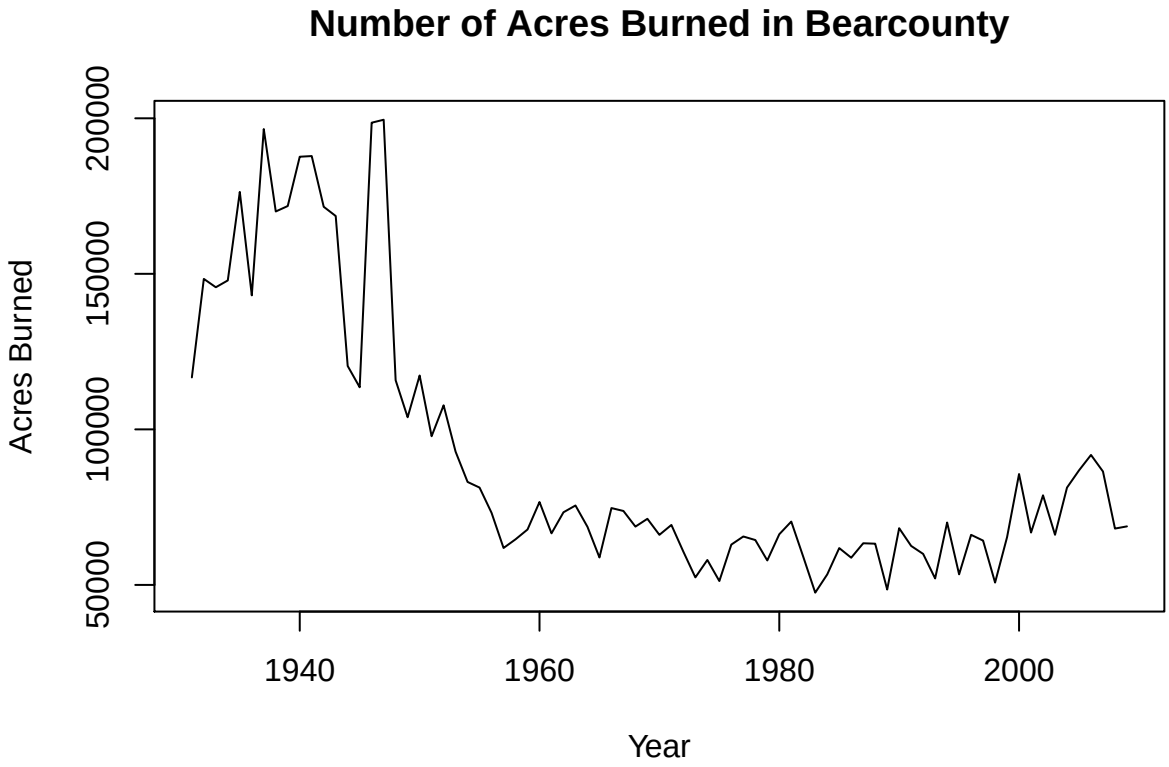
Stat 153 Final Project

1 Executive Summary

The governors and the fire-fighters of Bearcounty provided time series data of the number of acres burned in Bearcounty each year. In this analysis, I use an AR(2) to predict the number of acres burned in the next ten years to come. The required yearly allocation of budget and resources to limit the spread of these fires should be expanded based on my predictions, as my model predicts a rise in the number of acres burned in Bear County over the next 10 years.

2 Exploratory Data Analysis



(Figure 1): This figure plots the total number of acres of forest burned in Bearcounty from 1931-2009, where time is indexed by each year.

There is an overall downward trend in the number of acres burned in Bearcounty each year. The trend decreases exponentially with some asymptotic behavior around 50,000 acres burned per year. Perhaps the technology used to fight fires made substantial improvements throughout the 40s and 50s until technological progress plateaued off around the 60s, but this is conjecture not based in any research.

There is an interesting cyclical pattern, seemly happening every 4-6 years. By simply looking at the graph, the amplitude of these fluctuations varies across time. This leads us to the peculiar part of the time series. The time series is heteroskedastic with decreasing variance over time. This will make it quite difficult to

fit a parametric model based on sinusoids, as the number of acres burned may not repeat predictably over particular periods. However, as the time series trend becomes nearly flat after the 60s, seasonal effects will become easier to identify. Perhaps these seasonal effects are due to scheduled controled burns, natural fluctuations in humidity and temperature throughout each decade, or some combination of these and other factors.

3 Models Considered

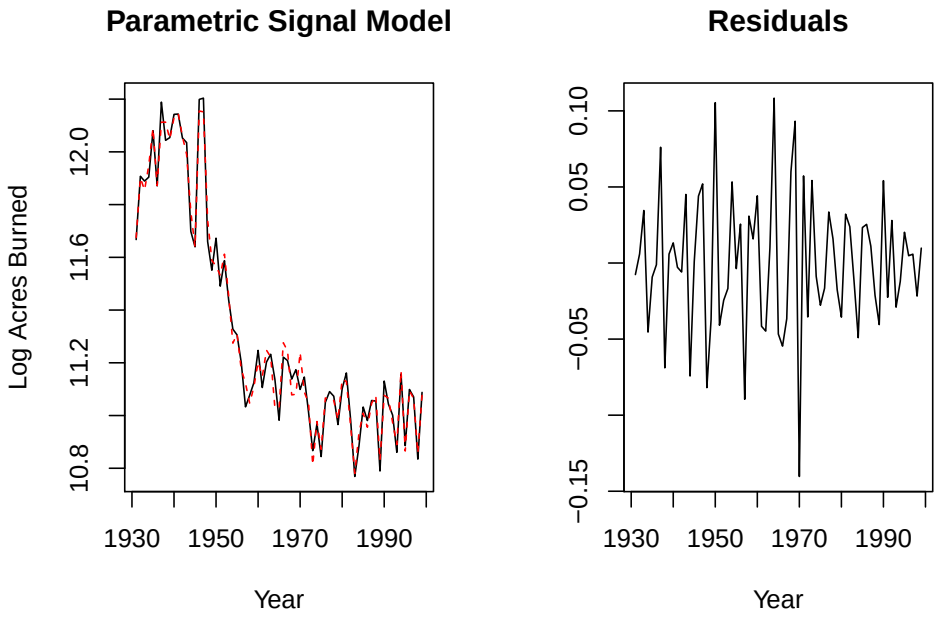
The signal of this data can be captured using parametric model and a differencing approach. ARMA models will be best equiped to handle the stationary "noise" left over in the latter portion of this report. I fit these models from years 1931-1999 so 10 years may be used as an "out-of-sample" consideration of how each model performs.

3.1 Parametric Signal Model

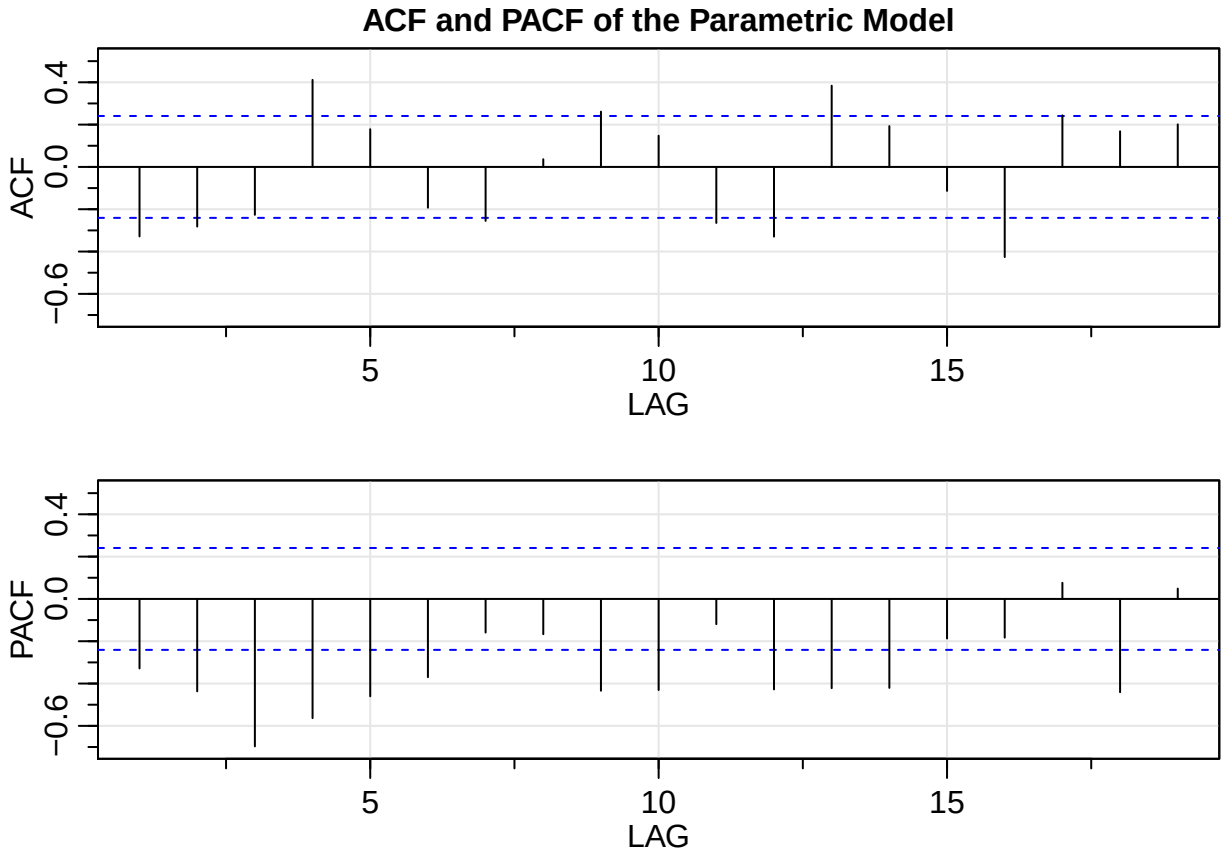
To address the natural downward trend of the number of acres burned, an exponential function with negative power may be fit to the data. The form of the exponential function is given by,

$$e^{-at}$$

such that t is the given year and a is a real number. Next, to remove relevant seasonal components of the data, a sinusoid with 6 different frequences may be fitted to the time series. These frequencies were determined by iteratively addressing the periodogram of the residuals and updating the model with a new sinusoidal component. Somewhat expectedly, relevent frequencies correspond to periods of 4, 5, and 6 years, as well as larger periods up to 25 years. To capture the decreasing fluctuations in the number of acres burned throughout each decade, each sinusoid interacts with the current year. Finally, heteroscedasticity of the data is addressed by taking the log of the data. (Figure 2) displays the model and the residuals below. (Figures 3 and 4) display the ACF and PACF of the residuals after the parametric model fit.



(Figure 2): The fit and residuals for the parametric model.



(Figures 3 and 4): ACF and PACF of the residuals after the parametric fit.

3.1.1 SARIMA(1,0,1,0,0,2)[2]

After examining the ACF and PACF, I decided it was plausible that p and q were greater than zero considering that both the ACF and PACF are nonzero for all lags. The original ACF and PACF were hard to diagnose, but after fitting an ARMA(1,1), I found that a SARIMA(0,0,2)[2] could be added as well. The residual plot did not show significant bars or trend, but in all my efforts I could not improve the Ljung-Box statistics as they were consistently less than the p-value cutoff and were considered insignificant.

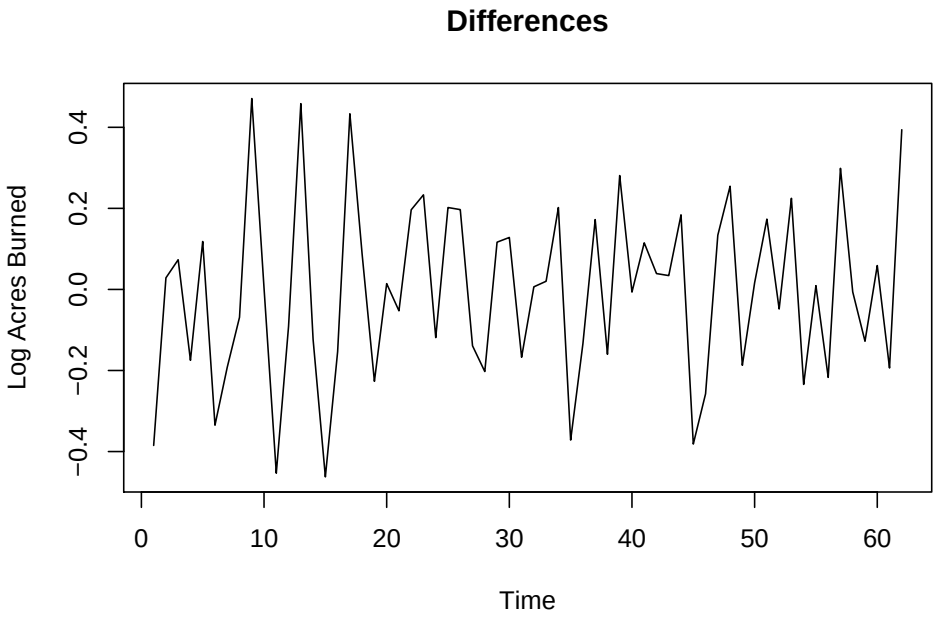
3.1.2 SARIMA(0,0,4,2,0,0)[4]

After diagnosing the ACF once more, I decided an MA(4) model was plausible, considering the last of the significant bars for short lags occurs at lag 4. In addition, after examining the ACF after this model fit, I moved forward with a SARIMA(0,0,4,2,0,0)[4]. The residual plot resembled a stationary process, but the Ljung-Box test did not indicate this was the case.

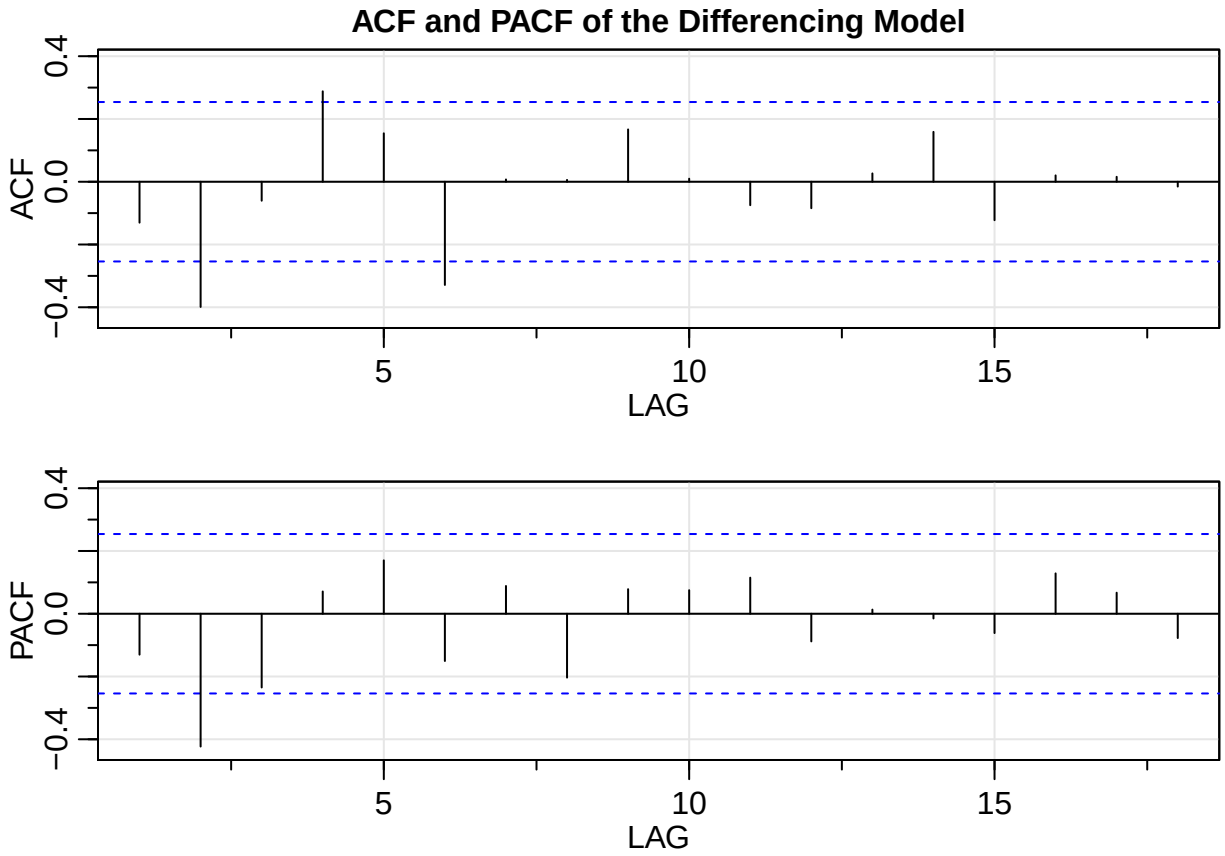
3.2 Differencing Model

As addressed in the above model, the periodogram revealed seasonal effects at periods of approximately 4, 5 and 6 years. This is an approximation due to leakage, but events like controlled burns and weather, humidity, and precipitation patterns could reasonably follow closely to these periods. To address the heteroskedasticity

of the number of acres burned per year, the log of the data set can be taken prior to differencing To remove the linear trend, I will difference once, and to remove any trend of higher order and also remove seasonal effects every 6 years, I have differenced once more at lag 6. (Figure 5) plots the result of these differencing operations and (Figures 6 and 7) displays the relevant ACF and PACF



(Figure 5): The result after second order differencing, once at lag-6, of the log data.



(Figure 6 and 7): ACF and PACF after second order differencing.

3.2.1 SMA(3)[2] on Second Differences.

Confronted with a strange ACF and PACF once more, I decided to get creative and fit a SARIMA(0,0,3)[2] considering the significant bars each spaced out by 2 in the ACF. The AR model had p-values above the cutoff for the Ljung Box test for less than 12. The residual plot displayed a significant bar for lag = 11, but I did not want to over complicate the model by trying to address this bar.

3.2.2 AR(2) on Second Differences.

On a similar note, I thought that the ACF revealed what seemed to be an exponentially decreasing pattern for bars separated by lag 2. After examining the ACF of the residuals of the SARIMA(1,0,0)[2] model, I also fit a AR(1) as well. This is simply an AR(2) model. The AR model had high p-values for the Ljung Box test for less than 12. The residual plot after fitting this model seemed to resemble white noise.

4 Model Comparison and Selection

In (Table 1) below, I displayed the MSE after comparing the log forecast to the log data for years 2000-2009. I also displayed the AICc and BIC for each model as a consideration of the complexity of each model. I also specified what model type was used to remove signal components of the time series, as the parametric models were considerably more complicated and the AICc and BIC are not necessarily the best indication of this.

Signal Model	SARIMA Fit	MSE	AICc	BIC	CV Score
Parametric Fit	SARIMA(1,0,1,0,0,2)[2]	0.663	-4.848	-4.668	NA
Parametric Fit	SARIMA(0,0,4,2,0,0)[4]	0.749	-5.209	-4.977	NA
Differencing	SAM(3)[2]	0.458	-0.365	-0.204	3.1734
Differencing	AR(2)	0.624	-0.236	-0.106	2.5314

Table 1: AICc, BIC, MSE and CV Score for each model. MSE displays the error after using the first 69 years to predict log acres burned from 1999-2009. The CV Score is calculated as the mean of the MSE when predicting 10 years ahead using all previous years, interating by 10 from 1931 to 1991.

The parametric models have better BIC and AICc scores, but the parametric fitting process is significantly more complicated than differencing. Considering the differencing models perfrom slightly better than the parametric fit models in their MSE measure, I discarded the parametric fit models because they were far too complicated and the differecing models seemed to be comprable. Furthermore, I calculated a cross-validation score for each differencing model as described in the caption. I found that the AR(2) had a CV score of 2.5314 and the SMA(3)[2] had a CV score of 3.1734. Because the AR(2) model had the lowest CV score out of the two models and comparable MSE "out of sample" (for years 2000-2009) to the other models, I moved forward with an AR(2) as my final model.

5 Results

Let Y be the original data, then

$$V_t = \nabla\nabla_6 \log(Y_t)$$

Where V is assumed to be a stationary process. Let W be a white noise. Mathematically, the AR(2) model is expressed as:

$$(1 - \phi_1 B - \phi_2 B^2)V_t = W_t$$

5.1 Model Parameters

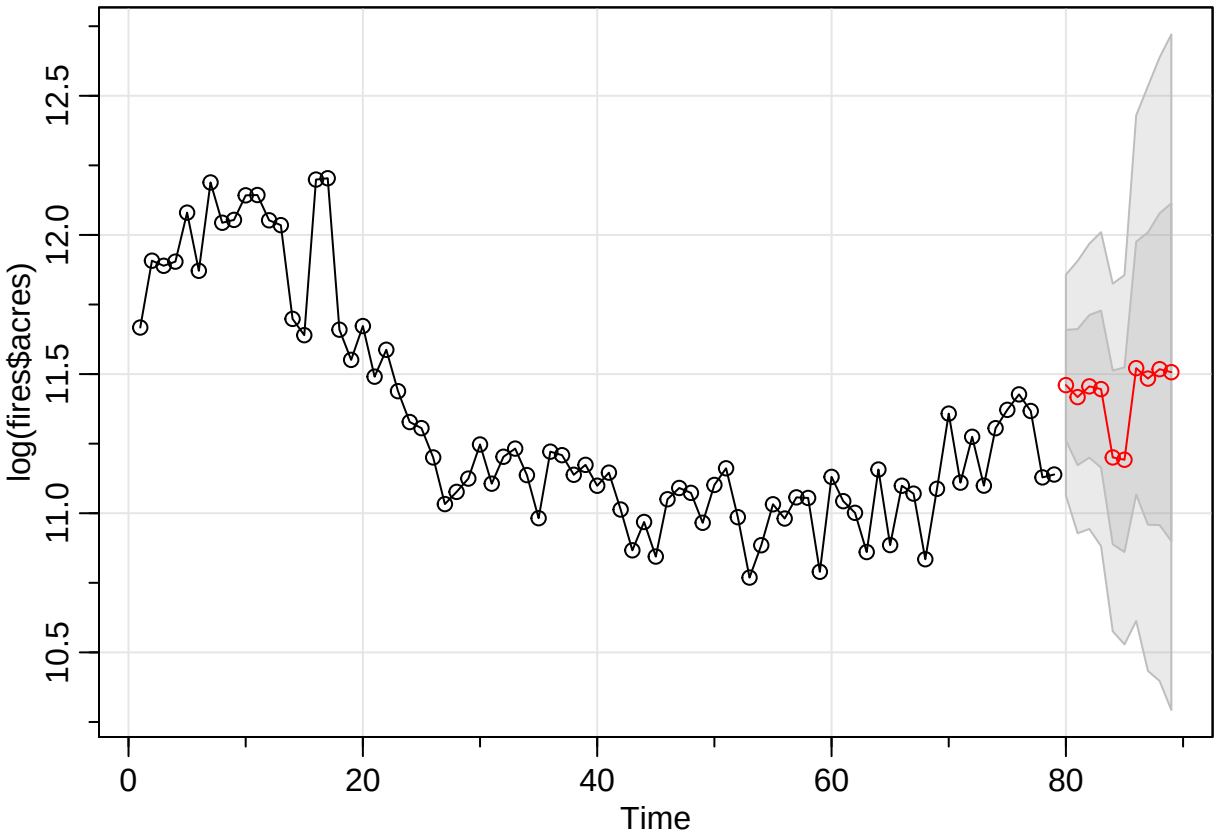
The model parameters and standard errors are estimated below in (Table 2).

Parameter	Estimate	Standard Error
ϕ_1	0.19495	0.1191
ϕ_2	0.45599	0.1172

(Table 2): The esimtated parameters of the AR(2) model on the stationary time series V.

5.2 Model Predictions

My model predicts a rise in acres burned by around 40,000 acres in 2010. This prediction would indicate to the firefighters of Bear County that the must expand and draw on all resources they can to prepare for the next year's fire season. by 2012, my model predicts a decrease in number of acres burned for a few years, then, another increase to around 100,000 acres burned. The firefighters of Bear County should expand their force and recruiting to keep the county safe. The predictions of my model are displayed below in (Figure 8).



(Figure 8): Final model predictions in log scale.