Stat 153 Final Project

Ishaan Srivastava

1 Executive Summary

In this report, we analyse time series data regarding forest fires in Bear County, California. Using past annual data, we estimate the number of acres projected to burn for the next ten years so that the Bear County Fire Department can allocate budget and resources accordingly. We choose a parametric signal model with an ARMA(0, 11) noise model and see that the amount of acres burned annually is projected to alternately increase and drop across the next 10 years, and thus Bear County must plan accordingly.

2 Exploratory Data Analysis

We begin by plotting the data below in Figure 1.

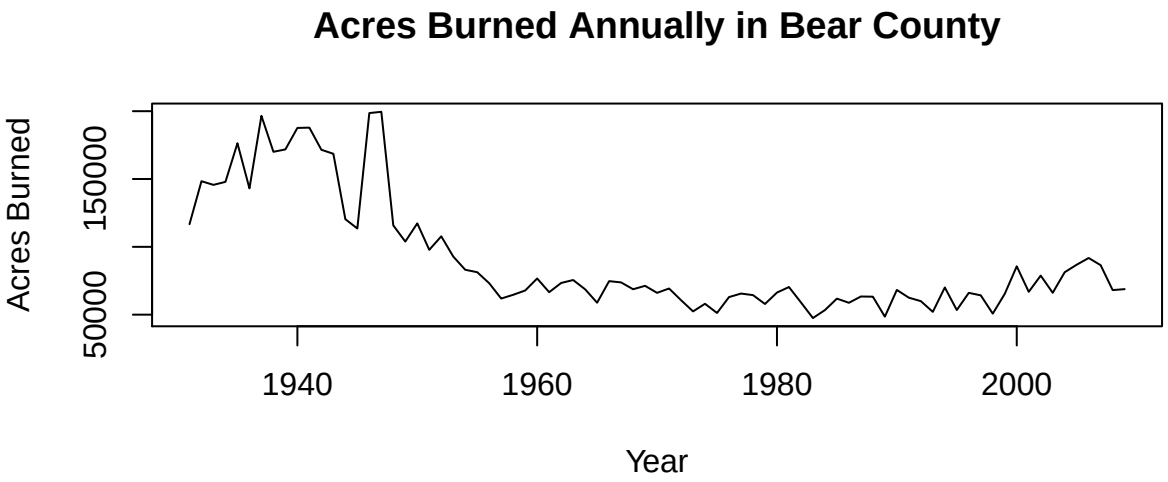


Figure 1: Acres burned annually in Bear County

Observing the plot, we see that the amount of land burned annually was consistently over 100,000 acres before a steep fall around the 1950s; since then the amount has been around 60,000-70,000 acres annually. Post the 1950s, the data has been broadly consistent in terms of variance ie. homoschedastic with some fluctuations especially in more recent years, while the trend has not followed any visually identifiable pattern. Note that since the data is recorded annually, there is no daily or monthly data that we may use to account for seasonality in terms of the wildfire season in a given year ie. intra-annual seasonality. The periodogram of the data (not shown in this report) shows no obvious seasonal frequencies across the years ie. inter-annual seasonality.

3 Models Considered

We construct two classes of models in order to model the signal in the data, namely non-parametric differencing models and parametric linear regression models. Each model is supplemented with two ARMA models to

account for the remaining noise, resulting in four final models included in this report.

3.1 Model 1: Differencing

We first employ differencing models. Before differencing, the data was transformed using a box-cox transformation with parameter λ such that the output most closely resembled a Gaussian AR process, with MLE of $\lambda = -0.4$. Hence the transformed data is $f(x) = \frac{x^{-0.4} - 1}{-0.4}$. Bearing in mind the El Niño Southern Oscillation (ENSO) that occurs every 2–7 years and can have a large impact on regular rainfall patterns and temperatures, I experimented differencing the data with a variety of lags, finally choosing lag 6. Beyond this, I also experimented with higher order differencing to find which models resulted in noise or residuals most closely resembling a stationary process.

With this heuristic in mind, the differencing model I finally chose was $\nabla_6 \nabla f(Fires)$, with $f(Fires)$ referring to the aforementioned Box-Cox transformation with $\lambda = -0.4$. The residuals are plotted below.

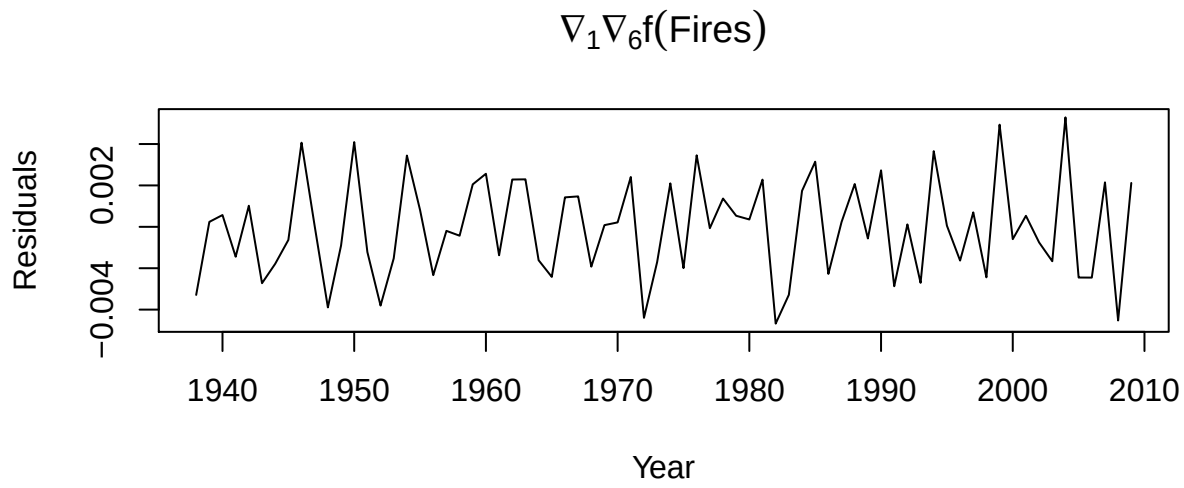


Figure 2: Residuals after differencing transformed fires data

There is admittedly some heteroschedasticity at the beginning and end of the differenced data, yet the residuals broadly resemble a stationary process. The differencing accounts for the peaks and troughs that seem to occur around 6 years, and also seek to eliminate the underlying trend.

In the plots of the autocorrelation function (ACF) and partial autocorrelation function (PACF), most values are within the blue bands corresponding to the 95% confidence interval constructed under the hypothesis of the stationary process being white noise for that lag.

3.1.1 ARMA(6, 0)

Based on the fact that the autocorrelations in the ACF plot are all within the blue bands after the 6th lag, I first modelled the residuals using an ARMA(6, 0) model with q = 6 and p = 0, otherwise known as an MA(6) model. This model treats the PACF values at lags 1 and 2 falling outside the blue bands as random chance, rather than an indication that there's some statistically significant autoregressive component in the residuals. We see that model ACF predictions track the track the true ACF values closely, with greater deviation from the true PACF values.

3.1.2 ARMA(1, 2)

For my second stationary process model, I used auto.arima with modified parameters, yielding an ARMA(1, 2) model. This model seems to resemble the true ACF and PACF values more closely than the ARMA(6, 0) model for the first few lags, but on the whole the ARMA(6, 0) tends to perform better for greater lags

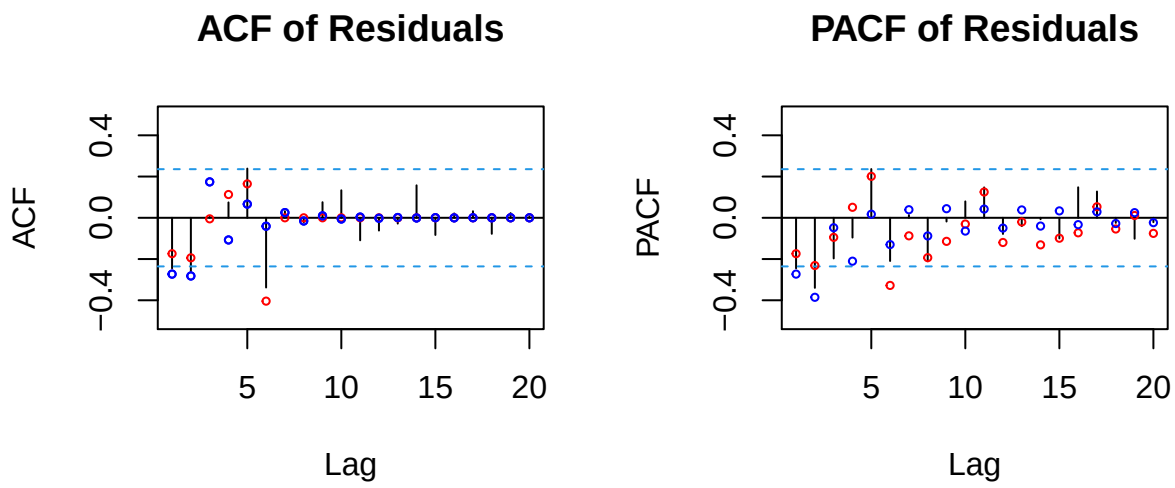


Figure 3: ACF and PACF of residuals. ACF and PACF values of the ARMA(6, 0) process are marked in red, while those of the ACF and PACF values of the ARMA(1, 2) process are marked in blue

and also in the cases where the true autocorrelation or partial autocorrelation is large. In both cases, we see satisfactory performance and thus move on to considering our next signal model.

3.2 Model 2: Parametric Modelling

When fitting parametric linear regression models to the data, I found the model resulting in non-stationary residuals due to discrepancies at either end of the data ie. the earliest data and the most recent data. As noted, the data before the 1950s seems to be substantially different from the rest of the data and since the primary goal is to forecast the data for the next 10 years, I exclude all data from before 1948, then produce a periodogram of the data to choose the corresponding Fourier frequencies for my model. Based on the periodogram, I chose the Fourier frequencies of 1/62 and 2/62, corresponding to the two peaks of the periodogram. The 62 comes from the number of observations used to construct the periodogram, and the 1 and 2 are the indices of the peaks. The final model is

$$\log(\text{y}) = \text{t} * sin(\frac{2\pi t}{62}) * cos(\frac{2\pi t}{62}) * sin(\frac{4\pi t}{62}) * cos(\frac{4\pi t}{62})$$

with * denoting interaction terms.

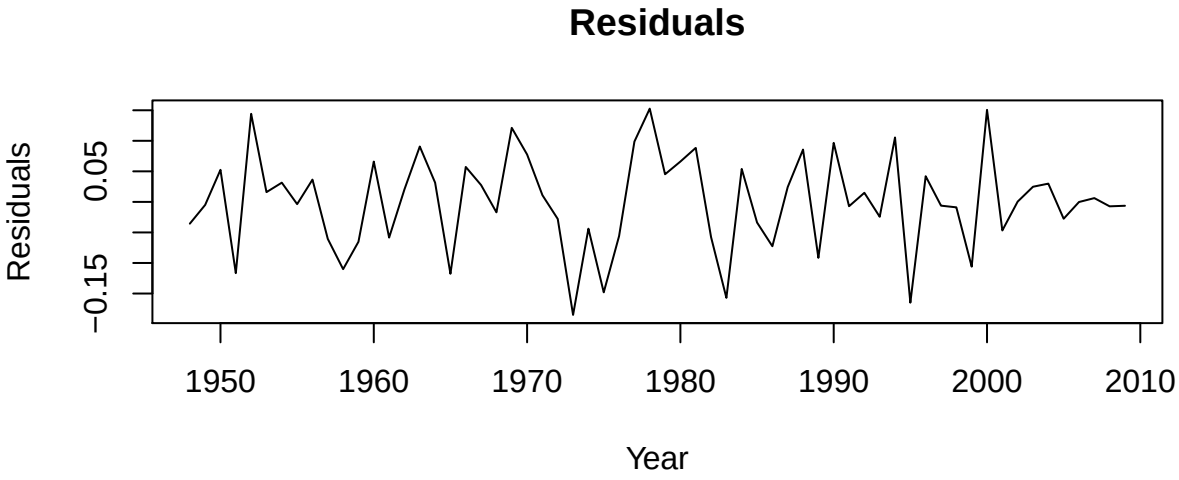


Figure 4: Residuals for parametric signal model

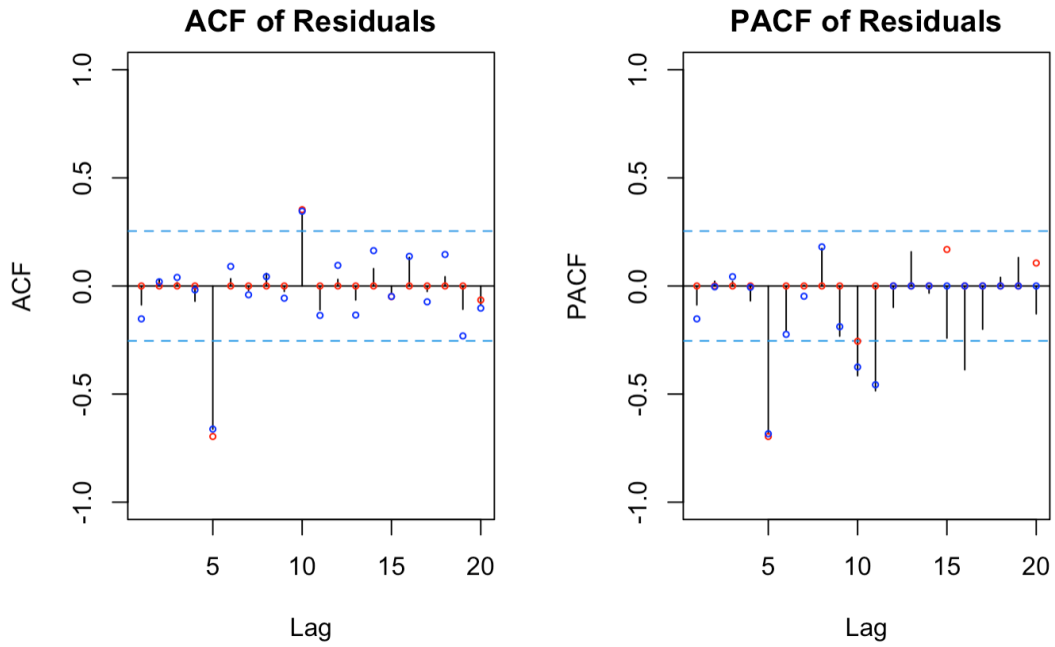


Figure 5: ACF and PACF of residuals. ACF and PACF values of the ARMA(2, 2)[5] process are marked in, while those of the ACF and PACF values of the ARMA(0, 11) process are marked in blue

The residuals, their autocorrelations, their partial autocorrelations, and corresponding ARMA models of the residuals are plotted above.

3.2.1 SARMA(2, 2)[5]

Based on the strong autocorrelations at lags 5 and 10 only in the ACF plot, I decided to use a seasonal model with S = 5 and Q = 2. Bearing in the mind the pattern of the statistically signifcant partial autocoorelations in the PACF plot, I also chose P = 2, finally ending up with an SARMA(2, 2)[5] model. Note that I experimented with multiplicative models and changing the values of P, Q, and S and found this to be the best at resembling the ACF and PACF of the data. Specifcally the model matches the ACF values almost closesly, but it's off the mark for certain PACF values, especially those at lags 11 and 16.

3.2.2 ARMA(1, 2)

As in the case of differencing, I used auto.arima with modified parameters for my second stationary process model, yielding an ARMA(0, 11) model, or simply an AR(11) model. As seen in the above figure, this model clearly outperforms the SARMA(2, 2)[5] model by having similar performnance on the ACF values and far better performance on the PACF values, although it also doesn't model the true PACF value at lag 16 accurately.

3.3 Model Selection

To select which model to use, I performed time series cross validation, with non overlapping test sets rolling through the past 10 years in the data in yearly segments. For the signal model, the entire data before the test set is used before differencing, but only data after 1948 is used for training the parametric model. Based on Root Mean Squared Prediction Error (RMSPE), the parametric signal model with the ARMA(0, 11) model was chosen. The RMSPE values themselves are available in the table above. We note that the parametric signal models have much lower PMSPE compared to the differencing models, which may raise concerns about

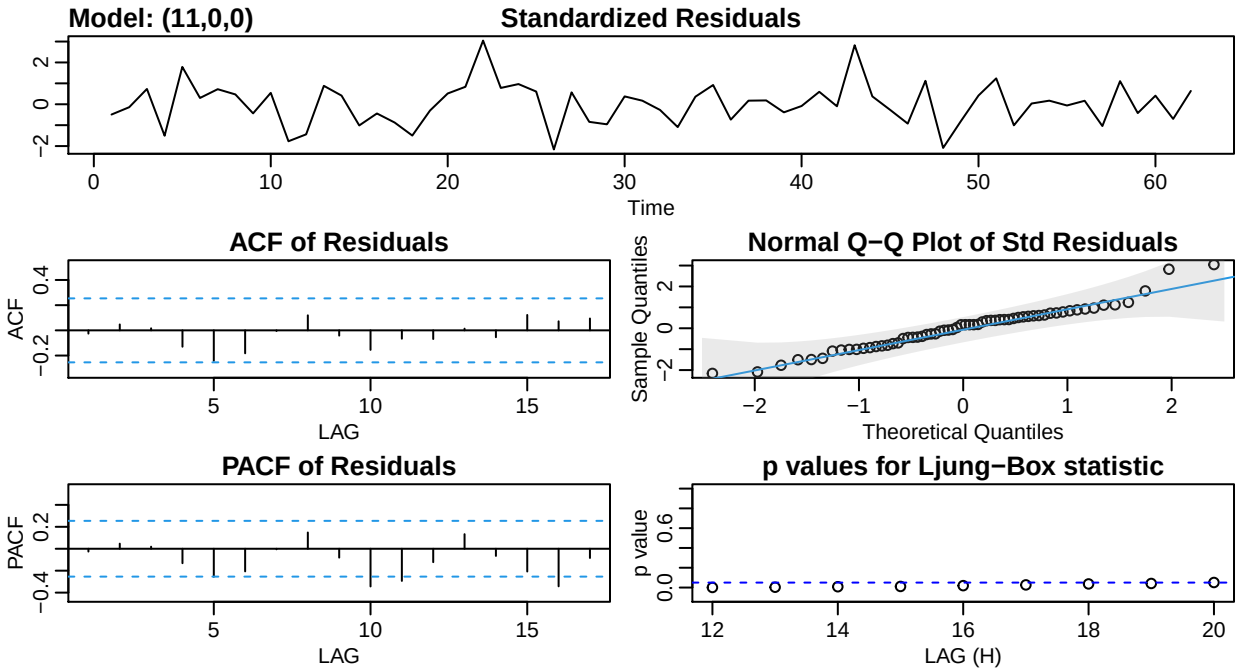
Table 1: Cross-validated out-of-sample root mean squared prediction error for the four models under consideration.

	RMSPE
Parametric Model + SARMA(2, 2)[5]	6877.935
Parametric Model + ARMA(0, 11)	6519.476
Differencing + MA(6, 0)	38126.514
Differencing + ARMA(1, 2)]	38126.510

overfitting. Yet given our use of cross-validation, this should have been safely mitigated.

3.4 Figures

The relevant diagnostic plots for our chosen ARMA model are provided below:



3.5 Results

The signal model is

$$\log(\text{y}) = \text{t} * sin(\frac{2\pi t}{62}) * cos(\frac{2\pi t}{62}) * sin(\frac{4\pi t}{62}) * cos(\frac{4\pi t}{62})$$

with * denoting interaction terms, y denoting the data and

$$X_t = W_t + \Sigma_{i=1}^{11} \phi_i X_{t-i}$$

where W_t is white noise and each ϕ is an estimated coefficient.

The noise model coefficients are given below:

	Estimate	SE
ar1	-0.5193	0.1130
ar2	-0.0701	0.1045
ar3	0.1984	0.1009
ar4	-0.2650	0.0943
ar5	-1.1500	0.1051
ar6	-0.7487	0.1650
ar7	-0.0707	0.1051
ar8	0.1925	0.0993
ar9	-0.2633	0.0953
ar10	-0.5340	0.1060
ar11	-0.4569	0.1207

There are too many parameters for the signal model to be summarised in a succinct table.

3.6 Predictions

My final forecasted values are given below.

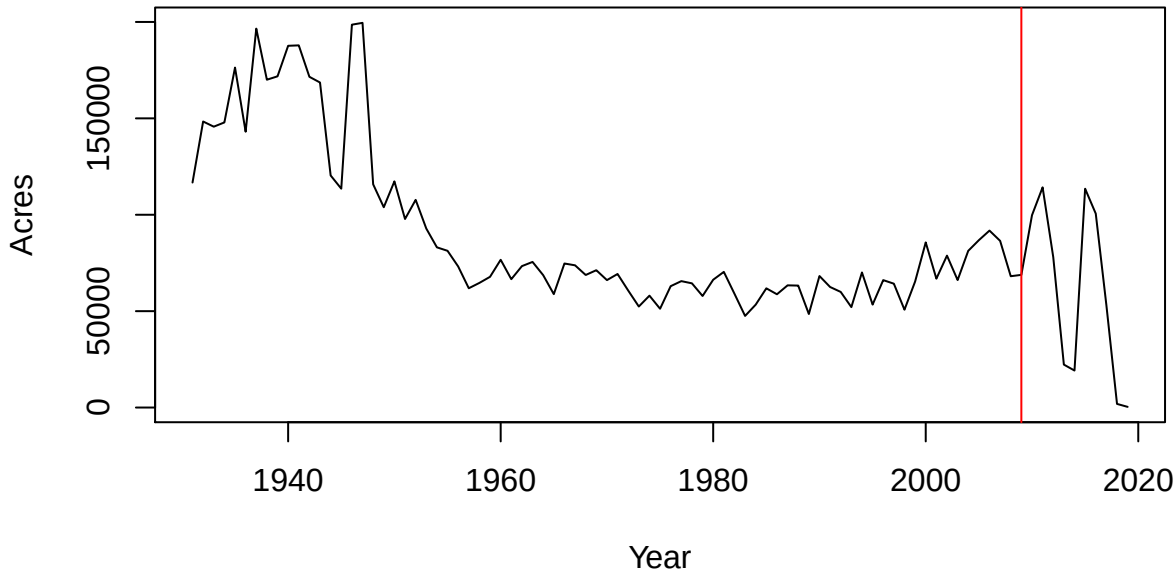


Figure 6: Number of acres bruned annually with projected acres after the red line

Note that my forecasts come after the red line and indicate an expected increase in forest fires in the coming years with alternating steep drops.