Melinda Etschman
Time Series Fall 2010

Time Series Analysis of Lake Lewisville Elevation

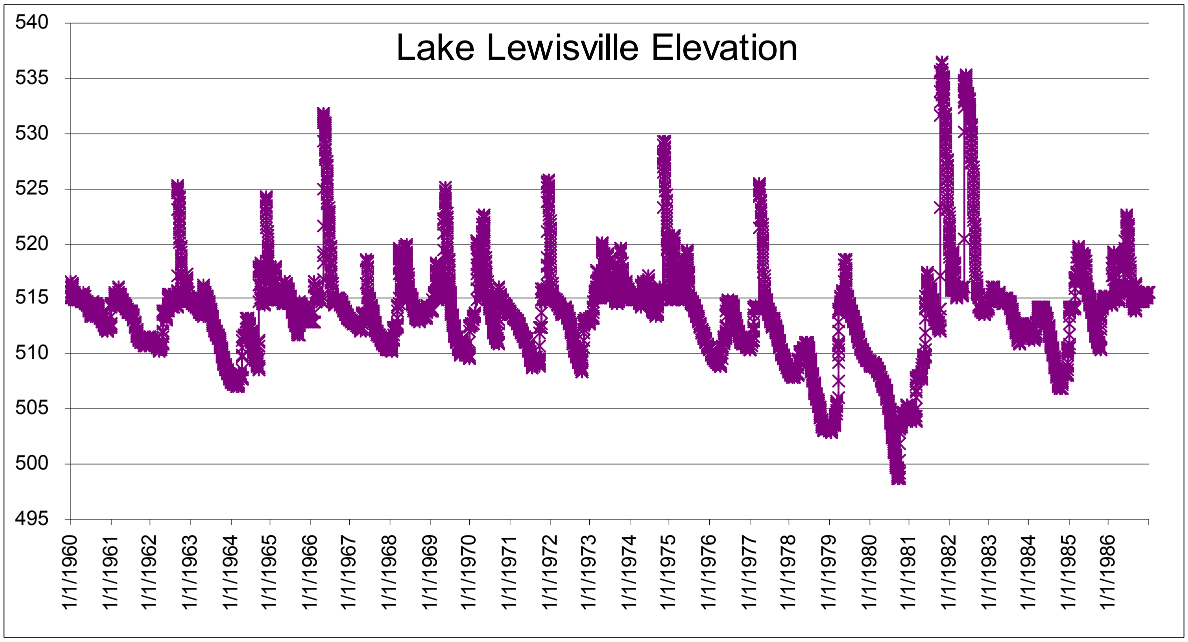
Introduction

This project will attempt to fit an ARIMA model to the time series of the daily elevation of Lake Lewisville. I am an avid boater and spend most of the spring, summer and fall days, when I'm not studying of course, on the lake. This also depends on the weather and lake levels and in my experience, I have seen lake levels vary seasonally as well as from year to year. Lake levels do not normally change drastically from one day to the next, but changes do occur, so this project will try to determine how the water elevation for any given day relates to past elevation.

Data

Data of elevation of Lake Lewisville was obtained from:
http://www.swf-wc.usace.army.mil/cgi-bin/rcshtml.pl?page=Reports
The data gathered consists of daily evaluations of lake elevation from 1960 to 2010, of which only 1960 – 1986 was used for the analysis and model build and produced over 9800 data points. The reason I selected this older data set is because there was a permanent increase of the conservation pool elevation from 515 feet above mean sea level to the current 522 feet above mean sea level. I didn't want this to be seen as a trend in the data and I didn't know how to otherwise adjust for this. I wanted to attempt an adjustment by adding 7 feet to my historic data if it was before the known date of change; however, I didn't know the exact date of change, but I do know the change occurred in 1988 due to the construction of a feeder lake by the name of Ray Roberts. Also, note that there was one adjustment made to a single data point for 12/13/1976. The original data source had 519.79, but this appears to be a typo based on the surrounding data and has been manually changed to be 510.79.

The graph below shows the lake elevation from 1960 – 1986.

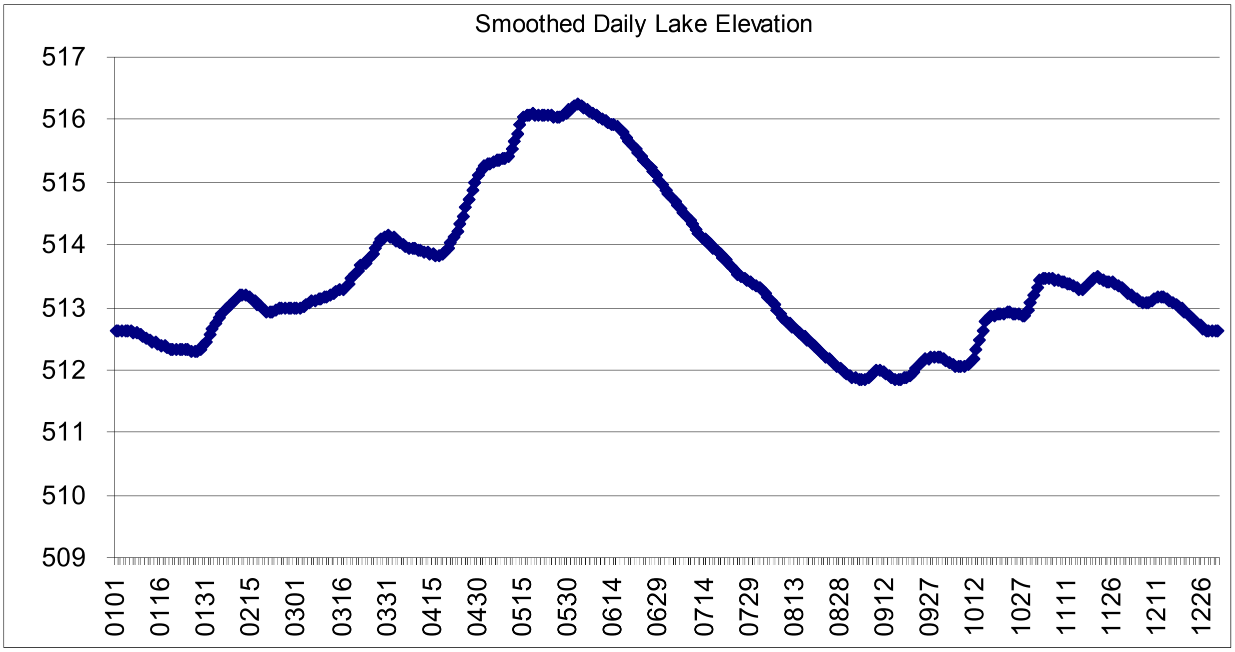


Clearly, there is a strong season pattern to the data. Although it is not always consistent from year to year, it can be seen that water elevation is normally highest just before the beginning of summer and gradually declines during the hot summer months. Lake levels historically regain a bit higher elevation during the fall, try to maintain during the winter and then again increases in elevation for a peak in the spring. This seasonality is to be expected for many reasons, one of which is that this lake serves as the principal water source for the city of Dallas, which presumably uses more water during the summer months. In addition, evaporation of the lake's water is high during the summer due to the lack of clouds/rain and extreme temperatures. Based on this knowledge, I will not examine the autocorrelation function for the raw data, but proceed to de-seasonalize the series.

Seasonal Adjustment

Although the seasonality of our data provides useful information for predicting the elevation for any given day, it obscures the relationship between elevation for consecutive days. In order to describe this relationship and construct an ARIMA model for our time series, we must remove the seasonal variation.

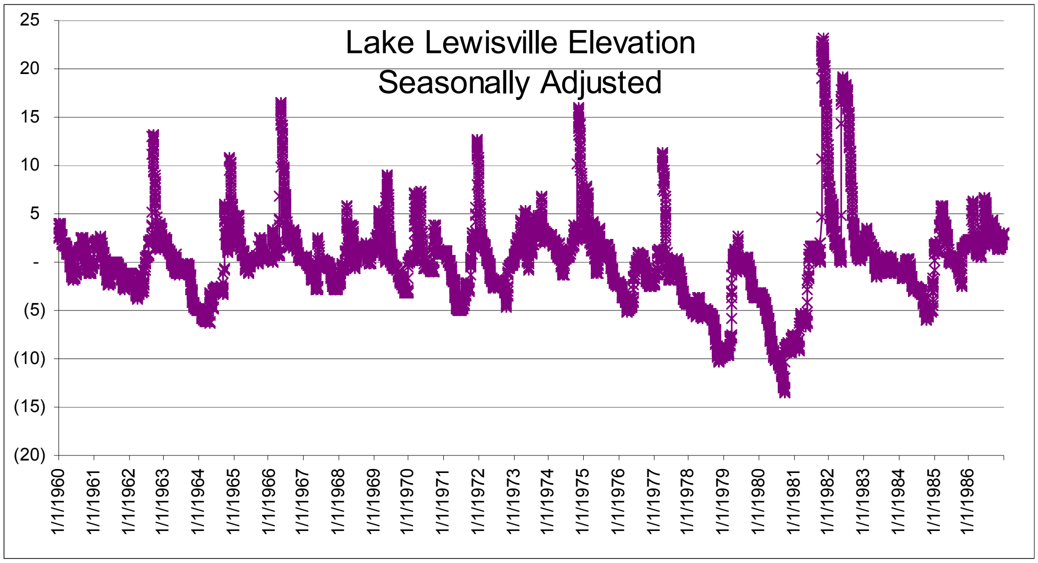
Before I de-seasonalized the data, I smoothed the data. The raw data, as seen above, formed a jagged up and down curve for some years and remained fairly flat for some years, but had the below general shape. A multi-year centered moving average was used. I concluded with an average of the daily elevation over 135 days (27 X 5) – a 27 year average of the day combined with two preceding and two following days. For example, the smoothed elevation for August 29th is the average of the temperatures from August 27 – 31 for 1960 – 1986.



Please note, I included 2/29 in my data, but adjusted for this by ignoring this day in my moving average (ie, for 2/28, I average the 26th, 27th, 28th of February with the 1st and 2nd of March). I then used a straight average of the moving average 2/28 and 3/1 for 2/29.

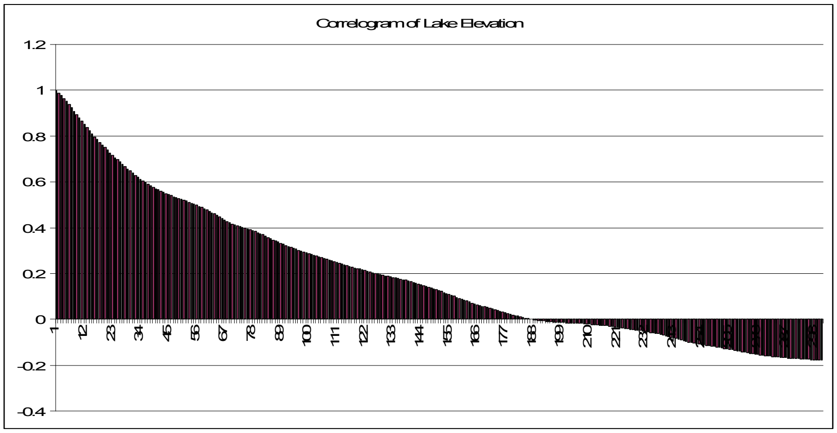
After I had smoothed daily elevations, I computed the seasonally adjusted elevation for each day by subtracting the day's elevation by the centered moving average corresponding to that day. I used an additive adjustment instead of a multiplicative adjustment because I wanted to assume the variance of the error term is constant instead of being proportional to the long-term average.

Please note the graph of the seasonally adjusted data doesn't seem all that different from the non-seasonally adjusted data. However, I believe I did this correctly after reading through lots of guidance.



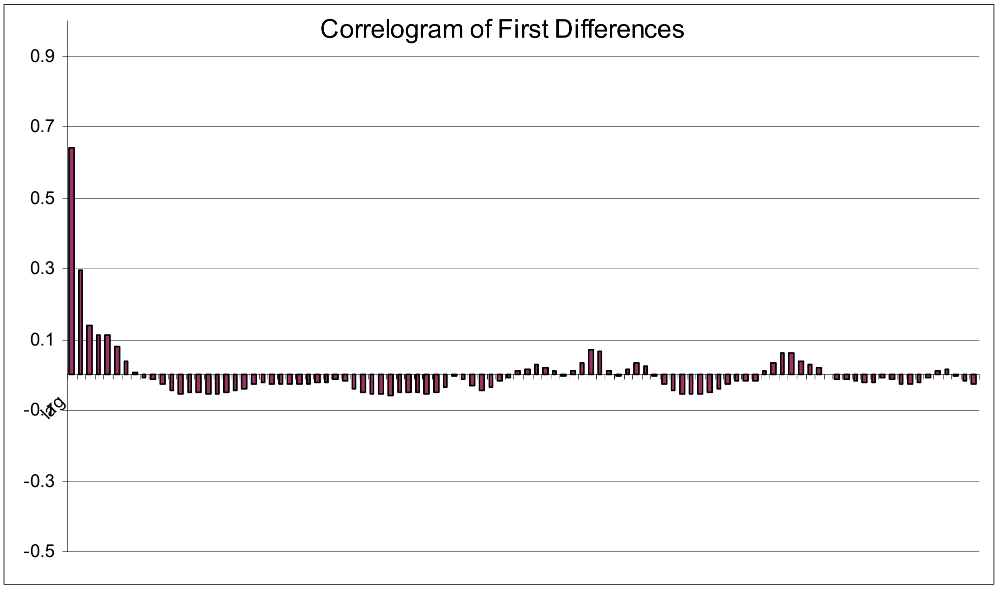
Model Specification & Fitting

Determine if the series is stationary by viewing the correlogram using the sample autocorrelation function. This function shows us how much interdependency there is between neighboring data points in the series. For a stationary series, the autocorrelation function must approach 0 as the displacement gets large. The sample autocorrelation function is an estimate of the autocorrelation function. The sample autocorrelation function for our data set is shown below for lags 0 to 300.



As you can see from the above graph, the sample autocorrelation function for our data set does not approach 0 quickly, so we conclude this series is not stationary.

We transform this series by taking first differences, and proceed by producing yet another correlogram of this new series as shown below for lags 0 to 100. This correlogram shows that the autocorrelation function does approach 0 quickly and from here we conclude that this series is stationary. In attempt to not over difference, we proceed with the first difference of lake elevation to model our time series.



The strong correlation of daily lake elevation on Lake Lewisville suggests that an autoregressive model may be the best fit. The next step was to fit autoregressive models of order 1, 2 & 3 and determine which model fit best.

First, I began with an AR(1) regression model. The AR(1) model forecasts that the current lake elevation is dependent on the immediately preceding day's elevation. Using the excel regression add-in, the following statistics were retrieved:

Regression Statistics	
Multiple R	0.64073
R Square	0.41054
Adjusted R Square	0.41037
Standard Error	0.19636
F Statistic	2542
Observations	3652

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	(0.00026)	0.00325	(0.07924)	0.93684
X	0.64073	0.01271	50.41893	-

These statistical results present a high enough R^2 and P-value suggesting this model is a good fit. In addition, the t-value and F statistic is very high indicating there is a relationship between the daily lake levels on consecutive days.

The AR(2) model forecasts that the current lake elevation is dependent on the previous two days' elevation.

Regression Statistics	
Multiple R	0.65776
R Square	0.43265
Adjusted R Square	0.43234
Standard Error	0.19270
F Statistic	1391
Observations	3651

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	(0.000315)	0.003189	(0.098627)	0.921440
X2	(0.193732)	0.016249	(11.922443)	0.000000
X1	0.764852	0.016244	47.084120	-

As the number of variables increase from 1 to 2, the F statistic drops, and the R^2, P-value, Standard Error, and t-value on the past days' temperature remain relatively the same.

And similarly, the AR(3) model forecasts that the current lake elevation is dependent on the previous three days' elevation.

Regression Statistics	
Multiple R	0.65947
R Square	0.43490
Adjusted R Square	0.43444
Standard Error	0.19236
F Statistic	935
Observations	3650

	Coefficients	*Standard Error*	*t Stat*	*P-value*
Intercept	(0.000287)	0.003184	(0.090188)	0.928143
X3	0.063137	0.016570	3.810381	0.000141
X2	(0.241920)	0.020565	(11.763539)	0.000000
X1	0.777036	0.016528	47.012689	-

Again, as the number of variables increase from 2 to 3, the F statistic drops, and the R^2, P-value, Standard Error, and t-value on the past days' temperature remain relatively the same.

We have seen above that adding the additional variables (day 2 and day 3) to the regression does not materially improve the results. Based on the principle of parsimony, we should choose the model that best fits with the least number of variables.

To further investigate each of the above autoregressive models, I produced the below Durbin-Watson statistics and compared them to the lower and upper critical values d_L and d_U. Please note, I used the critical values for sample size 2000 because I could not find tables that include sample sizes largest than this. I do not believe this will change the conclusion reached.

Significance = 5%

# variables	dL	dU
1	1.92548	1.92747
2	1.92447	1.92847
3	1.92347	1.92947

Significance = 2.5%

# variables	dL	dU
1	1.9114	1.9134
2	1.9104	1.9144
3	1.9094	1.9154

Significance = 1%

# variables	dL	dU
1	1.89505	1.89704
2	1.89405	1.89804
3	1.89305	1.89905

Model	DWS	
AR(1)	1.75173	-statistical evidence that the error terms are positively autocorrelated
AR(2)	1.97562	-statistical evidence that the error terms are **not** positively autocorrelated
AR(3)	2.00698	-statistical evidence that the error terms are **not** positively autocorrelated

I then produced the Box-Pierce Q statistic to test the null hypothesis that the data are independently distributed (i.e. the correlations in the population from which the sample is taken are 0, so that any observed correlations in the data result from randomness of the sampling process). The Box-Pierce Q statistic calculated to be 1500 at lag 1 and increased as the lag increased and was well beyond the critical region. Therefore, I further conclude that the AR(1) model is not white noise.

Conclusion

Based on the analysis until this point, the lake elevation for Lake Lewisville may best be forecast by applying an AR(1) model.

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + e$$

Therefore, based on the results the above, the following AR(1) model is formed:

$$Y_t = 0.000026 + 0.64073 (Y_{t-1}) + e$$

As expected the lake level for the current day is highly dependent on the elevation from the day before further demonstrating that lake levels are not random walks or white noise processes.

Although the results of this project provided a model to fit to historical lake elevations, it may have been better to look at historical precipitation, average temperature and/or water usage in the DFW area to better predict lake levels.

Analysis

See Time Series Analysis.xls for the data used and the calculations/analysis completed