California Forest Fires: Project Report

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 resources and plan accordingly. We choose a second-order nonparametric differencing signal model with an ARMA(1, 2) noise model and forecast that the amount of acres burned annually is projected to increase by ~50% in the coming years, suggesting an increase in fires in terms of severity and/or sheer quantity.

2 Exploratory Data Analysis

We begin by plotting the data below in Figure 1.

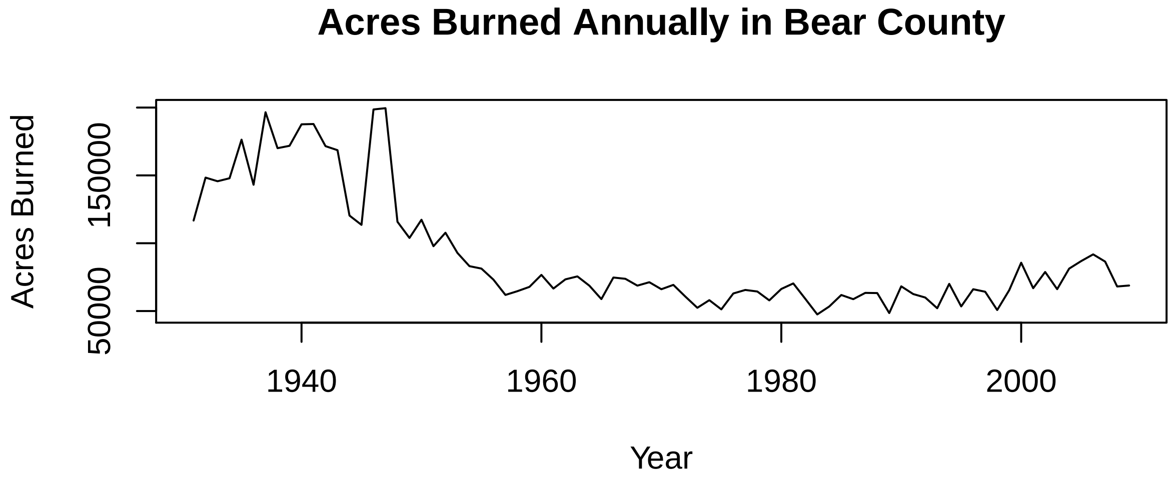

Figure 1: Acres burned annually in Bear County

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 have 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 are 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 for concision) 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 nonparametric 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 were 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 are $f(x) = \frac{x^{-0.4} - 1}{-0.4}$. Bearing in mind the El Niño Southern Oscillation (ENSO) that occurs every 2–7 years and can have a large impact on temperatures and regular rainfall patterns, I experimented differencing the data with a variety of lags, finally choosing a lag of 6 years. 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 $\nabla_1 \nabla_6 f(\text{Fires})$, with $f(\text{Fires})$ referring to the aforementioned Box-Cox transformed data with $\lambda = -0.4$. The residuals are plotted below.

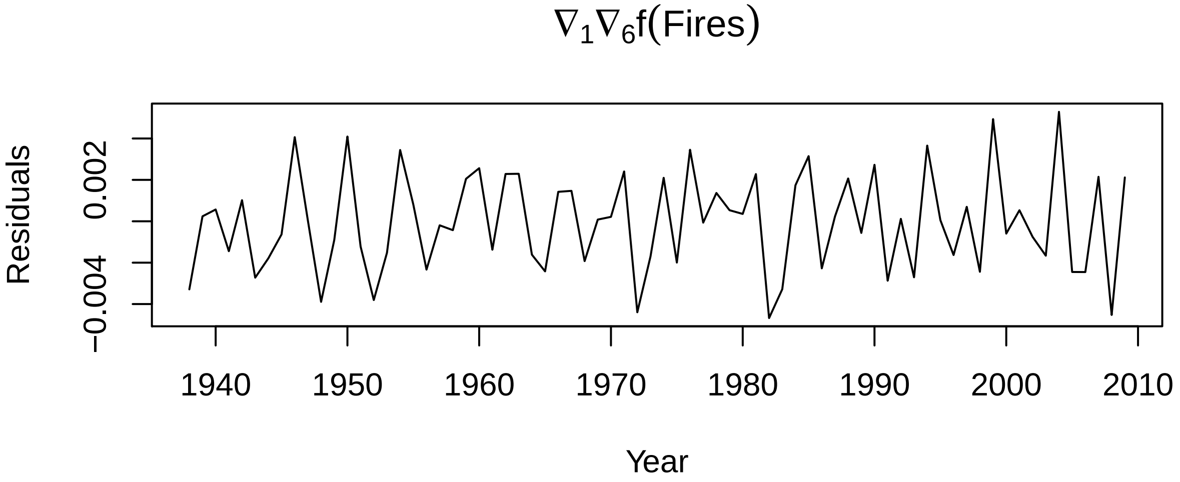

Figure 2: Residuals after differencing transformed fires data

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 every 6 years, and also serves to eliminate the underlying trend.

In the plots of the autocorrelation function (ACF) and partial autocorrelation function (PACF) in Figure 3, 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(0, 6)

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(0, 6) model with q = 6 and p = 0, otherwise known as a 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 true ACF values closely, with greater deviation in the case of model PACF predictions and values.

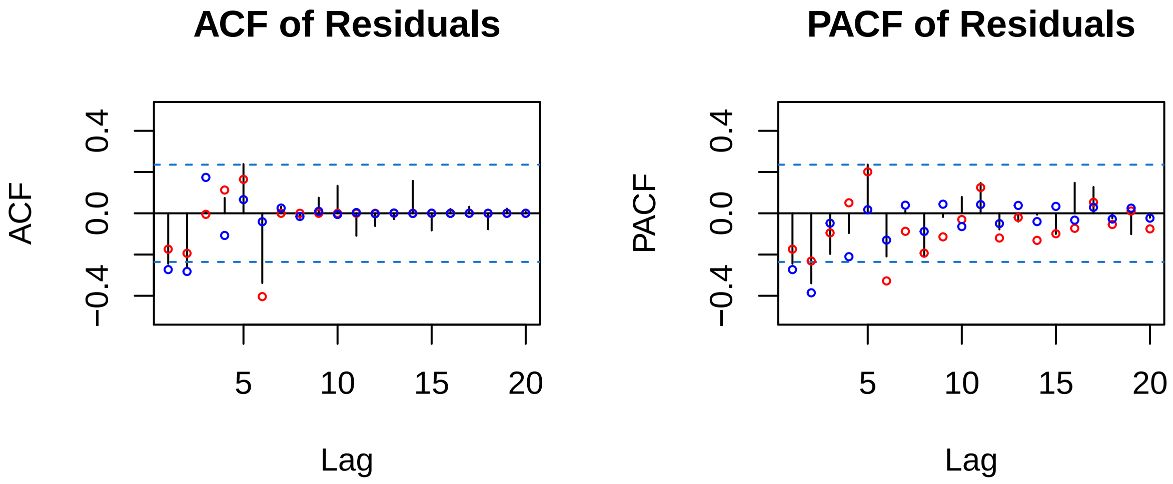

Figure 3: ACF and PACF of residuals. ACF and PACF values of the ARMA(0, 6) process are marked in red, while those of the ACF and PACF values of the ARMA(1, 2) process are marked in blue

3.1.2 ARMA(1, 2)

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

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

where y is the number of acres burned annually, t is the year, and * denotes interaction terms. Note that we take the log of the data as a variance stablising transform, hence we model log(y) instead of y directly.

The residuals are plotted below in Figure 4. Their autocorrelations and partial autocorrelations, and those of the corresponding ARMA models are plotted below in Figure 5.

3.2.1 SARMA(2, 2)[5]

Based on the strong autocorrelations at only lags 5 and 10 in the ACF plot (see Figure 5), I decided to use a seasonal model with S = 5 and Q = 2. Bearing in mind the pattern of the statistically signifcant partial autocorelations in the PACF plot, I also chose P = 2, finally ending up with a SARMA(2, 2)[5] model. Note that I experimented both with multiplicative models and changing the values of P, Q, and S, and found this to be the most effective at resembling the ACF and PACF of the data. Specifcally the model matches the ACF values closely, but is inaccurate for certain PACF values, especially those at lags 11 and 16.

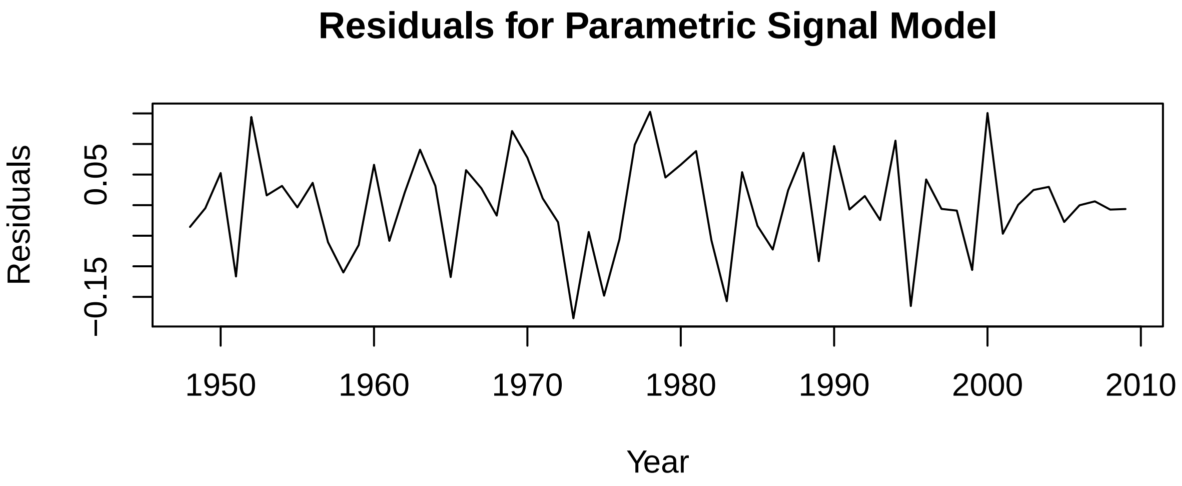

Figure 4: Residuals for parametric signal model log(y) = t $* sin(\frac{2\pi t}{62}) * cos(\frac{2\pi t}{62}) * sin(\frac{4\pi t}{62}) * cos(\frac{4\pi t}{62})$

3.2.2 ARMA(11, 0)

As in the case of differencing, I used auto.arima with modified parameters for my second noise model, yielding an ARMA(11, 0) model, or simply an AR(11) model. As seen in figure 5 below, this model clearly outperforms the SARMA(2, 2)[5] model by having similar performance 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.

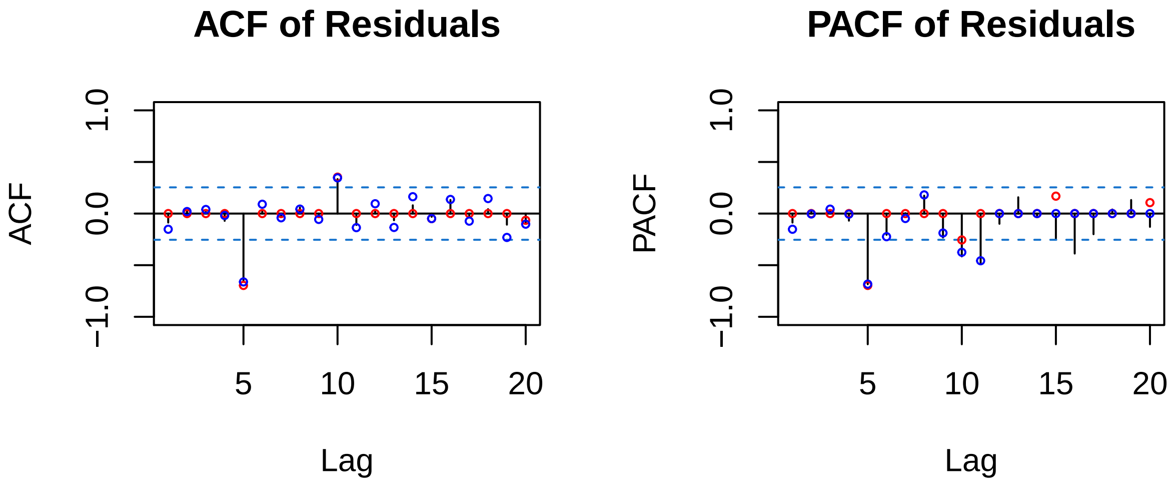

Figure 5: ACF and PACF of residuals. ACF and PACF values of the SARMA(2, 2)[5] process are marked in red, while those of the ARMA(11, 0) process are marked in blue

3.3 Model Selection

To select which model to use, I performed time series cross-validation, with validation sets rolling through the past 10 years in the data in yearly segments. The Root Mean Squared Prediction Error (RMSPE) values for each model are available in Table 1 below. We first note the large difference in the RMSPE values between the linear regression signal models and the second-order differencing signal models, which is likely casued by overfitting in the case of the linear regression signal model. We leave the exploration of other suitable parametric signal models for future reserch. Based on RMSPE values, we could choose either differencing model. Ultimately we choose the differencing model with the ARMA(1, 2) noise modelling process based on the agreement between the predicted and observed autocorelation and partial autocorrelations in Figure 3.

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

	RMSPE
Parametric Model + SARMA(2, 2)[5]	166686.59
Parametric Model + ARMA(11, 0)	164769.58
Differencing + ARMA(0, 6)	38126.51
Differencing + ARMA(1, 2)	38126.51

3.4 Diagnostic Plots

The relevant diagnostic plots for our chosen model are provided below.

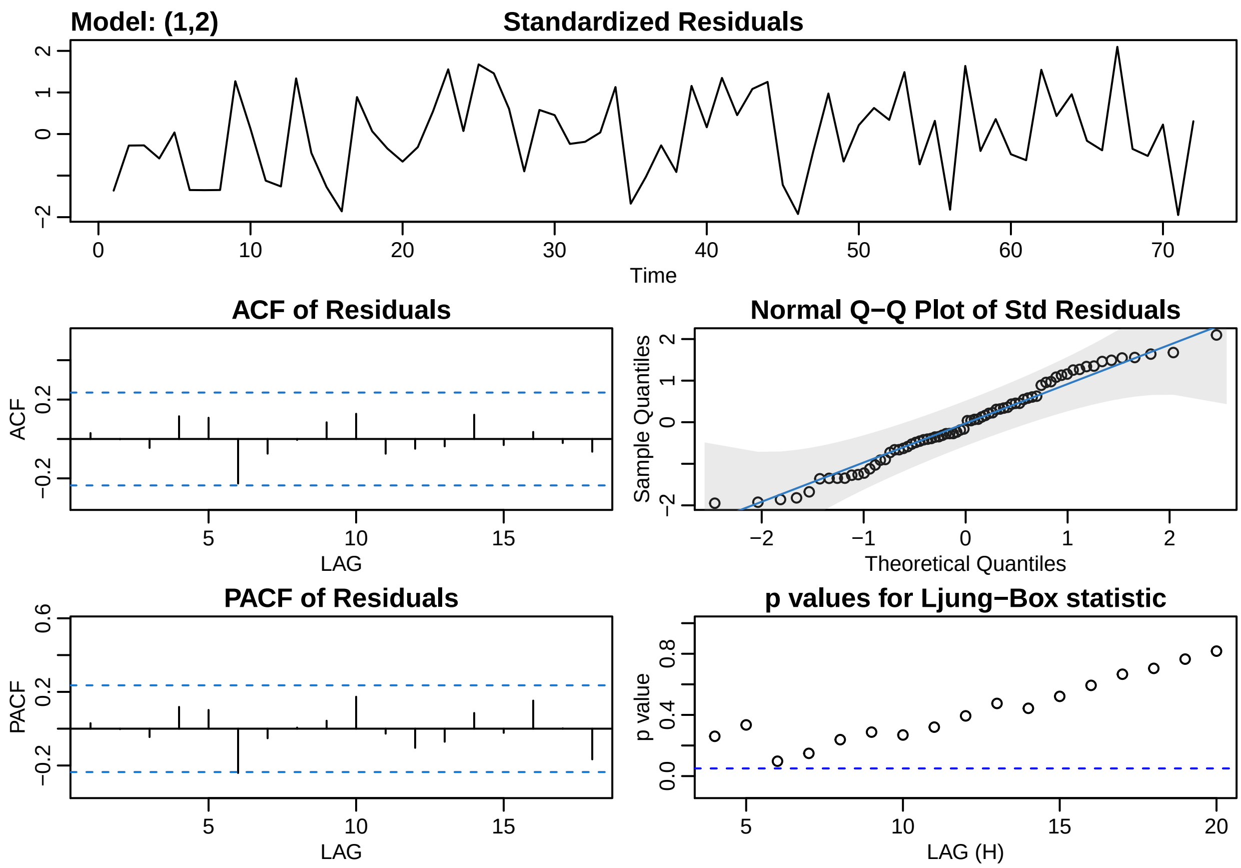

Figure 6: Diagnostic plots for final differencing + ARMA(1, 2) model

While the time series cross-validation was helpful in choosing the best model of the four being considered, the diagnostic plots in Figure 6 help us determine if the model actually appears to be a good fit for the data or not. Given that the residuals broadly appear homoschedastic, all the ACF values of the residuals are within the 95% confidence interval, all the standardised residuals are within the 95% confidence interval in the Normal Q-Q plot, all the p-values for the Ljung-Box statistics are greater than 0.05, and all but one of the PACF values of the residuals are within the 95% confidence interval, we conclude that the diagnostic plots do not indicate any major problems with our model while bearing in mind that this does not guarantee accuracy.

3.5 Results

The signal model is $\nabla_1 \nabla_6 f(\text{Fires})$ with $f(x) = \dfrac{x^{-0.4} - 1}{-0.4}$ and the corresponding noise model is $X_t - \phi X_{t-1} = W_t + \theta_1 W_{t-1} + \theta_2 W_{t-2}$, where W_t is white noise and $\phi_1, \theta_1, \theta_2$ are all coefficients to be estimated. The noise model coefficients are given below.

Table 2: Estimated coefficients for ARMA(1, 2) noise model

	Estimate
AR1	-0.6171
MA1	0.2486
MA2	-0.6295

3.6 Predictions

We plot the original data, along with the final forecasted values after the red line, in Figure 7 below.

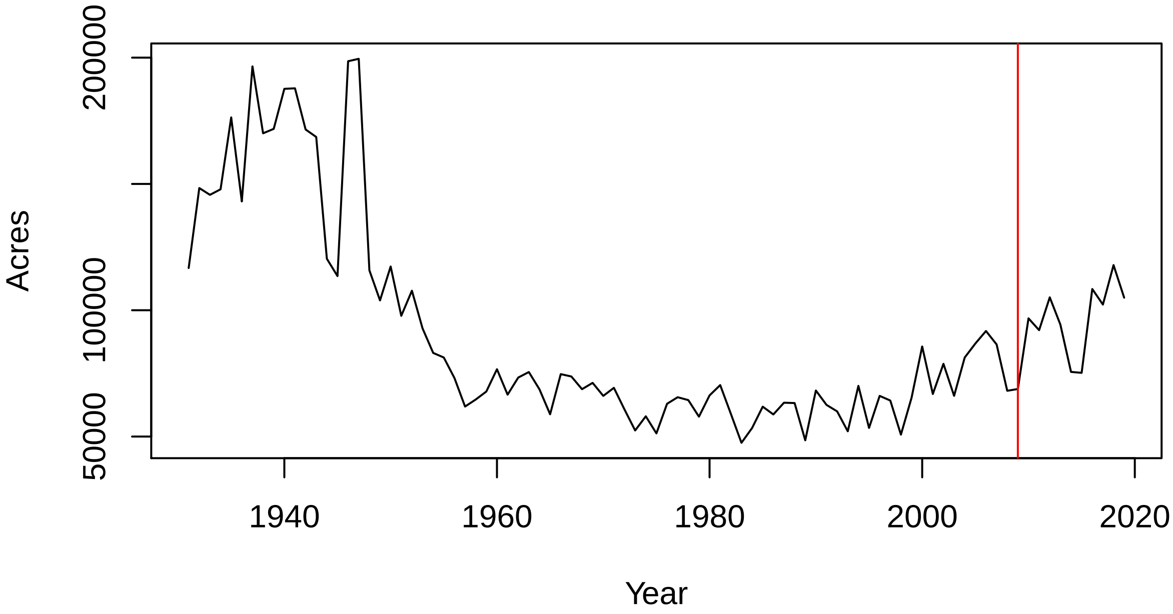

Figure 7: Number of acres burned annually with projected acres after the red line

Based on the forecasted values, we expect there to be a jump in the number of acres burned annually to ~95,000 for four years, followed by a lull where the values drop to ~75,000, slightly above the baseline value of ~68,000 in 2009 for two years, followed by another steep increase to ~105,000 acres burned annually for four years. In terms of actionable insight, this means the Fire Department should expand its capacity in terms of equipment and personnel in accordance with an expected ~50% increase in the number of acres burned in the coming years, which we expect to be correlated with an increase in fires in terms of severity and/or sheer quantity.