VEE Time Series Student Project Spring 2011 Heather Roth

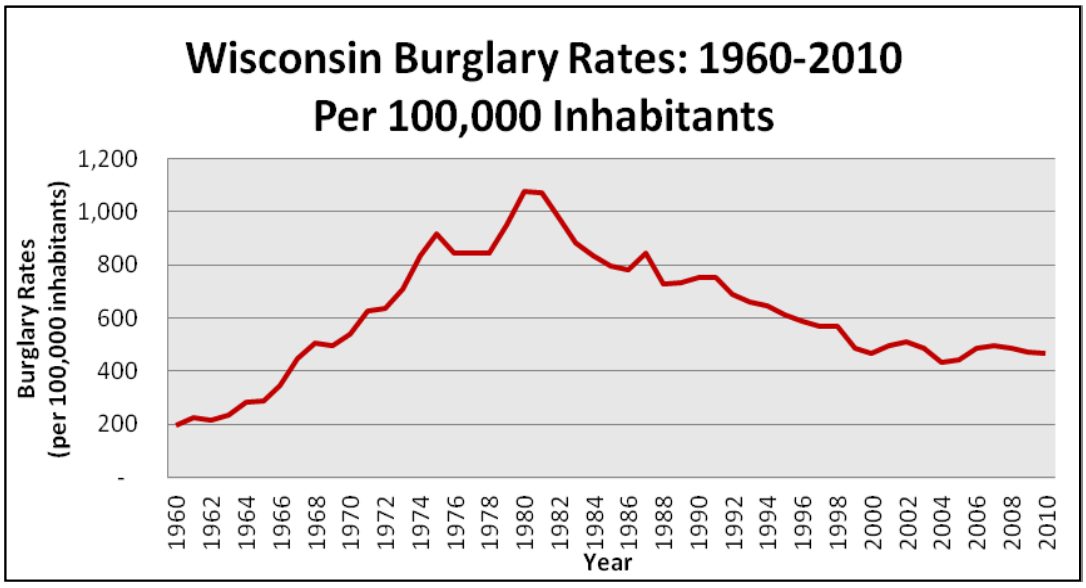
INTRODUCTION

Burglary is defined as the felony of breaking and entering the house of another with the intent of committing a crime. It is an offense against possession and habitation. As a first-time home owner, in a neighborhood I am not all-too-familiar with, the idea of a possible burglary is unsettling, to say the least. In 2010, 467 burglaries occurred in Wisconsin for every 100,000 inhabitants. Statistically, 0.467% is a small percentage, but a burglary could happen to anyone.

This project uses historical data to build an appropriate ARIMA model that can be used to project future burglary rates in Wisconsin. More specifically, the time series analysis uses annual burglary rates, per 100,000 inhabitants, in the state of Wisconsin between 1960 and 2010. A projection model is developed, and one-step-ahead forecasts are compared to actual data to determine if the model is reasonable.

DATASET

Historical data for the project was obtained from http://www.disastercenter.com/crime/wicrime.htm, a website maintained by The Disaster Center. The annual burglary rates from 1960 to 2005 are shown below.

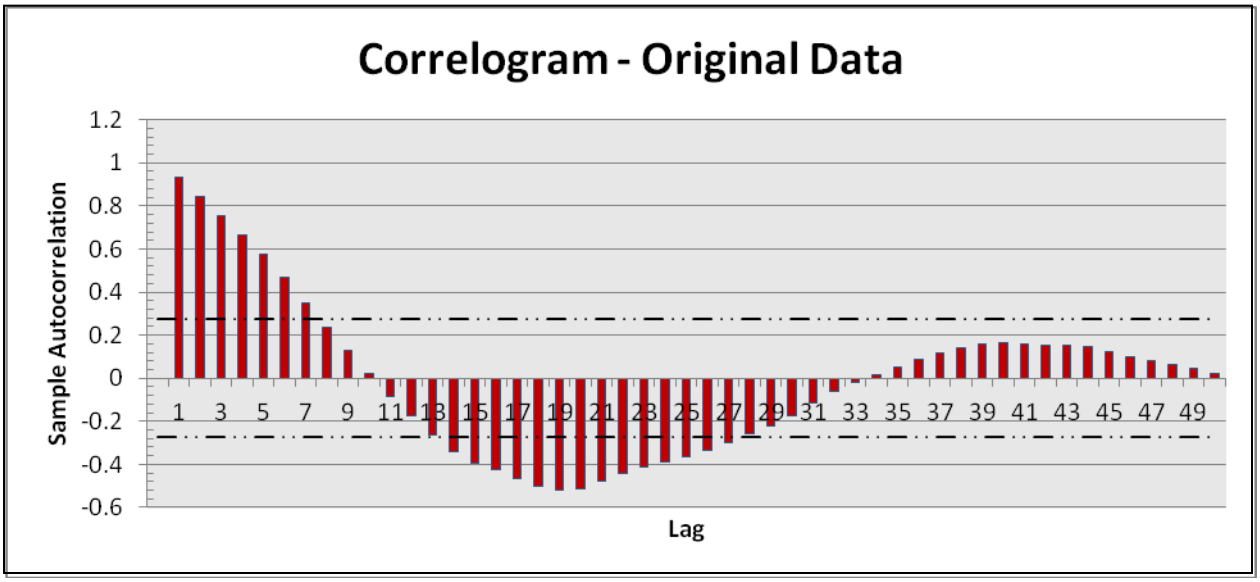


Before an ARIMA model can be developed, the data needs to be tested for stationarity. If the original series is not stationary, then transformations of the data (such as first and second differences) are observed and subsequently tested for stationarity.

TEST OF STATIONARITY

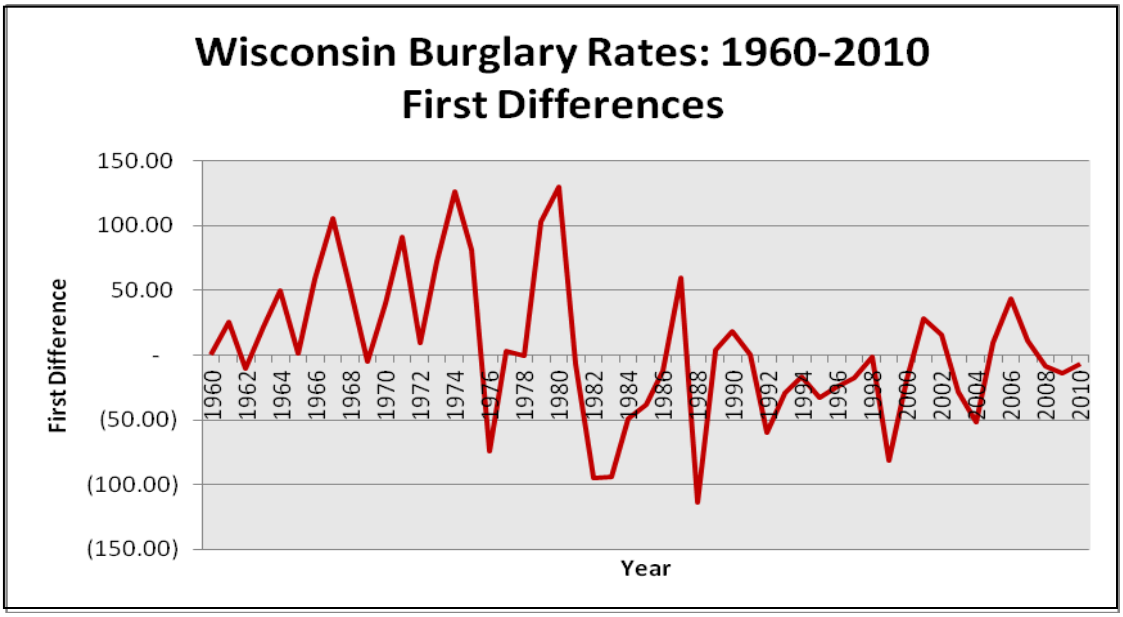
A process is considered to be stationary if its statistical parameters, such as mean and standard deviation, are constant over time. By looking at the original series above, it is clear that the mean is not constant over time. However, a more thorough investigation of determining stationarity is conducted by analyzing the process' sample autocorrelation function. The sample autocorrelation function describes the correlation

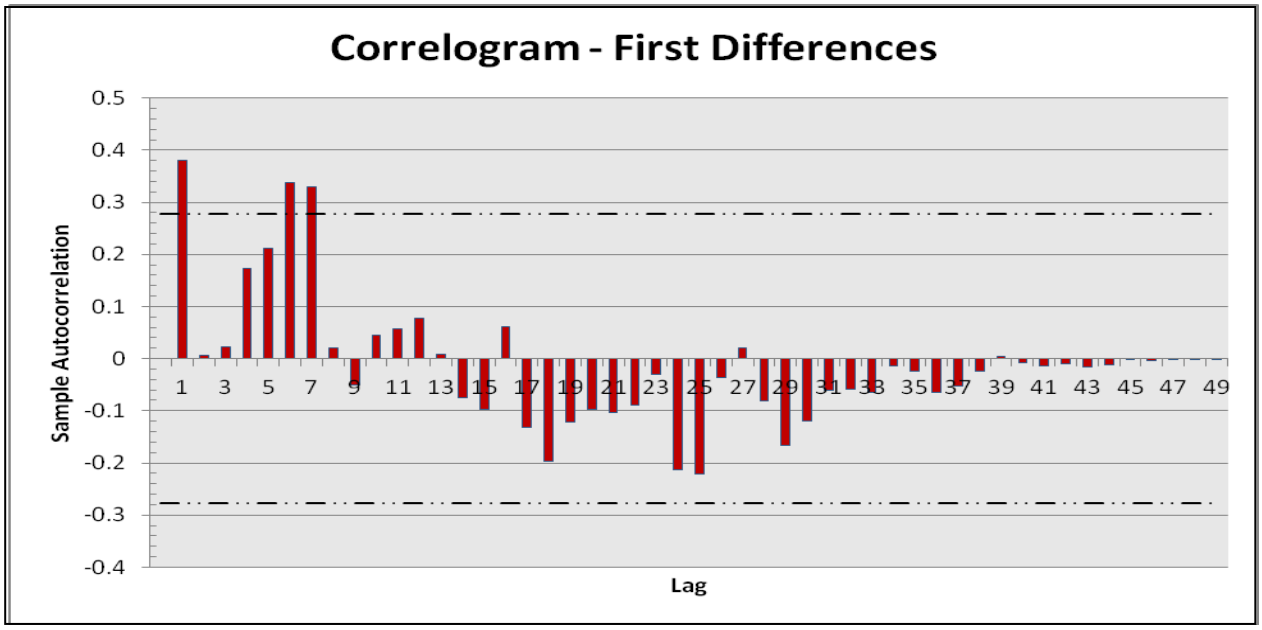
between values of a process at different points in time. It must be constant over time, with a zero mean, in order for the series to be considered stationary. The graphical representation used to study the sample autocorrelations is called a correlogram. Following, is the correlogram of the original data set.



First, notice that the sample autocorrelation decreases more linearly than exponentially, leading us to believe that this is an autoregressive process, as opposed to a moving average process. Second, the dashed horizontal lines, plotted at $\pm 1.96/\sqrt{(n)} = \pm 0.274$, give the 95% confidence interval and are intended to give critical values for testing whether or not the autocorrelation coefficients are significantly different from zero. If a process is stationary, approximately only 5% of the autocorrelation points will lie outside the 95% confidence interval. Here, over 40% of the points lie outside the confidence interval, leading us to conclude that the original data set is not stationary.

Next, we take first differences of the time series and examine whether or not the resulting data is stationary. A plot of the first differences, as well as its corresponding correlogram, is shown below.

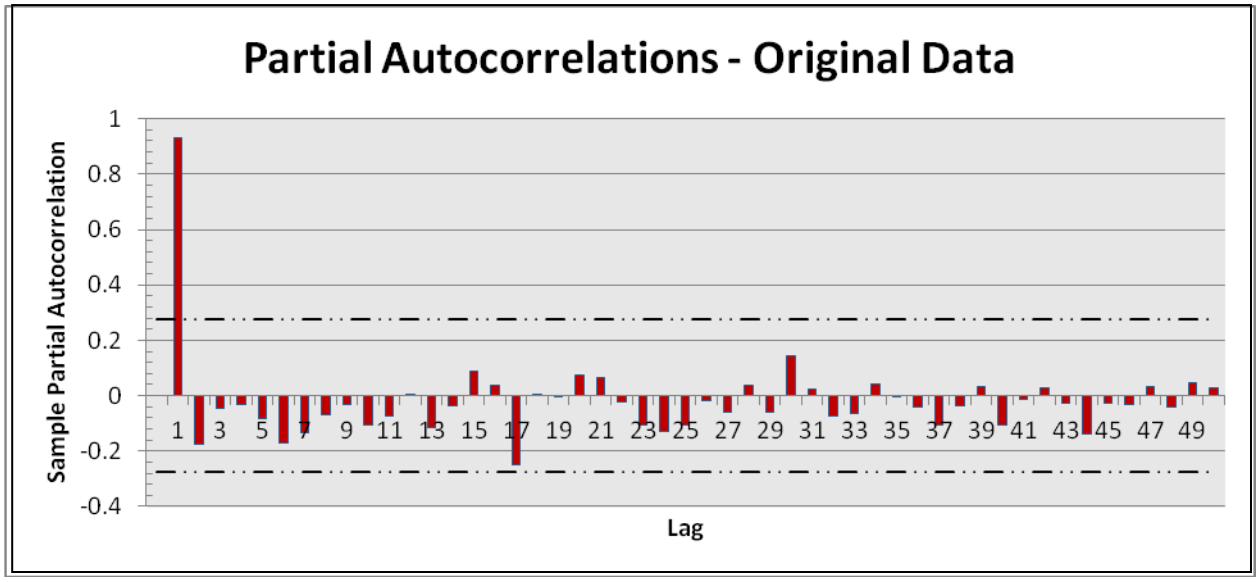




The plot of second differences indicates a model with a more consistent mean and variation, as necessary for stationarity. The autocorrelation function quickly reduces to zero and oscillates tolerably around zero thereafter. In addition, there is an acceptable amount of points that lie outside the 95% confidence interval, and so we conclude that the process is stationary.

MODEL SPECIFICATION

Since the sample autocorrelations of the original series decreased more linearly than exponentially, an autoregressive model should be assumed. We observe a graph of the sample partial autocorrelations to determine the appropriate order of the autoregressive model.



The partial autocorrelation at lag 1 is the only one that is significant (i.e. it lies outside the 95% confidence interval). This graph gives a strong indication that we should consider an AR(1) model for this time series. Nevertheless, we will entertain an AR(1), AR(2), and AR(3) model for comparison.

MODEL DIAGNOSTICS

The summary statistics obtained for the three autoregressive models are given below. Please note that all statistics have been rounded to five decimal places for simplicity.

			Coefficients			p-Values			Durbin Watson
Model	**Adjusted-R^2**	**Intercept**	**Y_{t-1}**	**Y_{t-2}**	**Y_{t-3}**	**Y_{t-1}**	**Y_{t-2}**	**Y_{t-3}**	**Statistic**
AR(1)	0.93822	43.23547	0.93863	.	.	0	.	.	1.23902
AR(2)	0.94257	41.40756	1.31812	-0.37981	.	0	0.00590	.	1.87127
AR(3)	0.93862	46.74599	1.37164	-0.57805	0.13870	0	0.01900	0.33760	1.96470

First, we observe the p-values. All the variables are significant at the 95% confidence level (i.e. $p<0.05$) except for Y_{t-3}. Hence, Y_{t-3} is not necessary for an accurate model.

Next, notice that the coefficients (ϕ_1, ϕ_2, and ϕ_3) meet the requirements of the stationarity conditions for all three models. The stationary condition for the AR(1) model requires that $|\phi_1|<1$; The stationary condition for the AR(2) model requires that $|\phi_2|<1$, $\phi_1+\phi_2<1$, and $\phi_2-\phi_1<1$; Finally, the stationary condition for the AR(3) model requirest that $|\phi_3|<1$ and $\phi_1+\phi_2+\phi_3<1$.

The Adjusted-R^2 statistic for all three models is also sufficient. Adjusted-R^2 describes the amount of variability that can be explained by the model. Thus, the higher the Adjusted- R^2 statistic, the more predictive power the model presents. Although the statistic increases slightly when moving from an AR(1) model to an AR(2) model, the increase is not significant. Thus, our intial reaction is to choose the AR(1) model.

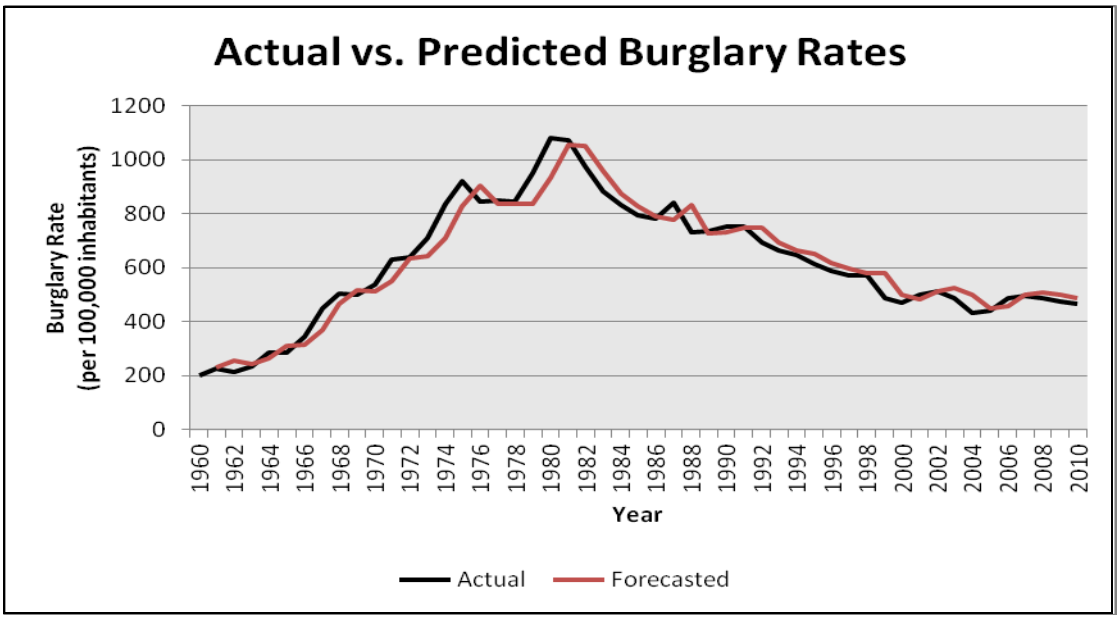
However, after observing the Durbin-Watson Statistic, an AR(2) model may be more accurate. The Durbin-Watson statistic ranges from 0 to 4 and is used to determine the presence of autocorrelation among the residuals. A statistic close to 2 indicates that autocorrelation is not present. If a Durbin-Watson statistic is substantially greater than 2, negative correlation is indicated, and vice versa. Notice that our AR(1) statistic may be cause for alarm since it suggests the presence of positive correlation among the residuals. The other two models, however, show a statistic very close to 2, signifying very little to no correlation.

Given everything we've discussed to this point, it would be reasonable to choose either the AR(1) or AR(2) model. If our objective were simplicity, the AR(1) would be more than effective. However, if our main focus were the accuracy of our predictions, the AR(2) model would be more precise. The "best model" is rather debatable.

By means of the principle of parsimony, which states that the model should use the least number of parameters necessary to adequately represent the time series, we move forward with the AR(1) model. This choice also confirms our initial reaction from studying the Partial Autocorrelations of the series.

FORECASTING

The least-squares estimation for the AR(1) model gives: $Y_t=0.93863*Y_{t-1} + 43.23547$. Based on this model, we calculate the one-step-ahead forecasts and compare those forecasts to actual data. A graph of this comparison is shown below.



As you can see, the model captures the general shape of the curve. Therefore, we conclude that it is reasonable and effective in forecasting the one-step-ahead burglary rates in WI.