Stat 153 Project: Acreage Burned in California Wildfires

Rowan Pan

12/14/2020

1 Executive Summary

Bear County is a heavily forested (fictional) county in California where many acres of land are burned due to wildfires every year. Using the data set of historical annual acres burned from the years 1931 to 2009, I will forecast the acres burned for the next ten years using a log first differences signal model with a AR(2) model for noise. According to this model's predictions, the impact of wildfires on acreage burned is on a decreasing trend and will slightly decrease year-over-year for the next decade. However, the county officials should still at the minimum sustain current firefighting budgets and take precautions with power lines, as the wildfire damages throughout the next ten years are still substantial and comparable to the recent years on record.

2 Exploratory Data Analysis

Looking more closely at the time series in the left panel of Figure 1, the overall trend of historical acres burned seems to be increasing from 1931 to 1940, followed by a rapid descent after that (with the exception of the massive upward spike at years 1946 and 1947), and then beginning a period of general stability from 1970 to 1990. More recently, in the 2000s, the acres burned from wildfires have picked up pace again.

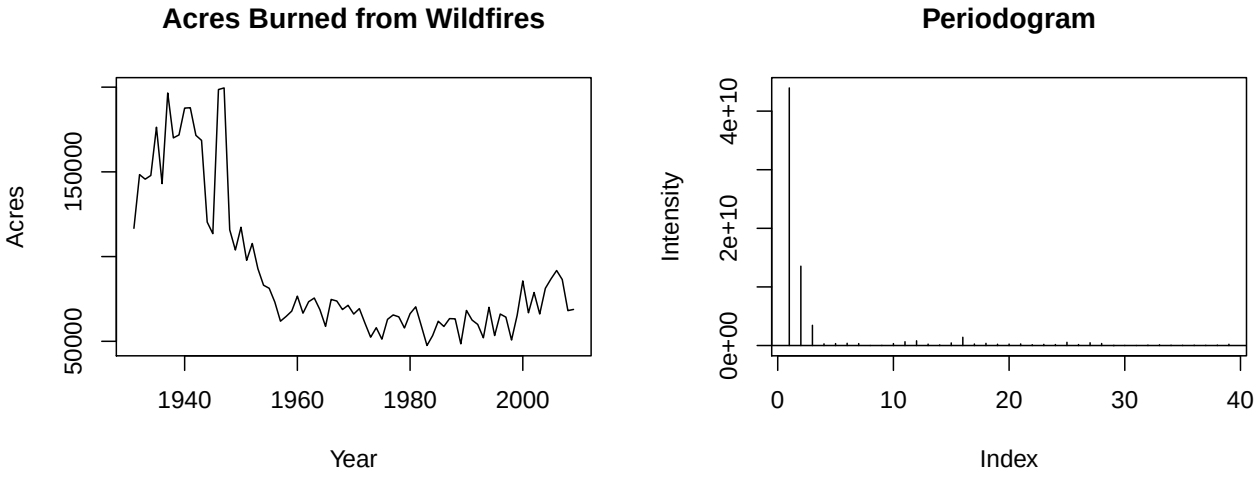

Figure 1: Acres burned each year in Bear County, and the resulting periodogram

Additionally, there exists heteroscedasticity in the time series, with higher variance in earlier years than middle or later years. Moreover, the variance increases with the mean, as the earlier years (left side) have higher mean and generally higher variance too. Lastly, there does not seem to be obvious seasonality or seasonal spikes in the time series.

The right panel presents the periodogram of acres burned, where the dominant index $j = 1$ corresponds to a frequency of $j/n \approx 0.0126$, where n is the number of observations, 79. There also exists two smaller bars at $j = 2$ and $j = 3$, which will be discussed more at length in the parametric model section.

3 Models Considered

To model the signal of the data, a parametric model and a differencing approach are employed. The noise that remain will be addressed using two ARMA models each, resulting in four total ARMA models.

3.1 First Differences Signal Model

First, I will pursue stationarity with lag-1 differencing, which is equivalent to annual differences, as each observation represents one year and there does not seem to be other strong seasonality present. As mentioned in the EDA section, the variance increases with the mean. To combat this larger variance in earlier years, the log is used as a variance stabilizing transform (VST). The log transform deals with heteroscedasticity better than the square root since the range of the data (i.e. standard deviation) increases linearly with the mean, which is equivalent to saying the variance increases quadratically with the mean. As seen in the left panel of Figure 2, the time series of the log differenced data looks stationary and homoscedastic for the most part, albeit containing a larger spike at years 1946 and 1947.

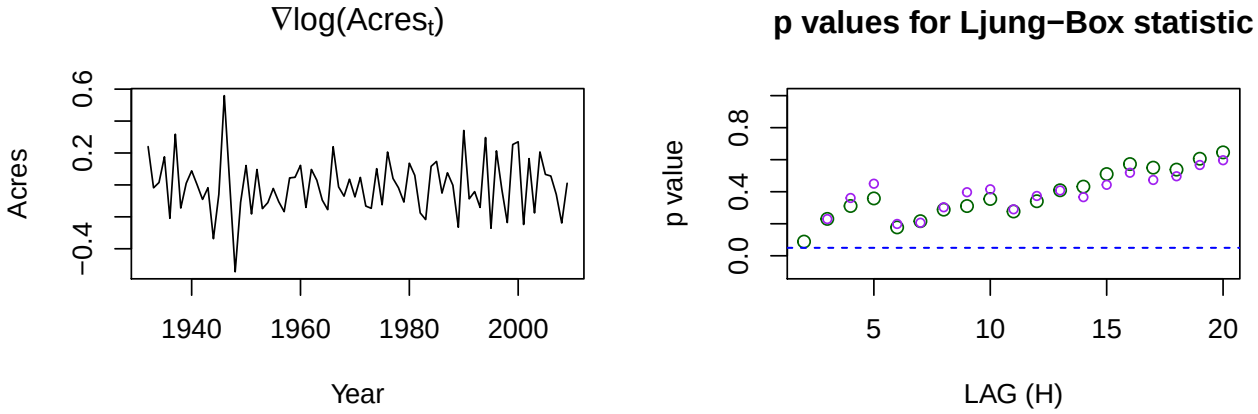

Figure 2: The left panel shows the differenced series of log acres burned. The right panel shows the Ljung-Box statistic plot for the differenced data with MA(1) model in dark green and the AR(2) model in purple.

3.1.1 First Differences with MA(1)

The `auto.arima()` function on the log differenced data recommends MA(1). This seemingly makes sense since the sample ACF in Figure 3 has one significant lag and the PACF seems to be tapering overall and mostly staying inside the blue confidence bands after lag 2. The fit of this model is shown in Figure 3 in dark green. The green circles show that the theoretical MA(1) specification fits well to the sample ACF and PACF, but does not overlay atop the significant autocorrelations super closely.

3.1.2 First Differences with AR(2)

The ACF of the log differenced data shows one significant autocorrelation at lag 1 and the PACF shows two significant autocorrelations at lags 1 and 2. As can be seen in the left plot of Figure 3, the excess in lag 1 of the sample ACF is quite minor. Additionally, the ACF exhibits a slight decreasing pattern – albeit not a super obvious decreasing rollercoaster that is nice to see for a clear-cut justification of an AR model. Combined with that fact that there are two significant lags, an AR(2) model will be used as the second noise model. The fit of this model is shown in red in Figure 3. The red dots for AR(2) indicate a better theoretical fit than the green dots, as the red dots follow lags 1 and 2 in both the sample ACF and PACF very closely.

Lastly, the green and purple circles in the Ljung-Box plot in Figure 2 shows that both models have insignificant p-values at all 20 lags shown, which is a good sign for the model fit, as we cannot reject the null hypothesis that the standardized residuals are independent. If the p-values were instead significant, that suggests there is some correlational structure to the residuals that the ARMA model was not able to account for.

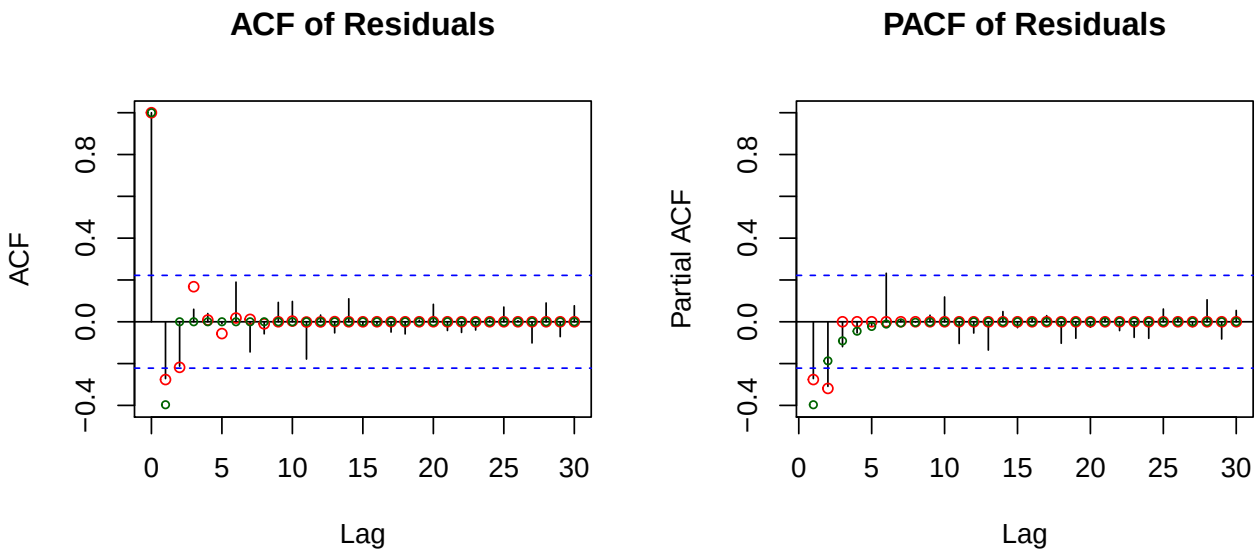

Figure 3: Autocorrelation function (ACF) and partial autocorrelation function (PACF) values for the log first differences. Red dots reflect the AR(2) model, while the green dots reflect the MA(1) model.

3.2 Parametric Signal Model

Secondly, I will consider a parametric model using the index from dominant sinusoids derived from the periodogram in Figure 1. As mentioned in the EDA, the periodogram has a significant spike at $j = 1$, which corresponds to a frequency of $j/n \approx 0.0126$. Although there are spikes at $j = 2$ and $j = 3$, I chose not to include those frequencies due to speculations about leakage, since they are not overwhelmingly sticking out after removing the dominant sinusoid at $j = 1$ (not shown in this report to conserve space).

The sinusoid parametric model can be written as follows, where $Acres_t$ is acres burned in year t and X_t is a stationary noise process.

$$log(Acres_t) = \beta_0 + \beta_1 t + \beta_2 cos\left(\frac{2\pi t}{79}\right) + \beta_3 sin\left(\frac{2\pi t}{79}\right) + X_t \tag{1}$$

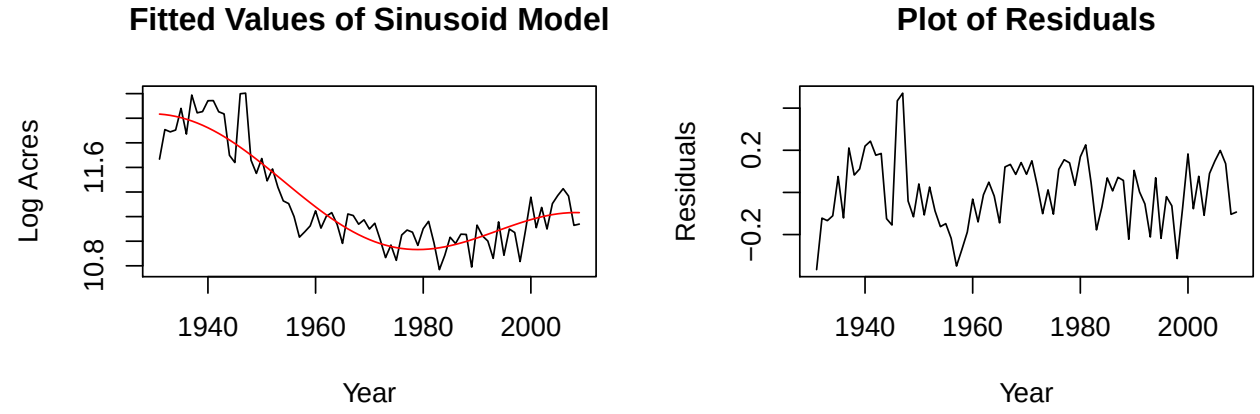

Figure 4: Parametric fit of acres burned. The left panel shows the model's fitted values in red, plotted on top of the acres burned data in black. The right panel shows the residuals of this model.

The fitted model in red tracks time series quite well. The residual plot in Figure 4 also appears stationary over time with generally constant variance, albeit with a tall spike in the years 1946 and 1947 again.

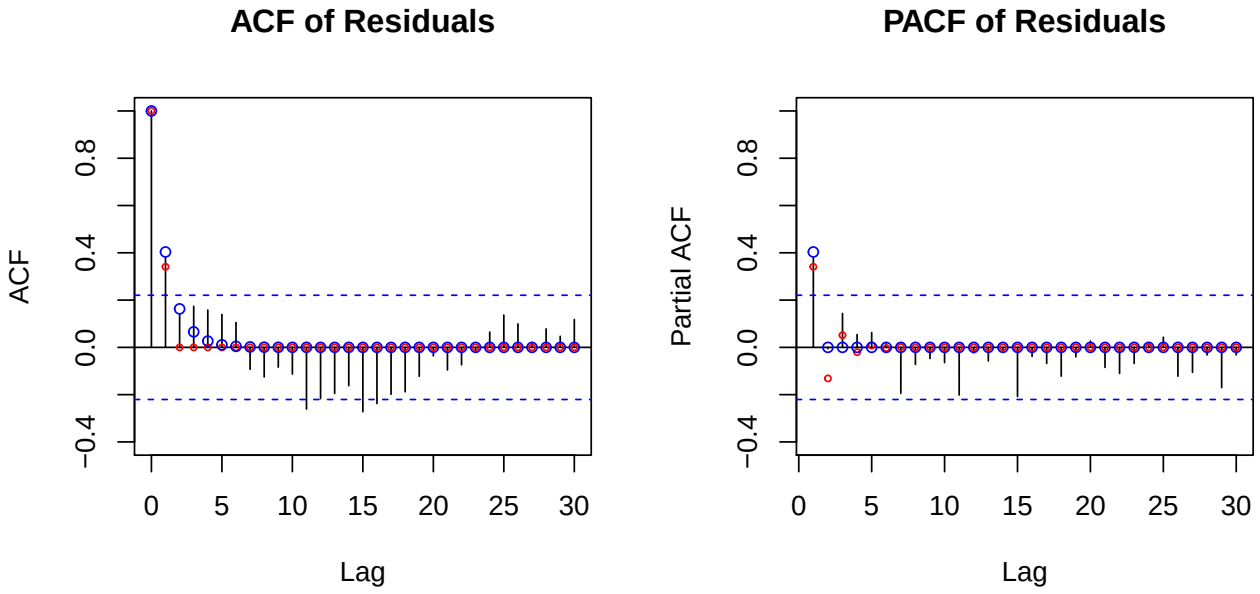

Figure 5: Autocorrelation function (ACF) and partial autocorrelation function (PACF) values for the parametric sinusoid model. Red dots reflect the MA(1) model, while the blue dots reflect the AR(1) model.

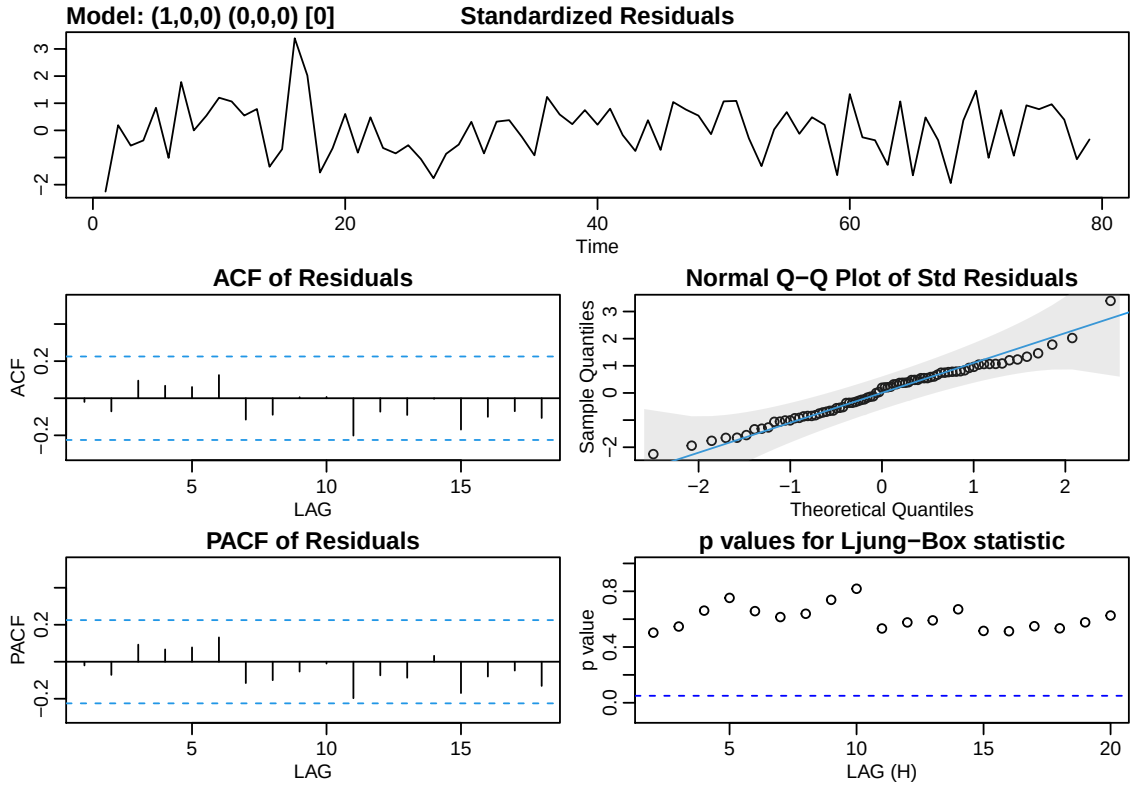

Figure 6: Diagnostic plots from sarima for Parametric Model with MA(1)

3.2.1 Parametric Model with AR(1)

The recommended model according to `auto.arima()` is ARIMA(1,0,0), which is AR(1). This makes sense since the sample PACF has one significant lag and the sample ACF has an decreasing rollercoaster-like pattern. The blue dots show that the theoretical fit of AR(1) is great, as it overlays and fits the sample ACF/PACF very well, especially at lag 1. Additionally, the entirety of `sarima` function's diagnostic plots are shown, where it shows that all lags in ACF/PACF are within the 95% confidence bands and all p-values in the Ljung-Box statistic are insigificant, both of which are good signs of model fit.

3.2.2 Parametric Model with MA(1)

As we can see in the sample ACF, there is only a significant sample autocorrelation at lag 1. Combined with the observation that the sample PACF has a slight decreasing pattern, it is most fitting to use another simple model, namely the MA(1), to compare with AR(1) from before. The theoretical model fit is shown in red in Figure 5, where it shows that the red dots do not follow the sample autocorrelations in the sample ACF/PACF as well as the blue dots.

4 Model Comparison and Selection

In this section, the best among the four candidate models will be selected for forecasting purposes.

4.1 Information Criterion

As we can see in the Table 1 below, according to all three information criterions, the two best performing ARMA models (one for each signal model) is First Differences with MA(1) and Parametric Model with AR(1), since these two have the lowest scores of AIC, BIC, and AICc. It is also important to note that the information criterion scores cannot be compared across different likelihood functions (i.e. signal models in this case), which is one reason why another metric of comparison is introduced in the next section.

Table 1: Information Criterion Diagnostics for Four Models Under Consideration

Model	AIC	BIC	AICC
First Differences with MA(1)	-0.7264380	-0.6357954	-0.7243867
First Differences with AR(2)	-0.7228388	-0.6019820	-0.7186808
Parametric Model with AR(1)	-0.9158332	-0.8258542	-0.9138346
Parametric Model with MA(1)	-0.9066414	-0.8166624	-0.9046428

4.2 Cross Validation

Time series cross validation is a method to help with selecting the model with the best out-of-sample fit. To perform this cross validation exercise, the entire data is split into training and testing sets. The non-overlapping testing sets act as a rolling window through the last 40 years in the data, 1970-2009, in 5 year segments (e.g. 1970-1974, 2005-2009). Each was forecasted using a training data set of all the data that occurred prior to the appropriate testing set. Then, the four candidate models' forecasting performances will be compared through the root mean squared prediction errors (RMSPE). In the end, the model with the lowest RMSPE chosen to forecast the next ten years.

Cross validation provides a more robust way to directly compare these models, as it gauges absolute out-of-sample performance using the RMSPE. As shown in Table 2 below, the best performing model is the First Differences with AR(2) model. This result is rather interesting, especially given how the First Differences with MA(1) model performed better in the information criterion comparison. However, since cross validation is a more comprehensive way to compare between all the models, I will consider this metric with more weight. By that metric, between the two differencing models, since the AR(2)'s RMSPE is a lot lower than MA(1)'s, I conclude the additional parameter is worthwhile and choose the more complex AR(2) model for forecasting.

Table 2: Cross-validated Out-of-sample Root Mean Squared Prediction Error

	RMSPE
Parametric Model with AR(1)	0.9128781
Parametric Model with MA(1)	0.9187643
First Differences with MA(1)	0.8168405
First Differences with AR(2)	0.7877398

5 Results

To forecast acres burned in the next ten years, a differencing model will be utilized. Let $Acres_t$ be the acres burned in year t, ∇ be the differencing operator, E be the expectation, and ϕ be the autoregressive parameters. X_t is a stationary process defined by AR(2), where W_t is white noise with variance σ_W^2.

$$log(Acres_t) = log(Acres_{t-1}) + E[\nabla log(Acres_t)] + X_t \tag{2}$$

$$X_t = \phi_1 X_{t-1} + \phi_2 X_{t-2} + W_t$$

5.1 Estimation of Model Parameters

Estimates of the forecasting model parameters can be found in the appendix section in Table 3. It is reassuring that both autoregressive parameters ϕ_1 and ϕ_2 are statistically significant and interesting that they also have the same standard error.

5.2 Prediction

Figure 7 shows the forecasted acres burned for the next ten years, from 2010 to 2019, in red. The predictions are made after "undoing" the log VST in the log first difference with AR(2) model. The model predicts a very slight downward trend in acres burned in the upcoming 10 years, with one exception being a slight upward bump in the prediction for year 2013. Although the trend is downward, the amount of acres burned is still substantial and comparable with recent years on record, not to mention the theoretical prediction interval aroud the forecasts (not shown in plot). It is in the best interest of Bear County to maintain (or increase) the level of funding and resources allocated to firefighting. Also, the county should anticipate roughly similar levels of wildfire damage as recent years and have precautions ready (e.g. shutting off power lines).

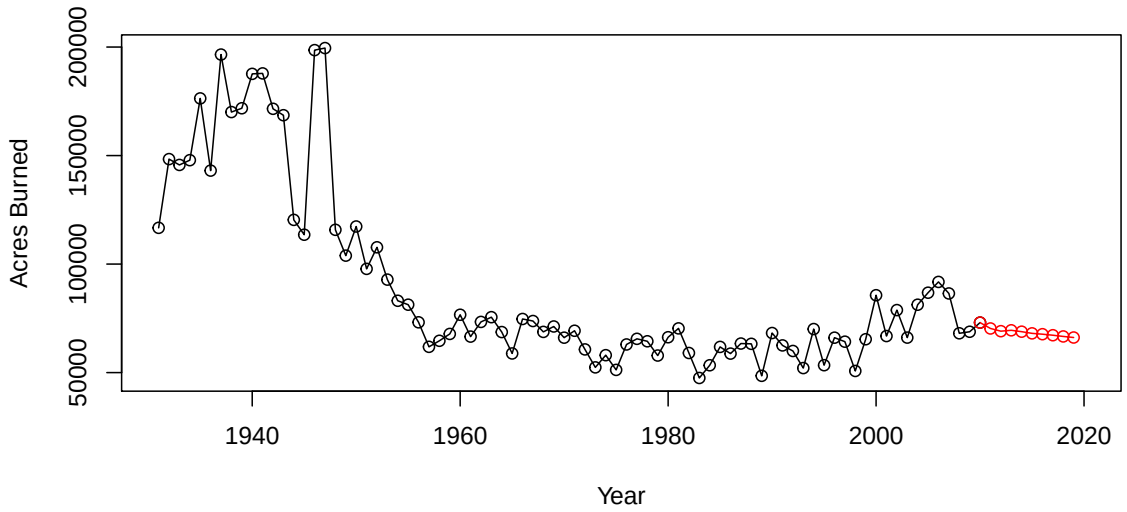

Figure 7: Model forecast, next 10 years, in red. The black points are historical acres burned data. Prediction intervals are not included but it is worth noting that uncertainty is greater for predictions further out in time.

6 Appendix

Table 3: Estimates of the forecasting model parameters in Equation (2), with their standard errors (SE) and descriptions.

Parameter	Estimate	SE	Coefficient Description
ϕ_1	-0.3647	0.1084	AR coefficient 1
ϕ_2	-0.3189	0.1084	AR coefficient 2
$E[\nabla log(Acres_t)]$	-0.0076	0.0108	Mean of Log Differences
σ_W^2	0.0255		Variance of White Noise