Optimizing ARIMA Models for Midwest Fire Forecasting: A 10-Year Analysis of NASA FIRMS Data and Environmental Factors

Caitlyn Feldpausch[1]

[1]University of Michigan, Department of Climate and Space Sciences and Engineering.

Corresponding author: Caitlyn Feldpausch (feldpaw@umich.edu)

Key Points:

- The optimal dynamic and static autoregressive moving average models for non-seasonal forecasting has first order differenced monthly aggregated MODIS and VIIRS SNPP fire detection data with an autoregressive term, and with no moving average term, which is primarily determined based on model error analysis.
- Upon fitting a seasonal term to the optimized model by assuming harvest season trends occur each year, a forecast created by synthetic data depicts convergence to about 3700 to 3800 detections after 10 months.
- Due to this convergence, the optimized autoregressive moving average model with a seasonal term added is best for short-term forecasting when utilizing the error analysis preferred monthly aggregated fire detection data.

Abstract

Midwestern U.S. fires can be overlooked in favor of the West. Current methods of fire detection prediction can be time-consuming, especially those involving digitization. Determining the best ARIMA model for non-seasonal time-series forecasting of Midwest fire detections and then including a seasonal term enacts quicker timespan predictions based on historical data and increases accuracy through error analysis. A time-series analysis of fire detections, landcover types, vegetation health, and relative humidity during general harvest season months of May to October for 1 January 2009 to 1 January 2019 provides model variable overviews. A Pearson Correlation Coefficient determines linear strength between fire detections and the other variables. A level of differencing of monthly, yearly, or monthly and yearly aggregated fire detection data according to the Augmented Dickey-Fuller test, with one-to-one mapped environmental variables, achieves stationarity. A Durbin-Watson test and autocorrelation function plots determine aggregated dataset autocorrelation to add autoregressive and moving average model terms. Autocorrelation assumptions change when model structure is ascertained. The optimal time-series aggregated fire detection dataset for either dynamic or static ARIMA forecasting is determined based on Root Mean Square Error percentages and Mean Absolute Percentage Error scores. The optimal, static and dynamic non-seasonal ARIMA is the first order differenced monthly aggregated fire detection data with an autoregressive term.

Plain Language Summary

We use NASA Fire Information for Resource Management (FIRMS) MODIS and VIIRS-SNPP fire detections, which has detected burns as point data in UTC. VIIRS-SNPP spatial resolution is 375m taken twice daily and available since 2012. MODIS spatial resolution is 250m and available since 2000. Relative humidity data is included as lower relative humidity can lead to greater chances of burns. We include NCEP/DOE Reanalysis II daily average relative humidity data as low relative humidity can increase burn chances. The Landsat-derived USDA National Agricultural Statistics Service Cropland Data Layer with a high spatial resolution of 30m is used for burn landcover types. It's a geo-referenced, annual landcover raster from satellite imagery and in-situ data. Normalized Daily Vegetation Indices are obtained from the Google Earth Engine catalog's MODIS Terra

Surface Reflectance data, and provide landcover health during times of fire detections. Map and analysis products use the WGS 1984 mercator projection.

1.1 Introduction

We take an Autoregressive Integrative Modeling Average model approach as vegetation burn predictions involve complex atmospheric and environmental variable interactions. Predictive methods have trade-offs between timeliness and accuracy. For example, utilizing spatial digitizing methods, whereby individual burns are analyzed in satellite imagery, is incredibly time consuming as burns can occur over multiple dates, or come from emissions sources. However, the process is accurate as it involves comparing RGB band composites over multiple days and burn periods, fully assessing the extent of the fire that occurred. Then utilizing that information to make burn area estimates and burn landcover analysis. Other algorithm, index, and model methods which take less time to analyze large point source datasets can be less accurate if trained ineffectively and require error analysis.

Through this analysis, a time-series forecasting autoregressive integrative moving average model is optimized by error analysis to determine the best aggregated fire detection dataset and model parameters to make future forecasts. Increasing accuracy through model validation with Mean Absolute Percentage Errors (MAPE) and Root Mean Square Error (RMSE) scores. Taking less time to make detection amount predictions over time spans by implementing a time-series forecasting model instead of manual detection digitization and analysis in satellite imagery.

This method was already successful when landcover changes, climate events, and remote sensing in the Atlantic Forest region was associated to fire risk using the ARIMA model (Paper 1). In the analysis, historical landuse and landcover change were detected using Landsat satellite imagery and weather variables included relative humidity, rainfall, maximum and minimum air temperature, and wind. Predictive data was obtained from an ARIMA model whose statistics were analyzed by simple linear correlation, a coefficient of determination, and standard error of estimation. We take a similar approach to optimal ARIMA model construction for non-seasonal forecasting of aggregated fire detection data.

1.2 The ARIMA Model Explained

The model utilized will be an ARIMA model, or Autoregressive Integrated Moving Average model which can forecast time-series non-seasonal trends using historical values and three terms within the model. The three terms are the AR autoregressive, I integrative, and MA moving average terms.

An integrated series for the integrative I term is a version of the original time series that needs differencing to become stationary (Nau, 2020a). The lags of stationary time series are known as the autoregressive AR terms in the forecasting equation and the lags of forecast errors are known as the moving average MA term (Nau, 2020a). Lags are delays between an observed data point and prior values. These term components are separate parameter values within the ARIMA model, noted as $ARIMA(p, d, q)$ where p is the number of autoregressive terms, d is the number of non-seasonal differences to achieve stationarity, and q is the number of moving average terms or lagged forecast errors (Nau, 2020a). Random-walk, autoregressive, and exponential smoothing models are special cases of the ARIMA model (Nau, 2020a).

A random-walk model of $ARIMA(0,1,0)$ can be fitted to the data if only including the integrative term. This is the simplest ARIMA model form for a nonstationary data series. The form for this equation is:

$$\hat{Y}_t = Y_{t-1} + \alpha$$

which can be fitted as a regression model with no intercept where the dependent variable is the first difference of Y (Nau, 2020a). Hence, $ARIMA(0,1,0)$ is a nonseasonal differenced model with or without a constant (Nau, 2020a). If the mean step size alpha is nonzero, then the model has drift. A model without alpha assumes the series takes a random step away from its last recorded position at each time step and that the mean value of those steps is zero (Nau, 2020c).

Autoregressive models denoted $ARIMA(1,0,0)$ with no differencing or MA term is denoted as:

$$\hat{Y}_t = \alpha + \theta_1 Y_{t-1}$$

and if the coefficient theta equals 1, then it is removed and we end up with the random-walk model above, with growth (Nau, 2020b).

Simple exponential smoothing ARIMA models have a first order differenced integrative term and a moving average term represented as $ARIMA(0,1,1)$ or also as:

$$\hat{Y}_t = Y_{t-1} - \theta_1 e_{t-1}$$

which do not include a constant term as $\theta_1 = 1 - \alpha$ (Nau, 2020a, 2020b). This exponential smoothing corrects autocorrelation errors of random-walk models where instead of using "the most recent observation as the forecast for the next observation, it is better to use an average of the last few observations in order to...more accurately estimate the local mean" (Nau, 2020a). To do this, moving averages that are exponentially weighted based on past values are used.

These model types will come up in later trial-and-error for model optimization. Models may or may not have both autoregressive, moving average. An integrative term is unnecessary if the data series is already nonstationary.

For this analysis, a static and dynamic ARIMA model will be optimized. The dynamic model will utilize walk-forward validation. This dynamic ARIMA model with walk-forward validation is a time-series validation technique where the temporal order of the forecast is preserved, unlike with cross-validation. At each iteration the training set is divided into a train and test set as long as the test set is always following the training set (Zhang, 2022). The ARIMA model method is switched to static to see if the optimal ARIMA(p,q,r) model changes. Static models are time invariant regression models and assume parameters associated with predictors are fixed with time (Crozier et al., 2005).

2 Materials and Methods

2.1 Time-Series Analysis of Variables

Midwestern U.S. daily fire detections, daily average Normalized Difference Vegetation Index (NDVI) values indicating vegetation health, daily average relative humidity, and yearly landcover types are graphically analyzed from 1 January 2009 to 1 January 2019 taken during general harvest season months of May to October. This provides historical seasonal trend context to variables going into model statistics and tests calculated to determine the ARIMA model terms. Average relative humidity is included as lower relative humidity can increase vegetation dryness, making it appear dead or unhealthy by NDVI indices, and increasing burn chances.

2.2 Pearson Correlation Coefficient

A Pearson Correlation Coefficient is used to determine linear strength between fire detections and the other environmental or weather variables. A Pearson Correlation Coefficient can assess linear strength between variables. Coefficient values closer to 1 are more strongly positively correlated and coefficient values closer to -1 are more strongly negatively correlated. Values close to zero indicate no linear correlation. A matrix is constructed between daily fire detections, daily average relative humidity percentages, and daily average NDVI values. Through the process of forming this matrix, a one-to-one mapping is done. Each row representing a fire detection has an associated daily NDVI value, daily average relative humidity percentage, and landcover type.

2.3 Differencing for Dataset Stationarity

The differencing of monthly, yearly, and both monthly and yearly aggregated fire detections datasets achieves stationarity and determines the ARIMA model integrative "I" term by removing seasonality. Any nonlinear transformations such as logging, deflating, or raising to a power can be used to convert the data series to a form with consistent and generally symmetrical appearance over time (Nau, 2014).

Differencing of the aggregated fire detection datasets is done until the datasets are stationary according to the Augmented Dickey-Fuller (ADF) test. Within Tables 2 and 3, when the p-value is below the chosen significance level of 0.05 or 5%, and the ADF statistic is far below the critical values, then the null hypothesis is rejected for the ADF test and the time-series is stationary. Else, the data series is non-stationary until an order of differencing is applied and a unit root is present (Nau, 2020b).

If significant under- or overdifferencing of the dataset has occurred, then a unit root is in the AR or MA model term coefficients and higher-order differencing of the time-series is needed (Nau, 2020b). If the sum of AR term coefficients is about one, then AR terms should be reduced by one and an order of differencing added. If the sum of MA term coefficients is about one, then one MA term and one order of differencing is removed (Nau, 2020b).

2.4 Dataset Autocorrelation Determined to Address Model AR and MA Terms

The dataset must be checked for autocorrelation errors, which means some number of autoregressive and moving average terms are required in the model (Nau, 2020a). Autocorrelation at a given lag is the "the correlation between the variable and itself lagged by k periods" (Nau, 2020c). If data series values are random or appear statistically independent in plotting, then there is no autocorrelation, and estimated values are close to zero (Nau, 2020c). Aggregated dataset autocorrelation is determined by a Durbin-Watson test and Autocorrelation Function (ACF) plots to implement either an AR or MA model term.

First, a multiple linear regression model is fitted to the monthly aggregated, stationary dataset. This considers the one-to-one mapped environmental variables associated with fire detection likelihoods and information. Then we determine whether the aggregated fire detection datasets are autocorrelated via a Durbin-Watson statistic test for lower order differenced data, which would be the monthly aggregated data only. An obtained statistic value determines the autocorrelation of multiple linear regression model residuals (Durbin–Watson, 2008).

A multiple linear regression model with p-1 independent variables can be expressed as:

$$Y = X_{\beta} + \varepsilon$$

where Y is the number of observations related to the independent variable as vector $(n \times 1)$, β is a vector $(p \times 1)$ of estimated parameters, ε is a vector $(n \times 1)$ of the errors, and X is the matrix $(n \times p)$ associated with independent variables (Durbin–Watson, 2008). Regression model residuals are expressed as:

$$e = Y - \hat{Y}$$

and statistic d, according to The Concise Encyclopedia of Statistics (2008) is represented as:

$$d = \left(\frac{\sum_{t=2} (e_t - e_t - 1)^2 \,)^T}{(\sum_{t=1}^{T} e_t^2)} \right).$$

It comes from assuming correlation of successive errors (Simon et al., 2024). A statistic less than two indicates positive autocorrelation, a statistic of 2 indicates none, and a statistic from two to four means negative autocorrelation (Simon et al., 2024). According to a Duke Fuqua School of Business online page, if the autocorrelation is positive, this generally means an AR term can be added to the model, and if the autocorrelation is negative, this generally means an MA term can be added instead (Nau, 2020b). Later, this assumption will not always be true when optimizing the model for lowest MAPE and RMSE.

After monthly aggregated data is checked for autocorrelation, an ACF plot is used to further determine positive or negative or no autocorrelation present for the other aggregated datasets of higher order differencing. A fitted multiple linear regression model is done again to the other time-series aggregated datasets. All coefficients falling within a graphical light blue confidence band and coefficients at lag-0 are considered not statistically significant (Romer, 2024a). If aggregated dataset series required first order differencing and the differenced autocorrelations are ruled non-significant, then a random-walk model of the original series is completed (Romer, 2024a).

2. 5 Optimal Dynamic and Static ARIMA Model Selected by Error Analysis

Optimal dynamic and static ARIMA models for forecasting are found primarily through differencing knowledge applied to the integrative term, some trial and error, and by Root Mean Square Error scores and Mean Absolute Percentage Errors. The values for p, d and q within the ARIMA models take values of zero, one, or two. This is due to issues arising when increasing these values further for relatively smaller, time-series aggregated fire detection data sets.

2.6 Re-introducing Seasonality to Forecast Fire Detections Based on Optimal Model

The non-seasonal and optimal static ARIMA models are adjusted to include the optimal seasonal term by error analysis. The static $SARIMA(0,0,0,12)$ model is used to make forecasts on first order differenced monthly aggregated data and non-differenced monthly aggregated data. This seasonal term of 12 comes from inferring seasonal patterns occur every twelve months, as the harvest season repeats every year. The seasonal model has its own P, D, and Q parameters, which for this analysis equate to zero (Romer, 2024b). This is because ARIMA p, d, and q terms are already optimized.

3 Data

NASA Fire Information for Resource Management or FIRMS MODIS and VIIRS fire detection points are obtained for the continental U.S. region at the Archive Download page via https://firms.modaps.eosdis.nasa.gov/download/. VIIRS has 375 m sensor resolution, 250 m imagery resolution that is collected twice daily (VIIRS). MODIS has 250 m Terra data available from November 2000 and Aqua data available from 20 January 2002, and VIIRS Suomi (VIRRS SNPP) 375 m data available from 20 January 2012, all to present time (FIRMS FAQ). NASA FIRMS supports an open data policy encouraging use of data and graphics with correct citations.

NCEP/DOE Reanalysis II daily mean relative humidity NetCDF data files at different pressure levels for the continental U.S. and based out of Boulder, Colorado, are taken from the NOAA Physical Sciences Laboratory website via https://downloads.psl.noaa.gov/Datasets/ncep.reanalysis2/Dailies/pressure/. There are no access restrictions or caveats associated with using the data as long as proper citations are implemented.

Landsat-derived geospatial landcover data layers for 1 January 2009 to 1 January 2019 nationally and at 30 m are taken from the United States Department of Agriculture National Agricultural Statistics Service (USDA NASS) website's Research and Science Cropland Data Layer page and from the customer service staff via https://www.nass.usda.gov/Research_and_Science/Cropland/Release/index.php. This data is for public access, under public domain, and free to redistribute with no USDA NASS warrant on any conclusions made from the data.

MODIS Terra Normalized Daily Vegetation Indices at 500m are obtained from the Google Earth Engine Data Catalog via https://developers.google.com/earth-engine/datasets/catalog/MODIS_MOD09GA_006_NDVI. There are no restrictions on the subsequent use, sale, or redistribution of MODIS data and products acquired through the LP DAAC.

MODIS Combined 16-Day NDVI at 250m are obtained from the Google Earth Engine Data Catalog where There are no restrictions on the subsequent use, sale, or redistribution of MODIS data and products acquired through the LP DAAC and it is available via https://developers.google.com/eartengine/datasets/catalog/MODIS_MCD43A4_006_NDV.

Missing average daily NDVI values for one or two dates per year, except 2013, became linearly interpolated based on assuming a linear trend between the two nearest known data values. Those dates requiring linear interpolation of missing daily NDVI values were: 26 August and 7 September 2009, 11 August 2010, 19 January 2011, 11 May 2012, 14 March and 12 December 2014, 28 October 2015, 18-27 February 2016, 24 April 2017, and 5

December 2018. Midwest averaged 16-Day MODIS Terra daily NDVI values filled in the missing NDVI value for 18 February 2016.

Version 3.3.2 of ArcGIS Pro is used for assigning environmental and atmospheric variables to each fire detection point, as well as for clipping all data to the Midwest region. It is preserved at https://pro.arcgis.com/en/pro-app/latest/get-started/download-arcgis-pro.htm, available via basic, standard, or advanced license levels and single or concurrent use licensing provided by creator, professional, or professional plus user types and developed openly at https://developers.arcgis.com/.

Code files, an ArcGIS project file, and some data files are available via the following. Feldpausch, C. OptimizingARIMAModels, GitHub with README, open-source license, 308cait2019, https://github.com/308cait2019/OptimizingARIMAModels-.git

4 Results

4.1 Historical Analysis of Model Variables

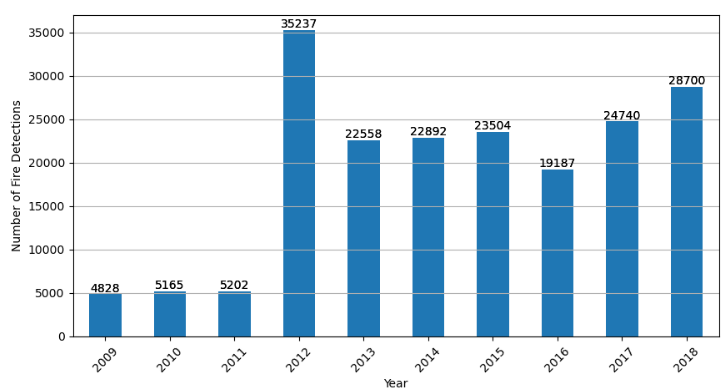

Figure 1. The number of Midwestern U.S. MODIS (Aqua and Terra) and VIIRS SNPP fire detections per year from 1 January 2009 to 1 January 2019.

The highest number of fire detections are recorded in 2012, with a significant jump in data availability from 2012 onwards, likely resulting from VIIRS temporal coverage beginning in January of 2012.

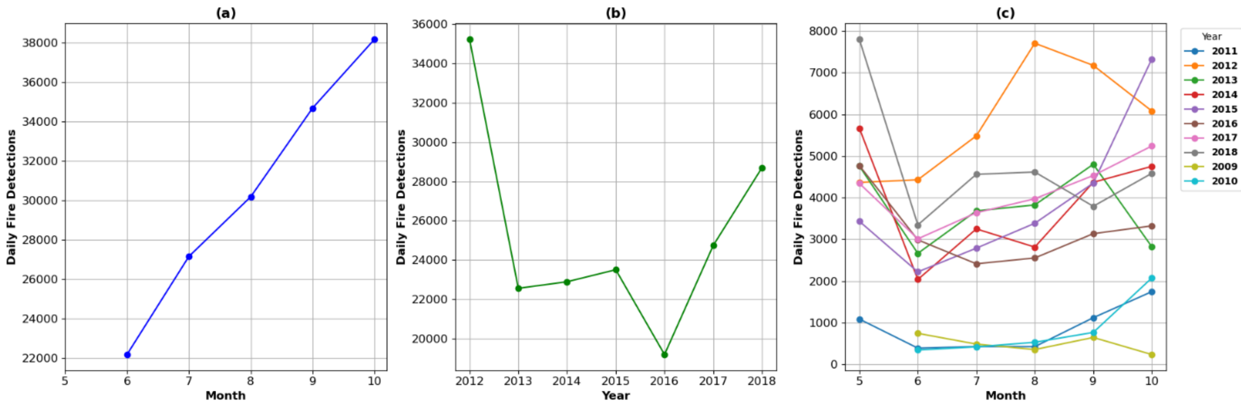

Figure 2. At left for (a), total MODIS/VIIRS SNPP daily fire detections recorded per month. At middle for (b), total MODIS/VIIRS SNPP daily fire detections recorded per year. At right for (c), total MODIS/VIIRS SNPP monthly fire detections recorded per year, all over 1 January 2009 to 1 January 2019.

May and October record the most fire detections, with the least amount in June by Figure 2 (a). There is minimal variation in the number of detections between 2009-2011 and 2013-2015 by Figure 2 (b). All the years show a drop in detections from May to June. By high seasonal variation between the far-right and -left plots, data is nonstationary by initial time-series inspection in Figure 2.

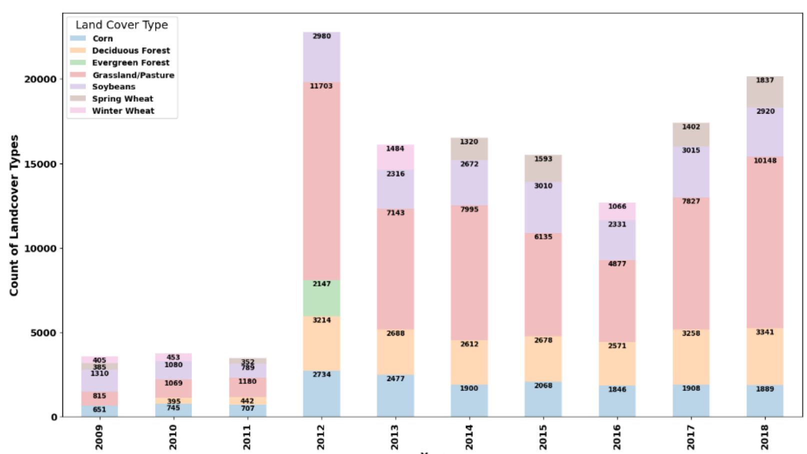

Figure 3. Counts of the top five Midwestern U.S. USDA NASS Cropland Data Layer landcover types per year from 1 January 2009 to 1 January 2019.

A jump in the number of landcover types recorded from 2012 onwards can in part be attributed to Landsat 8 data introduced since 2013 (USDA Crops). Before 2012 there was minimal variation in landcover types present and after 2012 grassland and pastures become the most recorded types with deciduous forest, soybeans, and corn.

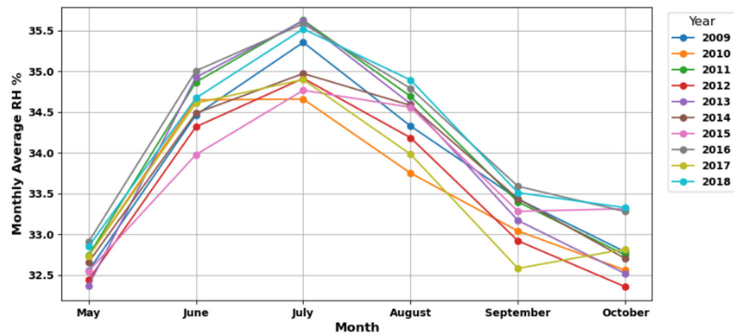

Figure 4. Harvest season monthly average relative humidity percentages (RH %) for 1 January 2009 to 1 January 2019.

There is a trend of higher monthly average relative humidity in July for all years. The start and end of the harvest season experience the lowest average monthly relative humidity percentages. The monthly average relative humidity was within 32% to 36%. The lower the relative humidity, the higher the likelihood of fire.

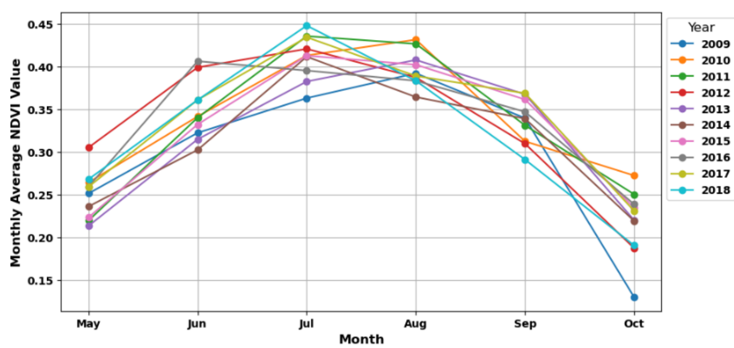

Figure 5. Harvest season monthly average MODIS Normalized Difference Vegetation Index (NDVI) values for 1 January 2009 to 1 January 2019.

All daily NDVI values were averaged over the Midwest region, and then averaged per day and month. A general trend sees indices increase from May to July, peak, and then decrease from August to October.

4.2 Linear Strength Between Fire Detections and Environmental, Weather Variables

Table 1. A Pearson Correlation Coefficient matrix of linear strength between variables. Coefficient values closer to one between variables signify stronger positive linear relationships.

	Daily Fire Detections	**Daily Average RH %**	**NDVI Value**
Daily Fire Detections	1.0	1.0	1.0
Daily Average RH %	1.0	1.0	0.561
NDVI Value	1.0	0.561	1.0

As a one-to-one mapping was done where each row representing a fire detection has an associated daily NDVI value, daily average relative humidity percentage, and landcover type, this results in a directly linear relationship between daily fire detections and the other variables. A one-to-one mapping resulting in a linear transformation is known as an injection or injective function. Despite the above matrix not including landcover types, they were also mapped so each fire detection had an associated yearly landcover type. Hence, landcover types similarly have a direct linear relationship to daily fire detections.

4.3 Achieving Non-Stationary Aggregated Fire Detection Data (with Mapped Environmental and Atmospheric Variables)

Table 2. Augmented Dickey-Fuller test statistics with associated p-values and critical values for monthly, yearly, and both monthly and yearly aggregated fire detection data.

	ADF Test Statistic	**P-Value**	**Critical Values at Significance Percentages**
Monthly Aggregated Data ADF Test Results	0.31680	0.97810	1%: -7.35544 5%: -4.47436 10%: -3.12693
Yearly Aggregated Data ADF Test Results	-2.00029	0.28644	1%: -4.47313 5%: -3.28988 10%: -2.77238
Monthly and Yearly Aggregated Data ADF Test Results	-1.15927	0.69086	1%: -3.57147 5%: -2.92262 10%: -2.59933

In Table 2, as the p-values are all above a chosen significance level of 0.05 or 5%, this suggests a failure to reject the null hypothesis, and the aggregated datasets are still non-stationary. Thus, the aggregated time-series data is non-stationary by earlier historical plots and this ADF test. This test is repeated until the ADF test indicates the aggregated data series are stationary.

Table 3. Augmented Dickey-Fuller test statistics with associated p-values and critical values for monthly, yearly, and both monthly and yearly aggregated fire detection data after orders of differencing have been applied to achieve stationarity.

	Monthly Aggregated Detection Data with Orders of Differencing to Achieve Stationarity	**Yearly Aggregated Detection Data with Orders of Differencing to Achieve Stationarity**			**Monthly & Yearly Aggregated Detection Data with Orders of Differencing to Achieve Stationarity**	
	1st Order Differencing	**1st Order Differencing**	**2nd Order Differencing**	**3rd Order Differencing**	**1st Order Differencing**	**2nd Order Differencing**
ACF Test Statistic	-28.965	-2.277	-1.061	-18.132729	-3.303	-5.624716
P-Value	0	0.179	0.730	0	0.014	0
Critical Values at Significance Level Percentages	1%: -7.355, 5%: -4.474, 10%: -3.127	1%: -4.939, 5%: -3.478. 10%: -2.844	1%: -6.045, 5%: -3.929. 10%: -2.987	1%: -6.045, 5%: -3.929. 10%: -2.987	1%: -3.571, 5%: -2.923. 10%: -2.599	1%: -3.581, 5%: -2.927. 10%: -2.602

After first order differencing, monthly aggregated detection data is stationary. For yearly aggregated detection data this requires third order differencing by the ADF test. For both monthly and yearly aggregated detection data this requires second order differencing. An interesting point is that the critical values at given significance percentages are the same for second and third order differenced yearly aggregated data.

4.4 ARIMA Model Terms Initially Identified

With a Durbin-Watson statist greater than 2 at about 3.474, the first order differenced monthly aggregated fire detection data series is negatively autocorrelated.

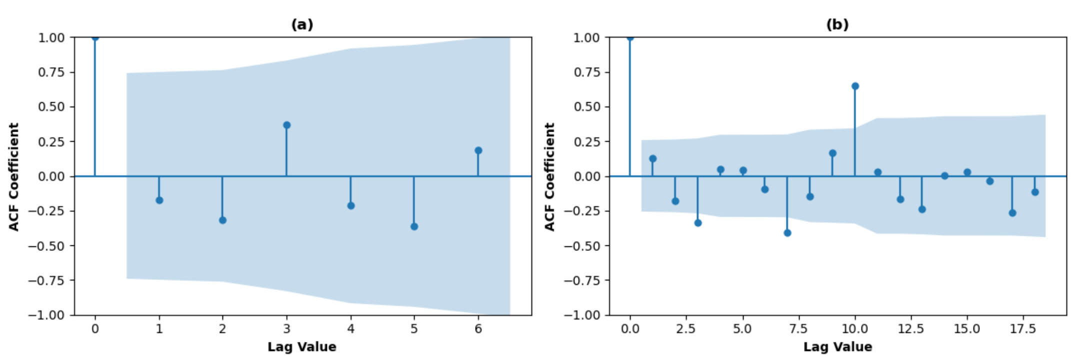

Figure 6. Autocorrelation function (ACF) plots for stationary and higher order differenced yearly and both monthly and yearly aggregated fire detection data. ACF coefficients are plotted at different lag values with a blue confidence band.

Since the autocorrelation function plot coefficients all fall within the blue confidence bands in Figure 6 (a) and (b), this presumes there is likely no autocorrelation in the yearly aggregated and both monthly and yearly aggregated fire detection datasets of higher order differencing. It is evident for now that all our time-series aggregated datasets are stationary by applied differencing and not autocorrelated by the Durbin-Watson test and ACF plots. There will initially be no moving average or autocorrelation terms in the statistically obtained ARIMA model as we have initially determined no autocorrelation present in the aggregated data series.

4.5 The Optimal Non-Seasonal Dynamic and Static ARIMA Models

Table 3. Aggregated data series' MAPE and RMSE scores for non-seasonal ARIMA models as determined by previously decided upon autoregression, differencing, and moving average terms by statistics tests. Aggregated data series' optimal non-seasonal ARIMA model MAPE and RMSE scores based on trial and error to get the lowest MAPE and RMSE combination possible.

Monthly aggregated data ARIMA models	**Statistically obtained ARIMA model**	ARIMA(0,1,0)	MAPE: 30.81% RMSE: 1261.551
	Optimized statistically obtained ARIMA model	**ARIMA(1,1,0)**	**MAPE: 11.26%** **RMSE: 440.727**
Yearly aggregated data ARIMA models	**Statistically obtained ARIMA model**	ARIMA(0,3,0)	MAPE: 2060.83% RMSE: 165281.946
	Optimized statistically obtained ARIMA model	ARIMA(0,1,0)	MAPE: 110.72% RMSE: 6435.568

Monthly and yearly aggregated data ARIMA models	**Statistically obtained ARIMA model**	ARIMA(0,2,0)	MAPE: 842.33% RMSE: 14671.846
	Optimized statistically obtained ARIMA model	ARIMA(0,1,1)	MAPE: 122.10% RMSE: 2657.964

After some trial and error, the optimal dynamic ARIMA model for monthly and yearly aggregated data ends up being $ARIMA(0,1,1)$, a basic random walk model with exponential smoothing. A lower MAPE and RMSE error was achieved when changing the model to first order differencing instead of the second order used to make the dataset stationary and also by adding an MA term. The yearly aggregated data is also transitioned back to first order differencing for significantly better error scores with a dynamic random walk model. Yet, both aggregated dataset lowest MAPE and RMSE scores achieved are still significantly high in the hundreds or thousands, and thus still not ideal for forecasting. For this reason, the monthly aggregated data with lowest ARIMA error MAPE and RMSE scores is the most optimal for non-seasonal, dynamic forecasting.

Again, through some trials with series differencing knowledge and error analysis, the optimal monthly aggregated data from the dynamically validated ARIMA model is also the optimal for static modeling. There are most optimal errors of a MAPE at about 5% and an RMSE at about 299 for the static $ARIMA(1,1,0)$ model.

4.6 Re-introducing Seasonality and a Forecast

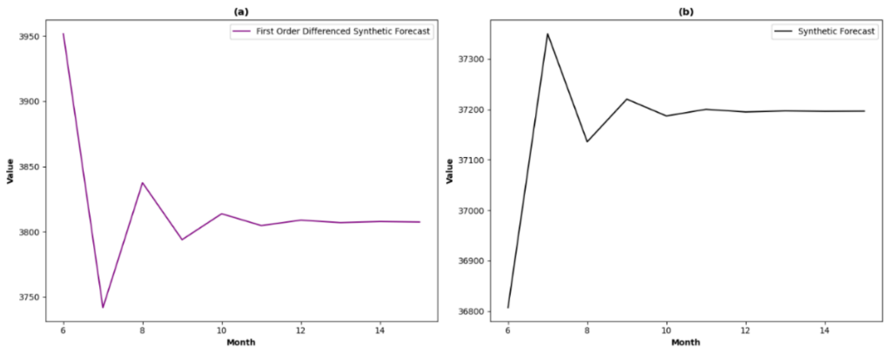

Figure 7. The SARIMA first order differenced monthly aggregated synthetic data forecast is (a) an the SARIMA monthly aggregated synthetic data forecast is (b), both forecasts for 10 months into the future.

5 Conclusions

Since an MA term is added to the optimal aggregated monthly and yearly fire detection non-seasonal ARIMA model, autocorrelation was previously misinterpreted as nonexistent for that data series. Due to differencing levels changing to first order for the yearly aggregated and both monthly and yearly aggregated fire detection data series, overdifferencing to achieve stationarity was done earlier in the statistics and term determination process.

The best dataset ends up being the monthly aggregated fire detection data series with a first order difference to achieve stationarity and an autoregressive term, denoted as $ARIMA(1,1,0)$. It accounts for possible negative correlation found by the Durbin-Watson test. The MAPE and RMSE are the lowest out of all the aggregated dataset ARIMA models.

The ARIMA dynamic and static models do exhibit errors when working with smaller datasets and trying to increase parameters past a value of 3. For this reason, data was not split to before and after 2012, when VIIRS data became available causing a spike in detections.

The 10-month forecast of synthetic monthly aggregated fire detections with $ARIMA(1,1,0)$ and a seasonal term of 12 applied to first order differenced aggregated data and the aggregated data separately shows interesting trends. The number of detections for both plots appear to be converging towards a potential stabilization or equilibrium in the underlying trend. There are significant fluctuations in the first couple of forecast months. Differencing appears to have inversely affected the first few forecasts, showing how stationarity affects dataset predictions. Thus, detection forecasts past 10 months will be at about 3700 to 3800 consistently due to convergence, and therefore this optimized model with a fitted seasonal term is best for short-term forecasting.

- Code files, an ArcGIS project file, and some data files are available via the following. Feldpausch, C. OptimizingARIMAModels, GitHub with README, open-source license, 308cait2019, https://github.com/308cait2019/OptimizingARIMAModels-.git
- Version 3.3.2 of ArcGIS Pro is used for assigning environmental and atmospheric variables to each fire detection point, as well as for clipping all data to the Midwest region. It is preserved at https://pro.arcgis.com/en/pro-app/latest/get-started/download-arcgis-pro.htm, available via basic, standard, or advanced license levels and single or concurrent use licensing provided by creator, professional, or professional plus user types and developed openly at https://developers.arcgis.com/.

6 References

Crozier, W. W., Schön, P-J, & Prévost, E. (2005, June). Static versus dynamic model for forecasting salmon pre-fishery abundance of the River Bush: a Bayesian comparison. ScienceDirect. *Fisheries Research*, *71*(1-2). https://doi.org/10.1016/j.fishres.2005.01.002

Dodge, Y. (2008). Durbin–Watson Test. Springer, New York, NY. *The Concise Encyclopedia of Statistics*. https://doi.org/10.1007/978-0-387-32833-1_122

Nau, R. (2020b, August 18). Identifying the numbers of AR or MA terms in an ARIMA model. ARIMA Models for Time Series Forecasting; Duke University Fuqua School of Business. https://people.duke.edu/~rnau/411arim3.htm

Nau, R. (2020a, August 18). Introduction to ARIMA: nonseasonal models. ARIMA Models for Time Series Forecasting; Duke University Fuqua School of Business. https://people.duke.edu/~rnau/411arim.htm

Nau, R. (2014, October 30). Notes on nonseasonal ARIMA models. Statistical forecasting: notes on regression and time series analysis; Duke University Fuqua School of Business. https://people.duke.edu/%7Ernau/Notes_on_nonseasonal_ARIMA_models--Robert_Nau.pdf

Nau, R. (2020c, August 18). Random walk model. Simple forecasting models; Duke University Fuqua School of Business. https://people.duke.edu/~rnau/411rand.htm

Romer, M. (2024a). Lesson 3: Identifying and Estimating ARIMA models; Using ARIMA models to forecast future values. Applied Time Series Analysis | STAT 510; Pennsylvania State University Eberly College of Science. https://online.stat.psu.edu/stat510/lesson/3

Romer, M. (2024b). 4.1 Seasonal ARIMA models. Applied Time Series Analysis | STAT 510; Pennsylvania State University Eberly College of Science. https://online.stat.psu.edu/stat510/lesson/4/4.1

Simon, L., Heckard, R., Wiesner, A., Young, D., Pardoe, L. (2024). T.2.3 - Testing and Remedial Measures for Autocorrelation | STAT 501; Pennsylvania State University Eberly College of Science. https://online.stat.psu.edu/stat501/lesson/t/t.2/t.2.3-testing-and-remedial-measures-autocorrelation

Zhang, X. [Xiaomeng], & Zhang, X [Xinyu]. (2022, March 20). Optimal model averaging based on forward-validation. ScienceDirect. *Journal of Economics*, *237*(2), 1-26. https://doi.org/10.1016/j.jeconom.2022.03.010