THE VARAIBILITY CLIMATE CHANGE IS FESPONSIBLE FOR IN VEGETATION LOSS IN GHANA

Quantifying The Status of Galamsey With Time Series Analsis

Kalong Boniface   Fugah Seletey Mitchell

Matti Meikäläinen

2022-09-05

Table of Contents

List of Figures

List of Tables

1 CHAPTER ONE

1.1 INTRODUCTION

One would anticipate that the majority of emerging nations, which are still in the early stages of economic development and growth, would have a high forest cover and little deforestation. This, however, has not been the case. Ghana is a lower-middle-income nation that is still working toward middle-income classification. However, it has already begun to see a deforestation rate that is comparable to that of middle-income countries. The rapid population expansion, clearing of feild for Galamsey operation,increased domestic need for wood for things like fuel, furniture, construction, and timber exports have all contributed to this trend, Bush fires in the 1980s, climate change, and lax law enforcement have all had an impact.

The purpose of this paper is to establish an understanding in time series analysis on remotely sensed data. Which will introduced us to the fundamentals of time series modeling, including decomposition, autocorrelation and modeling historical changes in **Vegetation Loss** in Ghana as a result of **Galamsey Operation** and the **Variability Climate Is Responsible For**, the Cause, Dangers and it's Environmental impact.

Galamsey also known as "*gather them and sell*",Owusu-Nimo et al. (2018) is the term given by local Ghanaian for illegal small-scale gold mining in Ghana . The major cause of Galamsey is unemployment among the youth in Ghana Gracia (2018) . Young university graduates rarely find work and when they do it hardly sustains them. The result is that these youth go the extra mile to earn a living for themselves and their family.

Another factor is that lack of job security. On November 13, 2009 a collapse occurred in an illegal, privately owned mine in Dompoase, in the Ashanti Region of Ghana. At least 18 workers were killed, including 13 women, who worked as porters for the miners. Officials described the disaster as the worst mine collapse in Ghanaian history "Women Die in Ghana Mine Collapse" (2009) .

Illegal mining causes damage to the land and water supplym (Ansah (2017)) . In March 2017, the Minister of Lands and Natural Resources, Mr. John Peter Amewu, gave the Galamsey operators/illegal miners a three-week ultimatum to stop their activities or be prepared to face the law Allotey (2017) . The activities by Galamseyers have depleted Ghana's forest cover and they have caused water pollution, due to the crude and unregulated nature of the mining process Gyekye (n.d.) .

Under current Ghanaian constitution, it is illegal to operate as galamseyer.That is to dig on land granted to mining companies as concessions or licenses and any other land in search for gold. In some cases, Galamseyers are the first to discover and work extensive gold deposits before mining companies find out and take over. Galamseyers are the main indicator of the presence of gold in free metallic dust form or they process oxide or sulfide gold ore using liquid mercury.

Between 20,000 to 50,000, including thousands from China are believed to be engaged in Galamsey in Ghana.But according to the Information Minister 200,000 and nearly 3 million people, recently are now into Galamsey operation and rely on it for their livelihoods "Gold, Guns and China: Ghana's Fight to End Galamsey" (2017) . Their operations are mostly in the southern part of Ghana where it is believe to have substantial reserves of gold deposits, usually within the area of large mining companies Barenblitt et al. (2021) . As a group, they are economically disadvantaged. Galamsey settlements are usually poorer than neighboring agricultural villages. They have high rates of accidents and are exposed to mercury poisoning from their crude processing methods. Many women are among the workers, acting mostly as porters for the miners.

1.1.1 Background of The Study

As Galamsey is considered an illegal activity, they operations are hidden to the eyes of the authorities.So locating them is quite tricky ,but with satellite imagery ,it now possible to locate their operating and put an end to it. One of the features of Google Earth Engine is the ability to access years of satellite imagery without needing to download, organize, store and process this information. For instance, within the Satellite image collection, now it possible to access imagery back to the 90's, allowing us to look at areas of interest on the map to visualize and quantify how much things has changed over time. With Earth Engine, Google maintains the data and offers it's computing power for processing.Users can now access hundreds of time series images and analyze changes across decades using GIS and R or other programming language to analyze these dataset.

1.1.2 Problem Statement

The Footprint of Galamsey is Spreading at a very faster rate, causing vegetation loss.Other factors accounting to vegetation loss may largely include climate change,urban and exurban development, bush fires. But not much works or research has been done to tell the extent to which Galamsey causes vegetation loss and the **Variability Climate Change Is Responsible For**. This research attempts to segregate the variability climate is responsible for in vegetation loss so as to attribute the residual variability to Galamsey and other related activities such as bush-fires etc.

1.1.3 Research Questions

To address the challenge of the vegetation variability in this work, the following several statements were formed:

- Are there any changes in vegetation cause by Galamsey and Climate change in Ghana?

- Is there any relationship between vegetation and land surface temperature in Ghana?

1.1.4 Research Objectives

The purpose is to establish an understanding in time series analysis on remotely sensed data. We will be introduced to the fundamentals of time series modeling, including decomposition, autocorrelation and modeling historical changes.

- Perform time series analysis on satellite derived vegetation indices

- Estimate the extent to which Galamsey causes vegetation loss in Ghana.

- Dissociate or single out the variability climate is responsible for in vegetation loss

1.1.5 Significance Of The Study

There have been significant changes in vegetation cover in Ghana over the past 30 years, and these dynamics are related strongly to climatic factors such as temperature and other factors. In this study, we want to examine the effects of climatic change on Ghana's vegetation during these thirty years.

This study allows us to explore climatic differences and climate-related drivers. Additionally, it offers a chance to research how climatic variability affects the ecosystem and human health. By merging climate and vegetation variation utilizing NDVI, LST, and EVI data to understand the relationship between vegetation and climate change under tropical climate conditions, it closes research gaps in Ghana. This study explores historical and projected vegetation and climate data, by sector, impacts, key vulnerabilities and what adaptation measures can be taken. It also explores the overview for a general context of how climate change is affecting **Ghana.**

1.1.6 Scope of The Study

1.1.7 Limitation Of The Study

The goal of time series modeling is to employ the simplest model feasible to account for as much data as possible while still developing an explanatory model of the data that does not over-fit the issue set.

Remote sensing data has additional limits that make this more difficult when dividing time series data into component pieces. It is almost certain that data from distant sensing will not provide the same level of precision.

Additionally, atmospheric factors can distort the visual findings, causing the vegetation's color to shift dramatically from image to image as a result of atmospheric factors (fog, ground moisture, cloud cover).

1.1.8 Organization of The Study

2 CHAPTER TWO

2.1 LITERATURE REVIEW

The distribution of plant species, the richness and composition of plant communities, the structure of the vegetation (such as biomass and leaf area), and how the ecosystem uses water, nutrients, and carbon are all predicted to change as a result of climate change. These plant responses to climate change will be the outcome of numerous lower-level plant reactions, such as adjustments in net plant carbon uptake, adjustments in plant water usage, adjustments in plant growth and biomass allocation, competitive interactions, and reactions to disturbances. It is challenging to predict prospective plant reactions to future climatic changes based only on theory or on laboratory and field studies due to the complexity of climatic impacts on vegetation and the length of time it takes for the responses to become obvious.

To project vegetation responses to changing climates, computer simulation models that integrate theory and experimental results are frequently used, and the following are some studies that have been done previously where we reclassifies the drivers into human activity, and climate change for an empirical review and None Parametric Test followed by Time Series Analysis for the theoretical review .

2.1.1 Empirical Review

According to studies, there is now significant change in vegetation on the earth than there was thirty years ago, and it is distributed differently.

More than half of the changes they found are attributed to the consequences of a warmer climate, with people only being responsible for about a third. Perhaps surprisingly, they are unable to definitively link approximately 10% of the changes to either the climate or us.@alex2013

Several models and hypotheses have been established in the environmental literature to explain the relationship between human behaviour, and environmental (forest) deforestation or depletion. Recent environmental and energy economics literature focuses on the energy consumption choices made by businesses and people in developing countries Gertler et al. (2016) . Africa's energy supply is made mainly of fuel wood and charcoal to a degree of about 58%. Specht et al. (2015) . Before other demands for forest goods

like furniture and paper, the need for fuel wood for cooking and heating is frequently identified as the main driver of deforestation.

The causes of tropical forest decline are unclear, according to DeFries et al. (2010) . However, scientists were able to pinpoint the two primary causes of deforestation in the 21st century using information from satellite-based estimations in 41 different countries. The authors found a favorable association between forest loss and increases in urban population as well as agricultural exports using two methods of regression analysis. The same proof, however, was not discovered in the case of the increase in rural population. This suggests that forest loss is unavoidable in regions with high levels of human activity.

2.1.2 Theoretical Review

This study review, will follow narrative approach to gain insight into research topics. A time series is a set of observations, each being recorded at a particular time and the collection of such observation is referred to as time series data. The data is analysed to extract statistical information, characteristics of the data and to predict the output. As the data might tend to follow a pattern in time series data, the Machine Learning model finds it difficult to predict appropriately hence time series analysis and its approaches have made it simpler for prediction.

2.1.2.1 What is Time Series Analysis?

A time series in mathematics is a collection of data points that have been listed, graphed, or indexed according to time. is a series of photos taken at successive, evenly spaced intervals of time. Time series are utilized in many areas of applied research that use temporal data, including statistics, signal processing, pattern identification, econometrics, mathematical finance, weather forecasting, and earthquake prediction. Time series analysis refers to techniques for deriving useful statistics and other aspects of time series data through analysis. Time series forecasting is the process of using a model to forecast future values based on values that have already been observed. Regression analysis is frequently used to assess correlations between one or more different time series, however this method of analysis is not without its limitations.

In 1987 "Confidence Intervals: An Empirical Investigation of the Series in the m-Competition - ScienceDirect" (n.d.) , Makridakis and Hibon, time series analysis experts, held the M-Competition, where participants may submit their forecasts on 1001 time series data drawn from economics, industry, and demographics. The competition revealed four key findings, including:

- Statistically sophisticated or complex methods do not necessarily provide more accurate forecasts than simpler ones.

- The relative ranking of the performance of the various methods varies according to the accuracy measure being used.

- The accuracy when various methods are being combined outperformed, on average, the individual methods being combined and does very well in comparison to other methods.

- The accuracy of the various methods depends upon the length of the forecasting horizon involved.

The time series data is visualized and analyzed to find out mainly three things, trend, seasonality, and Heteroscedasticity.

Trend: It can be characterized as the observation of a rising or escalating pattern throughout time. While in normal time series the mean is an arbitrary function of time, in stationary time series the mean of the data must be constant across time.

Seasonality: This term describes a cycle of events. a pattern that, after some time, keeps happening.

Heteroscedasticity: It is also referred to as level, and it is described as the non-constant variance from the mean computed over time.

Few approaches do not perform well when trends are present in the data, and even fewer do not perform well when the data is seasonal. In order to choose the optimal statistical method for forecasting, trends, seasonality, and heteroscedasticity must be taken into account.

2.1.2.2 Time Series Forecasting Using Stochastic Models

The selection of a proper model is extremely important as it reflects the underlying structure of the series and this fitted model in turn is used for future forecasting. A time series model is said to be linear or non-linear depending on whether the current value of the series is a linear or non-linear function of past observations.

In general models for time series data can have many forms and represent different stochastic processes. There are two widely used linear time series models in literature.

Autoregressive (AR) and *Moving Average (MA)* models, combining these two, the Autoregressive Moving Average (ARMA) and *Autoregressive Integrated Moving Average (ARIMA)* models have been proposed in many literature. The *Autoregressive Fractionally Integrated Moving Average (ARFIMA)* model generalizes ARMA and ARIMA models. For seasonal time series forecasting, a variation of ARIMA. The *Seasonal Autoregressive Integrated Moving Average (SARIMA)*] model is used.

ARIMA model and its different variations are based on the famous Box-Jenkins principle Hipel and McLeod (1994) and so these are also broadly known as the Box-Jenkins models.

Linear models have drawn much attention due to their relative simplicity in understanding and implementation. However many practical time series show non-linear patterns. For example, as mentioned by R. Parrelli in , non-linear models are appropriate for predicting volatility changes in economic and financial time series. Considering these facts, various non-linear models have been suggested in literature. Some of them are the famous Autoregressive Conditional Heteroskedasticity (ARCH) model and its variations like Generalized ARCH (GARCH) , Exponential Generalized ARCH (EGARCH) etc., the Threshold Autoregressive (TAR) model, the Non-linear Autoregressive (NAR) model, the Non-linear Moving Average (NMA) model, All the methods consider either of trend, seasonality, or heteroscedasticity to predict the future output. Time series data must be decomposed based on the findings from data analysis. Based on the findings from analysis data must be broken into trend or seasonality. Zhang, Eddy Patuwo, and Y. Hu (1998)

2.1.2.2.1 Exponential Smoothing Models:

Time-series data relies on the assumption that the observation at a certain point of time depends on previous observations in time . The previous observations are given weights as they contribute to the future prediction. The process of weighting is done using a notation called '**Theta**'. To find the best possible value for theta, we must perform sum of squared errors between the actual versus predicted value of the previous observation. Using this process, we can predict the next value but to predict more than one value this process does contribute much as the prediction as going to be same as the previous value.

To understand the methods and to evaluate different models, few concepts like *stationarity* and *differencing* must be understood. Both these concepts help in making the core concepts of the methods easy to interpret.

2.1.2.2.2 Stationarity:

Stationarity alludes to an irregular process that creates a time-series which has mean, and distribution to be constant through time. Distribution only depends on time and not location in time (Manuca, R. and Savit, R., 1996). If the distribution is same over different time windows is strong stationarity and if only mean and variance are similar, then it is weak stationarity. Irrespective of strong or weak, stationarity helps build a class of models such Autoregression (AR), Moving Average (MA), ARIMA (Witt, A., Kurths, J. and Pikovsky, A., 1998).

An MA(q) process is always stationary, irrespective of the values the MA parameters [23]. The conditions regarding stationarity and invertibility of AR and MA processes also hold for an ARMA process. An ARMA(p, q) process is stationary if all the roots of the characteristic equation $\phi(L) = 0$ lie outside the unit circle. Similarly, if all the roots of the lag equation

$\theta(L) = 0$ lie outside the unit circle, then the ARMA(p, q) process is invertible and can be expressed as a pure AR process..

2.1.2.2.3 Differencing:

This concept is used to make trending and seasonal data stationary. Subtraction between current observation and previous observation is the process of differencing. It helps in making the mean constant (Dickey, D.A. and Pantula, S.G., 1987).

2.1.2.2.4 Autoregressive models (AR):

AR work on a concept called lags which is defined as the forecast of a series is solely based on the past values in the series (Cryer, J.D., 1986). Formula for Autoregression AR(1): $y_t = \omega + \phi Y_{t-1} + e_t$ is stationary when $|\_{1}| < 1$ with a constant mean $\mu = \dfrac{\omega}{1 - \phi_1}$ and constant variance $\gamma_o = \dfrac{\sigma^2}{1 - \phi_1^2}$ Where ; y_t = Target , ω = Intercept, ϕ = Coefficient, Y_{t-1}=Lagged target, e_t = Error\

It depends only on one lag in the past and also called AR model of order one (Shibata, R., 1976). Autoregressive models are also known as long memory models as they must keep the memory of all the lags until its initial start point and must calculate their value. If there is any shock incident in the past which must have led to fluctuations in the data, it will have its effect on the present value which makes the model quite sensitive to shocks (Shibata, R., 1976).

2.1.2.2.5 Moving Average (MA):

The moving average model forecasts a series based on the past error in the series called error lags. Hunter, J.S., Formula for moving average method is given as: $y_t = \omega + \theta e_{t-1} + e_t$

In (2), all the abbreviations are same to AR model formula except, = Previous error

There arises a question as this method uses the error for the previous value but when it reaches to the first point there will be no previous value, to overcome this the average of the series is considered as the value before the starting point. These are short memory models as the error in the past will not have much effect on the future value (Hunter, J.S., 1986).

2.1.2.2.6 Comparing AR method with MA method:

Let focus on the two methods which were used in the early years of time series forecasting and compare the performance of each model on a particular task. Testing against general autoregressive and moving average error models where the regressors include lagged dependent variables. (Godfrey, L.G., 1978) In their paper have explained the order of the error process under the alternate hypothesis using lagrange multiplier test (Silvey, S.D., 1959). As per the tests the errors of both the models were similar, but the constraints were different under which the tests were performed are also to be considered. As they have concluded in their paper stating that that the outcome of the model's performance depends on the estimate chosen to be null hypothesis or alternate hypothesis.

In addition, paper written by (Baltagi, B.H. and Li, Q., 1995), Demonstrates the comparison of AR and MA model using Burke, Godfrey, and Termayne test. To the error component model. They explained choosing of this test is because these are simple to implement as they only require within residual or OLS residual (Baltagi, B.H. and Li, Q., 1995). The outcome of the experiment was explained as when the test used within residual AR model performed well but had problems, if the test used OLS residual MA model performance was good. They have concluded stating that MA will performance much better when the parameters are changed.

The findings of both the paper were quite different but one cannot prove either of the model to be better as the performance depends on the parameters used in the model. Each model is unique to its use case, and it depends on the user to choose accordingly based on the data.

2.1.2.2.7 Autocorrelation and Partial Autocorrelation Functions (ACF and PACF)

To determine a proper model for a given time series data, it is necessary to carry out the ACF and PACF analysis. These statistical measures reflect how the observations in a time series are related to each other. For modeling and forecasting purpose it is often useful to plot the ACF and PACF against consecutive time lags. These plots help in determining the order of AR and MA terms. For a time series $x(t), t = 0, 1, 2, ...$ the Autocovariance [21, 23] at lag k is defined as:

μ is the mean of the time series, i.e. $\mu = E[x_t]$. The autocovariance at lag zero i.e. y_0 is the variance of the time series. From the definition it is clear that the autocorrelation coefficient p_{k} is dimensionless and so is independent of the scale of measurement. Also, clearly $-1 \leq p\_k \leq 1$. Statisticians Box and Jenkins [6] termed y_k as the theoretical Autocovariance Function (ACVF) and p_k as the theoretical Autocorrelation Function (ACF).

Another measure, known as the Partial Autucorrelation Function (PACF) is used to measure the correlation between an observation k period ago and the current observation, after controlling for observations at intermediate lags (i.e. at lags < k) [12]. At lag 1,

PACF(1) is same as ACF(1). The detailed formulae for calculating PACF are given in [6, 23].

Normally, the stochastic process governing a time series is unknown and so it is not possible to determine the actual or theoretical ACF and PACF values. Rather these values are to be estimated from the training data, i.e. the known time series at hand. The estimated ACF and PACF values from the training data are respectively termed as sample ACF and PACF [6, 23].

As given in [23], the most appropriate sample estimate for the ACVF at lag k is ACF plot is useful in determining the type of model to fit to a time series of length N. Since ACF is symmetrical about lag zero, it is only required to plot the sample ACF for positive lags, from lag one on-wards to a maximum lag of about N/4. The sample PACF plot helps in identifying the maximum order of an AR process.

2.1.2.3 Autoregressive Moving Average (ARMA) model:

ARMA model is a combination of AR and MA models. The equation of the AR model of order one, when it reaches to the starting point will have infinite moving average (Choi, B., 2012). In ARMA model p and q have to defined, where p = number of significant terms in ACF and q = number of significant terms in PACF.

To determine the optimal value for p and q there are two ways:

- Plotting patterns in correlation
- Automatic selection techniques

2.1.2.3.1 Plotting patterns in correlation:

2.1.2.3.1.1 Auto correlation factor (ACF):

It is the correlation between the observations at the current time stamp and observations at the previous time stamp (Hagan, M.T. and Behr, S.M., 1987).

2.1.2.3.1.2 Partial auto correlation factor (PACF):

The correlation between the observations at two different time stamps, assuming both observations are correlated to the observations at another time stamp (Hagan, M.T. and Behr, S.M., 1987).

2.1.2.3.2 Automatic selection techniques:

There are three commonly used techniques for automatic selection of time series model:

2.1.2.3.2.1

- *Minimum info criteria (MINIC):*This builds multiple combinations of models across a grid search of AR and MA terms. It then finds the model with lowest Bayesian information criteria (Stadnytska, T., Braun, S. and Werner, J., 2008).

- *Squared canonical correlations (SCAN):* It looks at correlation matrix of the data, then it compares it with its lags. It then looks at the eigen values from the correlation matrix to find the combination of AR and MA probably having SCAN as 0. It finds the pair as the best where the convergence is quickest (Stadnytska, T., Braun, S. and Werner, J., 2008).

- *The extended sample auto correlation function (ESACF):* As it is known that AR and MA are related. Essentially it filters out the AR terms until only MA piece is left. This process is repeated until fewest AR terms are left and maximum MA terms (Stadnytska, T., Braun, S. and Werner, J., 2008).

  It completely depends on the individual to choose from either of the methods helping them to find the optimal value of p and q for better performance of the model.

2.1.2.4 Autoregressive Integrated Moving average (ARIMA):

To understand ARIMA model, we need to understand ARMA model as this is just an extension to ARMA model. Essentially, we need to make data stationary to feed it to a machine learning model. It is done by through differencing. ARIMA models are mathematically written as $\boldsymbol{ARIMA(p,d,q)}$, where p and q are same as ARMA model but $\boldsymbol{d}$ = number of first differences (Yu, G. and Zhang, C., 2004, May).

2.1.2.5 Seasonal Autoregressive Integrated Moving Average (SARIMA):

SARIMA models were introduced to handle seasonality in the data. Seasonality is different from stationarity; however, seasonality can be handled using stationarity up to some extent, but seasonal correlations cannot be eliminated completely. SARIMA models are mathematically written as $SARIMA(p, d, q)(P, D, Q)^s$.

Where;

P = Number of seasonal AR terms, D = Number of seasonal differences, Q = Number of seasonal MA terms, s = Length of the season.

Removing seasonality will help the model to perform better but getting rid of seasonality in data is a difficult task to do.

2.1.2.6 Comparing ARIMA method with SARIMA method:

Library and Valipour (2015) investigated it on long-term runoff forecasting in the United States in contrast to ARIMA and SARIMA. The outcomes demonstrated that SARIMA models outperformed ARIMA models. However, it was discovered that SARIMA models were extremely sensitive, and even a small change in a parameter would have a negative impact on the model's performance. ARIMA and SARIMA models have been applied by (Wang, S., Li, C., and Lim, 2019) from the perspectives of Linear System Analysis, Spectra Analysis, and Digital Filtering. The researchers were obliged to go outside of these models for greater performance after it was established that ARIMA and SARIMA both had poor performance.Although they claimed the ARMA-SIN model was superior to the ARIMA and SARIMA models, they also acknowledged that it was more challenging to study and comprehend the ideas.

The results of Library and Valipour (2015) have demonstrated that SARIMA is superior; however, their assertion is incongruent when it is contrasted with the results of (Wang, S., Li, C. and Lim, A., 2019).

The choice of a particular approach must be based on the data; following analysis, it is known whether the data have a trend, in which case ARIMA must be used, or whether they are seasonal, in which case SARIMA would be beneficial.

2.1.2.7 ADVANTAGES AND DISADVANTAGES OF TIME SERIES FORECASTING

Advantages of time series forecasting:

- Time series forecasting is of high accuracy and simplicity.

- It can be used to analyze how the changes associated with the data point picked correlate with changes in other variables during the same time span.

- Statistical techniques have been developed to analyze time series in such a way that the factor that influences the fluctuation of the series may be identified and handled.

- It can give good output with less variables. As regression models fail with less variables, time series models will work better and effectively.

Disadvantages of time series forecasting:

- Time series models can easily be overfitted, which lead to false results.

- It works well with short term forecasting but does not work well with long term forecasting.

- It is sensible to outliers, if the outliers are not handled properly then it could lead to wrong predictions.

- The different elements that impact the fluctuations of a series cannot be fully adjusted by the time series analysis

2.1.2.8 Forecast Performance Measures

While applying a particular model to some real or simulated time series, first the raw data is divided into two parts (**Training Set and Test Set**). The observations in the training set are used for constructing the desired model. Often a small sub-part of the training set is kept for validation purpose and is known as the **Validation Set**. Sometimes a preprocessing is done by normalizing the data or taking logarithmic or other transforms. One such famous technique is the Box-Cox Transformation [23]. Once a model is constructed, it is used for generating forecasts. The test set observations are kept for verifying how accurate the fitted model performed in forecasting these values. If necessary, an inverse transformation is applied on the forecast values to convert them in original scale. In order to judge the forecasting accuracy of a particular model or for evaluating and comparing different models, their relative performance on the test dataset is considered.

Due to the fundamental importance of time series forecasting in many practical situations, proper care should be taken while selecting a particular model. For this reason, various performance measures are proposed in literature [3, 7, 8, 9, 24, 27] to estimate forecast accuracy and to compare different models. These are also known as performance metrics [24]. Each of these measures is a function of the actual and forecast values of the time series.

2.1.2.8.1 Description of Various Forecast Performance Measures

In each of the forthcoming definitions, y_t is the actual value, f_t is the forecaste value, $e_t = y_t - f_t$ is the forecast error and n is the size of the test set. Also, $\bar{y} = \dfrac{1}{n} \sum_{t=1}^{n} y_t$ is the test mean and $\sigma^2 = \dfrac{1}{n-1} \sum_{t=1}^{n} (y_t - \bar{y})^2$ is the test variance.

2.1.2.8.1.1 The Mean Forecast Error $MFE = \dfrac{1}{n} \sum_{t=1}^{n} e_t$

- It is a measure of the average deviation of forecasted values from actual ones.

- It shows the direction of error and thus also termed as the Forecast Bias.

- In MFE, the effects of positive and negative errors cancel out and there is no way to

know their exact amount.

• A zero MFE does not mean that forecasts are perfect, i.e. contain no error; rather it only

indicates that forecasts are on proper target.

• MFE does not panelize extreme errors.

• It depends on the scale of measurement and also affected by data transformations.

• For a good forecast, i.e. to have a minimum bias, it is desirable that the MFE is as close

to zero as possible.

2.1.2.8.1.2 The Mean Absolute Error $MAE = \dfrac{1}{n} \sum\limits_{t=1}^{n} |e_t|$

• It measures the average absolute deviation of forecast values from original ones.

-It is also termed as the Mean Absolute Deviation (MAD).

• It shows the magnitude of overall error, occurred due to forecasting.

• In MAE, the effects of positive and negative errors do not cancel out.

• Unlike MFE, MAE does not provide any idea about the direction of errors.

• For a good forecast, the obtained MAE should be as small as possible.

• Like MFE, MAE also depends on the scale of measurement and data transformations.

• Extreme forecast errors are not panelized by MAE.

2.1.2.8.1.3 The Mean Squared Error $MSE = \dfrac{1}{n} \sum\limits_{t=1}^{n} e_t^2$

• It is a measure of average squared deviation of forecast values.

• As here the opposite signed errors do not offset one another, MSE gives an overall idea of the error occurred during forecasting.

• It panelizes extreme errors occurred while forecasting.

• MSE emphasizes the fact that the total forecast error is in fact much affected by large

• individual errors, i.e. large errors are much expensive than small errors.

- MSE does not provide any idea about the direction of overall error.

- MSE is sensitive to the change of scale and data transformations.

- Although MSE is a good measure of overall forecast error, but it is not as intuitive and

- easily interpretable as the other measures discussed before.

2.1.2.8.1.4 The Root Mean Squared Error $RMSE = \sqrt{MSE} = \sqrt{\dfrac{1}{n}\sum\limits_{t=1}^{n} e_t^2}$

RMSE is nothing but the square root of calculated MSE.

All the properties of MSE hold for RMSE as well.

3 CHAPTER THREE

3.1 METHODOLOGY

Data from a time series is a set of observations made in a particular order over a period of time. There is a chance for correlation between observations because time series data points are gathered at close intervals. To help machine learning classifiers work with time series data, we provide several new tools. We first contend that local features or patterns in time series can be found and combined to address challenges involving time-series categorization. Then, a method to discover patterns that are helpful for classification is suggested. We combine these patterns to create computable categorization rules. In order to mask low-quality pixels, we will first collect data from Google Earth Engine in order to choose NDVI, EVI values and Climate Change data.

Instead of analyzing the imagery directly, we will summarize the mean NDVI and EVI values. This will shorten the analysis time while still providing an attractive and useful map. We will apply a smoothing strategy using an ARIMA function to fix the situation where some cells may not have NDVI and EVI for a particular month. Once NA values have been eliminated, the time series will be divided to eliminate seasonality before the normalized data is fitted using a linear model. We will go to classify our data on the map and map it after we have extracted the linear trend.

3.2 Research Design

In this study, the submission used a quantitative approach. Instead of using subjective judgment, findings and conclusions heavily rely on the use of statistical methods and reliable time series models.

3.2.1 Data Representation

The Republic of Ghana, a nation in West Africa, will serve as the location for the experimental plots for this study. It shares borders with the Ivory Coast in the west, Burkina Faso in the north, and Togo in the east. It borders the Gulf of Guinea and the Atlantic Ocean to the south. Ghana's total size is 238,535 km2 (92,099 sq mi), and it is made up of a variety of biomes, from tropical rainforests to coastal savannas. Ghana, which has a population of over 31 million, is the second-most populous nation in West

Africa, behind Nigeria.Accra, the nation's capital and largest city, as well as Kumasi, Tamale, and Sekondi-Takoradi, are other important cities.

3.2.2 Assumptions

```
# | label: tbl-Data Frame
# | tbl-cap: "Collected from Google Earth Engine"
Data_Frame <- read.csv("Data/Time_Series.csv")
Time_Serie <- read.csv("Data/Time_Series.csv")%>%
  select(year,NDVI,EVI,Precipitation,MinTemperature,MaxTemperature)%>%
  group_by(year)%>%
  summarise_each(funs(median))

kable(Time_Serie,longtable = T, booktabs = T)%>%
add_header_above(c(" ","Vegetation Indices" = 2,"Climate Change"= 3))%>%
  kable_styling(latex_options = c("repeat_header"))
```

	Vegetation Indices		Climate Change		
year	NDVI	EVI	Precipitation	MinTemperature	MaxTemperature
2001	0.7499792	0.5585050	92.40138	217.8009	339.8688
2002	0.6174545	0.4369971	128.81453	218.3945	335.7865
2003	0.6536485	0.4331280	126.15505	220.3344	335.7901
2004	0.8261204	0.5229186	111.66388	218.4170	330.1023
2005	0.7293837	0.4568775	97.24309	221.1652	331.5585
2006	0.7049664	0.4158985	94.90058	223.3435	328.1211
2007	0.6237489	0.4103359	141.95891	221.0357	332.8255
2008	0.7596742	0.4722161	115.86296	218.7837	326.2420
2009	0.7773603	0.5738872	125.95610	222.3701	333.1027
2010	0.6525374	0.4520815	106.98314	222.4885	334.3856
2011	0.6884439	0.4225565	93.91318	222.5436	333.8334
2012	0.6388747	0.4048055	120.05225	221.5151	329.7881
2013	0.6507667	0.4595734	112.69597	220.2121	332.4169
2014	0.7436867	0.5034097	113.95883	225.5163	330.4826
2015	0.7136898	0.4966929	94.67460	226.7593	333.6384
2016	0.7367146	0.4272637	41.50454	227.0843	342.1368
2017	0.7779291	0.5227510	48.82135	227.0180	339.5088
2018	0.5982892	0.3773405	112.13091	226.0173	336.0045
2019	0.7154516	0.4652327	99.74812	226.2367	343.0207
2020	0.7201148	0.4171044	87.38073	223.8555	342.4410

(continued)

| | Vegetation Indices | | Climate Change | | |
year	NDVI	EVI	Precipitation	MinTemperature	MaxTemperature
2021	0.7996552	0.5748818	61.39364	228.4829	344.2889

3.2.3 Exploratory Data Analysis (Summary statistics)

::: {#tbl-Summar Statistics .cell tbl-cap='Summary statistics for Climate Date and Vegetation Loss In Ghana'}

```
Describe <- describe(Time_Serie%>%select(-year))
kable(Describe,longtable = T, booktabs = T)%>%
  kable_styling(latex_options = "scale_down")
```

|	vars	n	mean	sd	median	trimmed	mad	n
NDVI	1	21	0.7084995	0.0636711	0.7154516	0.7080571	0.0916292	0.59828
EVI	2	21	0.4668789	0.0574933	0.4568775	0.4631496	0.0590303	0.37734
Precipitation	3	21	101.3435117	25.5929142	106.9831351	103.9479064	19.3762675	41.50454
MinTemperature	4	21	222.8273684	3.2814108	222.4885167	222.8007065	4.4890372	217.80092
MaxTemperature	5	21	335.0163498	5.1440040	333.8334396	334.9218019	4.9679154	326.24199

:::

```
mpg_list <- split(Time_Serie$EVI,Time_Serie$Precipitation)
disp_list <- split(Time_Serie$NDVI,Time_Serie$Precipitation )
inline_plot <- data.frame(year = c(2002, 2010, 2018), mpg_box = "", mpg_hist = "",
mpg_line1 = "", mpg_line2 = "",
mpg_points1 = "", mpg_points2 = "", mpg_poly = "")

inline_plot %>%
kbl(booktabs = TRUE) %>%
kable_paper(full_width = FALSE) %>%
column_spec(2, image = spec_boxplot(mpg_list)) %>%
column_spec(3, image = spec_hist(mpg_list)) %>%
column_spec(4, image = spec_plot(mpg_list, same_lim = TRUE)) %>%
column_spec(5, image = spec_plot(mpg_list, same_lim = FALSE)) %>%
column_spec(6, image = spec_plot(mpg_list, type = "p")) %>%
column_spec(7, image = spec_plot(mpg_list, disp_list, type = "p")) %>%
column_spec(8, image = spec_plot(mpg_list, polymin = 5))
```

year	mpg_box	mpg_hist	mpg_line1	mpg_line2	mpg_points1	mpg_points2	mpg_po
2002							
2010							
2018							

```
pairs(Time_Serie,bg = c("red", "green", "blue"),pch = 21)
```
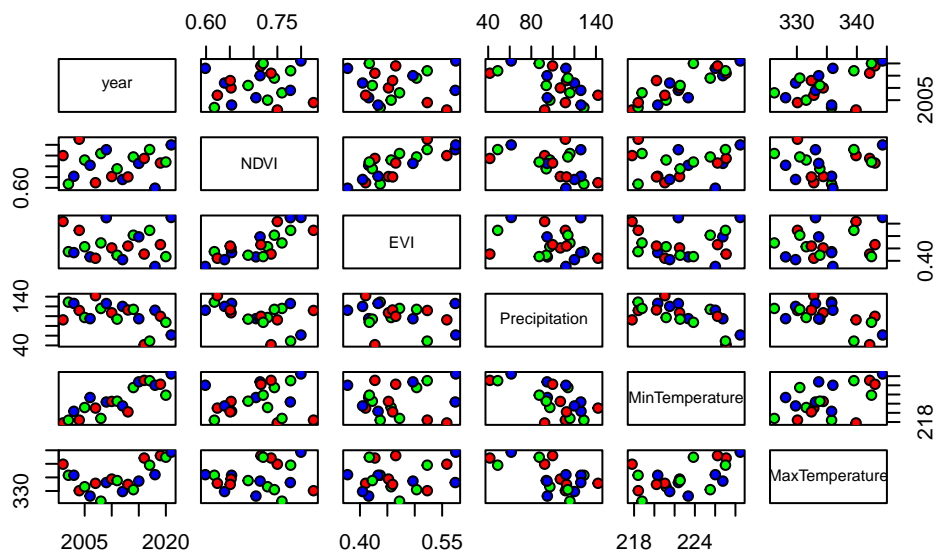

{#fig-

Pairs Plot fig-pos='H'}

```
summary(lm(EVI~Precipitation+MinTemperature+MaxTemperature,Time_Serie))
```

```
Call:
lm(formula = EVI ~ Precipitation + MinTemperature + MaxTemperature,
    data = Time_Serie)

Residuals:
     Min       1Q   Median       3Q      Max
-0.07335 -0.04589 -0.01146  0.03952  0.12389

Coefficients:
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Precipitation	1	0.0037387	0.0037387	1.0430853	0.3214201
MinTemperature	1	0.0010319	0.0010319	0.2878889	0.5985287
MaxTemperature	1	0.0004055	0.0004055	0.1131364	0.7407174
Residuals	17	0.0609334	0.0035843	NA	NA

```
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     0.8721110  1.5041062   0.580    0.570
Precipitation  -0.0006584  0.0007387  -0.891    0.385
MinTemperature -0.0031942  0.0053660  -0.595    0.560
MaxTemperature  0.0011142  0.0033124   0.336    0.741

Residual standard error: 0.05987 on 17 degrees of freedom
Multiple R-squared:  0.0783,    Adjusted R-squared:  -0.08436
F-statistic: 0.4814 on 3 and 17 DF,  p-value: 0.6996
```

::: {#tbl-Analysis of Variance Table .cell tbl-cap='ANOVA Table for Climate Date and Vegetation Loss In Ghana'}

```
lm<-lm(EVI~ Precipitation + MinTemperature +MaxTemperature,Time_Serie)
kable(anova(lm),booktab = T) %>%
  kable_styling(latex_options = c("repeat_header"))
```

:::

3.2.4 Non-Parametric Analysis

3.2.5 Time-series analysis
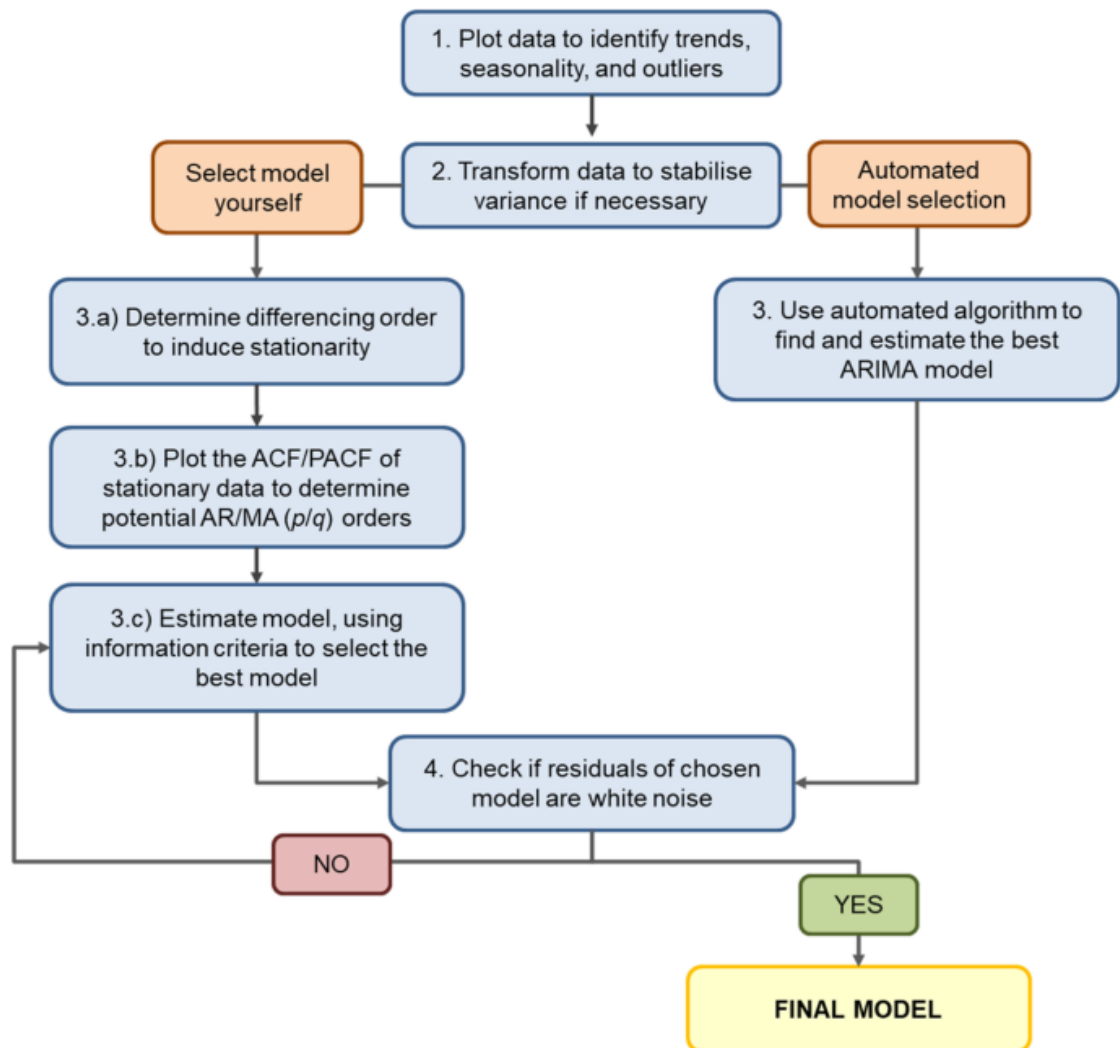
3.2.5.1 Steps involved in Box-Jenkins approach

Figure 3.1: FlowChart

4 CHAPTER FOUR

4.0.1 Analysis and Finding

Before building an ARIMA model we checked that if the series is stationary. That is, we needed to be determined that the time series is constant in mean and variance are constant and not dependent on time.Here, we look at a couple of methods for checking stationarity. If the time series is provided with seasonarity, a trend, or a change point in the mean or variance, then the influences need to be removed or accounted for. Augmented Dickey–Fuller (ADF) t-statistic test to find if the series has a unit root (a series with a trend line will have a unit root and result in a large p-value).

```
# The Stationary Signal and ACF
plot(Rt,col= "red", main = "Stationary Signal")
acf(Rt, lag.max = length(Rt),
    xlab = "lag", ylab = 'ACF', main = '')

#The Trend Signal anf ACF

plot(Tt,col= "red",main = "Trend Signal")
acf(Tt, lag.max = length(Tt),
    xlab = "lag", ylab = "ACF", main = '')
```

Discuss: Shows the initial ACF plot and we can see that before lag 25 almost all are significant and having no trend it needs to be differentiated before performing any analysis. Clearly the seasonality is visible even in the ACF plot.
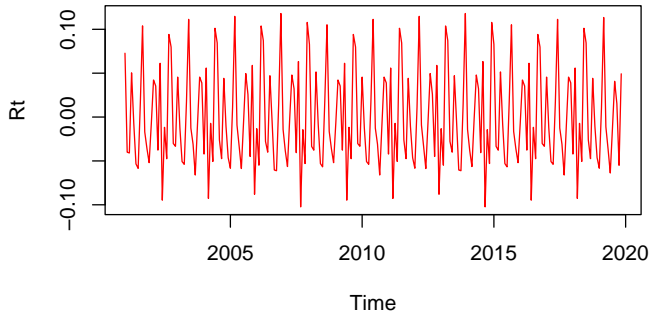
Dickey-Fuller Test and Plot
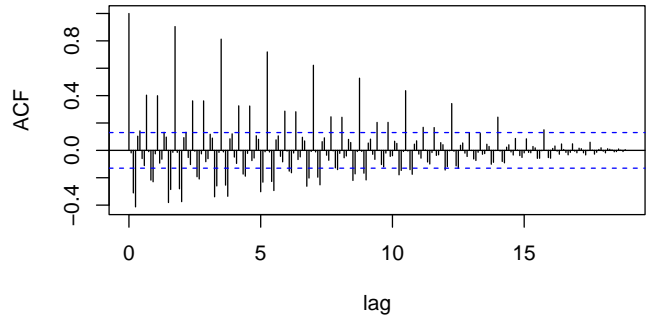
```
tseries::adf.test(Tt)
```

```
	Augmented Dickey-Fuller Test

data:  Tt
Dickey-Fuller = -19.266, Lag order = 6, p-value = 0.01
alternative hypothesis: stationary
```
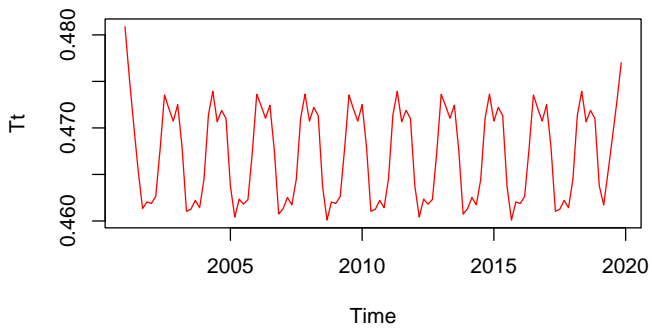
Stationary Signal
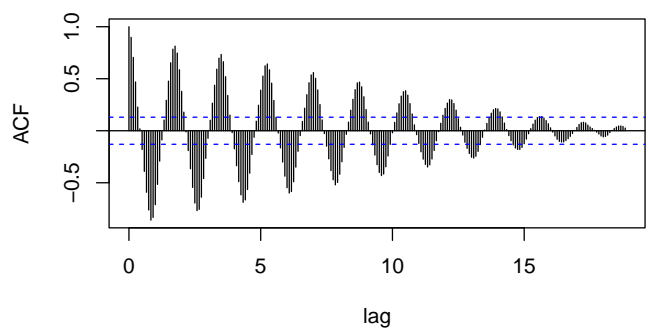

(a) Stationary Signal

(b) Trend Signal

Trend Signal

(c) Stationary Signal

(d) Trend Signal

Figure 4.1: ACF Plot and PACF plot analysis for sample between 2000 and 2020:

Discuss: The DF test confirms that it is stationary as p value < 0.05 and thus can be used for further analysis.This is after doing double differentiation.It is noteworthy that the stationary signal (top left) generates few significant lags that are larger than the ACF's confidence interval (blue dotted line, bottom left). In contrast, practically all delays in the time series with a trend (top right) surpass the ACF's confidence range (bottom right). Qualitatively, we can observe and infer from the ACFs that the signal on the left is steady (due to the lags that die out) whereas the signal on the right is not (since later lags exceed the confidence interval).

4.0.2 Specification of the Model

We can create the SARMA model as SARMA(,0,)X(0) based on the previous study

If there hasn't been any differentiation, we can label it as zero. With the first parameter being PACF and the second being ACF, the first component of multiplication is the non-seasonal part.

The SARMA model's seasonal component follows a similar approach. Since this model has a larger value than the prior SARMA model, it cannot be used. We can utilize the GARCH model and test to see whether the AIC value is better along with the ARMA model as well by omitting the seasonal element as the seasonality pattern is not guaranteed. Generalized Auto-Regressive Conditional Heteroskedasticity models, or GARCH models, can be abbreviated. The GARCH model is commonly used, to estimate value returns for stocks and other financial instruments where trends are unknown. In order to improved the AIC values, we are testing in our case study. Utilizing rugarch, we'll apply the seasonal ARMA-GARCH model.

Since the data is stationary, we go about finding the p and q values from ACF and PACF plots or use auto.arima() in R.

```
auto.arima(Time_Series)
```

```
Series: Time_Series
ARIMA(3,0,0)(2,0,2)[12] with non-zero mean

Coefficients:
          ar1      ar2      ar3     sar1     sar2     sma1     sma2     mean
      -0.1844  -0.4666  -0.2758   0.0626  -0.2229  -0.5593  -0.1545   0.4669
s.e.   0.0655   0.0739   0.0789   0.3384   0.1187   0.3338   0.2796   0.0004

sigma^2 = 0.001793:  log likelihood = 393.81
AIC=-769.62   AICc=-768.79   BIC=-738.8
```

Discuss: As we are not differencing the model we can consider ARMA(2,0,3) has the best model. Which is the best p and q value also found from the ACF and PACF plots.

Residual Analysis

Discuss: From the above time series plot we can conclude that, the trend within the year values for 1960,2016 and 2020 are similar. We can observe that during start of the year in January the unemployment rate increases and becomes constant during February, March and then decreases sharply post April. Then in mid of the year it increases to a certain level and attains constant until late/end of the year. Clearly we can see some pattern when we do time series plot within a single year. It can be concluded that unemployment rate is higher during winter months and decreased post April which is summer season. Thus the seasonal aspect can be clearly understood.

4.0.3 Modeling and Parameter estimation

Where the **ARIMA (PACF, Num_Diffrentation, ACF)** model have the below format for the parameters. Coefficients for various models:

Discuss: Based on the different models, we can see that ARIMA(2,2,5) had the least AIC value, sigma^2 being the least therefore is the best model for given time series. Find the below time series plot for the residuals.

4.0.3.1 Residual Analysis

4.0.3.2 Residual Plot

4.0.3.3 Shapiro Test

4.0.3.4 Ljung-Box

Time-series Forecasting

Discuss: The plot shows the forecasting to plot for the next 20 values which is shown by the blue region.

```
#| tbl-cap: "Linear regression model for predicting EVI from Time"
tdx.ns <- data.frame(time = c(1:length(Time_Series)), trend = Time_Series - tdx.d
summary <- summary(lm(formula = trend ~ time, data = tdx.ns))
summary
```

Table 4.1: **?(caption)**

```
Call:
lm(formula = trend ~ time, data = tdx.ns)

Residuals:
     Min       1Q   Median       3Q      Max
-0.09840 -0.04340 -0.01218  0.04358  0.11309

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 4.668e-01  7.423e-03  62.888   <2e-16 ***
time        1.503e-06  5.645e-05   0.027    0.979
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.05574 on 225 degrees of freedom
Multiple R-squared:  3.151e-06, Adjusted R-squared:  -0.004441
F-statistic: 0.000709 on 1 and 225 DF,  p-value: 0.9788
```
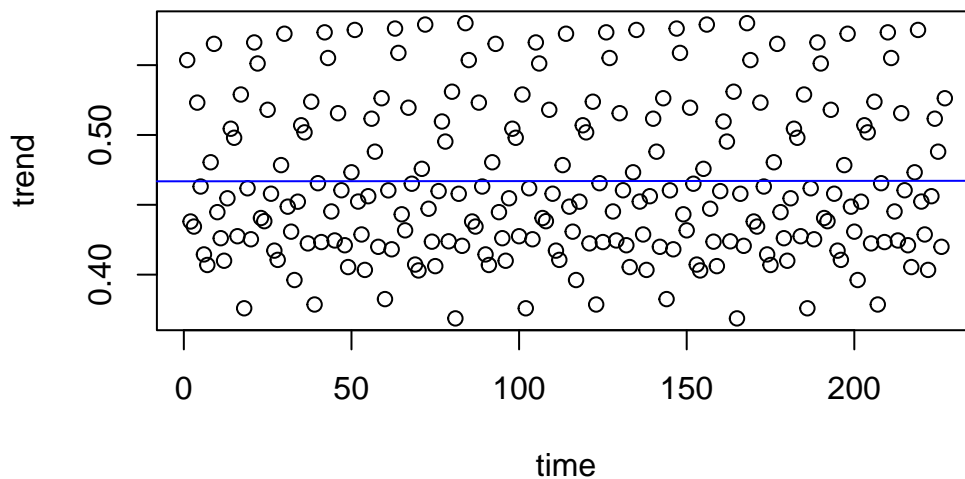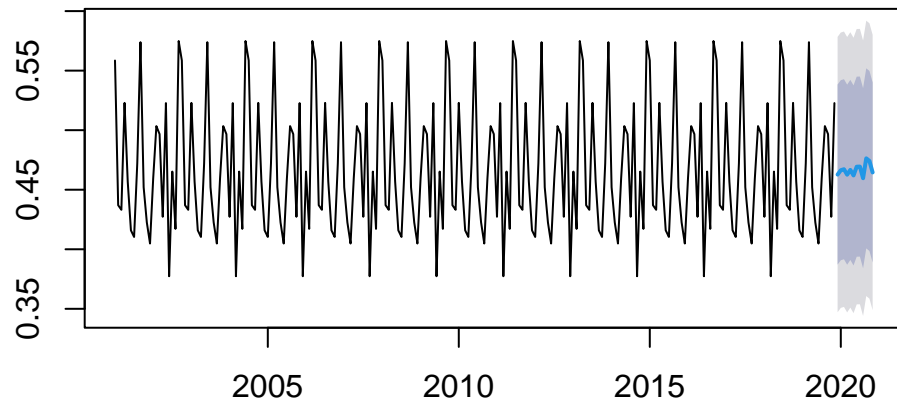
```
plot(tdx.ns)
abline(a = summary$coefficients[1,1], b = summary$coefficients[2,1], col = 'blue')
```

```
plot(evi.hw <- forecast::hw(y = Time_Series, h = 12, damped = T))
```

Forecasts from Damped Holt–Winters' additive method

5 CHAPTER FIVE

5.1 CONCLUSIONS AND RECOMMENDATIONS

5.1.1 Summary

Broadly speaking, in this study we have presented a state-of-the-art of the following popular time series forecasting models with their salient features:

- The Box-Jenkins or ARIMA models for linear time series forecasting.
- Some non-linear stochastic models, such as NMA, ARCH.
- SVM based forecasting models; LS-SVM and DLS-SVM.

5.1.2 Conclusions

It has been seen that, the proper selection of the model orders (in case of ARIMA), the number of input, hidden, output and the constant hyper-parameters (in case of SVM) is extremely crucial for successful forecasting. We have discussed the two important functions. AIC and BIC, which are frequently used for ARIMA model selection.

We have considered a few important performance measures for evaluating the accuracy of forecasting models. It has been understood that for obtaining a reasonable knowledge about the overall forecasting error, more than one measure should be used in practice. The last chapter contains the forecasting results of our experiments, performed on six real time series datasets. Our satisfactory understanding about the considered forecasting models and their successful implementation can be observed form the five performance measures and the forecast diagrams, we obtained for each of the six datasets. However in some cases, significant deviation can be seen among the original observations and our forecast values. In such cases, we can suggest that a suitable data preprocessing, other than those we have used in our work may improve the forecast performances.

5.1.3 Recommendations

Time series forecasting is a fast growing area of research and as such provides many scope for future works. One of them is the Combining Approach, i.e. to combine a number of different and dissimilar methods to improve forecast accuracy. A lot of works have

been done towards this direction and various combining methods have been proposed in literature [8, 14, 15, 16]. Together with other analysis in time series forecasting, we have thought to find an efficient combining model, in future if possible. With the aim of further studies in time series modeling and forecasting

References

Allotey, Godwin Akweiteh. 2017. "Stop Galamsey in 3 Weeks or Face the Law - Amewu." *Ghana News*, March. http://citifmonline.com/2017/03/29/stop-galamsey-in-3-weeks-or-face-the-law-amewu/.

Ansah, Marian Efe. 2017. "Galamsey, Pollution Destroying Water Bodies in Ghana - Water Company." *Ghana News*, March. http://www.leakxgh.com/2017/05/galamsey-pollution-destroying-water.html.

Barenblitt, Abigail, Amanda Payton, David Lagomasino, Lola Fatoyinbo, Kofi Asare, Kenneth Aidoo, Hugo Pigott, et al. 2021. "The Large Footprint of Small-Scale Artisanal Gold Mining in Ghana." *Science of the Total Environment* 781 (August). https://doi.org/10.1016/j.scitotenv.2021.146644.

"Confidence Intervals: An Empirical Investigation of the Series in the m-Competition - ScienceDirect." n.d. https://www.sciencedirect.com/science/article/abs/pii/0169207087900458.

DeFries, Ruth S., Thomas Rudel, Maria Uriarte, and Matthew Hansen. 2010. "Deforestation Driven by Urban Population Growth and Agricultural Trade in the Twenty-First Century." *Nature Geoscience* 3 (3): 178181.

Gertler, Paul J., Orie Shelef, Catherine D. Wolfram, and Alan Fuchs. 2016. "The Demand for Energy-Using Assets Among the World's Rising Middle Classes." *American Economic Review* 106 (6): 13661401.

"Gold, Guns and China: Ghana's Fight to End Galamsey." 2017. *African Arguments*, May. https://africanarguments.org/2017/05/30/gold-guns-and-china-ghanas-fight-to-end-galamsey/.

Gracia, Zindzy. 2018. "Causes and Effects of Galamsey in Ghana." *Yen.com.gh - Ghana News.*, January. https://yen.com.gh/104844-causes-effects-galamsey-ghana.html.

Gyekye, Joyce. n.d. "MD of Ghana Water Company Limited Says Fight Against Galamsey Is Being Lost." *Ghana Broadcasting Corporation.* http://www.gbcghana.com/1.11923484.

Hipel, Keith W., and A. Ian McLeod. 1994. *Time Series Modelling of Water Resources and Environmental Systems.* Elsevier.

Library, Wiley Online, and Mohammad Valipour. 2015. "Long-Term Runoff Study Using SARIMA and ARIMA Models in the United States." *Meteorological Applications* 22 (3): 592–98. https://doi.org/10.1002/MET.1491.

Owusu-Nimo, F., J. Mantey, K. B. Nyarko, Eugene Appiah-Effah, and A. Aubynn. 2018. "Spatial Distribution Patterns of Illegal Artisanal Small Scale Gold Mining (Galamsey) Operations in Ghana: A Focus on the Western Region." *Heliyon* 4 (2): e00534. https://doi.org/10.1016/j.heliyon.2018.e00534.

Specht, Maria Joana, Severino Rodrigo Ribeiro Pinto, Ulysses Paulino Albuquerque, Marcelo Tabarelli, and Felipe PL Melo. 2015. "Burning Biodiversity: Fuelwood Harvesting Causes Forest Degradation in Human-Dominated Tropical Landscapes." *Global Ecology and Conservation* 3: 200209.

"Women Die in Ghana Mine Collapse." 2009, November. http://news.bbc.co.uk/2/hi/africa/8356343.stm.

Zhang, Guoqiang, B. Eddy Patuwo, and Michael Y. Hu. 1998. "Forecasting with Artificial Neural Networks:: The State of the Art." *International Journal of Forecasting* 14 (1): 35–62. https://doi.org/10.1016/S0169-2070(97)00044-7.