Walter Reedy
Time Series Student Project
Spring 2009 (7-17-2009)

Daily Temperature in Yellowstone National Park, WY

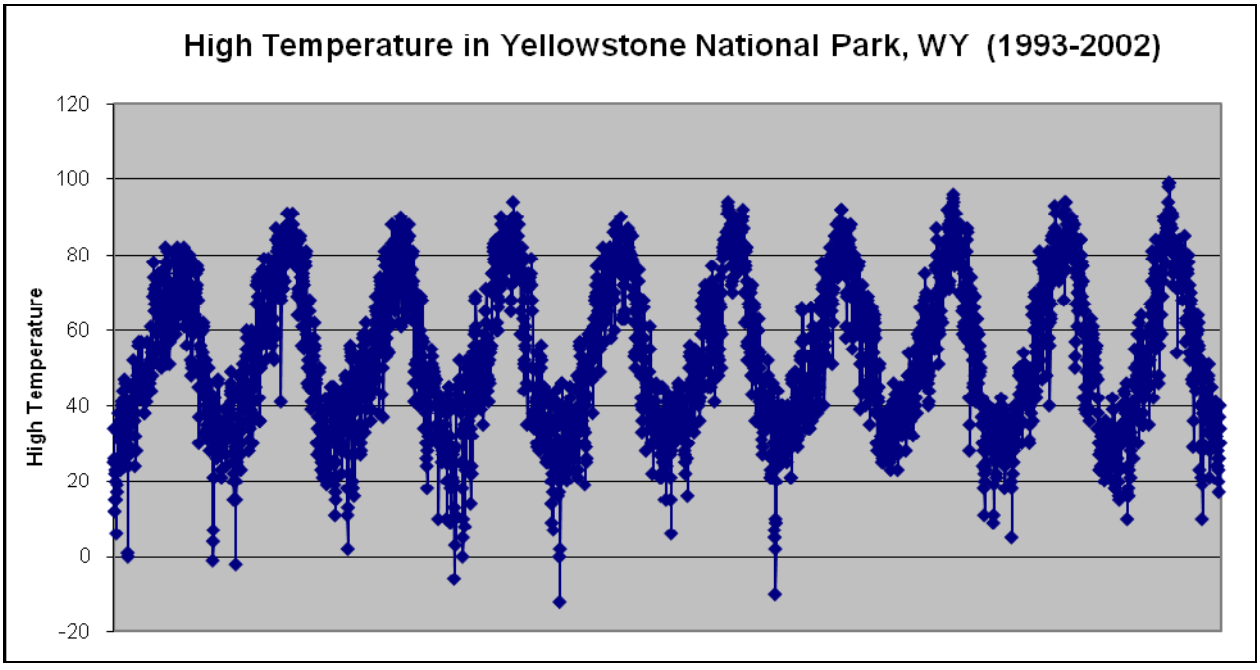
Objective

This project will attempt to create a model to forecast future high temperatures in Yellowstone National Park, WY by fitting an ARIMA time series model to the historic data.

Data

The data for this project was obtained from the NEAS website for Yellowstone National Park, WY (station 489905). I am using the dates 1/1/1993 to 12/31/2002 to fit an ARIMA time series. This 10 year span was selected for its sufficient data points (3,652 days) and for its completeness (no days were missing in the series.

The graph below shows the raw data. This graph shows that there is a strong seasonal pattern to the data, which is consistent with our intuition that the high temperature in the summer months is likely to be greater than the high temperature in winter months. We will need to de-seasonalize the data.



Seasonality Adjustment

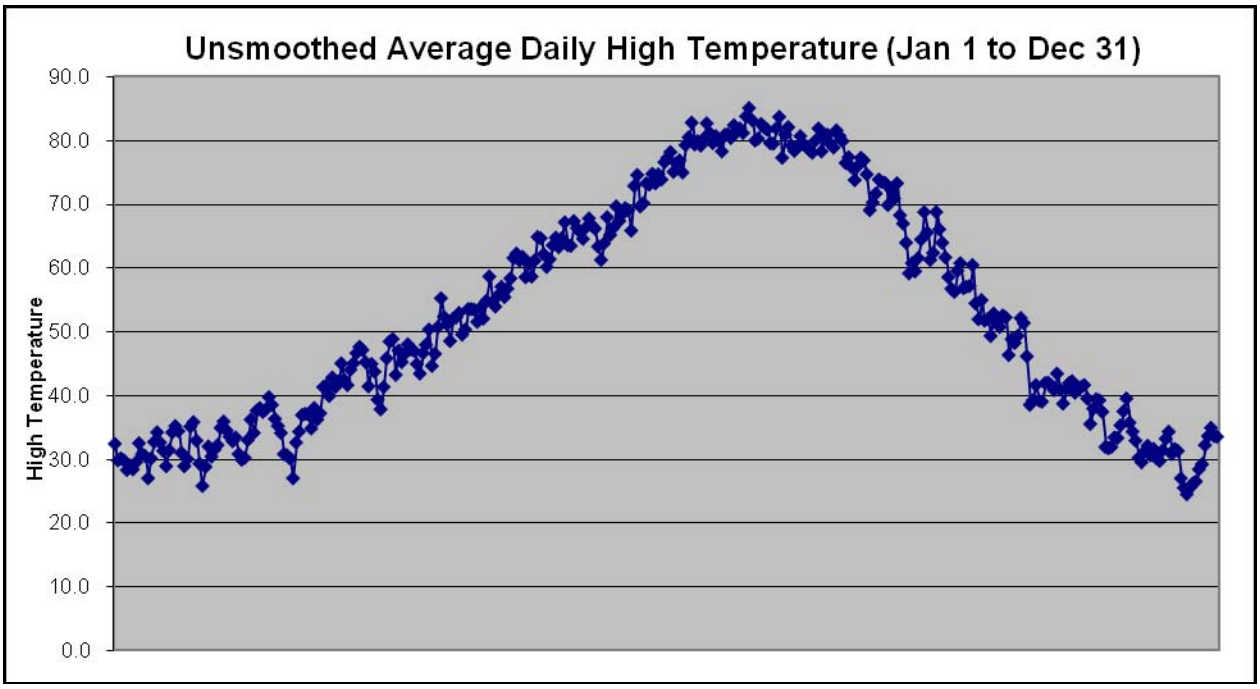
First we will calculate Mean Daily High Temperature, for example Mean Temperature for July 15 is the average of all temperatures on that date between 1993-2002 (82.6°F = [71 +71 +76 +86 +81 +84 +82 +92 +84 +99] / [10days]). It is not particularly intuitive that the average temperature calculated from the raw data for a single day (July 15) would be about 2 degrees higher than the days before and after it. Looking a little closer at the data over the ten year period, we see that July 15th's Minimum High is much higher than for July 13-14 and its Maximum High is much higher than for July 16-17. These 2 data points (July 15th's Min and Max High Temperatures) distort the calculated Mean Daily High Temperature to imply that July 15 is a significantly hotter day than preceding and following days. We would instead expect that the temperature to slowly increase from January to the summer months and then decline into the winter months. The volatiliy in the daily mean can be assumed to be due to the sparsity of observations for that particular day, which is an average of just 10 days.

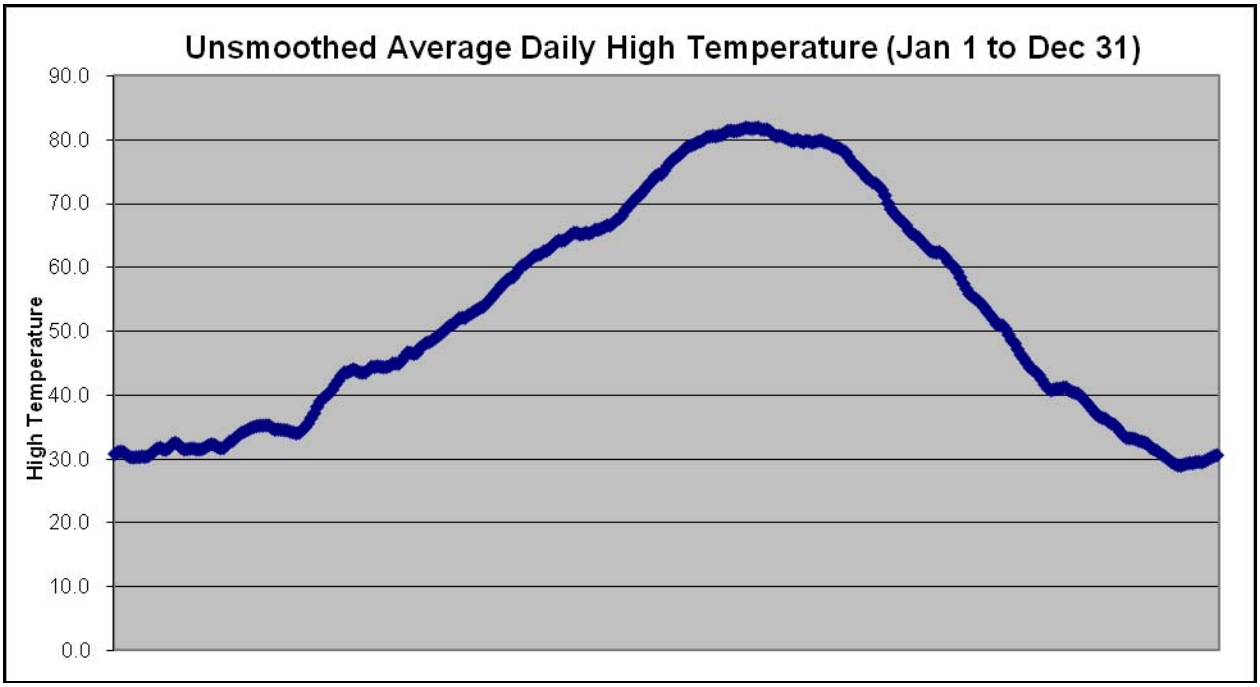
Date	Mean High Temperature	Difference July 15	Min High Temperature	Max High Temperature
July 13	79.1	-3.5	60	98
July 14	80.2	-2.4	63	99
July 15	**82.6**	**0.0**	**71**	**99**
July 16	81.1	-1.5	67	92
July 17	79.5	-3.1	65	91

To smooth the data, we will calculate the Smoothed Mean Daily High Temperature as equal to the 10 year average of the date and the five days before and after the date. For example, the Smoothed Mean Daily High Temperature for July 15 is the average of the 110 data points between July 10-20 for 1993-2002. Using this Smoothed Mean Daily High Temperature will reduce the distortion in our seasonality adjustment.

Date	Mean High Temperature	Smoothed Mean High Temperature
July 13	79.1	79.6
July 14	80.2	79.9
July 15	**82.6**	**80.3**
July 16	81.1	80.4
July 17	79.5	80.5

The graphs below show the “Unsmoothed” Raw Mean Daily High Temperature and the Smoothed Mean Daily High Temperature.





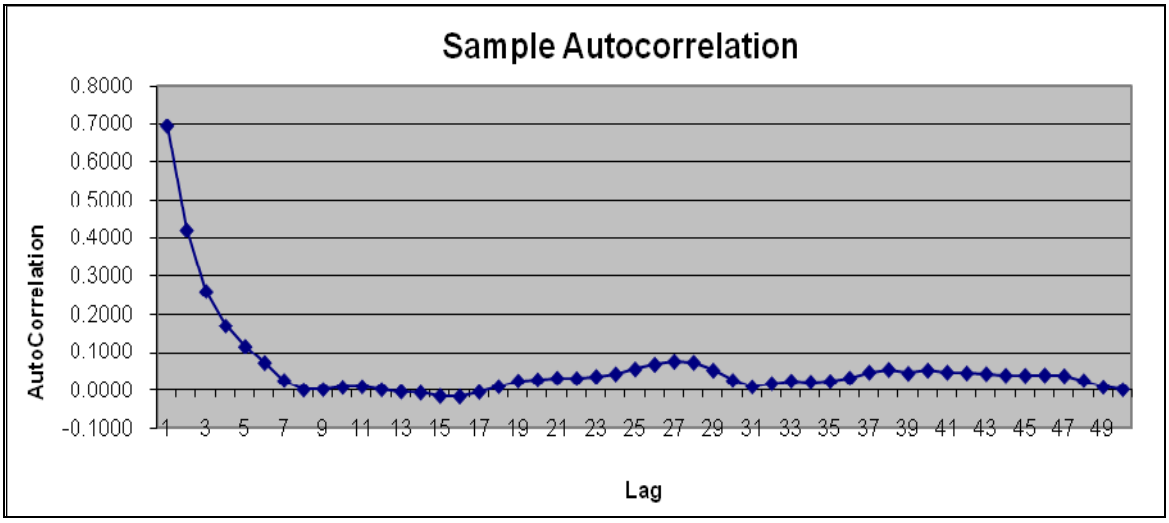
Next I calculate the seasonality adjusted time series by taking the difference in the original time series and the mean for each day. My average of these values over all the data points (0.0158) was slightly different than zero since the smoothed mean is slightly different than the simple average of the points I am taking a difference between. To correct for this, I normalize the series by subtracting each data point by 0.0158, which now makes the mean of this series equal to zero. This time series will represent how far the actual high temperature is from its expected value. It should be stationary with a mean of zero, which we will examine by graphing the sample autocorrelations.

Sample Autocorrelation

The sample autocorrelations will provide us with an estimate of the autocorrelation function. This will show us how dependant a value is on its surrounding data points. The equation we will use to calculate the sample autocorrelations is shown below.

$$\rho_k = \frac{\sum(y_t - \bar{y})(y_{t+k} - \bar{y})}{\sum(y_t - \bar{y})(y_t - \bar{y})}$$

The graph below shows the results for the first 50 lags (values of k).



The sample autocorrelations fall to zero as the lag gets larger. This indicates that the series is stationary. Also, the first few lags have a value significantly greater than zero, which suggests that the true autocorrelation coefficient is not zero and thus the series is not a white noise process. Since the sample autocorrelation does not fall to zero until about the 7th lag, this suggests that the ARIMA model is of an order less than 7. For this project, I will examine AR(1), AR(2), AR(3), and ARMA(1,1)

Model Specification and Estimation

AR(1): I ran a linear regression to fit the parameters of an AR(1) model,

where X Variable 1 = the high temperature 1 days ago

Regression Statistics	
Multiple R	0.69
R Square	0.48
Adjusted R Square	0.48
Standard Error	6.74
Observations	3651

	Coefficients	*Standard Error*	*t Stat*
Intercept	0.00	0.11	0.01
X Variable 1 **ϕ1 =**	**0.69**	0.01	57.92

AR(2): I ran a linear regression to fit the parameters of an AR(2) model,

where X Variable 1 = the high temperature 2 days ago

X Variable 2 = the high temperature 1 days ago

Regression Statistics	
Multiple R	0.70
R Square	0.49
Adjusted R Square	0.49
Standard Error	6.70
Observations	3650

	Coefficients	*Standard Error*	*t Stat*
Intercept	0.00	0.11	-0.01
X Variable 1 **ϕ2 =**	**-0.11**	0.02	-6.94
X Variable 2 **ϕ1 =**	**0.77**	0.02	46.89

AR(3): I ran a linear regression to fit the parameters of an AR(3) model,

where X Variable 1 = the high temperature 3 days ago

X Variable 2 = the high temperature 2 days ago

X Variable 3 = the high temperature 1 days ago

Regression Statistics	
Multiple R	0.70
R Square	0.49
Adjusted R Square	0.49
Standard Error	6.70
Observations	3649

	Coefficients	*Standard Error*	*t Stat*
Intercept	0.00	0.11	0.02
X Variable 1 **ϕ3 =**	**0.03**	0.02	1.66
X Variable 2 **ϕ2 =**	**-0.14**	0.02	-6.53
X Variable 3 **ϕ1 =**	**0.77**	0.02	46.80

ARMA(1,1): I used Yule Walker equations to fit the parameters of an ARMA(1,1) model,

where Sample $\rho 1 = 0.6920 = \phi 1 - \theta 1$

Sample $\rho 2 = 0.4195 = \phi 1 * (\phi 1 - \theta 1)$

Solve: **$\phi 1 = 0.6062$**

$\theta 1 = -0.0859$

Regression Analysis on AR(1), AR(2), and AR(3)

The t statistics are high for each model, indicating that there is a relationship between high temperatures on consecutive days. The R^2 improves only slightly from AR(1) to AR(2) and AR(3), while the standard error decreases only slightly. This implies that adding the 2nd and 3rd days to the regression improves the result, but the improvement may not be material enough to select the AR(2) or AR(3) model over the AR(1) model given the principle of parsimony.

Durbin-Watson Test

I will test the Null hypothesis that there is no serial correlation in the residuals of our selected model. Since all four Durbin-Watson statistics are close to 2.0000, I will not reject the Null Hypothesis that there is no serial correlation. It is noteworthy that the AR(2) model has a DW statistic that is much closer to 2.0000 than the other models.

Model	Durbin Watson Statistic
AR(1)	1.8413
AR(2)	1.9926
AR(3)	1.8411
ARMA(1,1)	1.8382

Box-Pierce Q Statistic

I will test the joint hypothesis that all of the autocorrelation coefficients are zero. This would indicate that the residuals of our selected model are the result of a white noise process. The results below are calculated using 200 lags (k=200) and varied the degrees of freedom as necessary.

The Box-Pierce Q statistic is significantly lower for AR(2) compared to AR(1) or ARMA(1,1). Further, the AR(1) model fails the test at the 10% critical level, although it would pass at the 5% critical level. Based on these results, I tend to favor the AR(2) over the AR(1) or ARMA(1,1).

AR(3) has a lower Q stat relative to the critical level than AR(2), but the improvement may not be material. Based on these results, I tend to favor the AR(2) over the AR(3) based on the principle of parsimony.

Model	Box-Pierce Q stat (k=200)	Critical 10% Level	Result	Critical 5% Level	Result
AR(1)	231.83	224.96	Reject	232.91	Accept
AR(2)	177.07	223.89	Accept	231.83	Accept
AR(3)	168.43	222.83	Accept	230.75	Accept
ARMA(1,1)	207.56	223.89	Accept	231.83	Accept

Conclusion

I conclude that am AR(2) time series is the most appropriate model for the High Temperature in Yellowstone National Park. My conclusion is supported by the following summary of results which were discussed in more detail above:

- Regression analysis shows an improvement in R^2 and Standard error for AR(2)
- Durbin-Watson statistic is very close to 2.0000 for AR(2)
- Box-Pierce Q statistic is significantly lower for AR(2)

Therefore, the equation for the model of how far the actual high temperature is from its expected value in Yellowstone National Park is:

$$(Y_t - \mu_t) = 0.77 (Y_{t-1} - \mu_{t-1}) - 0.11 (Y_{t-2} - \mu_{t-2})$$

Where: Y_t = High Temperature

μ_t = Smoothed Mean Daily High Temperature

$(Y_t - \mu_t)$ = Difference from the Smoothed Mean

Finally, we can also add back in the Average High temperature (which will vary by day) to get a forecast of the actual High Temperature for Yellowstone National Park. This will be equal to the long-term Expected Temperature + an adjustment based on the temperature of the last 2 days (which will be 0°F on average).

$$Y_t = \mu_t + 0.77 (Y_{t-1} - \mu_{t-1}) - 0.11 (Y_{t-2} - \mu_{t-2})$$