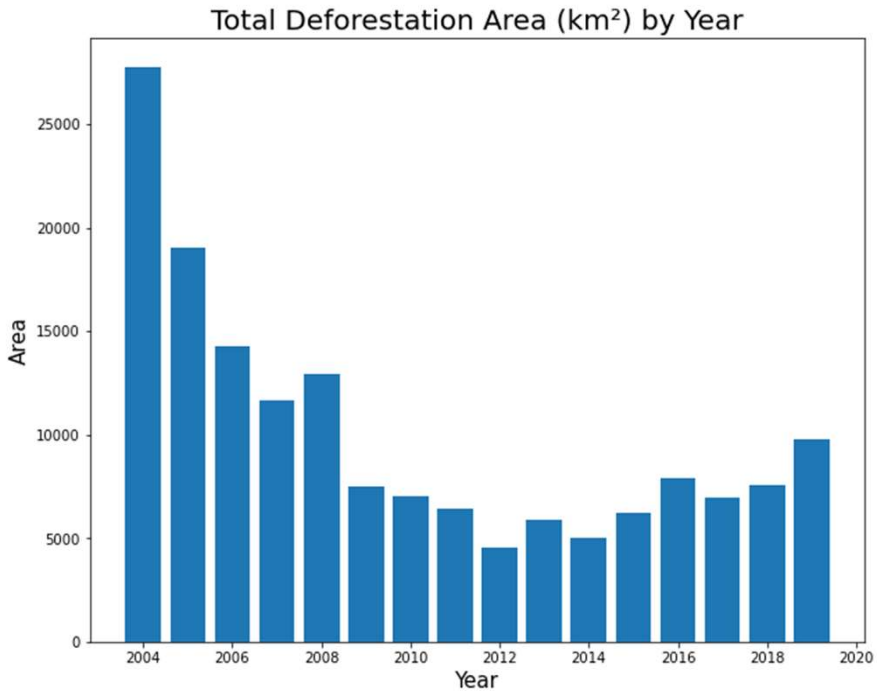
Rainforest Deforestation in Brazil: Analysis and Forecasting

by Dan Hoogasian

The Data

- Two datasets
  - Area lost: km^2 per year, by state and aggregate total
  - Fire spots (primary data): monthly frequency of burns in each state
  - After research, concluded time series data should be handled differently than regular data
  - Data can have cyclical nature
  - Date format needs to be converted to something scalar (e.g., seconds or hours)



Correlation: Downward Trend

- Correlations:

ACRE -0.01

AMAPA 0.02

AMAZONAS 0.15

MARANHAO -0.14

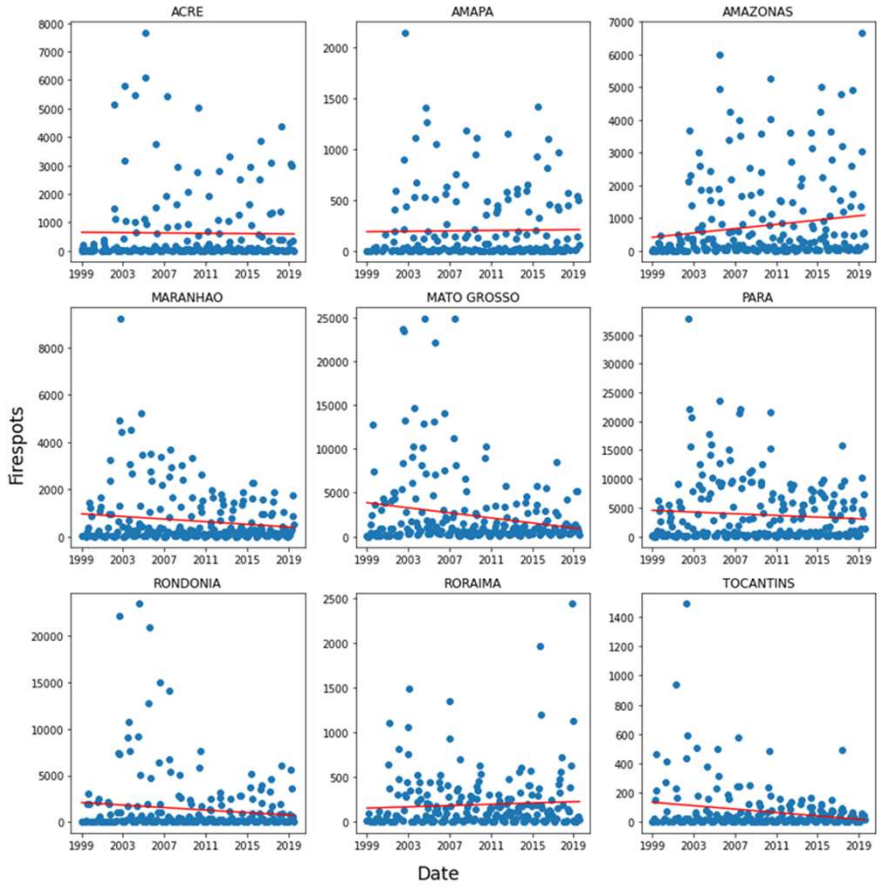
MATO GROSSO -0.21

PARA -0.08

RONDONIA -0.12

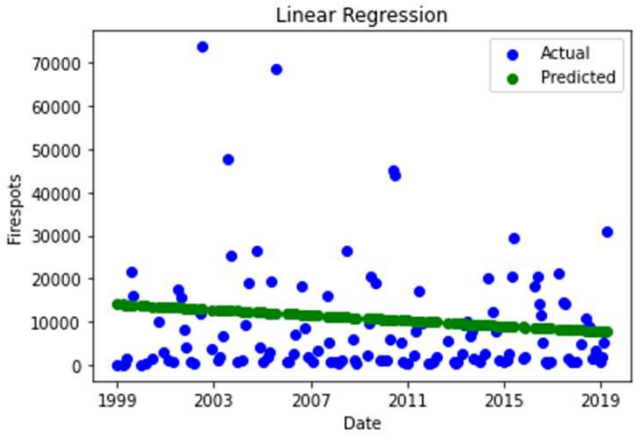
RORAIMA 0.07

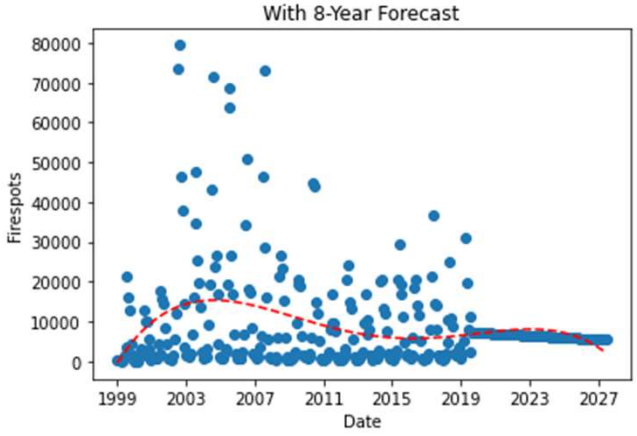
TOCANTINS -0.22



Run Supervised ML Algorithms

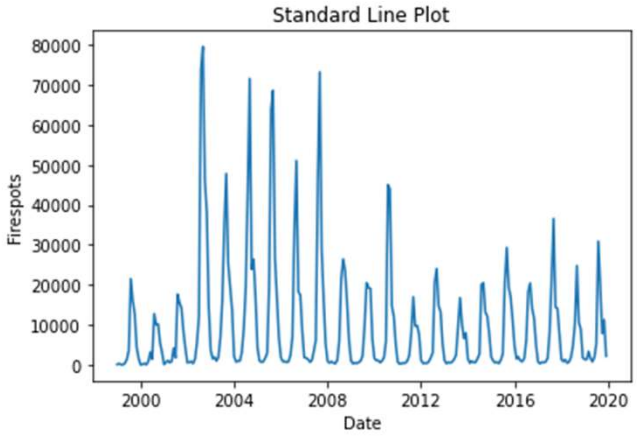
- Linear regression, quadratic regression, and k-NN
- Linear regression was best model, with RMSE of 12,653
  - Quadratic regression RMSE: 13,790
  - K-NN RMSE: 18,313
- 8-year forecast
  - This works with all three models, but there are other, possibly better options

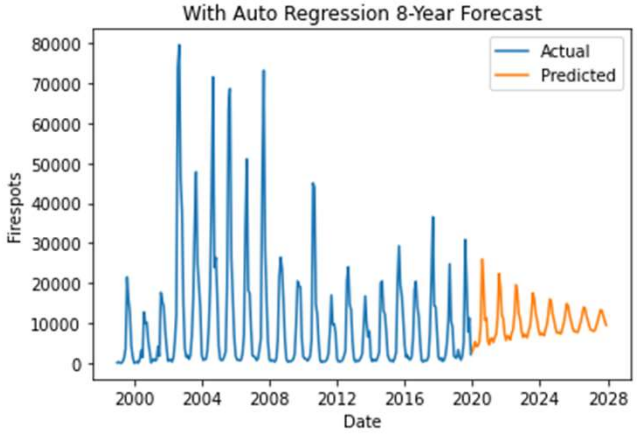




Time Series Forecasting Algorithms

- Standard line plot shows cyclical nature
- Moving average forecast
  - RMSE: 9,117
  - > my_data.rolling(window=10).mean()
- Auto regression
  - RMSE: 8,559
  - > model = AutoReg(train, lags=20)





More Advanced Model

- Seasonal Autoregressive Integrated Moving-Average (SARIMA) combines previous models and adds a seasonal component
- More complex than other models
  - Notation: $$SARIMA \underbrace{(p,d,q)}_{non-seasonal} \underbrace{(P,D,Q)_m}_{seasonal}$$
  - p and seasonal P: Number of autoregression terms (lags of stationary series)
  - d and seasonal D: Differencing that must be done to stationary series
  - q and seasonal Q: Number of moving-average terms (lags of forecast errors)
  - m: Seasonal length of data

Preliminary Steps for SARIMA

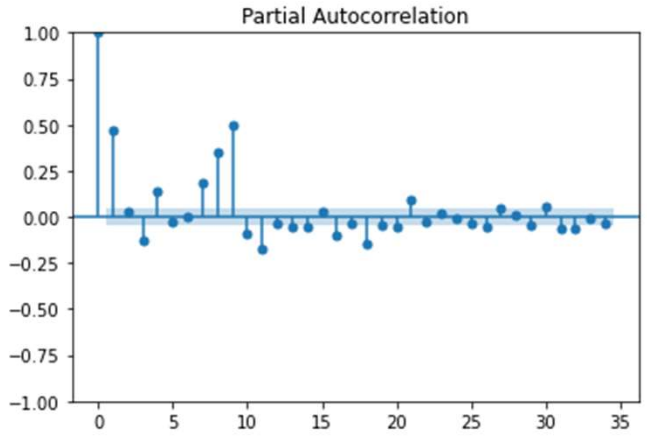
- First step is to determine if data is stationary (i.e., its statistical properties are stationary over time)
  - Can use tests like Augmented Dicky-Fuller test to test for stationarity
  - Augmented Dicky-Fuller test on data had p-value of 1.4e-11, which is small enough to statistically conclude stationarity

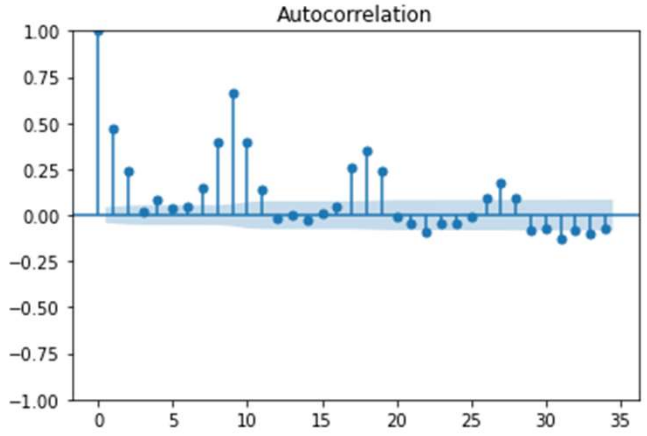
```
> from statsmodels.tsa.stattools import adfuller
> ad_fuller_result = adfuller(my_data['data'])
```

- If not stationary, use a transformation function (e.g., logarithm) to smooth data until desired p-value is achieved

Preliminary Steps for SARIMA (continued)

- Plot partial autocorrelation function (PACF) and autocorrelation function (ACF) graphs to inspect for seasonal variation
    - This is the seasonal component of the auto regression and moving-average models from before
    - Done to avoid having to find the parameter via tuning





Tuning SARIMA

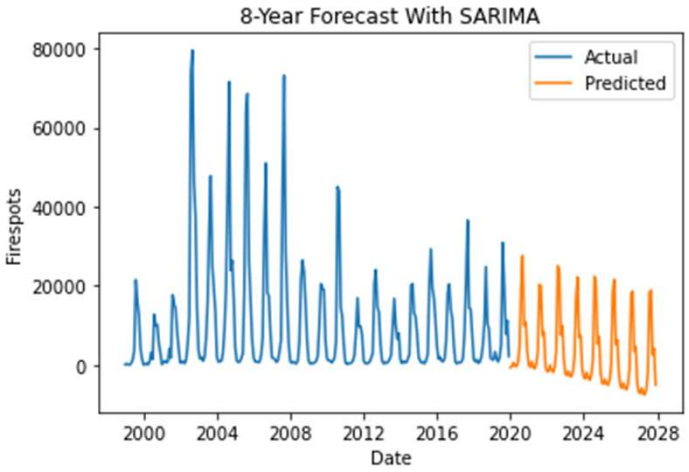
- Manual tuning is typically the preferred method
- Can be done automatically, but this is more computationally expensive and provides less options for tailoring
- Akaike's Information Criterion (AIC) is often used as performance metric
- Can experiment with different parameter selections on full data

	params	seasonal params	AIC
107	(0, 1, 0)	(2, 2, 2, 12)	1577.704587
485	(1, 2, 2)	(2, 2, 2, 12)	1608.860387
728	(2, 2, 2)	(2, 2, 2, 12)	1610.619538
161	(0, 1, 2)	(2, 2, 2, 12)	1610.847199
242	(0, 2, 2)	(2, 2, 2, 12)	1611.008585
...	...	...	...
243	(1, 0, 0)	(0, 0, 0, 12)	2727.479177
27	(0, 0, 1)	(0, 0, 0, 12)	2739.708120
405	(1, 2, 0)	(0, 0, 0, 12)	2751.150774
162	(0, 2, 0)	(0, 0, 0, 12)	2760.511250
0	(0, 0, 0)	(0, 0, 0, 12)	2850.025395

SARIMA Results

> SARIMAX(my_data, order=(1, 1, 1), seasonal_order=(2, 2, 2, 12))

- RMSE: 8,520
- Can see more consistent, cyclical nature than with other models
- Maintains more amplitude than with just auto regression



Notes

- Ideally, train/test splits on time series data would be done so each portion consists of consecutive data, where testing data is always chronologically after training data
- Ideally, this would be in cross-validation form, with first iteration taking a small portion of entire dataset, and each following iteration taking more training data until a split of the entire dataset (e.g., 70/30) is achieved
- SARIMA models are complex and running a basic tuning can take hours, so parameters in this example have limited range
- More investigation into parameter specifics and constraints would be needed to better fit the model

References

- Netto, M. B. (2019). *Brazilian Amazon Rainforest Degradation 1999-2019* [Data set].
- *Pythonic Cross Validation on time series*. (2014, September 16). Francescopochetti.com. http://francescopochetti.com/pythonic-cross-validation-time-series-pandas-scikit-learn/
- Brownlee, J. (2018, Aug. 6). Machinelearningmastery.com. Retrieved May 2, 2023, from https://machinelearningmastery.com/time-series-forecasting-methods-in-python-cheat-sheet/
- Wu, S. (2021, July 3). *Time series forecast in python*. Towards Data Science. https://towardsdatascience.com/how-to-predict-your-step-count-for-next-week-a16b7800b408
- Peixeiro, M. (2020, July 29). *Time series forecasting with SARIMA in python*. Towards Data Science. https://towardsdatascience.com/time-series-forecasting-with-sarima-in-python-cda5b793977b