Segregate The Variability Climate Is Responsible For In Vegetation Loss

Quantifying The Status of Galamsey With Time Series Analsis

Kalong Boniface Fugah Seletey Mitchell

2022-08-27

Table of Contents

1 CHAPTER ONE

1.1 INTRODUCTION

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 Galamsey Operation in Ghana, the Cause,Dangers and it's Environmental impact.

Galamsey also known as "gather them and sell",[7] 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 [5]. 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 [8] .

Illegal mining causes damage to the land and water supply [2] . 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 [1] . 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 [6] .

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 [4]. 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 [3] . 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 datasets.

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. 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 overfit 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

2.1.1 Theoretical Review

This literature review will follow narrative approach to gain insight into research topic. 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. The methods used and results from those methods

achieved by former researchers will be summarized including different methods on time series and comparing them with each other.

2.1.1.1 What is Time Series Analysis?

Makridakis and hibon, in time series analysis researchers have conducted a competition named M-competition in 1987 , Where participants could submit their forecasting on 1001 time series data taken from demography, industry, and economics. There were four main findings from the competition were:

- 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 defined as the observation of increasing or decreasing pattern over a period. In a stationary time series, mean of the time series data must be constant in time and whereas in general time series the mean is arbitrary function of time.

Seasonality: It refers to a cyclic happening of events. A pattern which repeats itself after a period.

Heteroscedasticity: It is also known as level; it is defined as the non-constant variance from the mean calculated at different time periods.

Few methods do not perform well in forecasting if the data is seasonal, and few do not perform well with trends in the data. Hence trends, seasonality and heteroscedasticity must be considered to select the best statistical method in forecasting.

2.1.1.2 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 [6, 8,12, 23] 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 [28], 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) [9, 28] model and its variations like Generalized ARCH (GARCH) [9, 28], Exponential Generalized ARCH (EGARCH) [9] etc., the Threshold Autoregressive (TAR) [8, 10] model, the Non-linear Autoregressive (NAR) [7] model, the Non-linear Moving Average (NMA) [28] 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.

2.1.1.2.1 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' (Cryer, J.D., 1986). 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.

2.1.1.2.2 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..

2.1.1.2.3 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).

2.1.1.2.4 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 $\mu = \frac{\omega}{1-\phi_1}$ and constant variance $\gamma_o = \frac{\sigma^2}{1-\phi_1^2}$
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).

2.1.1.2.5 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).

2.1.1.2.6 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.

2.1.1.2.7 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.

2.1.1.2.8 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

2.1.1.3 Plotting patterns in correlation:

There are two functions used for plotting patterns in correlation:

2.1.1.3.1 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).

2.1.1.3.2 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).

2.1.1.4 Automatic selection techniques:

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

2.1.1.4.1 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.

2.1.1.5 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 ARIMA(p,d,q), where p and q are same as ARMA model but d = number of first differences (Yu, G. and Zhang, C., 2004, May).

2.1.1.6 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.

2.1.1.7 Comparing ARIMA method with SARIMA method:

In comparison to ARIMA and SARIMA, (Valipour, M., 2015) investigated it on long-term runoff forecasting in the United States. The results have shown that SARIMA models have performed better than ARIMA model. However, it was seen that SARIMA models were very sensitive and a slight change in a parameter would result in poor performance of the model.

(Wang, S., Li, C. and Lim, A., 2019) have used ARIMA and SARIMA models from the perspective of Linear System Analysis, Spectra Analysis and Digital Filtering. It was shown that ARIMA and SARIMA both have not performed well and the researchers were forced to look beyond these models for better performance. They have mentioned that ARMA-SIN model was better but have also said it is relatively difficult to study and understand the concepts compared to ARIMA and SARIMA model.

The findings from the (Valipour, M., 2015) have proven SARIMA to better however, their claim contradicts when it was to be compared with the findings of (Wang, S., Li, C. and Lim, A., 2019). The use of a particular method must be based on the data, after the analysis it is known if that the data has trend, they must choose ARIMA and if the data has seasonality, choosing SARIMA would be helpful.

2.1.1.8 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

2.1.1.9 Time Series Forecasting Using Support Vector Machines

2.1.1.9.1 Concept of Support Vector Machines

Till now, we have studied about various stochastic and neural network methods for time series modeling and forecasting. Despite of their own strengths and weaknesses, these methods are quite successful in forecasting applications. Recently, a new statistical learning theory, viz. the Support Vector Machine (SVM) has been receiving increasing attention for classification and forecasting [18, 24, 30, 31]. SVM was developed by Vapnik and his co-workers at the AT & T Bell laboratories in 1995 [24, 29, 33]. Initially SVMs were designed to solve pattern classification problems, such as optimal character recognition, face identification and text classification, etc. But soon they found wide applications in other domains, such as function approximation, regression estimation and time series prediction problems [24, 31, 34].

Vapnik's SVM technique is based on the Structural Risk Minimization (SRM) principle [24, 29, 30]. The objective of SVM is to find a decision rule with good generalization ability through selecting some particular subset of training data, called support vectors [29, 31, 33]. In this method, an optimal separating hyperplane is constructed, after non-linearly mapping the input space into a higher dimensional feature space. Thus, the quality and complexity of SVM solution does not depend directly on the input space [18, 19].

Another important characteristic of SVM is that here the training process is equivalent to solving a linearly constrained quadratic programming problem. So, contrary to other networks' training, the SVM solution is always unique and globally optimal. However a major disadvantage of SVM is that when the training size is large, it requires an enormous amount of computation which increases the time complexity of the solution [24].

2.1.1.10 Forecast Performance Measures

While applying a particular model to some real or simulated time series, first the raw data is divided into two parts, viz. the Training Set and Test Set. The observations in the training set are used for constructing the desired model. Often a small subpart 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 forecasted 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 forecasted values of the time series.

2.1.1.10.1 Description of Various Forecast Performance Measures

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

The Mean Forecast Error (MFE)

This measure is defined as [24] $MFE = \frac{1}{n}\sum_{t=1}^{n} e_t$ The properties of MFE are:

2.1.2 Empirical Review

LULC data are records that documents to what extent a region is covered by wetlands, forests, agriculture, impervious surfaces, and other land and water forms. These water forms include open water or wetlands. Land use shows how people use landscape either for conservation, development, agriculture or mixed uses [6, 7]. Changes In land can be identified by analysing satellite imagery. However,

land use cannot be identified from satellite imagery. Satellite imagery give us information that helps in understanding the present landscape. Furthermore, to see changes through time, different years are needed. With this information, we can assess decades of data as well as insight into the possible effects of these changes that has occured and make better decisions before they can cause great harm. According to [10], five different types of LULC pattern were classified barren lands such as Galamsey Site, agricultural lands, urban lands, quarries, and free water bodies, to detect the 25years LULC change in the western Nile delta of Egypt. Supervised maximum likelihood classification (MLC) method together with landsat images were used in Erdas Imagine software. The finding shows a significant change in barren land changing into agricultural land continuously from 1984 to 2009.

Similarly, in [1] used the maximum likelihood algorithm (MLA) and Markov chain model (MCA) to study the LULC classification using ArcGIS and future prediction using Idiri respectively in Kathmandu city Nepal. Built-up, water body, forest area, open field and cultivated land classes classified. Results show built-up area significantly increased, and water body, forest area, open field and cultivated land decrease downward trend from 1976 to 2009. Furthermore, the Markov chain Analysis prediction for 2017 shows that in 2017 Urban area will increase to cover 72.24 % of the total land in Kathmandu and cultivated land remains only 20.90%. Waterbody and the open field will increase respectively by 0.59%, 0.19% whereas forest land decrease by 0.47%.

Furthermore, in [10], They made used of the Maximum likelihood classification (MLC), Change detection and spatial matrix analysis to analyse land cover change of fifty-year period (1954 to 2004) in Avellino Italy. The result shows 4 LULC classes, with urban land use increasing rapidly affecting the cultivated land mostly, while woodland and grassland cover decrease was at a lower rate. Moreover, in [11] studied the LULC changes and structure in Dhaka metropolitan, Bangladesh in a period of 1975 to 2005. Maximum Likelihood Classification (MLC) and transition matrix method were used for LULC classification and pattern of LULC. The Result shows six classes in LULC of the water body,

vegetation, bare soil, cultivated land built-up and wetland/lowland. Also, a significant increase in the built-up land, while cultivated land, vegetation and wetland decreased accordingly from 1975 to 2005.

Also, [12] used Maximum Likelihood Classification and comparison method to study the LULC classification and change respectively from 1976 to 2003 in Tirupati, India. The results show 6 LULC classes, agricultural land, built-up, dense forest, plantation, water spread and other land, a significant increase in built-up area, plantation forests and other land, while a decrease on the part of the waterbody, dense forest and agricultural land was noticed.

Moreover, in [13] studied the LULC change in Duzce plain Turkey. Supervised classification and the Corine land cover nomenclature methods used. The result shows 5 LULC classes as urban fabric, forest, heterogeneous agricultural land, inland wasteland and (Industrial, commercial, and transport) units with an accuracy assessment between 92.41% and 97.3% for LULC map 2010 and 1987 respectively. Also, a significant change in LULC was noticed with 11.2% increase in agricultural area and 335% decrease of forest land.

Also, a significant increase and a decrease of LULC were noticed between the years 1973, 1985, and 2000 within the classes.

```
library(openintro)  # for data
library(tidyverse)  # for data wrangling and visualization
library(knitr)      # for tables
library(broom)      # for model summary
library(imputeTS)
library(dplyr)
library(kableExtra)
library(forecast)
if(!require("pacman")){install.packages("pacman")}
pacman::p_load(char = c('rgee','reticulate','raster','tidyverse',
                'dplyr','sf','mapview','mapeddit','caret','forcats','reticulate',
                'rgee', 'remotes', 'magrittr', 'tigris', 'tibble', 'stars',
                'st', 'lubridate', 'imputeTS', 'leaflet', 'classInt',
                'ggplot2', 'googledrive', 'geojsonio', 'ggpubr','cartogram'),
               install = F, update = F, character.only = T)

library(rgee)

library(reticulate)
#ee_install()
ee_check()

ee_Initialize("kalong",drive = TRUE) # initialize GEE,
#this will have you log in to Google Drive
```

Now we will create a hexagonal grid over the study area

```
aoi.proj <- st_transform(aoi, st_crs(2392))
hex <- st_make_grid(x = aoi.proj, cellsize = 17280, square = FALSE) %>%
st_sf() %>%
rowid_to_column('hex_id')
hex <- hex[aoi.proj,]
plot(hex)
```

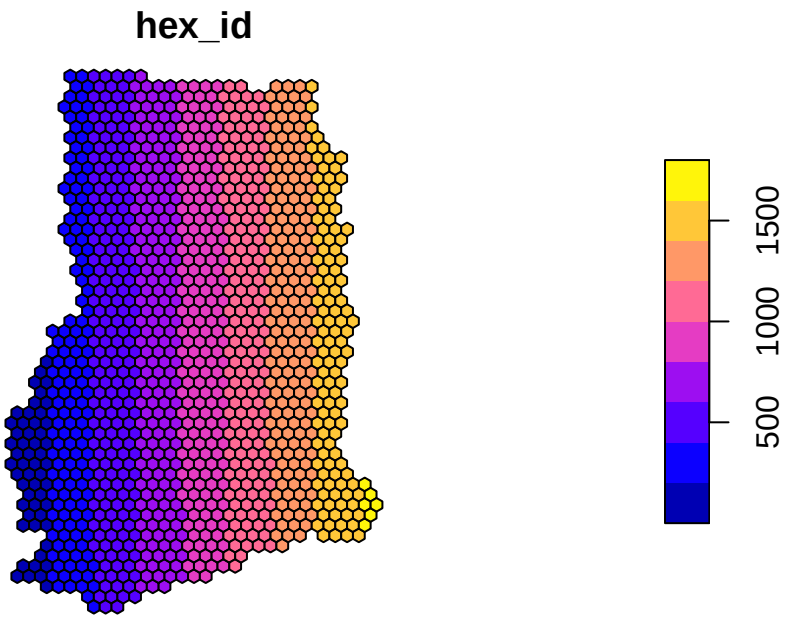


Figure 1: Now we will use the grid created above to extract the mean EVI values within each cell for the years 2000-2020.

Which we are going to perform a time series analysis on the data within each grid cell. But first, we will work through the procedure one step at a time.

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 Sentinel 2 data from Google Earth Engine in order to choose NDVI and EVI values.

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.

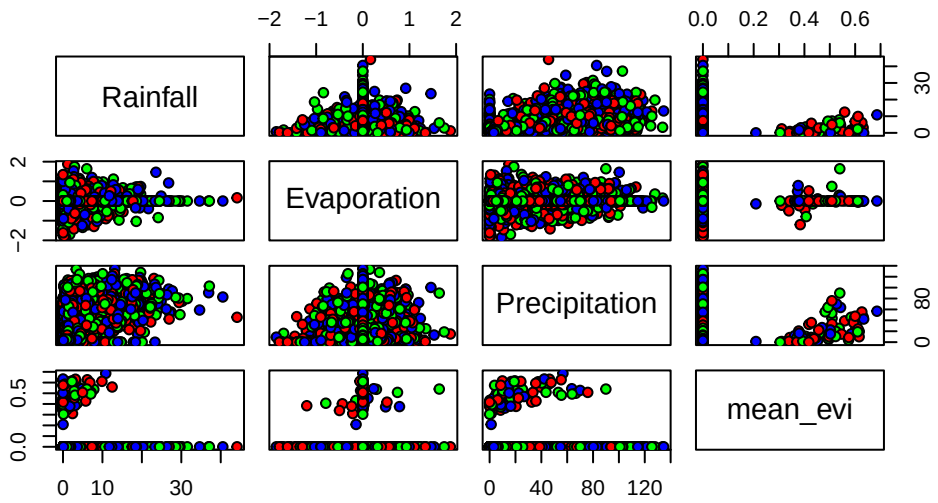
3.2.1 Data Representation

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.

```
head(Description)%>%
  kable(booktabs =  TRUE)
```

Date	Rainfall	Evaporation	Precipitation	mean_evi
2000-01-01	0.4869522	0	5.437918	0
2000-01-02	1.6640719	0	8.097965	0
2000-01-03	0.4059545	0	7.663450	0
2000-01-04	0.0001358	0	3.887136	0
2000-01-05	0.2125085	0	6.001542	0
2000-01-06	0.4051066	0	6.800775	0

```
# #| label: tbl-stats
# #| tbl-cap: "Summary statistics for Climate Date and Vegetation Loss In Ghana"
# library(psych)
# describe(Description) %>%
#   kable(booktads = TRUE)%>%
#   kable_styling(latex_options = "scale_down")

pairs.default(DF,main = "Corellation Between ",bg = c("red", "green", "blue"),pch = 21)
```



3.2.1.1 Time-series analysis

::: {#fig-Time Series And Decompostion .cell .column-page-right layout-ncol="2"}

```
# Convert data to time series.
tdx <- ts(data = tdx$mean_evi, start = c(2001, 1), end = c(2019, 11), frequency = 12)
plot(tdx)


tdx <- if(na.cnt > 0){imputeTS::na_kalman(tdx, model = "auto.arima", smooth = T)} else {
    tdx
}
plot(tdx)


tdx.dcp <- stl(tdx, s.window = 'periodic')
plot(tdx.dcp)

Tt <- trendcycle(tdx.dcp)
St <- seasonal(tdx.dcp)
Rt <- remainder(tdx.dcp)
# plot(Tt)
# plot(St)
plot(Rt)
```

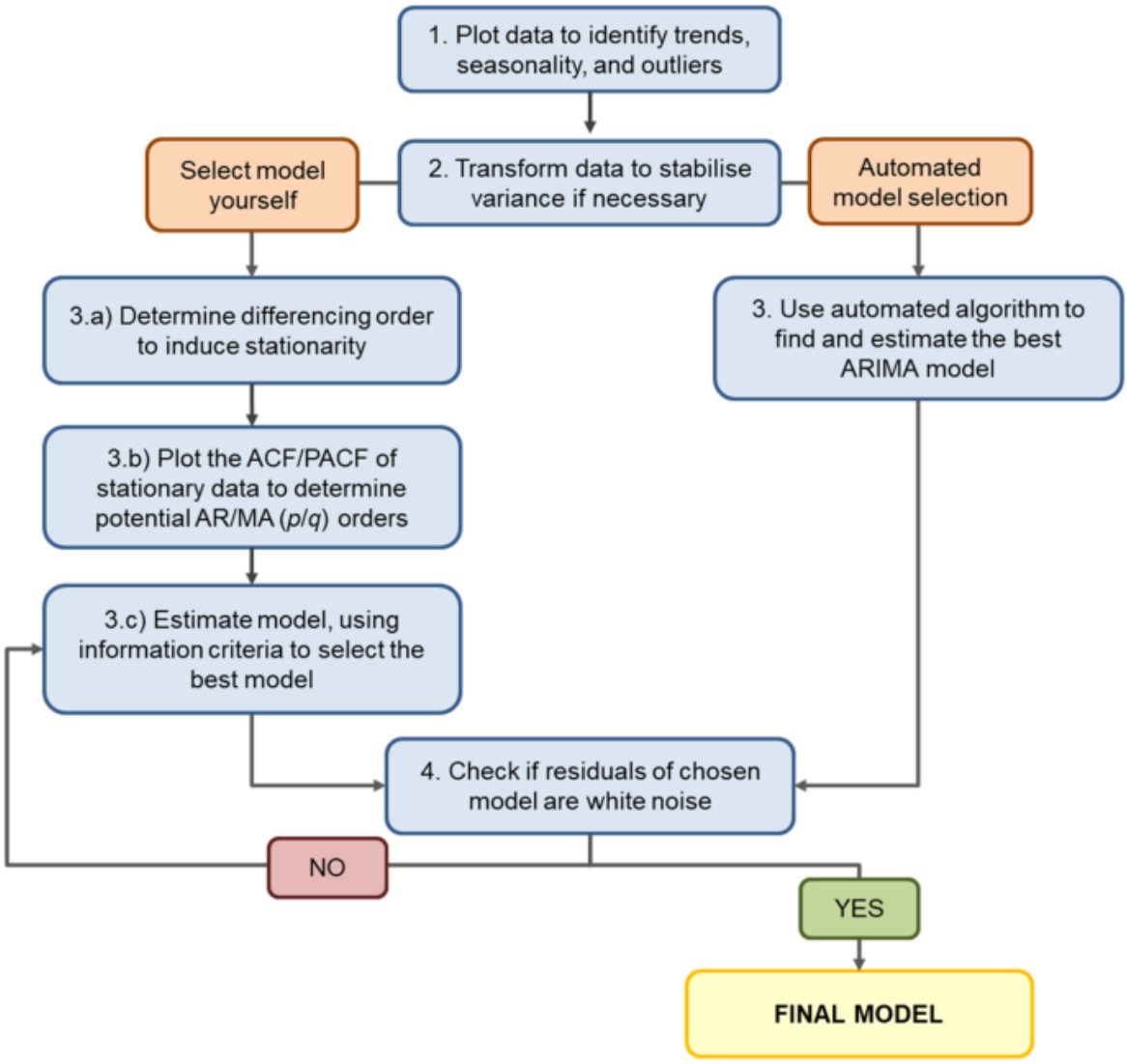
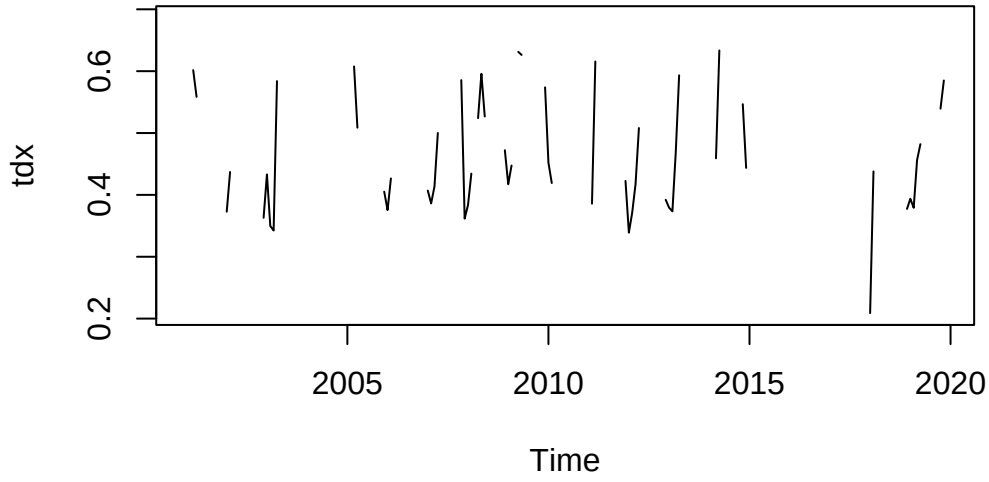
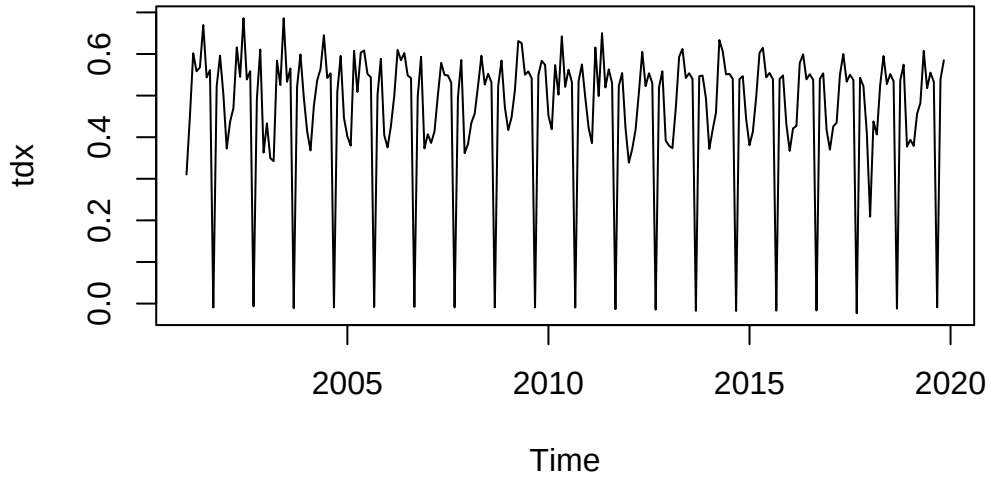


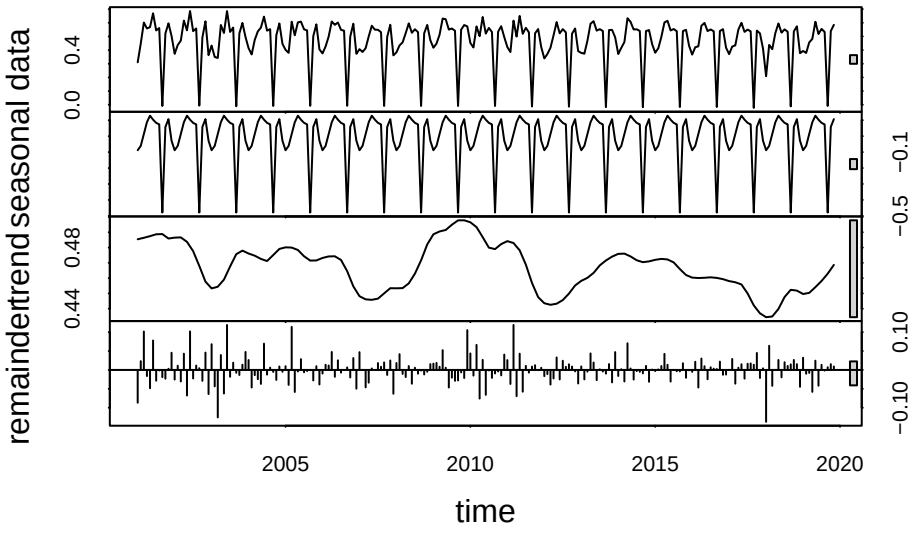
Figure 2: FlowChart



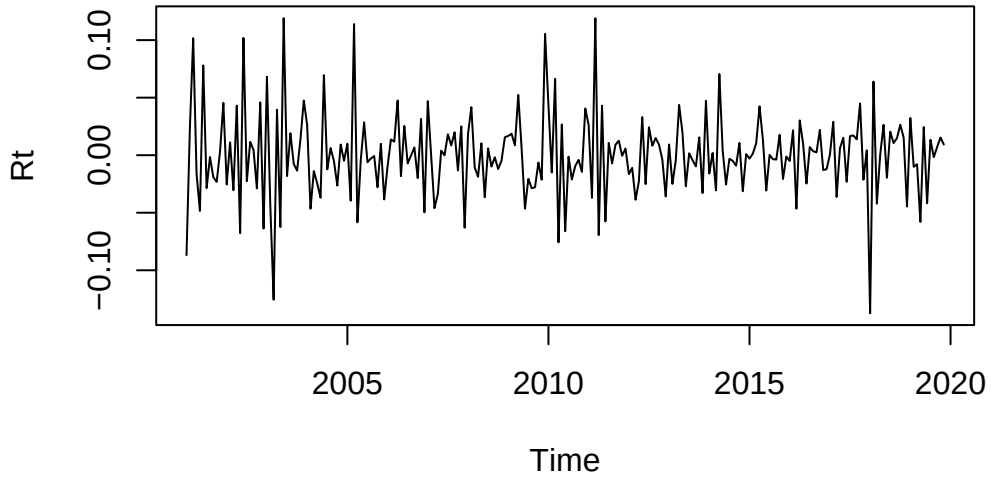
{#fig-Time Series And Decompostion-1}



{#fig-Time Series And Decompostion-2}



{#fig-Time Series And Decompostion-3}



{#fig-Time Series And Decompostion-4}

Time Series And Decompostion :::

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). ### Specification of the Model

Based on the above analysis we can form the SARMA model as, SARMA(,0,)X(,0,) Has no differentiation has been done we can mark it as zero. First part of multiplication is the Non-seasonal part with first parameter as PACF and second as ACF. Similarly its the same format for Seasonal part as well in SARMA model. From the above ACF, PACF analysis we can formulate the below models:
1.SARMA(2,0,3)X(1,0,2)
2.SARMA(3,0,1)X(1,0,1)

This model cannot be used, because it has higher value than the previous SARMA model. Also has the seasonality pattern is not certain we can use GARCH model and test to see if the AIC value is better along with ARMA model as well by ignoring the seasonal part. GARCH model can be abbreviated as Generalized Auto-regressive conditional Heteroskedasticity models. However GARCH model is usually used to estimate value returns for stocks and so on, where trends is not known. We are using to test in our use-case to look at better AIC values. We are going to apply the seasonal ARMA-GARCH model using rugarch.

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

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

GARCH

Result of GARCH Model with Specifications:
Shows the output of the GARCH model when ran on the dataset. Some of the observations we can see that and compare with the SARMA model. We can view various optimal parameters and their estimate and standard error as well. The LB test has values for p nothing less than 0.05 so we can say that null hypothesis is rejected and may assume correlation being present in the dataset. However LB test on Standardized squared residuals yield one of the p values closer to 0. The Arch LM tests has p values for lags 7 and 9 closer to zero. We can observe that the log-likelihood for SARMA is smaller than GARCH. Although, it is well noted that the higher the log-likelihood, the better the model. The GARCH model has higher log-likelihood when compared sarima but no much difference between their values. While the same is opposite for AIC value, where smaller the AIC value is the best model.

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.

3.2.1.2 ACF Plot analysis for sample between 2000 and 2020:

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.

PACF plot

3.2.1.3 Dickey-Fuller Test and Plot

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.

3.2.2 Modeling and Parameter estimation

ARIMA(1,2,1) ARIMA(1,2,4) ARIMA(1,2,5) ARIMA(2,2,1) ARIMA(2,2,4) ARIMA(2,2,5)
1.ARIMA(1,2,1)
2.ARIMA(1,2,4)
3.ARIMA(1,2,5)
4.ARIMA(2,2,1)
5.ARIMA(2,2,4)
6.ARIMA(2,2,5)
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.

3.2.2.1 Residual Analysis

3.2.2.2 Residual Plot

Discuss: Residual plot tells the points that are left after fitting the model. We can see that most points are closer to the line except at the middle of the plot. Now lets plot the ACF of residuals for the model to further understand its behavior.

3.2.2.3 ACF Residual Plot

Discuss: From the plot for ACF of residuals, we can clearly see that there is no statistically significant correlation for the data and every point is within the confidence interval.

Discuss: From the histogram we can see that, it slightly follow normal distribution if we ignore the outliers. But the plot is slightly right skewed in nature. For more understanding we need to perform quantile-quantile plot for the analysis.

Discuss: From the qqplot for the residuals we can say that, most of the points lie on the reference line, however they are few points towards the tail part of the plots that deviate slightly. QQ plot gives a better visual of the residuals how the sample quantiles are related to the theoretical quantiles. There are few tests which can be performed, to check the normality of the residuals one such is Shapiro test.

3.2.2.4 Shapiro Test

Discuss: The above figure shows the results of Shapiro-wilk test for the residuals of the model. If the value of p is equal to or less than 0.05, then the hypothesis of normality will be rejected by the Shapiro test. Here the p value is less than 0.05 so we can say that the residuals follow normal distribution.

3.2.2.5 Ljung-Box

Discuss: Ljung-Box test is next performed on the models to test the randomness of the data over the lags at the bigger perspective.The null hypothesis for LB test is that residuals are independently distributed if p values is less than 0.05. Based on that we can see that, independence has been captured by the data for the following model.

Time-series Forecasting

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

4 CHAPTER FOUR

4.0.1 The Analysis Of Variance (ANOVA) Method

4.0.2 The Empirical * Theory model

4.0.3 Assumptions

That looks better! Unfortunately though, there are a number of dates which don't have any evi value at all, let's figure out which ones these are.

```
library(tseries)
adf.test(Rt)
```

```
    Augmented Dickey-Fuller Test

data:  Rt
Dickey-Fuller = -8.639, Lag order = 6, p-value = 0.01
alternative hypothesis: stationary
```

```
adf.test(Tt)
```

```
    Augmented Dickey-Fuller Test

data:  Tt
Dickey-Fuller = -3.4545, Lag order = 6, p-value = 0.04798
alternative hypothesis: stationary
```

```
adf.test(tdx)
```

Table 1: **?(caption)**

(a)

term	estimate	std.error	statistic	p.value
(Intercept)	0	0.01	90.64	0.00
time	0	0.00	-2.78	0.01

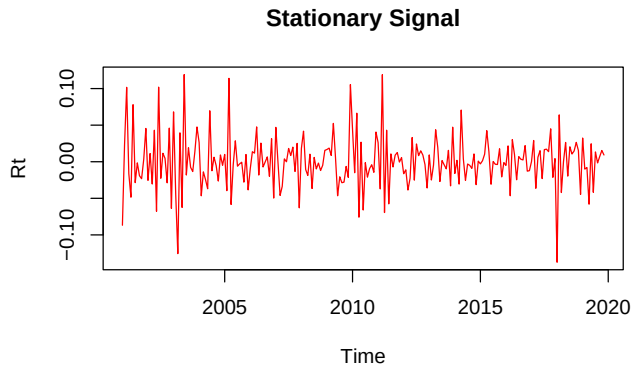
```
    Augmented Dickey-Fuller Test

data:  tdx
Dickey-Fuller = -8.2685, Lag order = 6, p-value = 0.01
alternative hypothesis: stationary
```

Autocorrelation Function (ACF) Identify if correlation at different time lags goes to 0

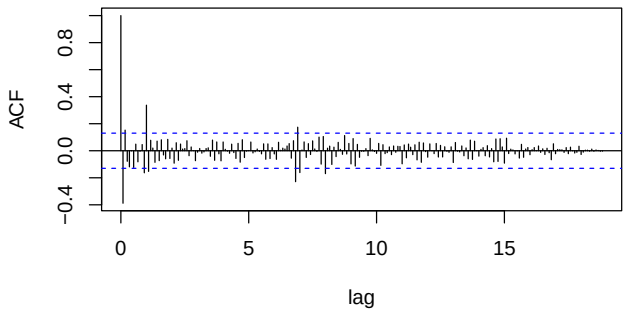
```
# The Stationary Signal and ACF
plot(Rt,col= "red", main = "Stationary Signal")
acf(Rt, lag.max = length(Rt),
    xlab = "lag", ylab = 'ACF', main = '')

#The Trend Signal anf ACF

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

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).

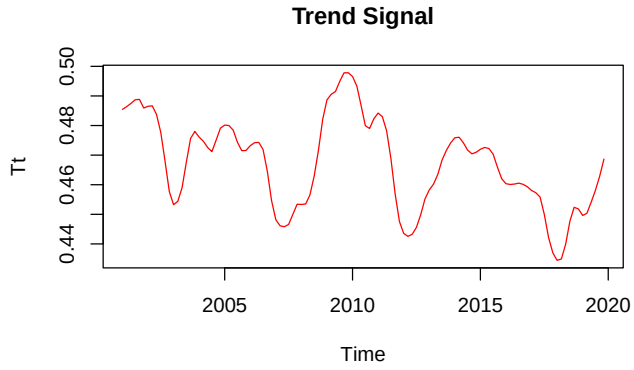
```
#| tbl-cap: "Linear regression model for predicting EVI from Time"
tdx.ns <- data.frame(time = c(1:length(tdx)), trend = tdx - tdx.dcp$time.series[,1])
summary <- summary(lm(formula = trend ~ time, data = tdx.ns))
summary %>%
  tidy() %>%
  kable(digits = c(0, 0, 2, 2, 2))

plot(tdx.ns)
abline(a = summary$coefficients[1,1], b = summary$coefficients[2,1], col = 'blue')
```



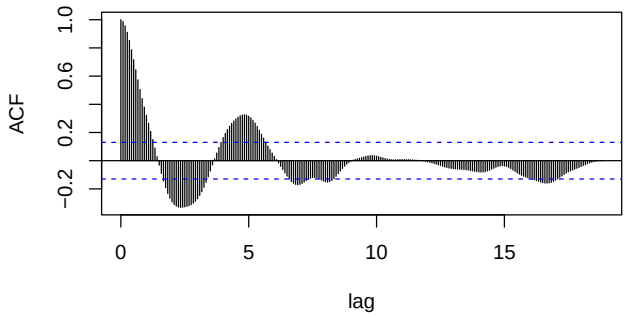
(a) Stationary Signal



(b) Trend Signal

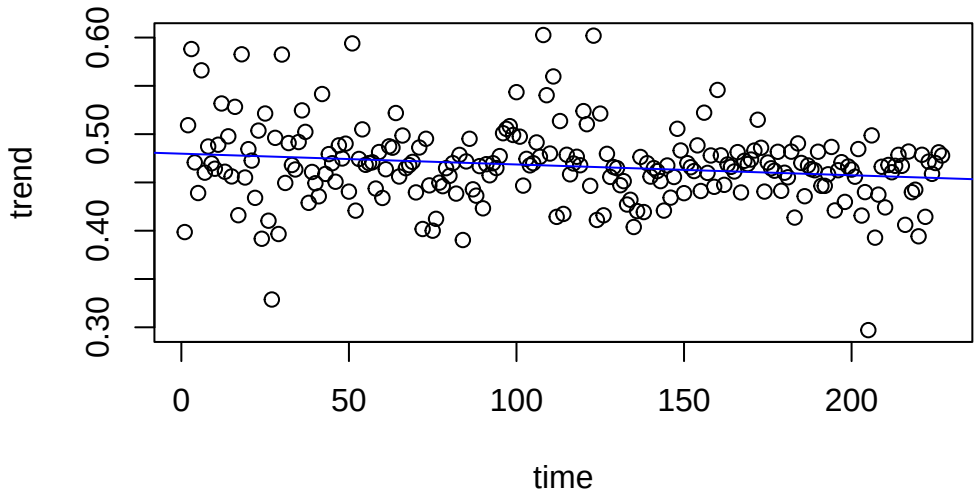


(c) Stationary Signal



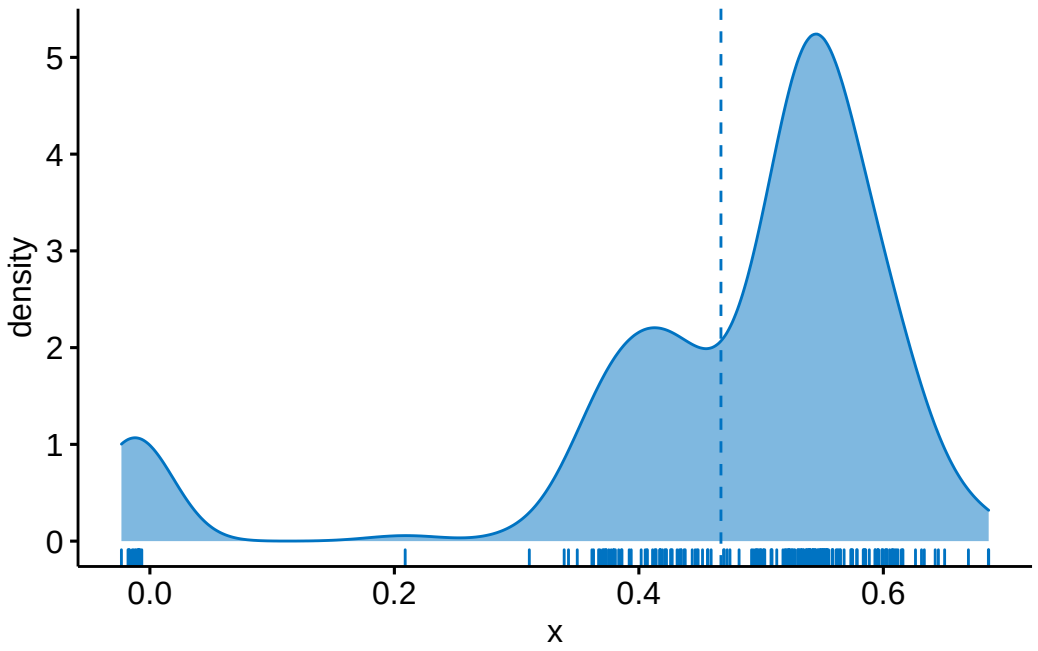
(d) Trend Signal

Figure 3: The Stationary Signal and ACF

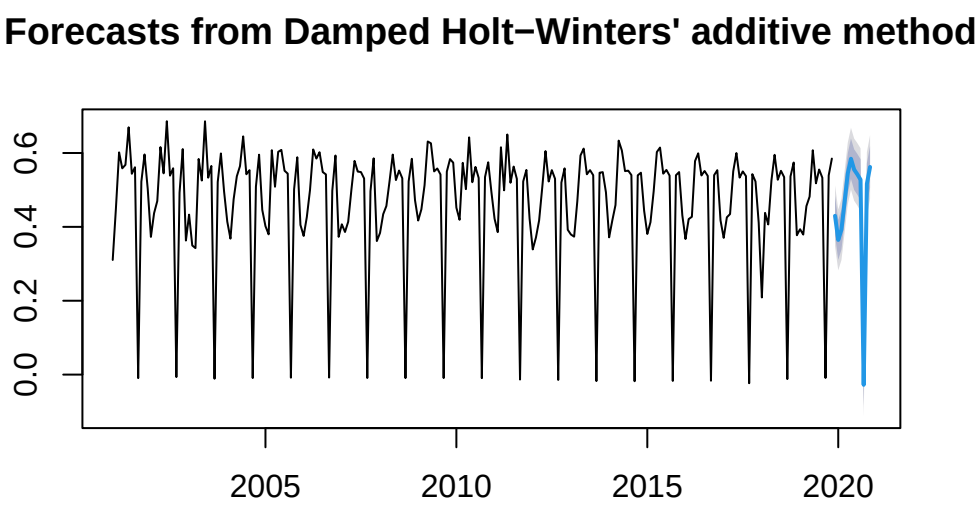



```
# # Count of na values to dataframe
# # Calculating Trend and Seasonal Strength
# evi.trend$NA_Values[i] <- na.cnt
# evi.trend$Trend[i] <- summary$coefficients[2,1]
# evi.trend$Trend_Strength[i] <- round(max(0,1-(var(Rt)/var(Tt+Rt))),1)
# evi.trend$Seasonal_Strength[i] <- round(max(0,1-(var(Rt)/var(St+Rt))),1)
# evi.trend$P_value[i] <- summary$coefficients[2,4]
# evi.trend$R_Squared[i] <- summary$r.squared
# evi.trend$Standard_Error[i] <- summary$sigma
# evi.trend[i,]

ggdensity(tdx,fill = "#0073C2FF",color ="#0073C2FF",add = "mean",rug = TRUE)
```

```
Don't know how to automatically pick scale for object of type ts. Defaulting to continuous.
```




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



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

5.2 References

References

[1] Godwin Akweiteh Allotey. “Stop galamsey in 3 weeks or face the law - Amewu”. In: *Ghana News* (Mar. 29, 2017). URL: http://citifmonline.com/2017/03/29/stop-galamsey-in-3-weeks-or-face-the-law-amewu/.

[2] Marian Efe Ansah. “Galamsey, pollution destroying water bodies in Ghana - Water Company”. In: *Ghana News* (Mar. 22, 2017). URL: http://www.leakxgh.com/2017/05/galamsey-pollution-destroying-water.html.

[3] Abigail Barenblitt et al. “The large footprint of small-scale artisanal gold mining in Ghana”. In: *Science of the Total Environment* 781 (Aug. 10, 2021). PMID: 33812105 Publisher: Elsevier B.V. DOI: 10.1016/j.scitotenv.2021.146644.

[4] “Gold, guns and China: Ghana’s fight to end galamsey”. In: *African Arguments* (May 30, 2017). URL: https://africanarguments.org/2017/05/30/gold-guns-and-china-ghanas-fight-to-end-galamsey/.

[5] Zindzy Gracia. “Causes and effects of galamsey in Ghana”. In: *Yen.com.gh - Ghana news.* (Jan. 31, 2018). URL: https://yen.com.gh/104844-causes-effects-galamsey-ghana.html.

[6] Joyce Gyekye. “MD of Ghana Water Company Limited says fight against galamsey is being lost”. In: *Ghana Broadcasting Corporation* (). URL: http://www.gbcghana.com/1.11923484.

[7] F. Owusu-Nimo et al. “Spatial distribution patterns of illegal artisanal small scale gold mining (Galamsey) operations in Ghana: A focus on the Western Region”. In: *Heliyon* 4.2 (Feb. 1, 2018). Publisher: Elsevier Ltd, e00534. DOI: 10.1016/j.heliyon.2018.e00534. URL: http://www.sciencedirect.com/science/article/pii/S2405844017325963.

[8] “Women die in Ghana mine collapse”. In: (Nov. 12, 2009). URL: http://news.bbc.co.uk/2/hi/africa/8356343.stm.