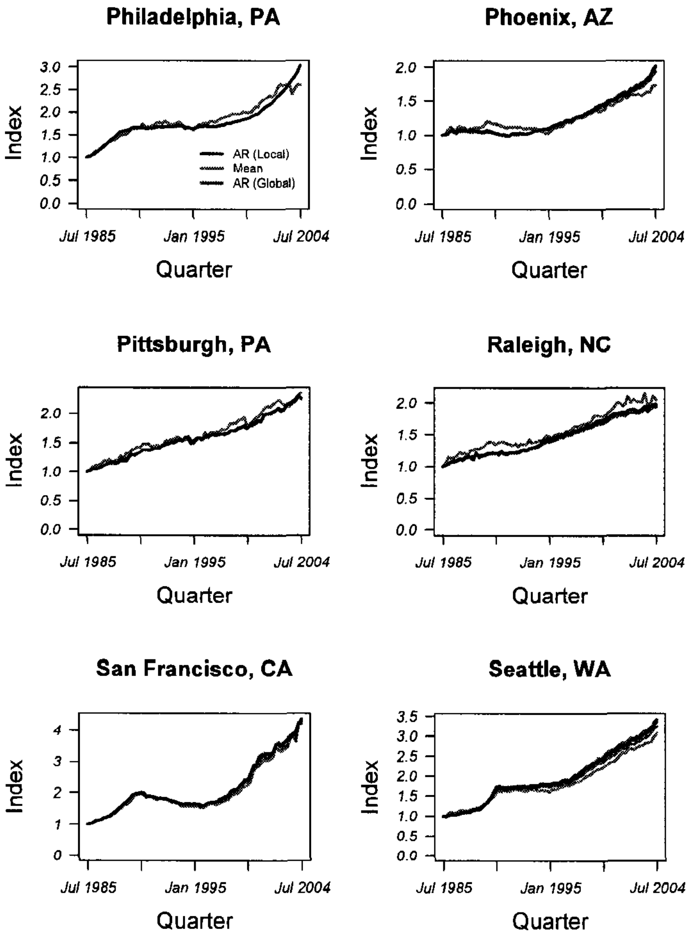
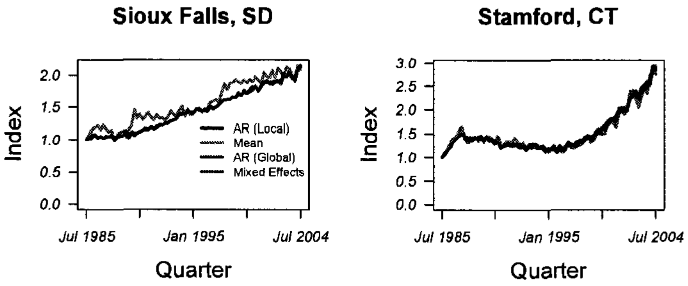


Figure F.8: Local Indices: Sioux Falls, SD-Stamford, CT



Bibliography

[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. Survey cities four california banks with possibly risky realty loans. *Los Angeles Times.* Dec. 30, 1989, 1.

[3] Beach, C.M., MacKinnon, J.G. (1978). A maximum likelihood procedure for regression with autocorrelated errors. *Econometrica.* **46** 51-58.

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

[5] Brown, L.D. (1986). *Fundamentals of Statistical Exponential Families–with Applications in Statistical Decision Theory.* Institute of Mathematical Statistics.

[6] Calhoun, C. (1996). OFHEO house price indices: HPI technical description. http://www.ofheo.gov

[7] Case, B., Pollakowski, H.O., Wachter, S. (1991). On choosing among house price index methodologies. *American Real Estate and Urban Economics Association Journal.* **19** 286-307.

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

[9] 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.

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

[11] Chamberlain, G. (1996). Lecture Notes 8–Panel Data II.

[12] Chatfield, C. (1993). Calculating interval forecasts. *Journal of Business and Economic Statistics.* **11** 121-135.

[13] Chatfield, C. (2001). Prediction intervals for time-series forecasting. *Principles of Forecasting: A Handbook for Researchers and Practitioners.* ed. Armstrong, J.S., Kluwer Academic Publishers, Massachusetts.

[14] Clapp, J.M., Giaccotto, C., Tirtiroglu, D. (1991). Housing price indices based on all transactions compared to repeat subsamples. *American Real Estate and Urban Economics Association Journal.* **19** 270-284.

[15] 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 Eocnomics and Sociology.* **57** 123-133.

[16] Gelfand, A.E., Ecker, M.D., Knight, J.R., Sirmans, C.F. (2004). The dynamics of location in home price. *Journal of Real Estate Finance and Economics.* **29** 149-166.

[17] Gelfand, A.E., Ghosh, S.K., Knight, J.R., Sirmans, C.F. (1998). Spatio-temporal modeling of residential sales data. *Journal of Business and Economic Statistics.* **16** 312-321.

[18] Gentle, J.E. (1998). *Numerical Linear Algebra for Applications in Statistics.* Springer-Verlag, New York.

[19] Glossary of postal terms–Publication 32 (1997).
http://www.usps.com

[20] Greene, W.H. (2003). *Econometric Analysis, 5th ed.* Prentice Hall, New Jersey.

[21] Harville, D.A. (1977). Maximum likelihood approaches to variance component estimation and to related problems. *Journal of the American Statistical Association.* **72** 320-338.

[22] Harville, D.A. (1997). *Matrix Algebra from a Statistician's Perspective.* Springer, New York.

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

[24] Kagarlis, M., Peterson, D., Eberlein, R., Fiddman, T. (2007). The Radar Logic^{TM}Daily Index.
http://www.radarlogic.com

[25] An improved National Price Index using Land Registry data (2006).
http://www.calnea.com

[26] LaMacchia, R.A, et al. (1994). *Geographic Areas Reference Manual.* U.S. Department of Commerce; Economics and Statistics Administrationg; Bereau of the Census.

[27] 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.

[28] Meissner, C., Satchell, S. (2007). A comparison of the Case-Shiller house price index methodology with the FT house price index methodology.
htp://www.acadametrics.co.uk

[29] Palmquist, R.B. (1982). Measuring environmental effects on property values without hedonic regression. *Journal of Urban Economics.* **11** 333-347.

[30] Pavlov, S.D. (2000). Space-varying regression coefficients: a semi-parametric approach applied to real estate markets. *Real Estate Economics.* **28** 249-283.

[31] Phillips, P.C.B. (1979). The sampling distribution of forecasts from a first-order regression. *Journal of Econometrics.* **9** 241-261.

[32] Prais, G.J., Winsten, C.B. (1954). Trend estimators and serial correlation. *Cowles Commission Discussion Paper No. 383.* Chicago, IL.

[33] Quigley, J.M. (1991). A simple hybrid model for estimating real estate price indexes. *Journal of Housing Economics.* **4** 1-12.

[34] Ravishanker, N., Dey, D.K. (2002). *A First Course in Linear Model Theory.* Chapman & Hall/CRC, Florida.

[35] Rice, J.A. (1995). *Mathematical Statistics and Data Analysis, 2nd ed.* Duxbury Press, California.

[36] Robinson, G.K. (1991). That BLUP is a good thing: the estimation of random effects. *Statistical Science.* **6** 15-32.

[37] Ross, S.M. (1996). *Stochastic Processes, 2nd ed.* John Wiley & Sons, Inc., New York.

[38] 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.

[39] Searle, S.R. (1982). *Matrix Algebra Useful for Statistics.* John Wiley & Sons, Inc., New York.

[40] 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.

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

[42] Shumway, R.H., Stoffer, D.S. (2006). *Time Series Analysis and Its Applications–With R Examples, 2nd ed.* Springer, New York.

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

[44] S&P/Case-Shiller® Home Price Indices (2007).
http://www.standardandpoors.com

[45] Stephens, W., Li Y., Lekkas, V., Abraham, J., Calhoun, C., Kimmer, T. (1995). Conventional mortgage home price index. *Journal of Housing Research.* **6** 389-418.

[46] Stock, J. (2002). Instrumental variables in statistics and econometrics. Entry in *International Encyclopedia of the Social Sciences.* 7577-7582. Elsevier, Amsterdam.

[47] Yeates, M.H. (1965). Some factors affecting the spatial distribution of Chicago land values, 1910-1960. *Economic Geography.* **41** 57-70.