Predicting Food Prices in Developing Countries

Springboard Capstone 2

The Problem

- Food insecurity affects 345 million people worldwide
  - Especially acute in sub-Saharan Africa and the Middle East
- Can future prices be predicted from past values?
- Conflict and climate change can exacerbate the issue
- Can these factors be incorporated into the analysis?

Who Would be Interested?

- International aid groups
- Local farmers
- Climate researchers?

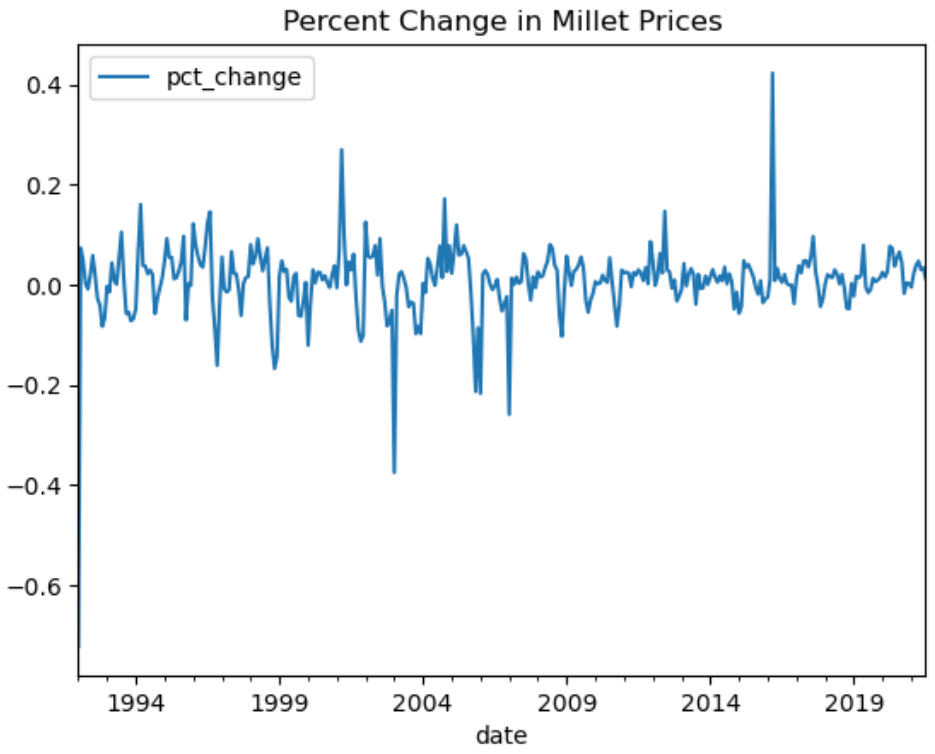
Data Wrangling

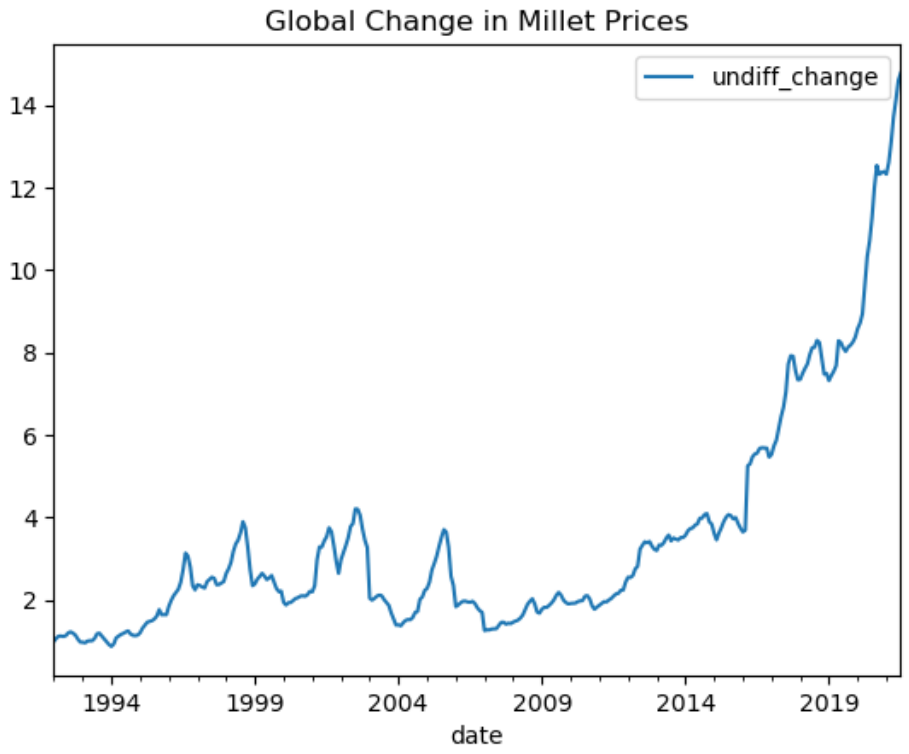
Data Wrangling

- Dataset gathered from Kaggle, originally from the World Food Programme
- Contains food prices from 98 countries for over 800 goods, but not complete!
- Further broken down by state and market
- Niger had longest dataset, beginning in 1990 (monthly)

Data Wrangling

- I want to use food prices from other countries
- Prices are tracked in the local currency, so average percent change was used instead
- Then converted to average global change in price

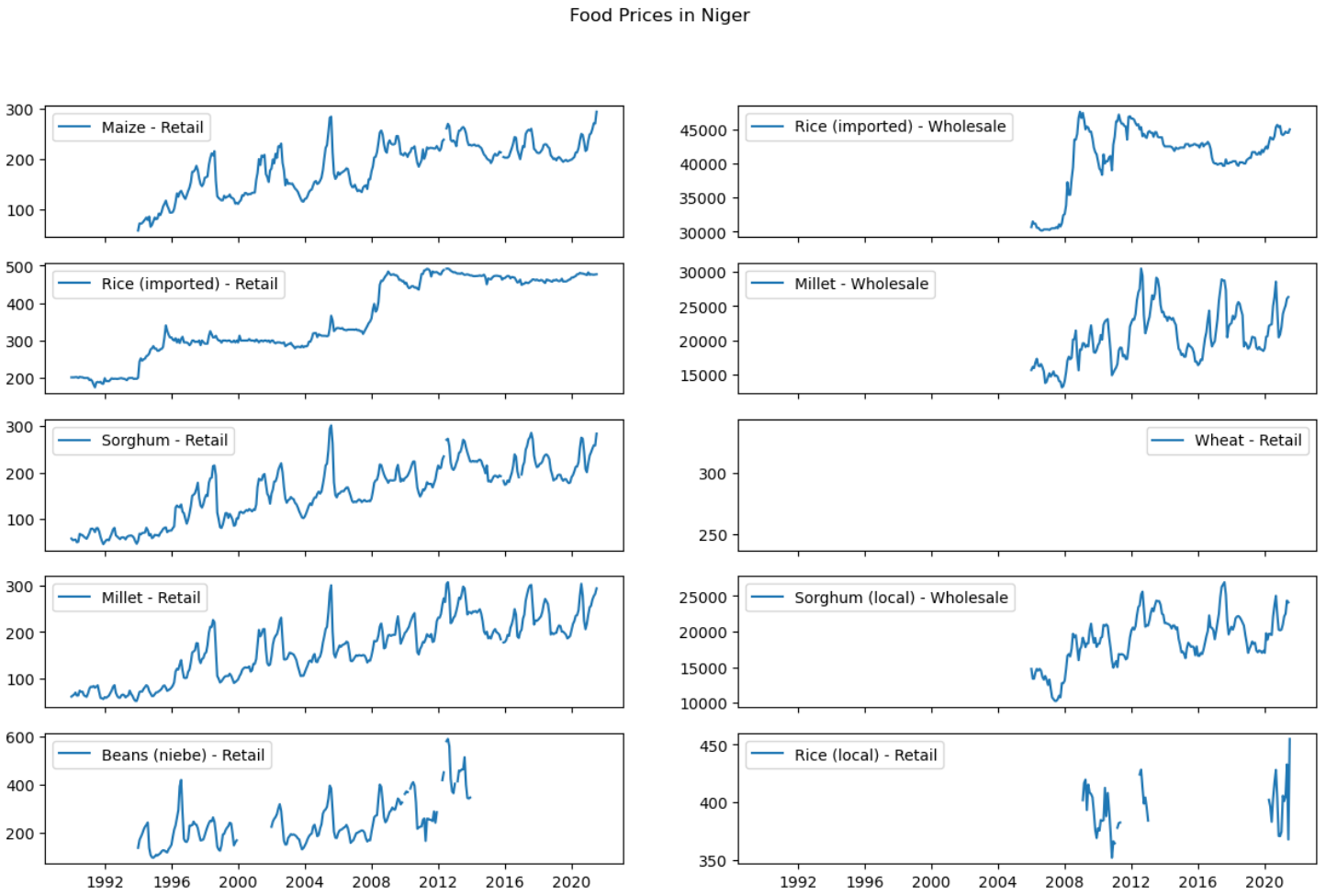




Exploratory Data Analysis

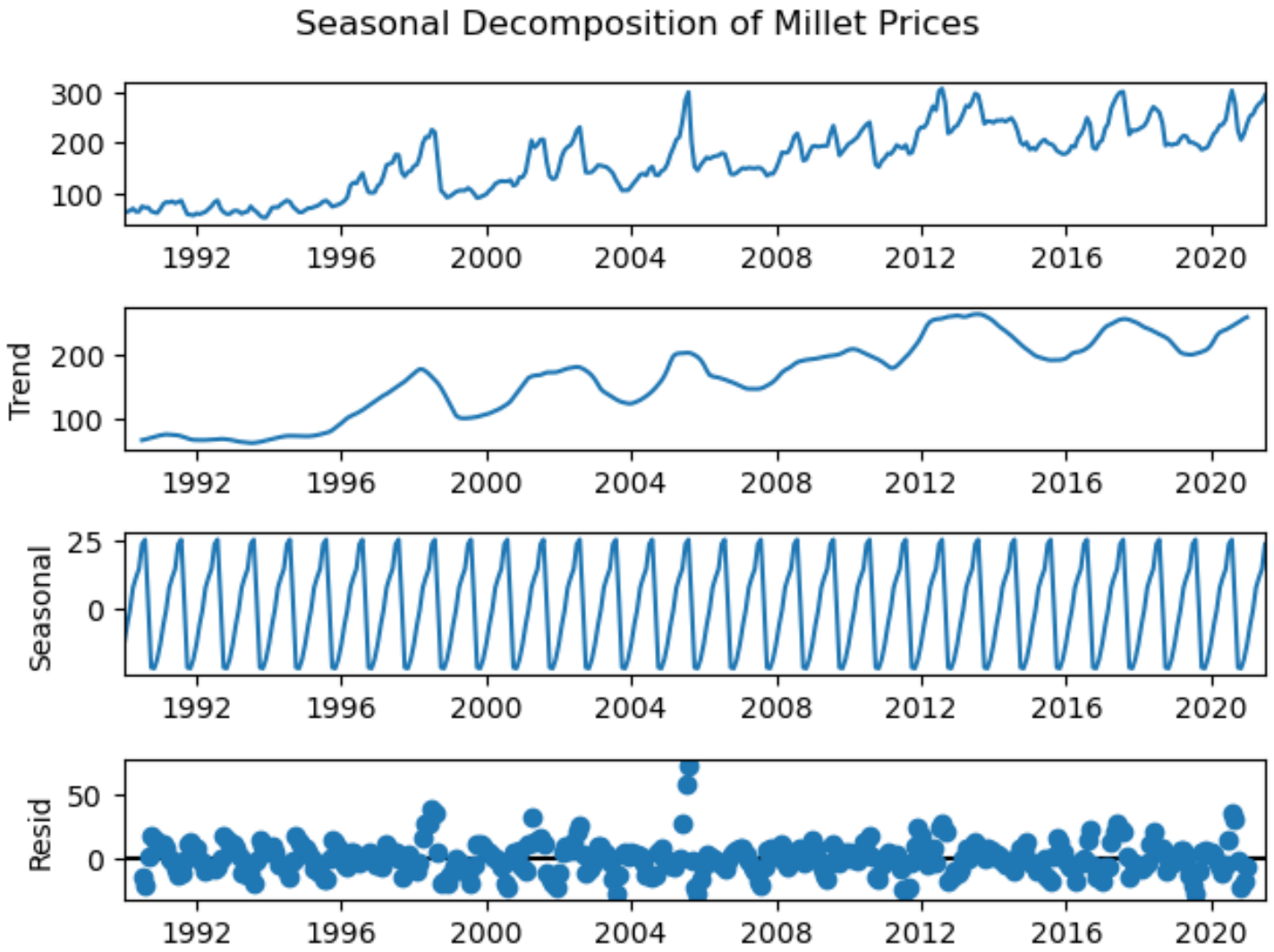
Exploratory Data Analysis

- Niger data “unpivoted” and prices averaged across the whole country
- Millet chosen for further analysis
- Difference in imported vs. locally grown crops?

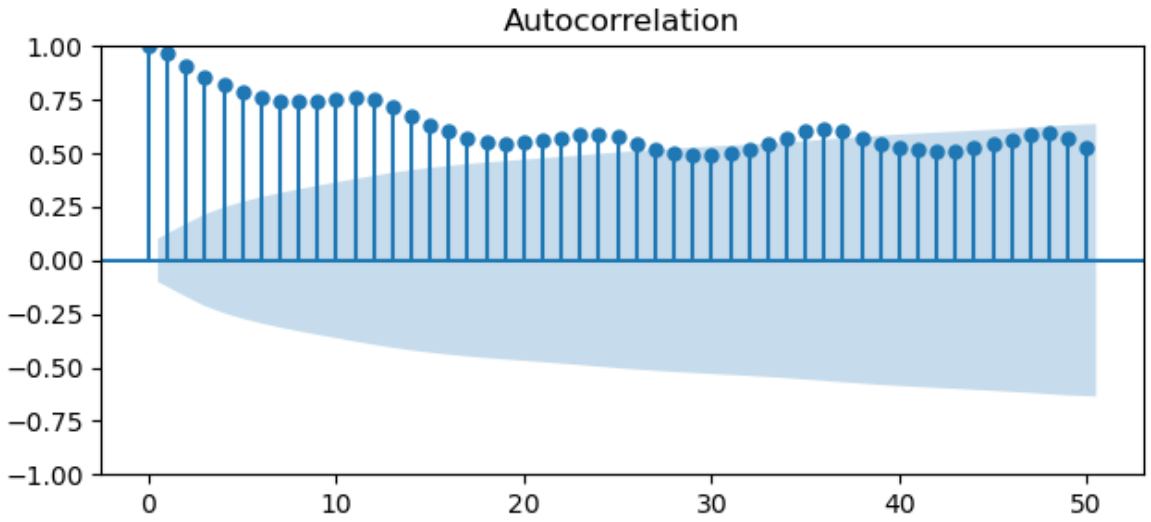


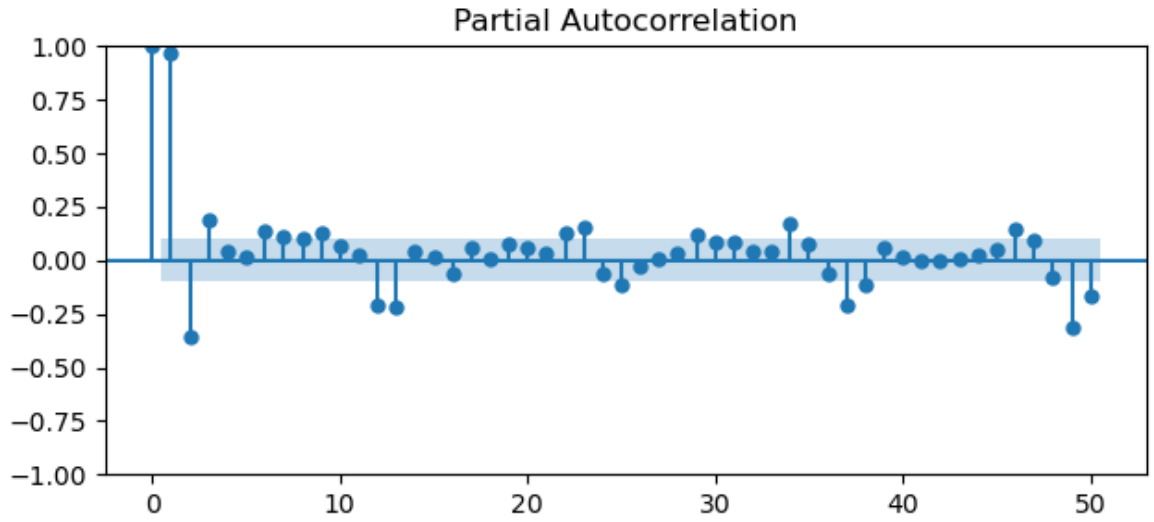
Exploratory Data Analysis

- Crop data is clearly seasonal!
- Seasonal decomposition can be used to pull out the components of a time series



Exploratory Data Analysis

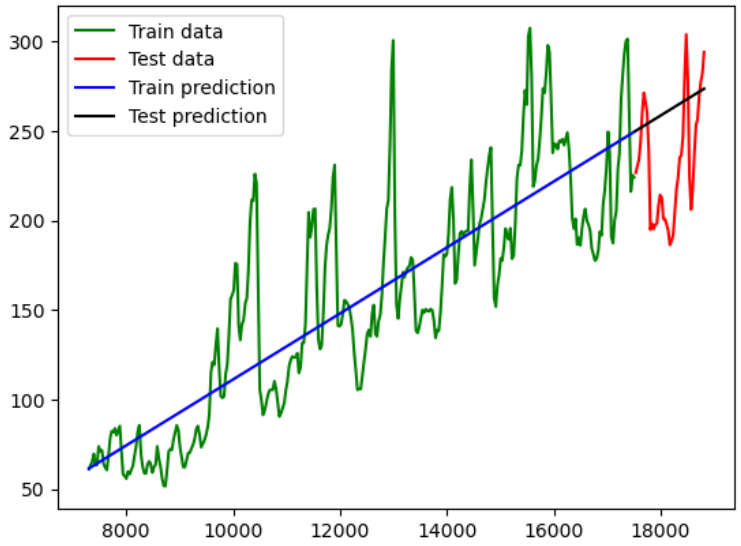


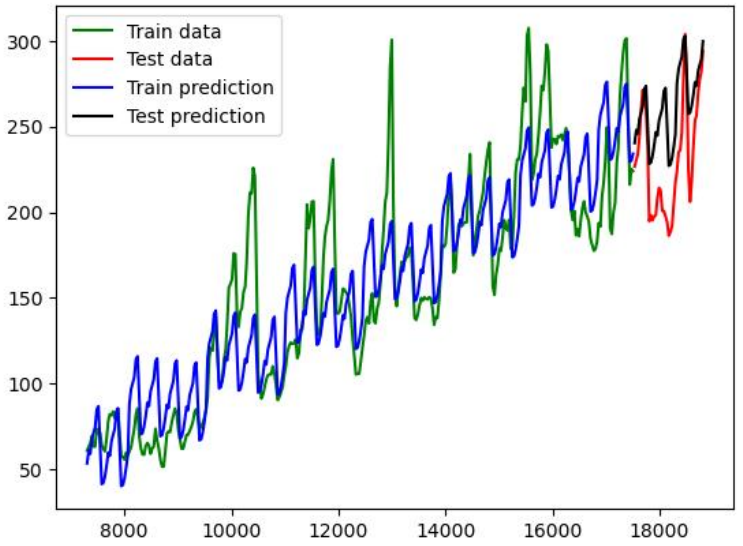


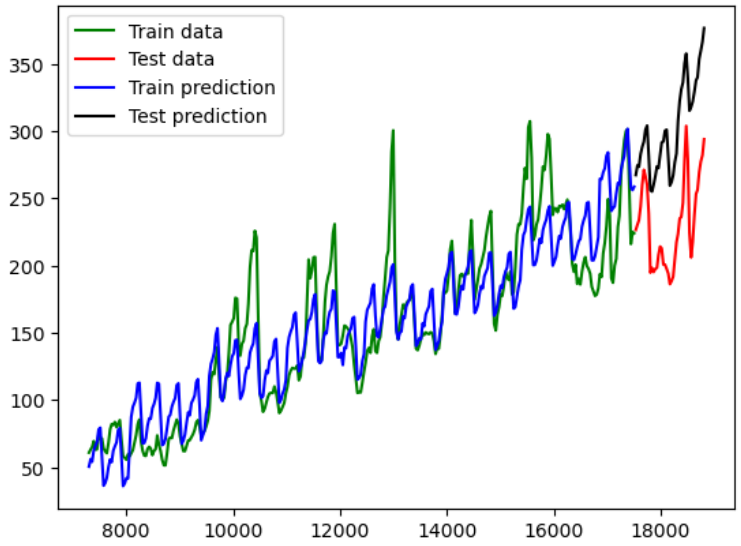
Preprocessing and Training

Preprocessing and Training

- Initial attempt used linear regression
- Adding month as categorical variable allowed for seasonality
- Adding worldwide millet price did not help!

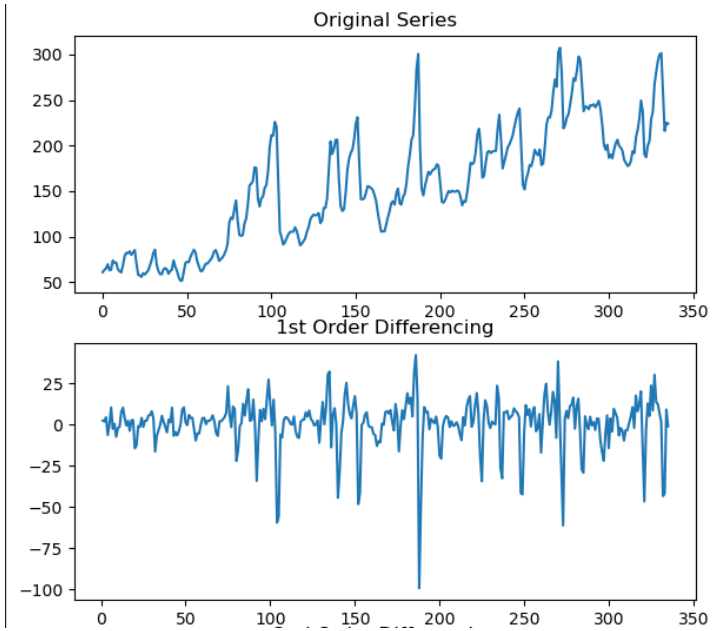


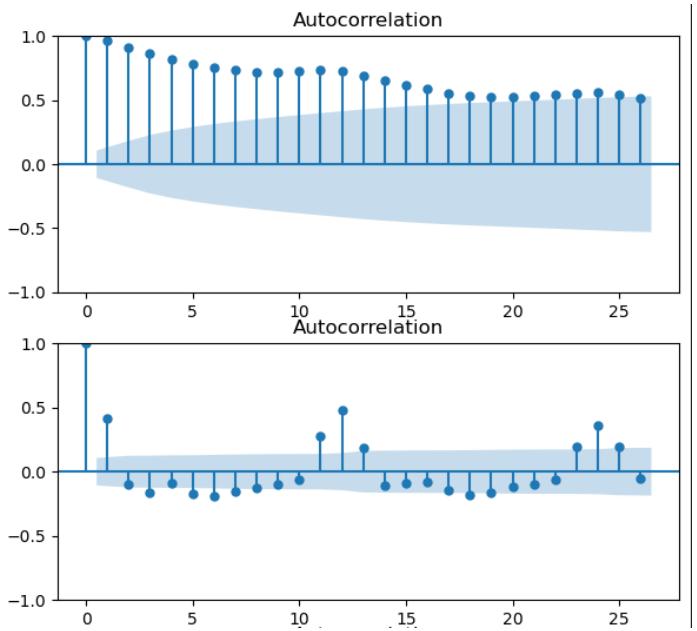




Preprocessing and Training

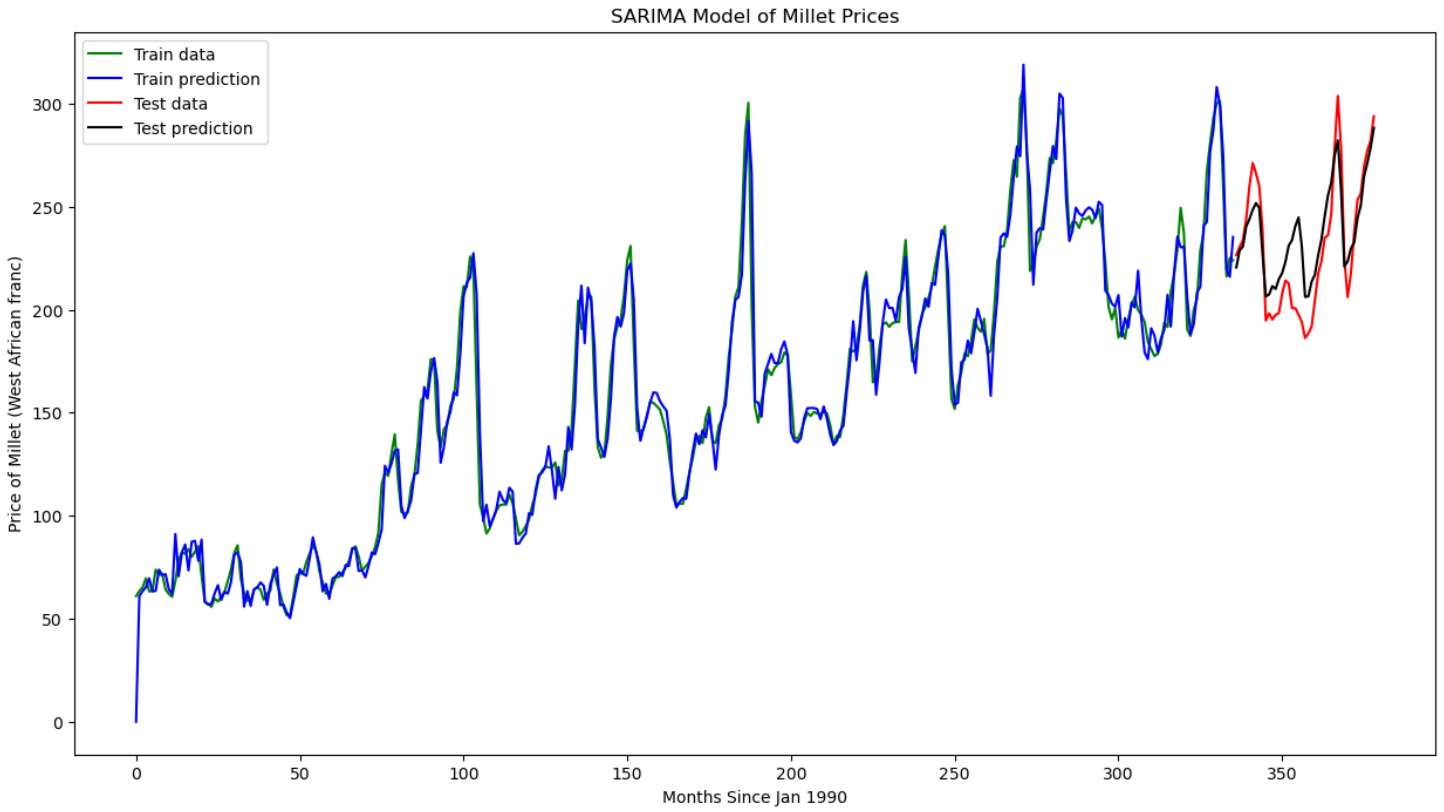
- Second attempt used ARIMA model
- Uses no features other than past values
- Has three parameters that can be estimated using plots of the auto-correlation, partial-autocorrelation, and difference
- Requires a stationary series at the end



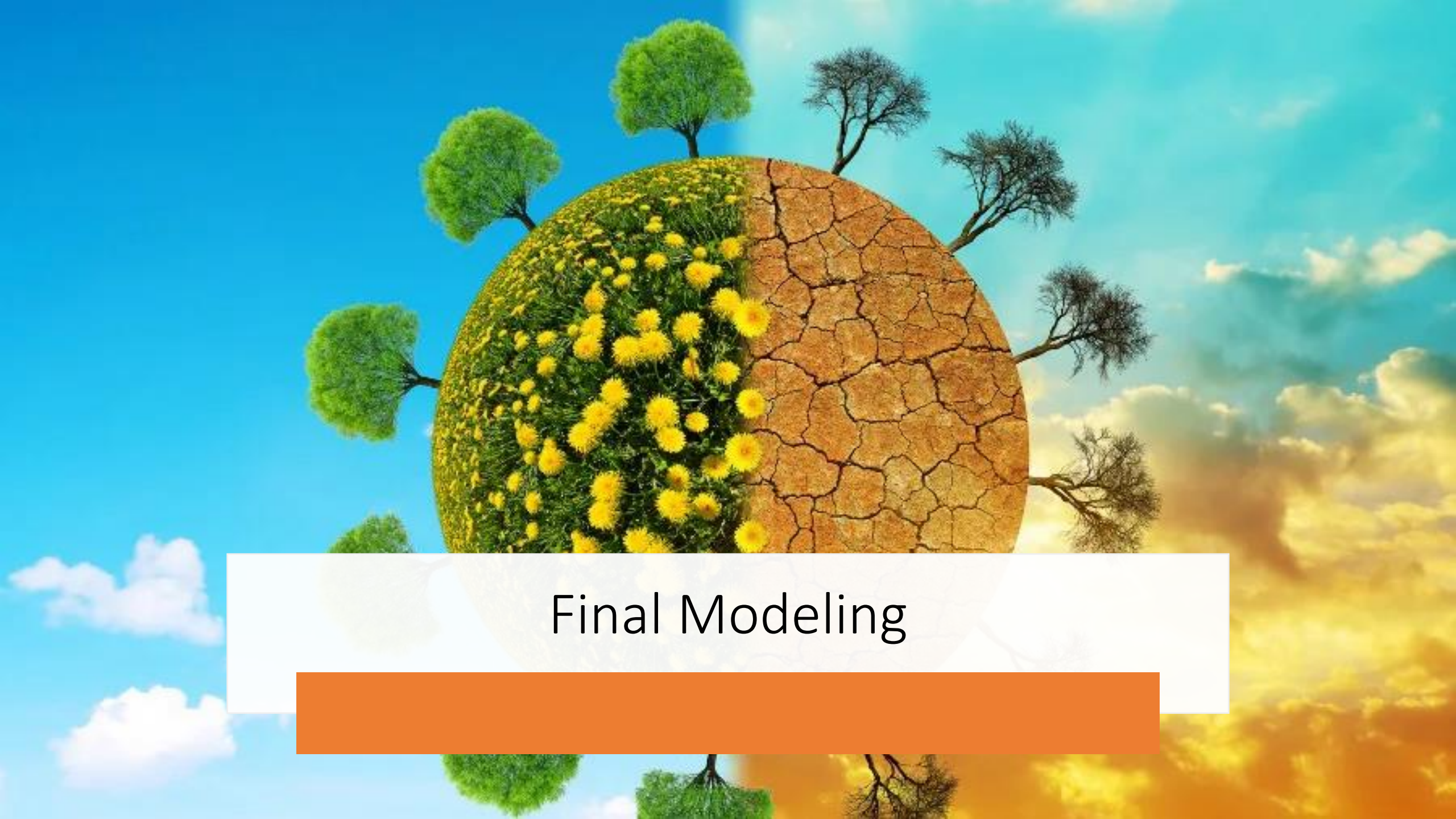


SARIMA

- Seasonal ARIMA can fix problems with previous approach
- Achieved mean absolute percentage error of 6.4%



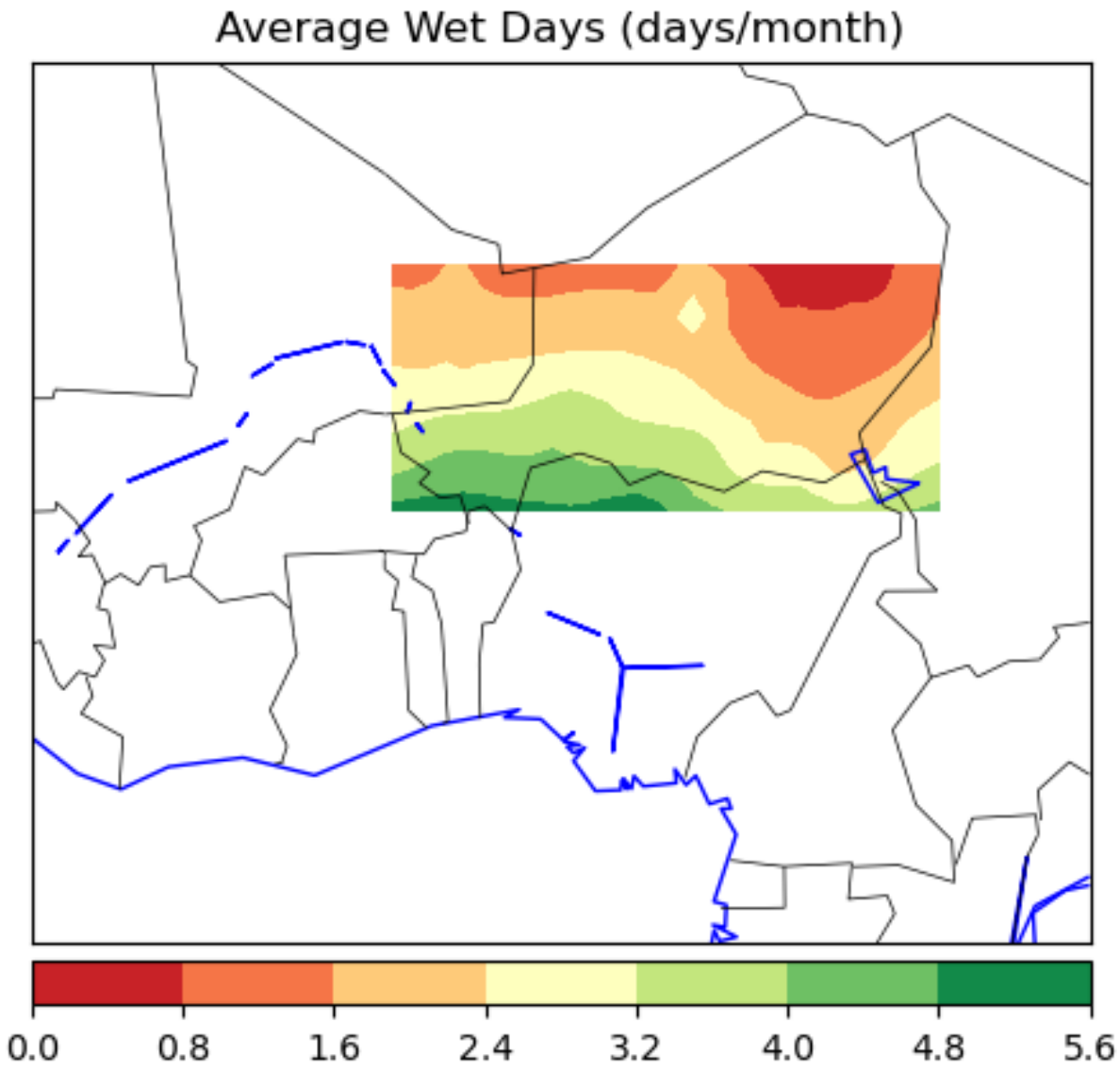
Cross Validating SARIMA



Final Modeling

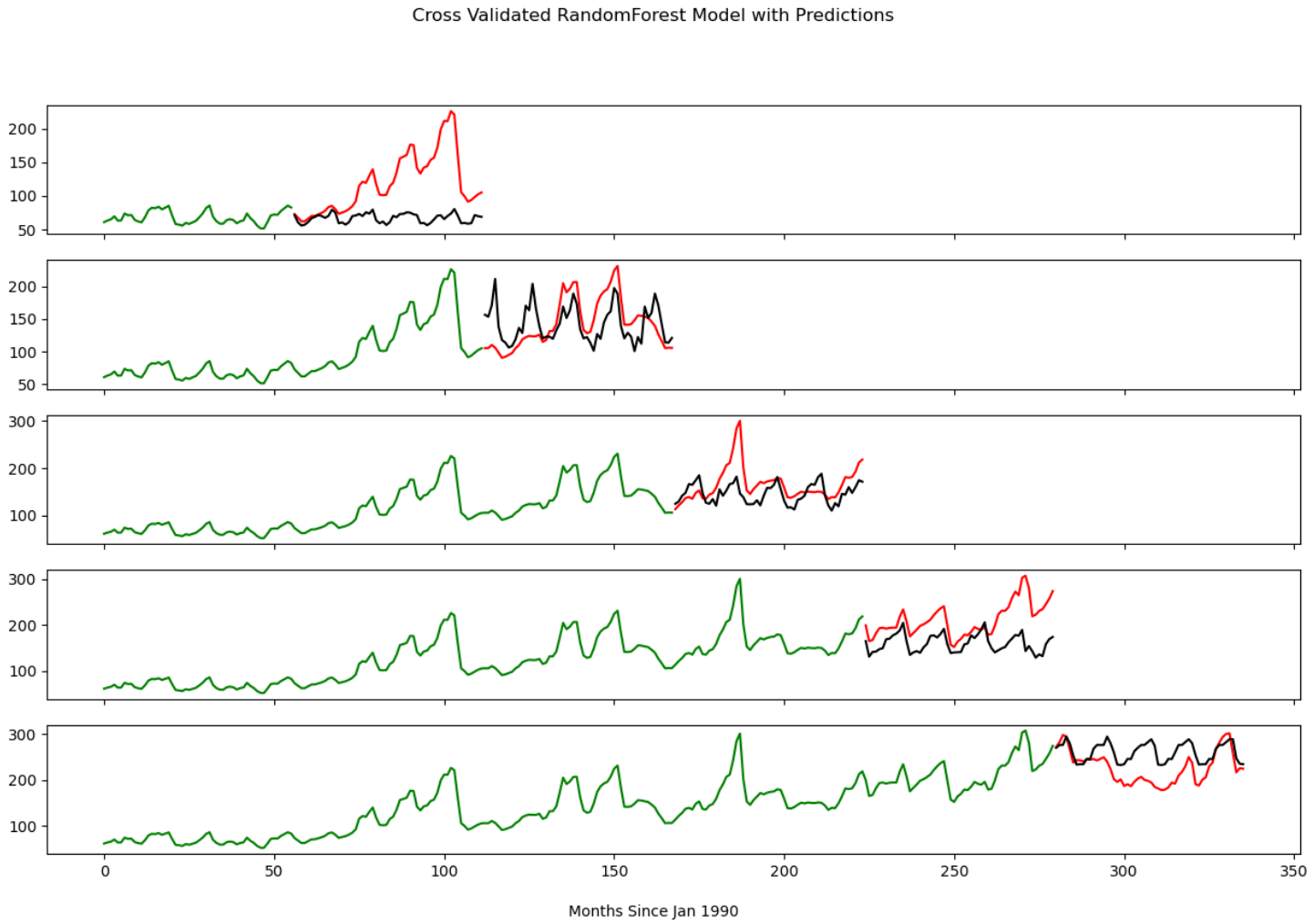
Final Modeling

- Climate data could aid in predictions
- Available from the Centre for Environmental Data Analysis
- Average wet days, average precipitation, and average temperature chosen



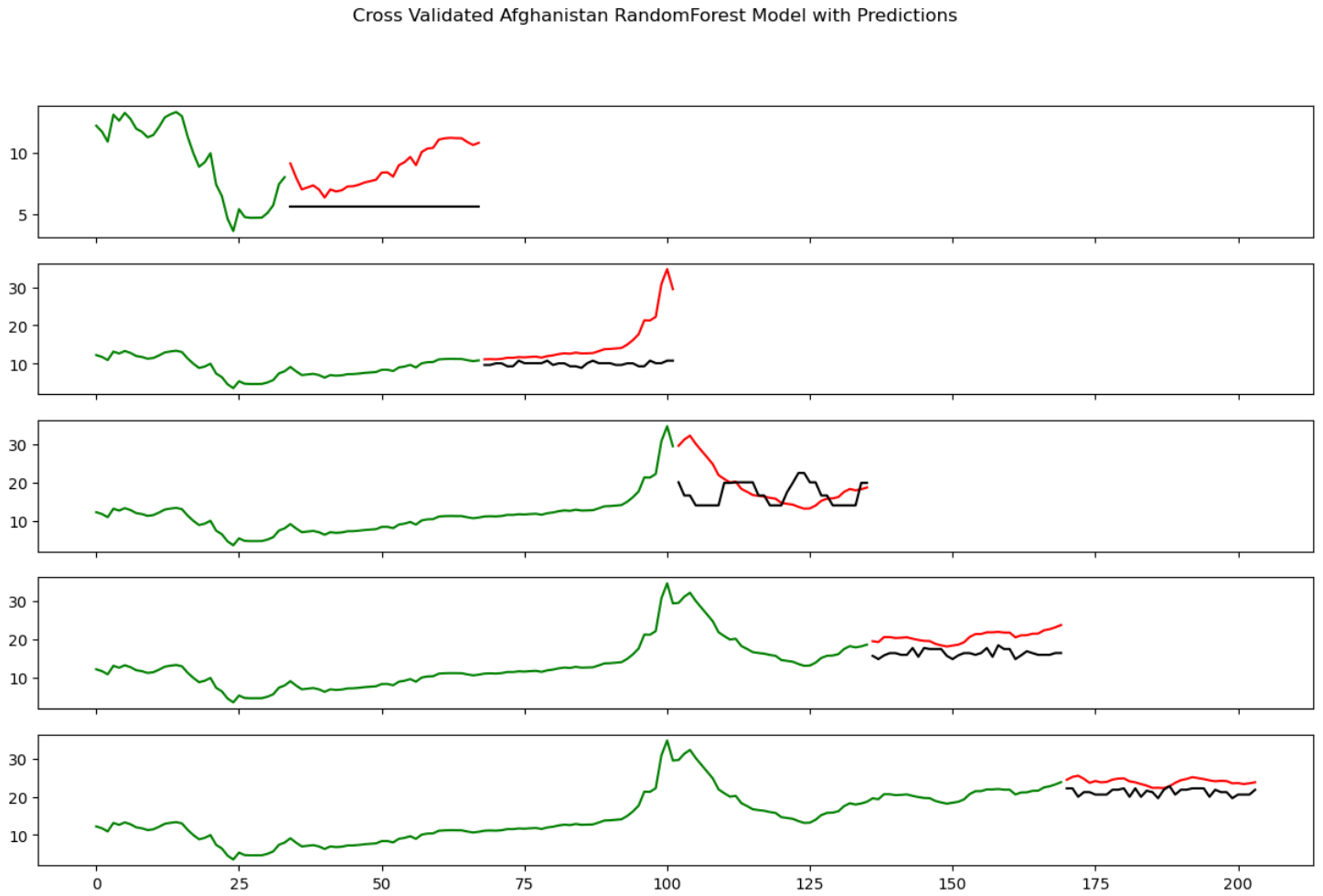
Final Modeling

- SARIMAX model still had problems on CV
- LinearRegressor and RandomForestRegressor tried with grid search
- Random forest had mape ~12%, but more robust

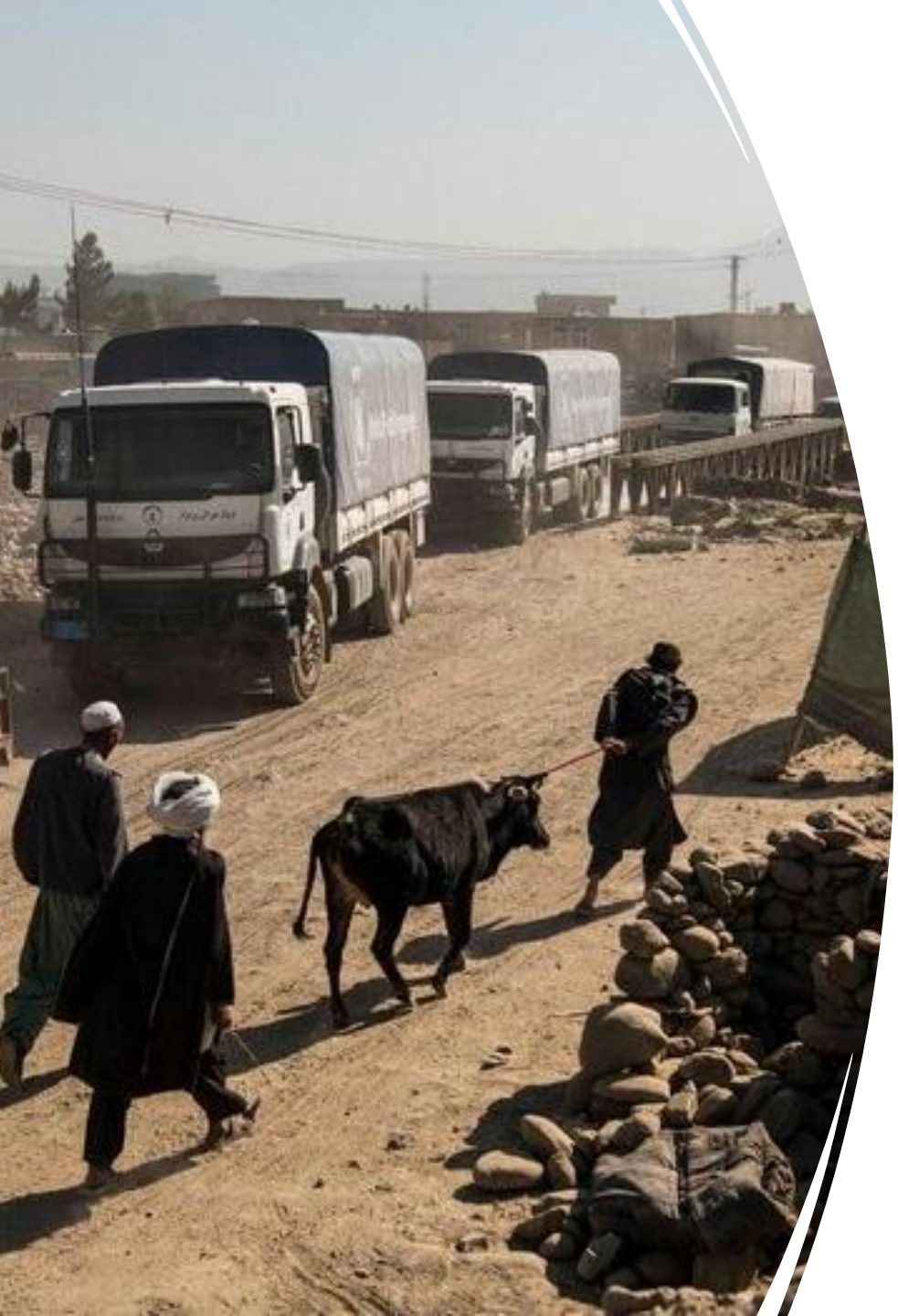


Afghanistan

- Repeated workflow, focusing on wheat prices
- Mostly imported, prices flatter and less seasonal
- Model achieved similar/slightly worse results (mape 11-25%)



Conclusion



Conclusion

- SARIMAX model achieved lowest mape
- Randomforest Regressor most robust model, probably most useful for an NGO
- Choice of good and context important to consider
  - Is it locally grown or imported?
  - How much data is available?

Next Steps

- Better predict for spikes, reframe around outlier detection?
- Climate data helped but not enough
  - Sentiment analysis from twitter to capture unrest
  - Insect and pest populations?