Chris Schaffer (Group 17)
DSC 425 Project
12/01/19

Forecasting Forest Fires

ABSTRACT

A forecast model was developed to help anticipate monthly forest fire counts in the Amazon Rainforest (as reported by the government of Brazil). The linear model's consistency helps characterize the severity of specific monthly totals by comparing them to a 'typical' number for that time of year. Analysis reveals that monthly fire counts are more common and varied in the Fall and occur most frequently in the state of Mato Grosso. These results, as well as some exploratory analysis, reveal a highly seasonal and stationary series. The linear model is able to adequately predict many months' worth of forest fires, but loses accuracy in the fall months.

Contents

INTRODUCTION

Interest in environmental monitoring and sciences has risen in the past decades as humans gain a better understanding of how we impact the planet and its ecosystems. In recent months, a large amount of international attention has been directed towards slash-and-burn activities taking place in the Amazon Rainforest. This biome alone accounts for over half of Earth's rainforests and boasts an unusually high amount of biodiversity [1]. The World Wildlife Foundation has commented on the recent Amazonian deforestation:

> *"These fires are destroying ecosystems, displacing wildlife, and jeopardizing the livelihoods of millions. Conserving the Amazon, and other areas like it, is essential to conserving our planet. As one of the world's most iconic forests burns, it's absolutely critical to consider how we are using this valuable resource and work to prevent the kind of disaster we are seeing today. That means deliberate conservation strategies that end deforestation and mitigate and adapt to climate change."*
>
> *– Kerry Cesareo, WWF Senior Vice President of Forests, August 2019*

While the Amazon covers much of the South American continent, a majority is within the borders of Brazil, whose government maintains records of Amazonian forest fires. This is the original source of the data used in this analysis. A summary of literature relating to forecasting fire outbreaks and predicting risk is presented in the following section.

LITERATURE REVIEW

The collection and compilation of data related to fires is not new. This is discussed in great detail by Field et al. in their paper *Development of a Global Fire Weather Database*. The researchers compiled information from dozens of international climate research agencies to create a detailed, comprehensive dataset for the entire world between the years of 1980 and 2012. Being a global study, South America, and Brazil specifically, are noted for having a particularly large number of fires [2].

A publication by the National Interagency Fire Center (NIFC) discusses how several environmental factors create favorable and even predictable conditions for the proliferation of wildfires. They concluded that humidity and temperature instability in the lower atmosphere are often precursors to natural wildfires [3]. While this data was not readily available for this analysis, historic weather data over the past two decades could be compiled and considered for higher-resolution modeling.

Forecasting techniques have been applied to environmental and ecological subjects before, including forest fire predictions. Diez et al. used meteorological conditions and historic information to forecast the number of fires occurring on a daily basis. Their model was developed specifically for a region in northern Spain and relied on information unavailable in this analysis (humidity, pressure, etc.). The researchers showed their model produces a good linear fit to actual data given certain weather conditions [4].

In a similar spirit, Mozny and Bares produced a classification model which defines a fire danger index. Their work was inspired by the danger forest fires present to open countryside in the Czech Republic. The model is maintained by the Czech Hydrometeorological Institute and uses weather conditions to summarize daily fire risks across the country. This is achieved by collecting weather information from localities around the Czech Republic. While their model is not meant for time series analysis, a superposition of the historic number of fires and their classifications demonstrates a strong relationship between their metric and the number of fires [5].

In 2015, Guatteri et al. were awarded a patent for their location- and weather-based approach to forecasting disasters. The method involves collecting real-time, remote sensor data and comparing this to typical weather conditions for a given region [6]. It's not known whether this system has been implemented in practice. Given the conclusions of the previously-cited publications, it's clear that an active monitoring system like this would be highly complementary to conservation efforts by indicating hazardous conditions before fires break out.

DATASET

The dataset used for developing a regression model spans the 20 years between 1998 and 2018. The data was originally compiled by the Brazilian government and is publicly available at the following address:

http://dados.gov.br/dataset/sistema-nacional-de-informacoes-florestais-snif

The dataset used for this analysis was downloaded from Kaggle in October 2019. The URL is provided below.

https://www.kaggle.com/gustavomodelli/forest-fires-in-brazil

The set contains four distinct features- *Year*, *Month*, *State*, and the target feature *Total* (number of fires for the month). Monthly totals were collected for the twenty-three Brazilian states that contain the Amazon Rainforest. A linear model, discussed later, makes use of the trend and seasonal components of the target feature.

Preprocessing

Some preprocessing was needed to address formatting issues in the records. These steps consisted of translating months to their English equivalents and addressing the decimal point formatting. Like many other countries, Brazil uses the decimal and comma in a manner opposite to the standard in the US. This subtle difference was not initially caught, and this made for an incorrect interpretation and analysis of the dataset. This formatting issue has since been rectified and the analysis in this paper reflects that.

The first five months were not included in this analysis. The reported number of fires nationwide for these months was entered as 0. This is probably indicative of a reporting failure or lack of information. Thus, the first entry of the series is for June 1998. Data was also not included for December of 2017, so the series ends at November 2017.

Table I – Description of Processed Dataset Features

	Year	Month	State	Total
Type	string	string	string	float
Unique Values	234	12	23	1,479

The time series was split into two partitions- one set on which the regression model was trained and another for testing. Forecasts are plotted alongside actuals at the end of this report for direct comparison. The portion reserved for testing covers the time between January 2013 and November 2017.

METHODOLOGY

The target feature is examined in two different ways before modeling- both as a distribution of values and as a time series. Descriptive statistics help describe the distribution of values and identify notable or extreme values. When treated as a time series, tests for autocorrelation and stationarity are applied to determine model type, parameters, and whether differencing is required.

RStudio, Spyder (a Python IDE), Tableau, and Excel were used to support preprocessing, exploration, and analyses.

Descriptive Statistics

The target feature for this regression analysis is the total monthly fire count. Table II provides statistics about the feature and the corresponding boxplot illustrates the distribution of values. The dataset clearly contains some extreme values. Under different circumstances, these values would either be removed or receive special treatment, but these were retained to help illustrate their severity when compared to model forecasts. Overall, these outlying values had a negligible impact on model development.

Table II – Target Feature Statistics

N	Min	Q1	Median	Mean	Q3	Max	Std. Deviation
234	415	2,156	6,238	15,594	22,408	110,988	20,069

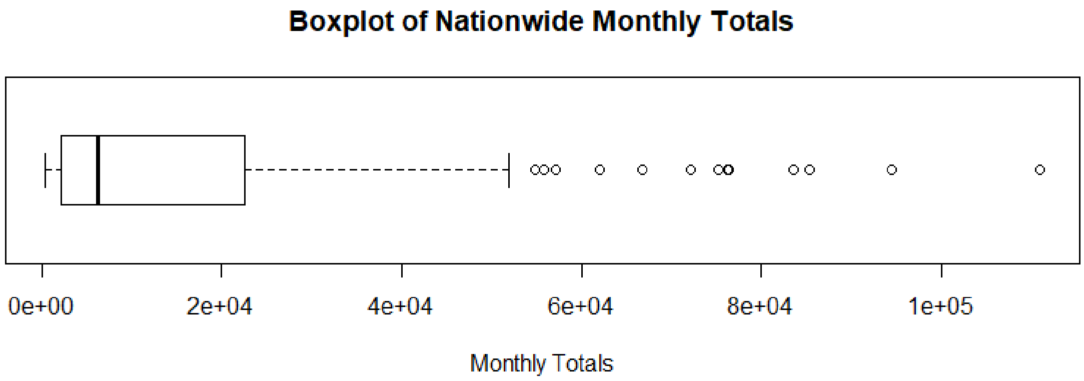


Figure 1 – Boxplot showing the distribution of the 20-year monthly fire totals.

Time Series Analysis

The following plot shows the national monthly fire count between June 1998 and November 2017. There appears to be a cyclical pattern in fires occurring on an annual basis. A visual inspection of the correlograms (Figure 2) corroborates the idea of strong seasonality in the time series.

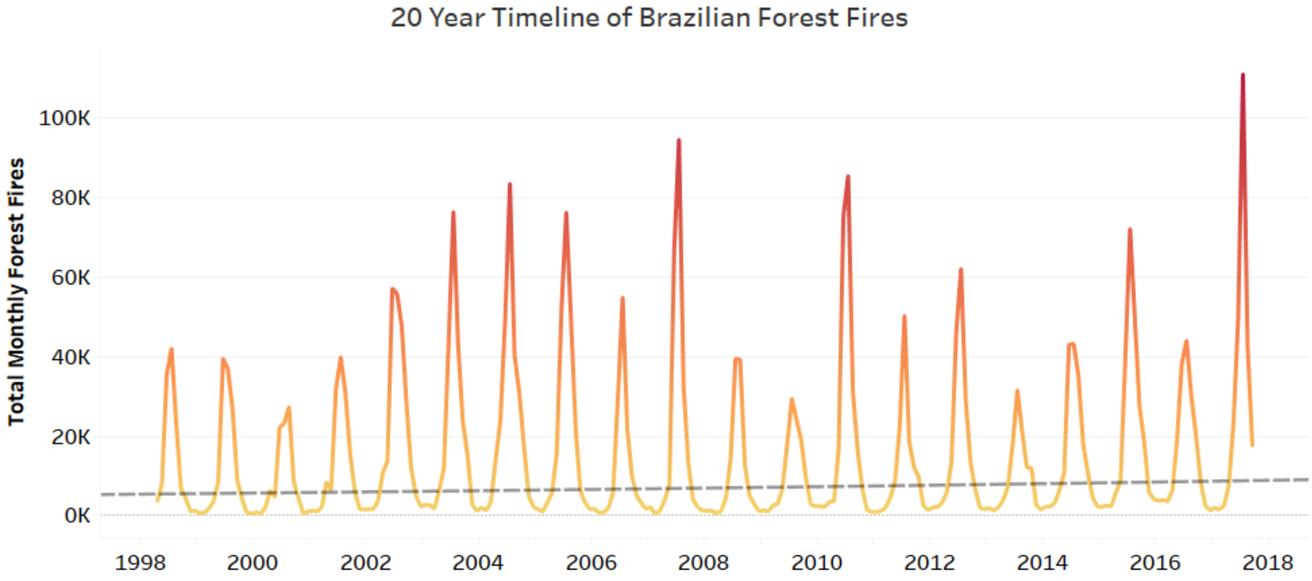


Figure 2 – Timeline of all fires, aggregated by month. The exponential trend line shows a very slight increase in the number of forest fires over time.

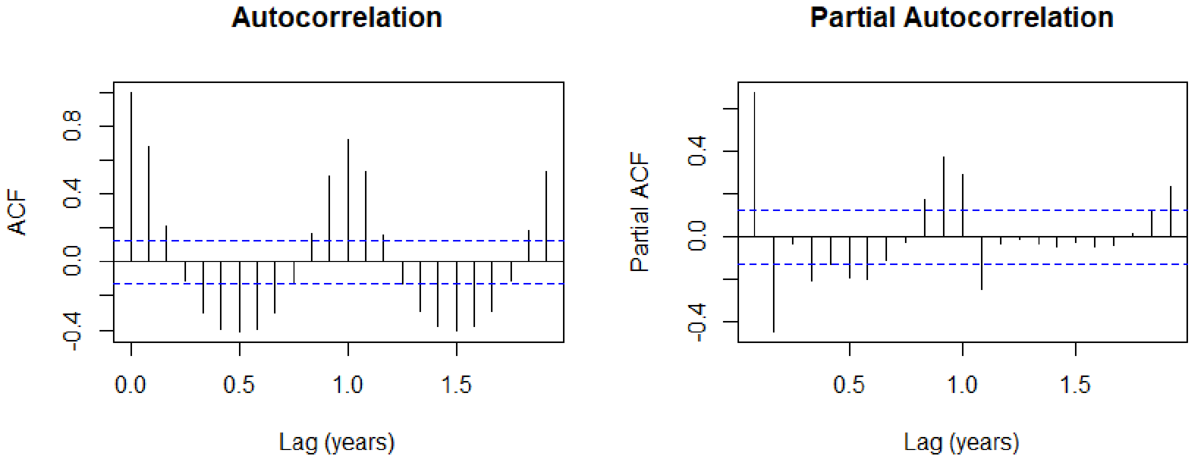


Figure 3 – Autocorrelation plots of the original time series.

The p-value of the Jarque-Bera test (<2.2e-16) indicates that the distribution of fires deviates from Normal by a notable amount. This is confirmed with a visual inspection of the boxplot. The Ljung-Box test p-value is incredibly significant (<2.2e-16), confirming that serial correlation exists in the signal. The ACF and PACF plots also show a great number of significant autocorrelations between the present value and the past lags.

Stationarity is important for a comprehensive understanding of the time series. The KPSS and Augmented Dickey-Fuller tests produce results characteristic of a stationary series. This implies that the series is suitable for modeling and forecasting.

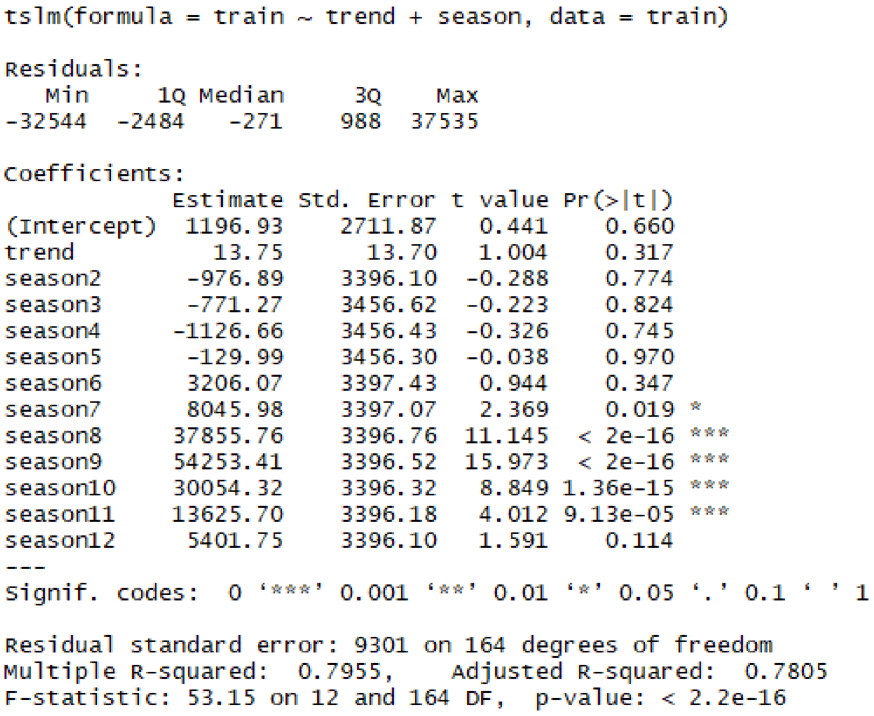
Table III – Target Series Test Statistics

	Jarque-Bera	Box-Ljung	KPSS	Aug. Dickey-Fuller
Statistic	290.01	108.06	0.0629	-9.8818
p-value	< 2.2e-16	< 2.2e-16	> 0.10	< 0.01

Regression Model

The signal is highly seasonal, and the fire counts only have a few extreme values. To best recreate this repeating series, a seasonal linear time series model was developed. The higher-order AR, MA, and ARIMA models needed to reproduce the signal had erratic predictions. Analysis also revealed the linear model to have the least-autocorrelated training residuals.

Figure 4 shows a printout summary of the TSLM model. As expected, the autumn months prove to be the most significant features in the model. The trend has little significance and is again no surprise given the behavior of the time series.



```
tslm(formula = train ~ trend + season, data = train)

Residuals:
   Min     1Q Median     3Q    Max
-32544  -2484   -271    988  37535

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  1196.93    2711.87   0.441    0.660
trend          13.75      13.70   1.004    0.317
season2      -976.89    3396.10  -0.288    0.774
season3      -771.27    3456.62  -0.223    0.824
season4     -1126.66    3456.43  -0.326    0.745
season5      -129.99    3456.30  -0.038    0.970
season6      3206.07    3397.43   0.944    0.347
season7      8045.98    3397.07   2.369    0.019 *
season8     37855.76    3396.76  11.145  < 2e-16 ***
season9     54253.41    3396.52  15.973  < 2e-16 ***
season10    30054.32    3396.32   8.849 1.36e-15 ***
season11    13625.70    3396.18   4.012 9.13e-05 ***
season12     5401.75    3396.10   1.591    0.114
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 9301 on 164 degrees of freedom
Multiple R-squared:  0.7955,    Adjusted R-squared:  0.7805
F-statistic: 53.15 on 12 and 164 DF,  p-value: < 2.2e-16
```

Figure 4 – Summary printout of the R TSLM function.

Model Residuals

A thorough understanding of the residuals is important for building faith in a model’s performance. According to the distribution, the model tends to slightly underestimate the number of fires on average. The following figures show a timeline and other information regarding the training residuals. The training set produces a series with a mean absolute percent error (MAPE) of 49%. A repeating, yet alternating signal is seen in the errors which corresponds to the large variations seen at annual extremes.

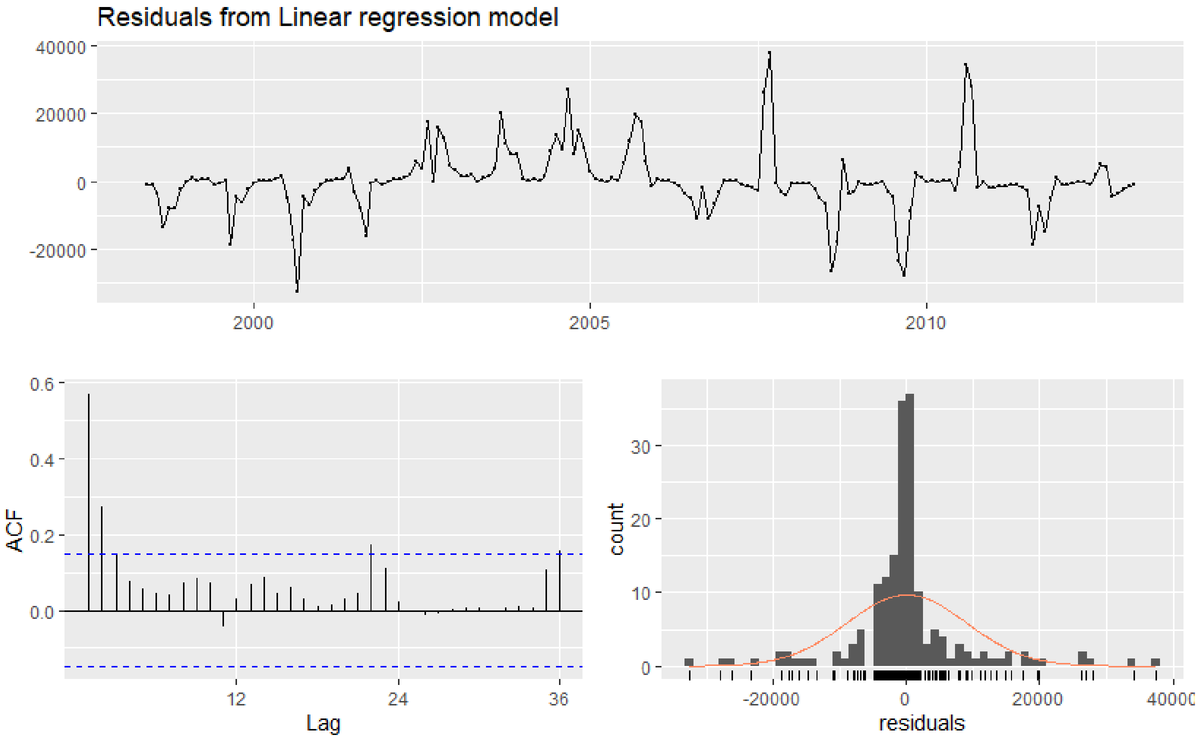


Figure 5 – Analysis of training residuals. The line graph (top) shows residual values over the training dataset's timeline. An ACF plot (lower left) is included for inspection of autocorrelations. The lower right shows a histogram of residual values.

Table IV – Residual Statistics

N	Min	Q1	Median	Mean	Q3	Max	Std. Deviation	RMSE	MAE	MAPE
177	-32,544	-2,484	-271	0	988	37,535	8,978	8,952	5,093	49.13%

Table V – Residual Series Test Statistics

	Jarque-Bera	Box-Ljung	KPSS	Aug. Dickey-Fuller
Statistic	177.86	58.799	0.21215	-4.4106
p-value	< 2.2e-16	1.743e-14	> 0.1	< 0.01

RESULTS

Forecasting

Approximately five years (January 2013 – November 2017) were set aside for testing. The following plots show the actual data (red) and predicted forecasts (blue). The shaded regions around the forecast represents the 1- and 2-sigma confidence intervals.

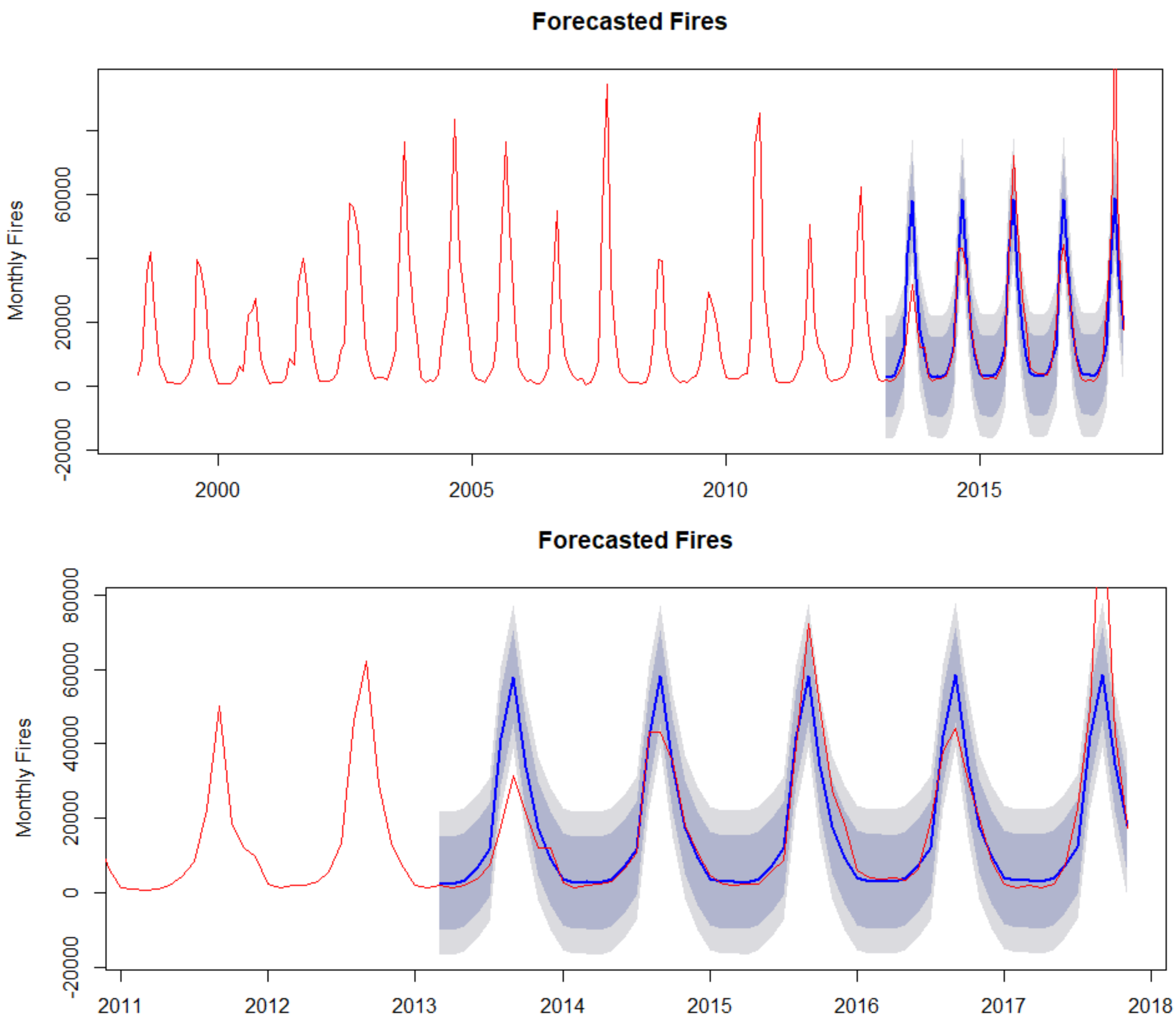


Figure 6 – Five-year forecasts (blue) of monthly fire totals. The actual data (red) is plotted alongside predictions for reference.

Forecast Residuals

As with the model residuals, an analysis of the forecast residuals is also included. Table VI and the corresponding boxplot describe the distribution of errors while Table VII provides statistics from the time series tests. Correlograms provide information about the residuals' autocorrelations.

Table VI – Forecast Residual Statistics

N	Min	Q1	Median	Mean	Q3	Max	Std. Deviation	RMSE	MAPE	MAE
59	-53,617	-19,953	462	4,032	31,707	57,584	29,567	29,583	1,275%	23,594

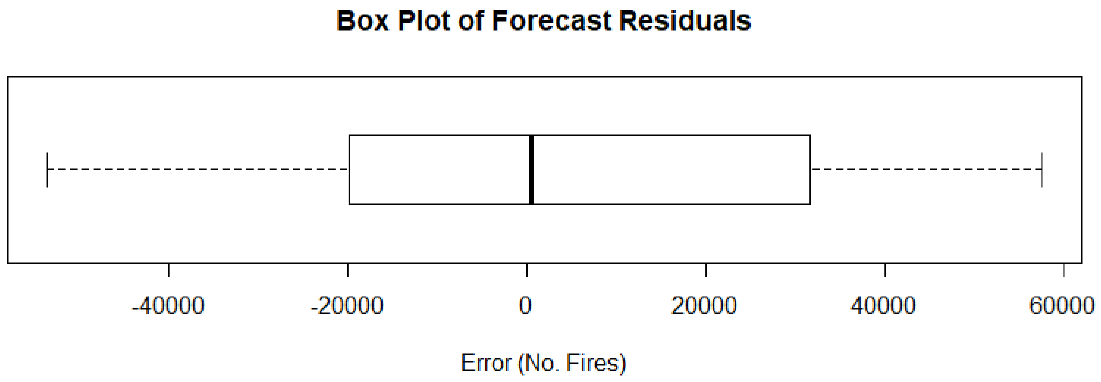


Figure 7 – Boxplot showing distribution of residuals.

Table VII – Forecast Residual Series Test Statistics

	Jarque-Bera	Box-Ljung	KPSS	Aug. Dickey-Fuller
Statistic	1.3589	36.518	0.0318	-5.8615
p-value	0.5069	1.513e-9	> 0.10	< 0.01

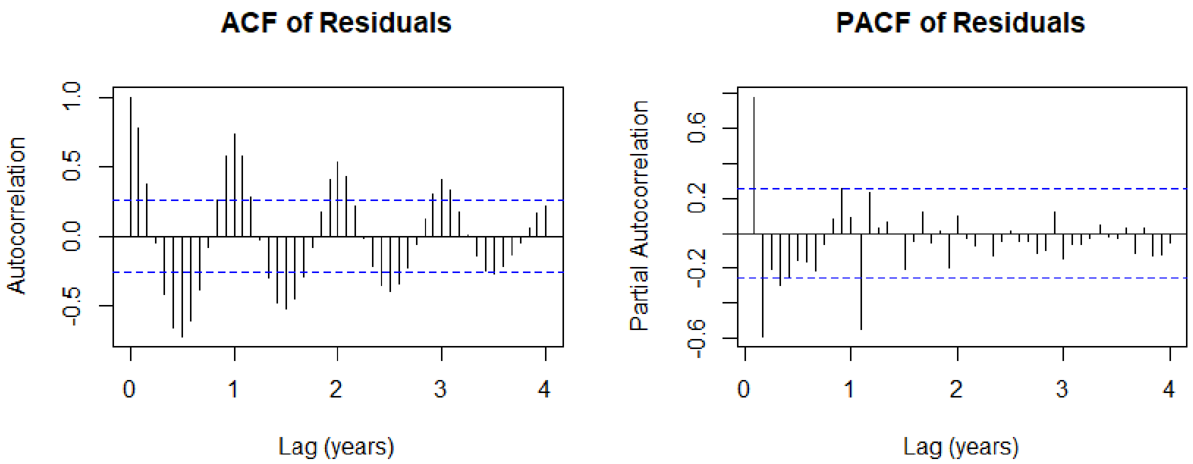


Figure 8 – Autocorrelation plots of the forecast residuals.

DISCUSSION

The forecasts show that the linear model produces very consistent periodic results. It is, however, unable to predict the unusually high number of fires in September 2017. This is not surprising since that occurrence is the single most extreme outlier in the series. In addition, the largest errors tend to occur at each year's maximum value. The factors underlying the extreme variance of these months are not entirely understood by considering this dataset alone. Drastic differences in these peaks might be caused by varying weather conditions or even public policy.

The forecast residuals have larger error statistics than the training set, but this is probably skewed somewhat by the extreme number of fires that occurred in 2017. Further analysis would be required to determine if the model is overfit to the training data or if the unusual values in the test set are mostly responsible for high errors.

The time series statistics and correlograms imply that the model is missing some useful information in the signal. The original PACF plot hinted that certain very large lag values had a correlation to the present value. These were initially thought of as being seasonal elements, but the addition of certain long lag values to the regression model may improve performance.

Since the TSLM function includes a band of the most likely values, it can be used as a benchmark for referencing just how severe a given month's total fire count is while still accounting for historic patterns. This can help researchers, activists, and public officials better characterize just how extreme a given occurrence is relative to an otherwise normal amount. Clearly more information would be needed to produce a more accurate model which captures the annual variations in autumnal fire counts.

CONCLUSION

A strong seasonal component in the time series makes for a relatively predictable series. The long-term trend is considered, but contributes very little to the overall pattern in monthly fire totals. This may suggest that while 2017 had an unusually high number of forest fires, it is not necessarily indicative of a permanent, drastic shift in the number of fires. More concerning than one particular event, however, is that the time series has a slightly better fit to an exponential trend model than a linear one. A non-linear growth in the number of fires would easily stand out against the repeating linear forecasts, suggesting that corrective policies be enacted.

This model could be suitable for using as a benchmark, but exhibits significant lack-of-fit as indicated by the Ljung-Box test. This makes accurate forecasting based solely off historic data very difficult. Data from 2018 and 2019 was not included in this set, but those 24 months would probably help verify whether the increase in 2017's fires was temporary or not. A new, persistent trend in forest fires would also require the development of an entirely new model and perhaps a differencing approach.

FUTURE WORK

This analysis did not consider the cause of forest fires, only their incidence. A much more accurate forecast model can probably be developed with the inclusion of economic and atmospheric data. Further data exploration may reveal that shifts in economic behavior translate to changes deforestation from farming, logging, and construction. Atmospheric and weather data are also highly recommended for future analyses. Several of the cited references suggests that changes in climate and weather greatly impact the potential for forest fires.

With regards to regression modeling, it may be possible to create an aggregate model composed of region or state-specific forecast models. This would require developing analytical models for each state, but the benefit may be a more accurate national monthly forecast. The multi-model approach could also help direct resources at the national and state levels by indicating where fires will take place.

Exploring more complex regression models could also result in better forecasts. A Long/Short-Term Memory (LSTM) model might, for example, be able to replicate some of the smaller nuances and unique fluctuations better than the linear model used for this analysis.

Finally, there are several different ways to convey fire prevention information to the public and various levels of government. Researchers have developed fire risk indices and forecast models to better understand natural and man-made fires. A more advanced model could probably leverage factors like condition-based probability and risk classification to summarize fire hazards to a non-technical public audience.

REFERENCES

[1] Amazon – Facts. https://www.worldwildlife.org/places/amazon. World Wildlife Fund. 1250 24th St., NW, Washington, DC 20037. Published and accessed November 2019.

[2] Field, R. D., et al. "Development of a global fire weather database." *Natural Hazards and Earth System Sciences* 15.6 (2015): 1407-1423.

[3] Davis, Rollo T. "Atmospheric stability forecast and fire control." *Fire Management Today* (1969): 56.

[4] Garcia Diez, A., L. Rivas Soriano, and E. L. Garcia Diez. "Medium-range forecasting for the number of daily forest fires." *Journal of Applied Meteorology* 35.5 (1996): 725-732.

[5] Možný, Martin, and Daniel Bareš. "Forecast danger of vegetation fires in the open countryside in the Czech Republic." *Rožnovský, J., Litschmann, T., eds.: Mendel a bioklimatologie. Brno, 3.–5. 9. 2014. Masarykova univerzita, Brno, 275* 285 (2014).

[6] Method and System for Automated Location Dependent Natural Disaster Forecast. Guatteri et al. United States Patent. US 9196145 B2. November 24, 2015.