USING PROC ARIMA TO MODEL TRENDS IN US HOME PRICES

Ross Bettinger, Analytical Consultant, Seattle, WA

ABSTRACT

We demonstrate the use of the Box-Jenkins time series modeling methodology to analyze US home prices. A selected set of regional market home prices was included as exogenous predictors to extend the analysis beyond a single time series. We used PROC ARIMA to build models with publicly-available estimates of national US home prices (Zillow Zestimate®) and regional market home prices distributed across the nation. Prewhitening of the exogenous predictors was performed to remedy the potential complication of autocorrelations in the predictors. We found that the resulting model satisfied the assumptions of the methodology and that the forecasts produced closely fit the actual data.

KEYWORDS

Dynamic regression, ARMA, ARMAX, transfer function, time series, model building, exogenous predictor, autoregressive, moving average, Box-Jenkins methodology, augmented Dickey-Fuller unit root test, differencing, prewhitening, nonstationarity, Zillow Zestimate

INTRODUCTION

The goal of this effort is to use the Box-Jenkins time series modeling methodology to build a time series model of US housing prices. PROC ARIMA is used to analyze the Zillow Zestimate home valuation metric for national US and selected regional housing markets. A time series model that includes exogenous predictors is called a *dynamic regression* model, and special care must be taken to ensure that the internal dynamics of each exogenous predictor is not confounded with the dynamics of any other predictor. This is equivalent to the assumption of independence among predictors in OLS regression.

The Box-Jenkins time series modeling methodology consists of three steps: identification of the lag structure of the series, estimation of model coefficients, and forecasting of future values. Models that consist of past response values only are called *autoregressive* (AR), models that contain past error terms only are called *moving average* (MA), and models that contain a mixture of autoregressive and moving average terms are denoted using the abbreviation *ARMA*. An ARMA model that includes e**X**ogenous predictors, i.e., variables equivalent to independent predictors in an OLS regression model, is called an ARMA**X** model. Special consideration must be given to ensure the independence of the lagged values of the exogenous predictor so that serial autocorrelation is not a confounding factor in the analysis. A *prewhitening* process is required to remove systematic autocorrelation from each exogenous predictor before it is included in modeling the response time series that is to be identified, estimated, and forecast.

When there are multiple exogenous predictors $X_{i,t}$ that influence a response variable Y_t, the same steps must be followed for the case in which there is only one X_t. Each exogenous predictor is a time series in itself that must be prewhitened independently of the others so that the identification of the Y_t response can be performed in the presence of the $X_{i,t}$. Once the response model has been identified, the estimation and forecasting steps follow directly.

ZESTIMATES® OF US HOME PRICES, APRIL 1997-APRIL 2013

Estimates of home prices for the time interval April 1997 to April 2013 were downloaded from the Zillow (www.zillow.com) website[1] for the US nationwide and the following metropolitan areas of the continental United States: Chicago, Los Angeles, New York, and Seattle. These regional markets were chosen because they span the US and because they vary in size and composition. The home value estimate is called a Zestimate® and is based on such attributes as number of bedrooms and bathrooms, square footage of the home, and other characteristics of a residence.

[1] http://www.zillow.com/howto/DataCoverageZestimateAccuracy.htm

Figure 1 shows the response, US Zestimates, and the exogenous predictors, for the time period April 1997 to April 2013.

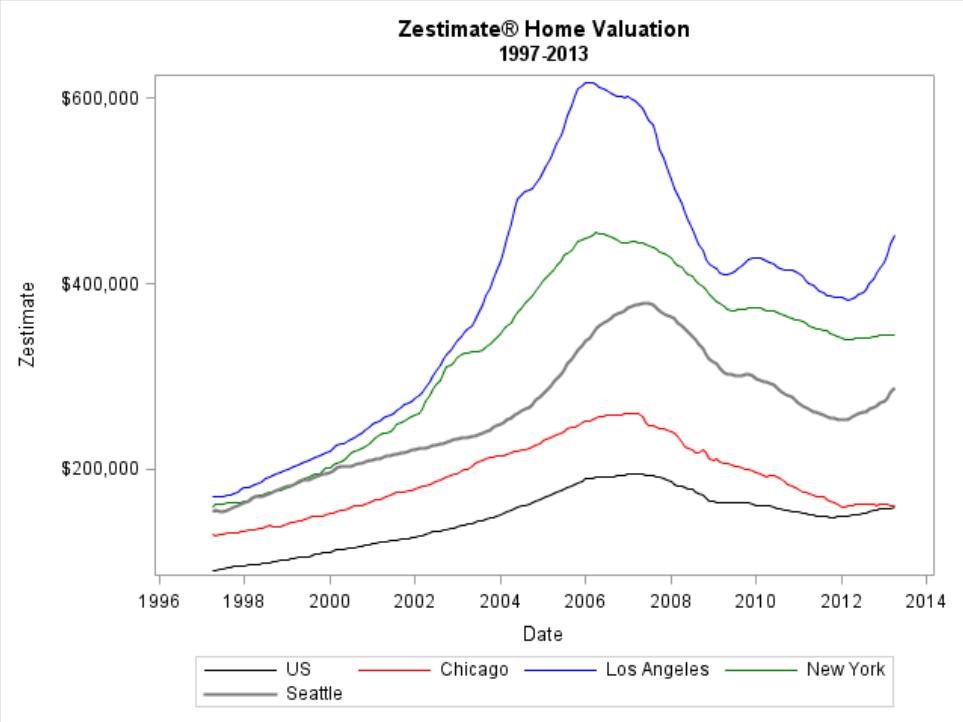

Figure 1: Zestimate Home Valuation, 1997-2013

MODELING EXOGENOUS PREDICTORS

We must build a time series model for each exogenous predictor for the purpose of prewhitening it. The prewhitening process removes the autocorrelation that may be present in each series so as to provide a series of independent observations to use in regression involving the response variable, US Zestimate.

CHICAGO

We begin the modeling process by identifying the appropriate model for the Chicago Zestimate time series with the following statements:

```
proc arima data=zestimate ;
      identify var=Chicago ; run ;
```

We are looking for evidence of nonstationarity[2] in the Chicago data that will require remediation in the form of differencing the observations, e.g., creating a new time series $Y_t'=\nabla Y_t = Y_t - Y_{t-1} = (1-B)Y_t$ where $B(\cdot)$ is the backshift operator with the property that $B(Y_t) = Y_{t-1}$. If we see a set of slowly-declining lags in the autocorrelation function (ACF), we may assume that there is nonstationarity in the errors due to the persistence of prior errors. In this case, differencing usually removes the persistence in the lags and creates a stationary time series.

The `identify` statement produces a number of tables and plots to facilitate analysis of the time series.[3]

[2] For a time series to be stationarity, the time series of the residuals must have a constant mean (independent of time) and its autocovariance function must also be independent of time (no significant autocorrelations at lags other than 0).

[3] Plots and tables are frequently omitted to conserve space.

The mean of the undifferenced series is 191,210, its standard deviation is 38,500, and there are 193 observations. Differencing ought to produce a series with a reduced standard deviation and a mean of 0.

There are significant autocorrelations between lags, so we must consider formulating an ARIMA model. The "I" in "ARIMA" indicates that the response series has been integrated, or differenced.

The augmented Dickey-Fuller unit root test indicates that at least one root of the characteristic polynomial of the time series is equal to 1 for up to two lags. The failure to reject the hypothesis at the p<.05 level of significance signifies that there may be at least one root equal to 1. A unit root is characteristic of nonstationary models. Differencing may be necessary to remove the nonstationarity in the data.

The observational plot in Figure 2 shows an upward trend followed by a downward trend. The slow decline in the ACF spikes indicates that past errors at high lags are carried along into the present. This property is characteristic of nonstationarity in the process. Differencing ought to result in a time series with rapidly-declining lags and values that approach the mean of the series.

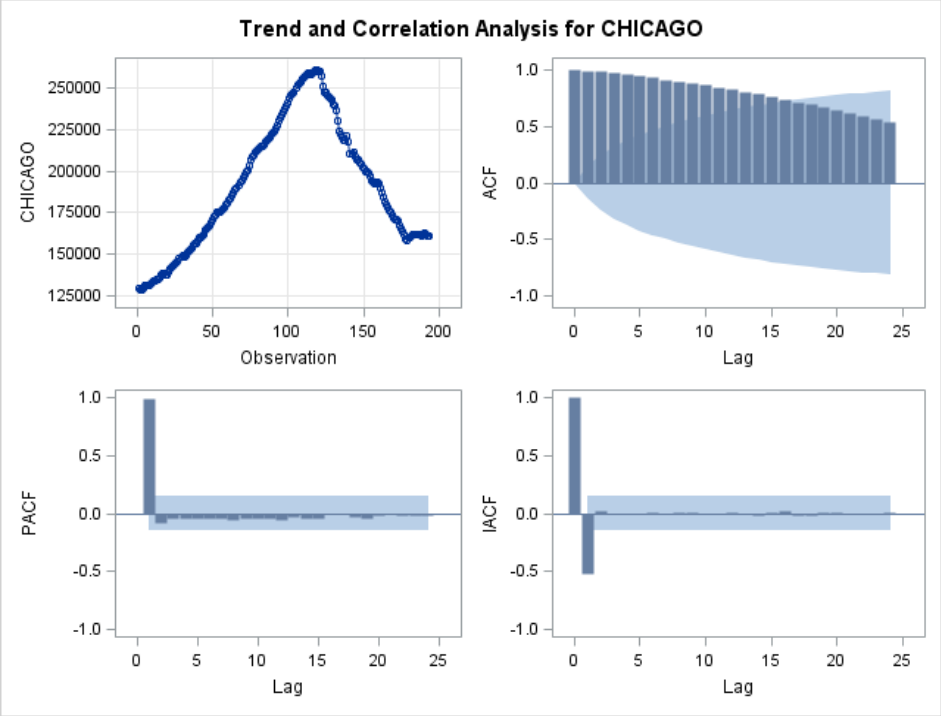

Figure 2: Identification Plots for Original Chicago Time Series

We submit the following code to obtain diagnostics of the first-differenced series:

```
Identify var=Chicago( 1 ) ; run ;
```

The mean is 165 for the first-differenced series, but theoretically it ought to be 0. The standard deviation is 1692, down from 38,500. This is a pronounced decrease, so we may assume that we took appropriate action by performing differencing. There are 192 observations: one was lost to differencing, but the total number is still sufficient for analysis.

Autocorrelations are reduced somewhat from the undifferenced series, but are still high, so the assumption of independence of errors ϵ_t is not valid.[4]

[4] In OLS regression, it is assumed that errors are uncorrelated. If this assumption is false, then the estimates of variance of predictors are too low and parameter estimates are inflated. They are said to be *inefficient*. In time series analysis, the error term ϵ_t is the source of randomness and is assumed to have the following characteristics:

$E(\epsilon_t) = 0$	Mean of errors = 0
$E(\epsilon_t^2) = \sigma^2$	Homoscedasticity, or constant variance
$E(\epsilon_t \epsilon_s) = 0$ for all $t \neq s$	Independence of errors

The hypothesis of a unit root is rejected at the p<.0001 level for the zero mean and the single mean models at lags 0 and 1. However, the fact that it was not rejected for the single mean or trend models at lag 2 at the p<.05 level may indicate the presence of a unit root in the characteristic polynomial of an AR(p) model of the response series. We must still check the ACF for exponentially-decreasing spikes at higher-order lags.

In Figure 3, we see that there is still slow decay in the moving average terms due to the influence of past errors. The augmented Dickey-Fuller unit root test for single mean or trend models at two lags did not disprove the presence of a unit root. A first difference does not remediate the nonstationarity in the data. We must apply a second difference and see if the nonstationarity is removed. We will also use tentative order selection algorithms (ESACF, MINIC, SCAN) to assist us in identification of the appropriate ARIMA model.

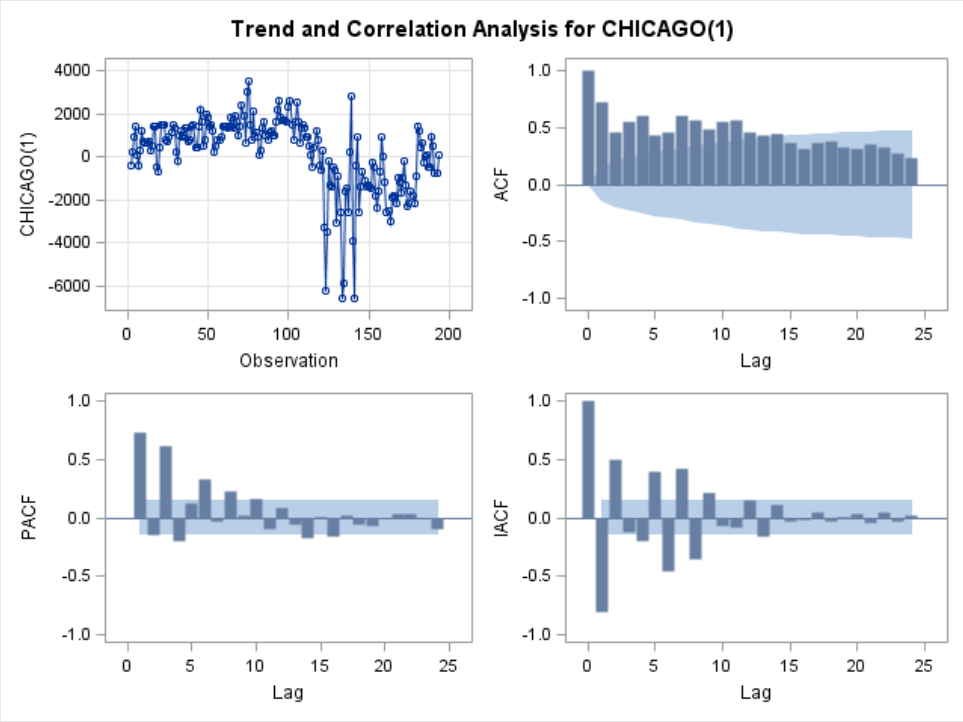

Figure 3: Identification Plots for First-Differenced Chicago Time Series

We submit the following code to obtain diagnostics of the second-differenced series:

```
identify var=Chicago( 1,1 ) esacf minic scan
p=( 1:10 ) perror=( 1:10 ) q=( 1:10 )
stationarity=( adf=( 0,1,2 )) ;
run ;
```

The mean of the first-differenced series has decreased from 165 to approximately 3, and the standard deviation has decreased from 1692 to 1258 after differencing twice. The mean is now much closer to 0, in agreement with theory, and the continued decrease in the variation in the series from the first difference statistic suggests that differencing the original series twice was appropriate. We have lost two degrees of freedom due to differencing, but there are still an adequate number with which to conduct the analysis.

The autocorrelations in the series are statistically significant at the p<.0001 level but we will assume that they are too small to be of practical importance.

The results of the unit root tests suggest that there are no unit roots in the characteristic polynomial of an AR(p) model fitted to the Chicago data at lags 0, 1, or 2. We may assume that the data are stationary as a result of differencing the original series twice.

In Figure 4, we see that second differencing has stabilized the model and removed the trend. The differenced series is centered around a mean of about 2.6 (close to 0), the ACF shows significant but rapidly decaying lags up to order 9, the partial autocorrelation function (PACF) shows significant but decaying sinusoidal lags up to order 9, and the inverse autocorrelation function (IACF) shows significant but decaying sinusoidal lags up to order 11.

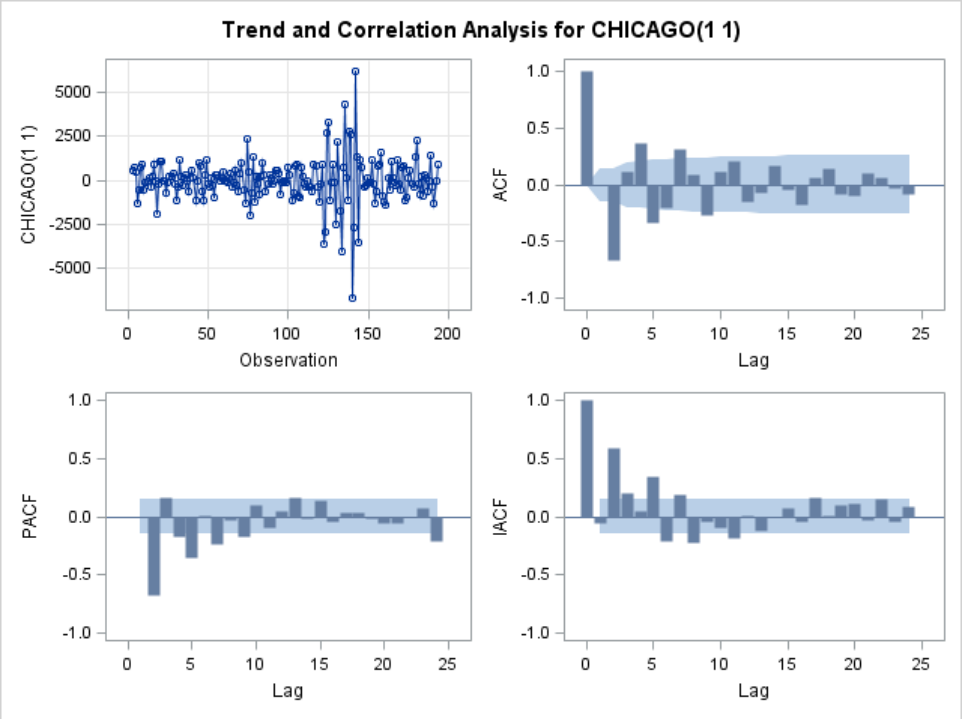

Figure 4: Identification Plots for Twice-Differenced Chicago Time Series

We will attempt to identify the lag structure using the squared canonical correlation (SCAN) algorithm, the extended sample autocorrelation function (ESACF), and the minimum information criterion (MINIC). These are tentative model order selection tests whose purpose is to inform the analyst of potential models to investigate. To conserve space, we will not include the tables produced by each test, but only report the summary results.

The combined results of the three model identification techniques are shown below. Note that the "p+d" combination includes the number of differences in the series being identified so that for the first SCAN entry in the table below, the candidate model to be estimated would be ARIMA(3,2,3).

ARMA(p+d,q) Tentative Order Selection

SCAN			ESACF		
p+d	q	BIC	p+d	q	BIC
3	3	13.64244	5	2	13.56627
5	2	13.56627	4	5	13.70996
2	5	13.66313	8	4	13.64668
7	1	13.61674	1	7	13.8519
			2	7	13.70159

(5% Significance Level)

We will investigate these models according to increasing BIC value.

We submit the following code to estimate the parameters and produce diagnostics for the ARIMA(5,2,2) model:

```
Estimate p=5 q=2 method=ml ; run ;
```

We see that the mean is not significant, nor are the parameter estimates for Y_{t-1}, Y_{t-4} or ϵ_{t-1} at the p<.05 level of significance.

Maximum Likelihood Estimation					
Parameter	**Estimate**	**Standard Error**	**t Value**	**Approx**	**Lag**
MU	-2.23914	19.68238	-0.11	0.9094	0
MA1,1	-0.10792	0.12933	-0.83	0.4040	1
MA1,2	0.50667	0.12877	3.93	<.0001	2
AR1,1	-0.06200	0.11877	-0.52	0.6017	1
AR1,2	-0.28060	0.11950	-2.35	0.0189	2
AR1,3	-0.24416	0.09648	-2.53	0.0114	3
AR1,4	0.17890	0.09443	1.89	0.0582	4
AR1,5	-0.46087	0.06595	-6.99	<.0001	5

We note that this model has a standard error estimate of 837 and SBC statistic of 3149. We will reestimate the model without the insignificant mean term to see if there is any improvement in performance.

We submit the following statement:

```
Estimate p=5 q=2 method=ml noint ; run ;
```

There is essentially no change in the values of the parameter estimates to three significant digits, and the t-statistics are only marginally improved from the previous case. The standard error estimate has decreased to 835 and the SBC has decreased to 3144. We will omit the mean in the sequel since the model is improved by its absence.

Maximum Likelihood Estimation					
Parameter	**Estimate**	**Standard Error**	**t Value**	**Approx**	**Lag**
MA1,1	-0.10819	0.12902	-0.84	0.4017	1
MA1,2	0.50645	0.12847	3.94	<.0001	2
AR1,1	-0.06211	0.11847	-0.52	0.6001	1
AR1,2	-0.28070	0.11921	-2.35	0.0185	2
AR1,3	-0.24422	0.09624	-2.54	0.0112	3
AR1,4	0.17880	0.09419	1.90	0.0577	4
AR1,5	-0.46086	0.06577	-7.01	<.0001	5

The correlations among parameter estimates shows that there may be cause for concern regarding violation of the independence of the AR terms from each other and from the MA terms.

Correlations of Parameter Estimates							
Parameter	**MA1,1**	**MA1,2**	**AR1,1**	**AR1,2**	**AR1,3**	**AR1,4**	**AR1,5**
MA1,1	1.000	0.047	0.838	0.096	0.741	0.125	0.116
MA1,2	0.047	1.000	0.011	0.845	0.084	0.735	0.126

Correlations of Parameter Estimates							
Parameter	**MA1,1**	**MA1,2**	**AR1,1**	**AR1,2**	**AR1,3**	**AR1,4**	**AR1,5**
AR1,1	0.838	0.011	1.000	0.113	0.701	0.142	-0.049
AR1,2	0.096	0.845	0.113	1.000	0.212	0.739	0.189
AR1,3	0.741	0.084	0.701	0.212	1.000	0.254	0.234
AR1,4	0.125	0.735	0.142	0.739	0.254	1.000	0.228
AR1,5	0.116	0.126	-0.049	0.189	0.234	0.228	1.000

However, the insignificant autocorrelation amongst the residuals shows no need for concern:

Autocorrelation Check of Residuals									
To Lag	**Chi-Square**	**DF**	**Pr > ChiSq**	**Autocorrelations**					
6	.	0	.	0.015	-0.046	0.002	0.005	0.004	0.042
12	9.84	5	0.0798	-0.086	0.109	-0.018	0.105	0.112	-0.033
18	16.72	11	0.1164	0.104	-0.009	-0.139	-0.030	-0.041	0.002
24	24.69	17	0.1018	-0.064	-0.082	0.094	-0.070	0.042	-0.102
30	28.36	23	0.2026	-0.044	-0.026	0.056	-0.072	0.055	-0.047
36	33.93	29	0.2420	0.053	-0.034	0.080	-0.018	-0.027	-0.110

The diagnostic plots for the residual correlations in Figure 5 put us at ease because they show no significant lags except for the IACF at lag 11, and the white noise plot indicates only two spikes above p<.05 at lags 8 and 15. Given the dynamic nature of the housing market during the time period under analysis, we may attribute these lags to spurious correlations in the data.

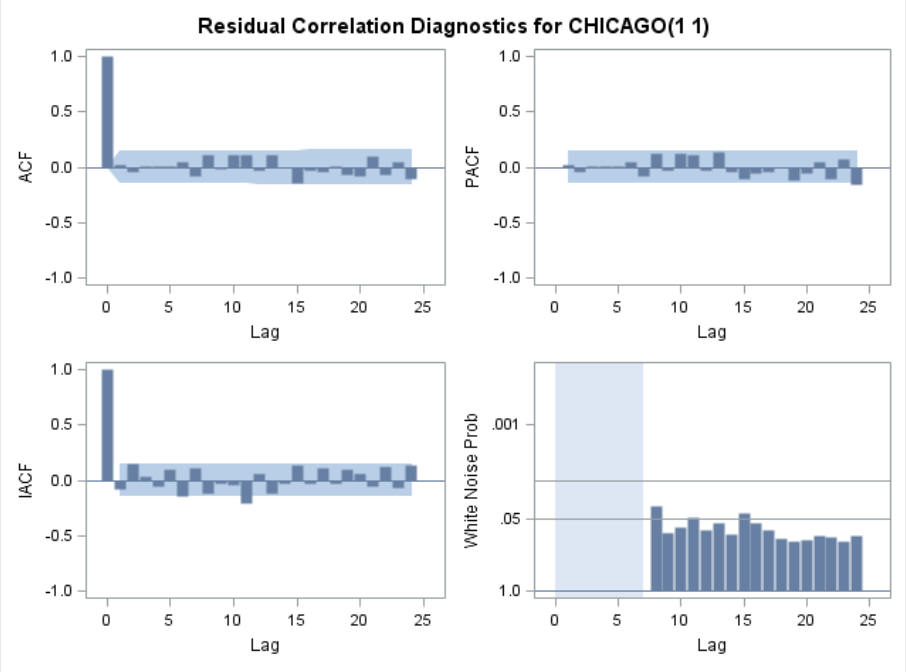

Figure 5: Residual Correlation Diagnostics for CHICAGO(1 1)

We conclude that the ARIMA(5,2,2) model of the Chicago exogenous predictor appears to be valid and representative of the data. The model is

$$(1 + 0.06211B + 0.2807B^2 + 0.24422B^3 - 0.01788B^4 + 0.46086B^5)(1 - \mathrm{B})^2 Chicago_t =$$

$$(1 + 0.10819B - 0.50645B^2)\epsilon_t$$

With only a little extra effort, we can fit the model ARIMA({2,3,5},2,{2}) to omit the non-significant parameter estimates of the ARIMA(5,2,2) model.

We submit the following statement:

```
Estimate p=( 2,3,5 ) q=( 2 ) noint ; run ;
```

and we observe that all of the parameter estimates are significant at the p<.005 level, the standard error estimate has risen slightly from 835 to 839.3898, and the SBC statistic has dropped from 3144 to 3133.229. The model is

$$(1 + 0.45243B^2 + 0.18431B^3 + 0.43498B^5)(1 - \mathrm{B})^2 Chicago_t = (1 - 0.34354B^2)\epsilon_t$$

We built models for the (p+d,q) pairs recommended by the SCAN, ESACF, and MINIC algorithms. We found that the BIC statistic was a good predictor of performance. The first model built, ARIMA(5,2,2), had good performance. After a little fine-tuning, its performance improved. The other models tested did not show equivalent performance in terms of accuracy (standard error estimate) or agreement with theory (residuals were not reduced to white noise).

We will use the ARIMA({2,3,5},2,{2}) model to prewhiten the Chicago series.

LOS ANGELES

Following the same methodology used to develop the Chicago time series model, the model for the Los Angeles time series is

$$(1 - 0.27371B^4 + 0.33823B^5 - 0.15301B^6 + 0.1231B^7 + 0.25513B^{12})(1 - \mathrm{B})^2 LosAngeles_t =$$

$$(1 + 0.75409B)\epsilon_t$$

NEW YORK

Similarly, the model for the New York time series is

$$(1 + 0.59998B^2 - 0.21227B^6 + 0.18804B^{10} + 0.18804B^{12})(1 - \mathrm{B})^2 NewYork_t =$$

$$(1 + 0.40388B)\epsilon_t$$

SEATTLE

Likewise, the model for the Seattle time series is

$$(1 + 0.97805B + 0.90857B^2 + 0.36664B^3 + 0.38711B^4)(1 - \mathrm{B})^2 Seattle_t =$$

$$(1 + 1.52536B + 0.93885B^2)\epsilon_t$$

MODELING US HOUSING TRENDS (RESPONSE VARIABLE: US)

We begin the modeling process by identifying the appropriate model for the US national home prices response variable with the following statements:

```
proc arima data=zestimate ;
       identify var=US( 1,1 ) stationarity=( adf=( 0,1,2 )) ;
       run ;
```

and observe that the twice-differenced series has been detrended. The mean of the twice-differenced series data is 3 and the standard deviation is 454.

There are small, but significant autocorrelations. The largest one in magnitude is -0.493 at lag 2. The smallest is -0.002 at lag 19.

The hypothesis of unit roots is rejected at the p<.0001 level for zero mean, single mean, and trend models at lags 0, 1, and 2.

The observational plot in Figure 6 shows no apparent trend. The data points of the twice-differenced series lie around 0. There is a significant spike in the ACF at lag 2, so we expect to see MA components in the ARIMA model that we will fit. The PACF shows significant spikes at lags 2, 5, and 7. The IACF shows significant spikes at lags 1, 2, 4, 5, 6, and 9.

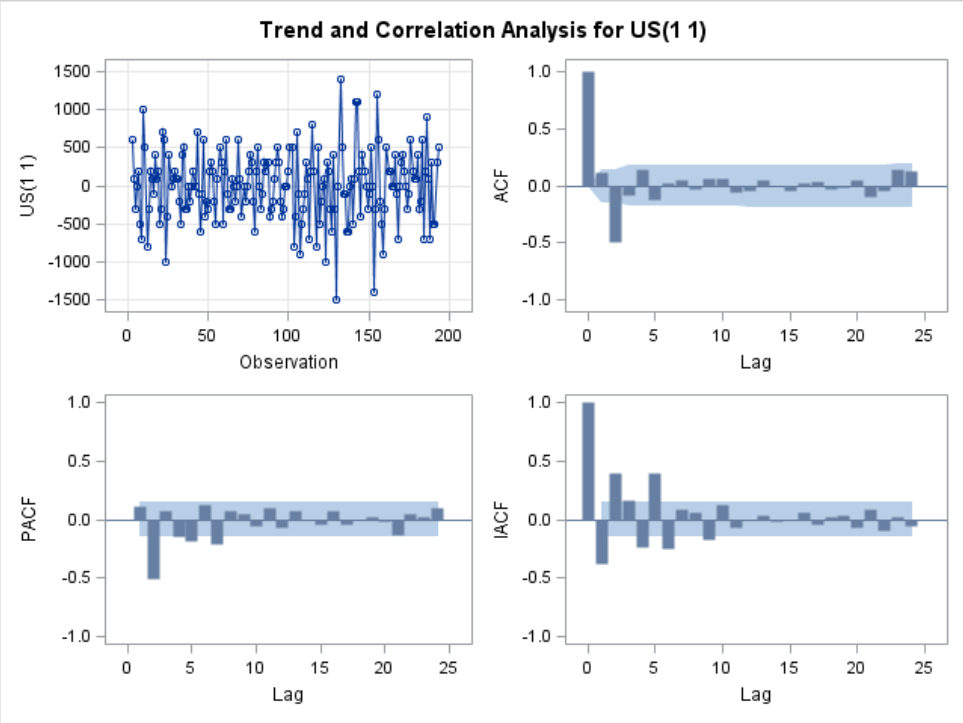

Figure 6: Trend and Correlation Analysis for US(1 1)

We built several models to reveal the most appropriate lag structure. We found that an ARMA({2,6,10,12},2,{1}) model had the best characteristics.

We submit the following estimation statement:

```
estimate p=(2,6,10,12) q=1 method=ml noint ; run ;
```

The Trend and Correlation Analysis for the US(1 1) plot shows rapidly-decaying sinusoidal spikes in the ACF, PACF, and IACF plots.

Before we estimate parameters, we must prewhiten the exogenous predictors so that their internal lag dynamics do not complicate the estimation process. This requirement is equivalent to the assumption that independent variables in OLS regression are uncorrelated. Each predictor is identified and estimated before it is used in estimating the US national home prices.

We submit the following identification statements:

```
identify var=US( 1,1 )
    crosscorr=(
        Chicago(1,1) Los_Angeles(1,1) New_York(1,1) Seattle(1,1)
    ); run ;
```

The crosscorrelation plots in Figure 7 indicate that the estimation statement must include time delays. The New York crosscorrelation plot shows the first significant spike at lag 2; the Seattle crosscorrelation plot shows the first significant spike at lag 6. These delays are included in the estimate statement.

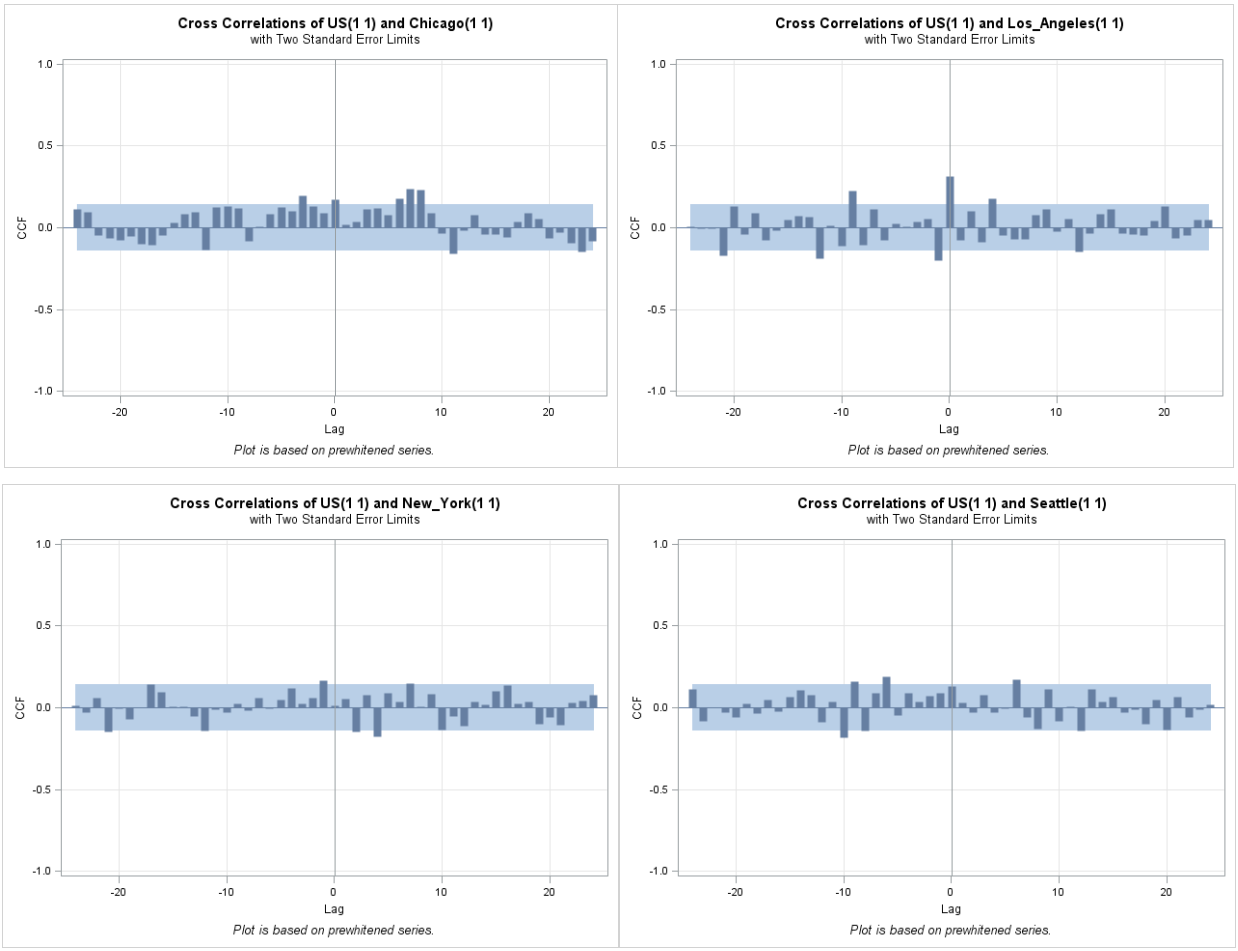

Figure 7: Cross Correlations of US(1 1) and Exogenous Predictors

We submit the following estimation statements:

```
estimate input=
    (       /( 6 11 ) Chicago
            /( 4    ) Los_Angeles
        2 $ /( 2    ) New_York
        6 $           Seattle
    )
    method=ml noint plot
    p=( 1 2 3 5 6 7 )
    ;
run ;
```

All parameter estimates are significant at the p<.05 level except for ϕ_3 in the US model, which is marginally insignificant.

The SBC statistic is 2635.314 and the standard error estimate is 309.6881.

The magnitude of the largest correlation between parameter estimates is 0.571 between ϕ_3 and ϕ_4.

The residual correlation diagnostics in Figure 8 indicate that there are no significant autocorrelations in the residuals at any lag for the p<.05 level of significance. The diagnostics indicate that the model has produced only uncorrelated white noise, which is characteristic of a properly-specified model.

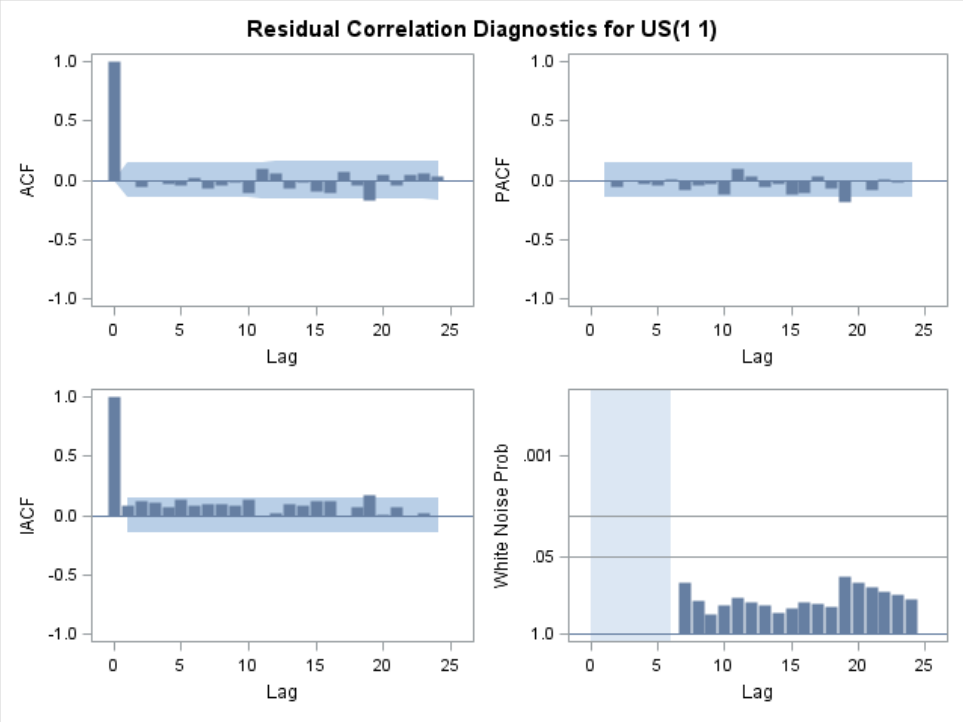

Figure 8: Residual Correlation Diagnostics for US(1 1)

The residual normality diagnostics in Figure 9 show that the distribution of the residuals is approximately normally distributed, with only relatively few observations in the tails of the distribution departing from normality. This is another characteristic of a properly-specified model.

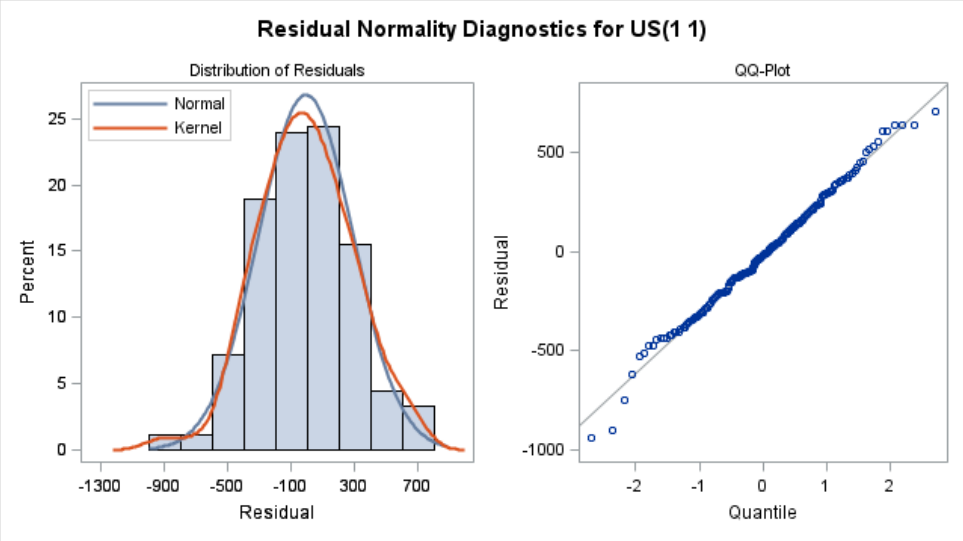

Figure 9: Residual Normality Diagnostics for US(1 1)

We conclude that the ARIMA({1,2,3,5,6,7},2,0) model of the US home price time series appears to be valid and representative of the data. The dynamic regression ARMA model can be written in the form

$$\nabla^2\Phi_{US}(B)Y_t = \nabla^2\beta_1(B)Chicago_t + \nabla^2\beta_2(B)LosAngeles_t + \nabla^2\beta_3(B)NewYork_{t-2} + \nabla^2\beta_4(B)Seattle_{t-6} + N_t$$

where

$$\Phi_{US}(B) = (1 - 0.22848B + 0.58874B^2 + 0.16173B^3 + 0.39672B^5 - 0.18814B^6 + 0.27404B^7)$$

$$\beta_1(B) = 0.92269\frac{1 - 0.34354B^2}{1 + 0.45243B^2 + 0.18431B^3 + 0.43498B^5}$$

$$\beta_2(B) = 0.096447\frac{1 + 0.75409B}{1 - 0.27371B^4 + 0.33823B^5 - 0.15301B^6 + 0.1231B^7 + 0.25513B^{12}}$$

$$\beta_3(B) = -0.06446\frac{1 + 0.40388B}{1 + 0.59998B^2 - 0.21227B^6 + 0.18804B^{10} + 0.1814B^{12}}$$

$$\beta_4(B) = 0.078132\frac{1 + 1.5236B + 0.93885B^2}{1 + 0.97805B + 0.90857B^2 + 0.36664B^3 + 0.38711B^4}$$

and N_t is an additive white Gaussian noise process.

We submit the following statement to compare original US home prices with their forecasted values:

```
forecast lead=0 out=forecast id=date interval=month ; run ;
```

and plot the results to produce Figure 10 of original US home prices and the forecast home prices created by the ARIMA model superimposed upon them. Note that the fit is so close that there is no benefit in adding confidence intervals (the standard error was only 310).

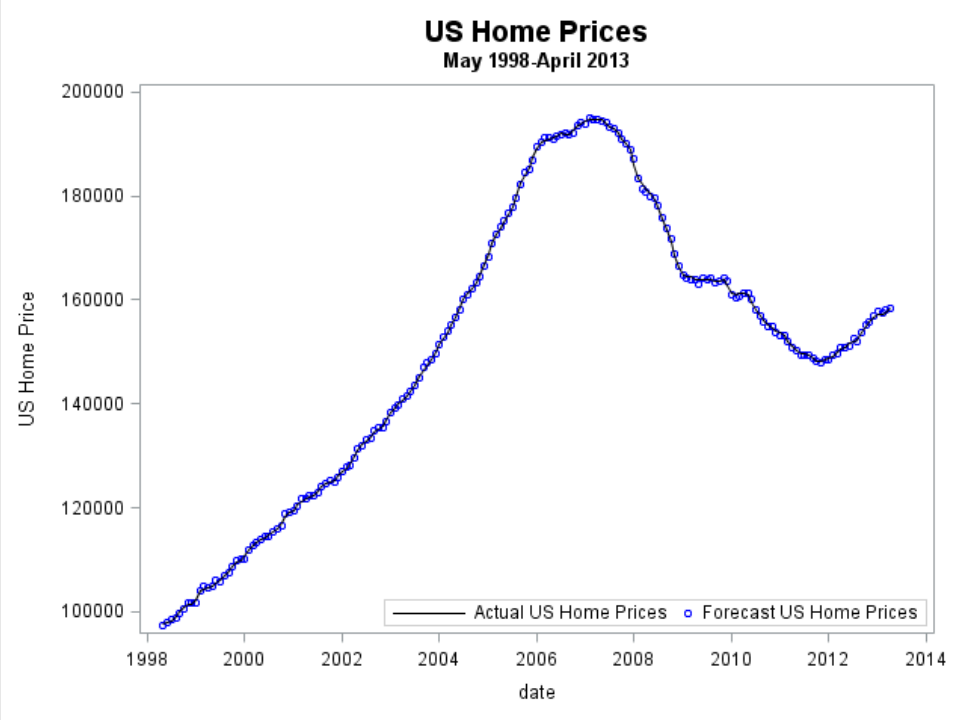

Figure 10: Original and Forecast US Home Prices

SUMMARY

We applied the Box-Jenkins time series modeling methodology to a set of US and regional home prices with the goal of building a dynamic regression model using ARIMA modeling with exogenous predictors. We used several different approaches to model building, including unit root tests and tentative order tests. We observed that inspection of the ACF, PACF, and IACF plots may provide more information than simply using the tentative order tests. We prewhitened the exogenous predictors to remove the confounding effects of their autocorrelations upon the response variable. We built an ARIMA model that closely fits the original data and satisfies the theoretical assumptions underlying the Box-Jenkins methodology.

It is critically important to accurately represent the lag structure of each time series involved in the modeling process because the lag structures, based on past data, determine the values of future predictions. The several time series, including the response and the exogenous predictors, were thoroughly scrutinized using the ACF, PACF, and IACF plots, and the iterative nature of these investigations was necessarily omitted due to limitations on space. If an analyst focuses on such measures of forecast accuracy as mean average percent error (MAPE) or root mean square error (RMSE) alone without thorough investigation into the lag structures of the constituent time series involved, then

forecasts based on future inputs may be inaccurate since the generative characteristics of the series may not have been captured in the model building process.

REFERENCES

1. Dickey, David and Terry Woodfield (2011). *Forecasting using SAS® Software: A Programming Approach*, SAS Institute, Inc. Cary, NC, USA.
2. Montgomery, Douglas C., Cheryl L. Jennings, and Murat Kulahci (2008). *Introduction to Time Series Analysis and Forecasting*, John Wiley & Sons, Inc. Hoboken, NJ.

BIBLIOGRAPHY

1. Brocklebank, John C., and David A. Dickey (2003). *SAS(R) for Forecasting Time Series, 2nd Ed,* SAS Institute, Inc., Cary, NC, USA.
2. Hyndman, Rob J. (2010). *The ARIMAX Model Muddle,* http://www.r_bloggers.com/the-arimax-model-muddle
3. http://support.sas.com/documentation/cdl/en/etsug/65545/HTML/default/viewer.htm#etsug_arima_toc.htm

ACKNOWLEDGMENTS

We thank Sam Iosevich, Joseph Naraguma, and Nitzi Roehl who reviewed preliminary versions of this paper and contributed their helpful comments.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author at:

Ross Bettinger

rsbettinger@gmail.com