

Contents

1	CHAPTER ONE	3
2	INTRODUCTION	3
2.1	Background of The Study	5
2.2	Problem Statement	6
2.3	Research Question	6
2.3.1	Research Objectives	6
2.4	Significance Of The Study	7
2.5	Scope of The Study	7
2.6	Limitation Of The Study	7
2.7	Organization of The Study	7
3	CHAPTER TWO	7
4	LITERATURE REVIEW	7
4.1	Theoretical Review	7
4.1.1	WHAT IS TIME SERIES ANALYSIS?	8
4.1.2	VARIOUS METHODS USED IN TIME SERIES FORE- CASTING AND COMPARING FEW METHODS	9
4.1.3	ADVANTAGES AND DISADVANTAGES OF TIME SE- RIES FORECASTING	17

4.1.4	18
4.1.5	18
4.2 Empirical Review	18
5 CHAPTER THREE	18
6 METHODOLOGY	18
6.1 Research Design	20
6.2 Specification of the Model	20
6.2.1 Data Representation	20
6.2.2 The Analysis Of Variance (ANOVA) Method	20
6.2.3 The Empirical * Theory model	20
6.2.4 Assumptions Underlying EBCT Model 1	20
6.2.5 Parameter Estimation	20
7 CHAPTER FOUR	20
8 ANALYSES AND FINDINGS	20
8.0.1 Summary Statistics	20
8.0.2 Distribution of Claim Amounts.	20
8.0.3 Time Series Trend Of Insurance Claim In Ghana Data	20
8.0.4 The Analysis Of Variance (ANOVA) Model Estimates.	20
8.0.5 The Empirical * Model Approach	20

8.0.6	Comparing ANOVA model and * model Premium E . . .	20
8.0.7	Expected Claims versus Actual Claims	20
8.0.8	Goodness of Fit Test	20
9	CHAPTER FIVE	20
10	CONCLUSIONS AND RECOMMENDATIONS	20
10.0.1	Summary	20
10.0.2	Conclusions	20
10.0.3	Recommendations	20
11	References	20

1 CHAPTER ONE

2 INTRODUCTION

All thing change, but how we respond to change is our responsibility, to fare it or embrace it. Resisting change leads to one fiat. Our own extinction. Time is a smybole of freedom and peace

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 its Environmental impact. Galamsey ("gather them and sell"), (Owusu Nimo 2018) is the term given by local Ghanaian for illegal small-scale gold mining in Ghana (David Yaw Danquah 2019). 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 Dompase, 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 (News 2009).

Illegal mining causes damage to the land and water supply (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 2021).

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 (Burrows 2017). Their operations are mostly in the southern part of Ghana where it is believed to have substantial reserves of gold deposits, usually within the area of large mining companies (Barenblitt 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.

2.1 Background of The Study

As Galamsey is considered an illegal activity, their operations are hidden to the eyes of the authorities. So locating them is quite tricky, but with satellite imagery, it is 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.

2.2 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.

2.3 Research Question

2.3.1 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
- Dissociate or single out the variability climate is responsible for in vegetation loss

2.4 Significance Of The Study

2.5 Scope of The Study

2.6 Limitation Of The Study

2.7 Organization of The Study

3 CHAPTER TWO

4 LITERATURE REVIEW

4.1 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.

4.1.1 WHAT IS TIME SERIES ANALYSIS?

Makridakis and Hibon, in time series analysis researchers have conducted a competition named M-competition in 1987 (Makridakis, S., Hibon, M., Lusk, E. and Belhadjali, M., 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. According to Cryer, J.D., 1986. 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.

4.1.2 VARIOUS METHODS USED IN TIME SERIES FORECASTING AND COMPARING FEW METHODS

There are 11 classical time series forecasting methods, they are:

- Autoregression (AR)
- Moving Average (MA)
- Autoregressive Moving Average (ARMA)
- Autoregressive Integrated Moving average (ARIMA)

- Seasonal Autoregressive Integrated Moving Average (SARIMA)
- Seasonal Autoregressive Integrated Moving Average with Exogenous Regressors (SARIMAX)
- Vector Autoregression (VAR)
- Vector Autoregression Moving Average (VARMA)
- Vector Autoregression Moving Average with Exogenous Regressors (VARMAX)
- Simple Exponential Smoothing (SES)
- Holt's Winter Exponential Smoothing (HWES)

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.

A. 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 (Cryer, J.D., 1986). 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.

B. Stationarity: Stationarity alludes to an irregular process that creates a time-series which has mean, and distribution to be constant through time. It is an important aspect to consider for building machine learning models. 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).

Both the papers have explained the stationarity to have weak stationarity and strong stationarity. However, which performs the best is not discussed in either paper. However, both have stated that the concept of stationarity helps in machine learning models.

C. 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).

D. Autoregressive models (AR): Autoregressive models 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 is given as:
(1) $In(1) = Target = Intercept + Coefficient \times Lagged\ target + Error$ Model that depends only on one lag in the past is called as 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).

E. 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:

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).

F. 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.

G. 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 be 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

1) Plotting patterns in correlation:

There are two functions used for plotting patterns in correlation: **a) 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). **b) 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) Automatic selection techniques:

There are three commonly used techniques for automatic selection of time series model: **a) 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).

b) 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).

c) 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.

H. 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).

I. 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 and 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.

J. 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.

4.1.3 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

4.1.4

4.1.5

4.2 Empirical Review

5 CHAPTER THREE

6 METHODOLOGY

Time series data is the collection of observations made sequentially at different points in time. Because data points in time series are collected at adjacent time periods there is potential for correlation between observations. we propose some new tools to allow machine learning classifiers to cope with time series data. We first argue that, time-series classification problems can be solved by detecting and combining local properties or patterns in time series. Then, a technique is proposed to find patterns which are useful for classification. These patterns are combined to build interpretable classification rules. First, we will pull Sentinel 2 to select NDVI and EVI data from Google Earth Engine, applying a quality filter to mask poor quality pixels. Instead of performing our analysis on the imagery itself, we will be summarizing the mean NDVI and EVI value, this will allow the analysis to take less time while producing a visually appealing and informative map. Some cells may not contain NDVI and EVI for a given month,

to correct this, we will apply smoothing method using an ARIMA function. Once NA values are remove, we will decompose the time series to remove seasonality and fit a linear model to the normalized data. Once we have extracted the linear trend, we will then make a move to classifier our dataon the map and map it.

6.1 Research Design

6.2 Specification of the Model

6.2.1 Data Representation

6.2.2 The Analysis Of Variance (ANOVA) Method

6.2.3 The Empirical * Theory model

6.2.4 Assumptions Underlying EBCT Model 1

6.2.5 Parameter Estimation

7 CHAPTER FOUR

8 ANALYSES AND FINDINGS

8.0.1 Summary Statistics

8.0.2 Distribution of Claim Amounts.

8.0.3 Time Series Trend Of Insurance Claim In Ghana Data

8.0.4 The Analysis Of Variance (ANOVA) Model Estimates.

8.0.5 The Empirical * Model Approach

8.0.6 Comparing ANOVA model and * model Premium E

8.0.7 Expected Claims versus Actual Claims

22

8.0.8 Goodness of Fit Test

9 CHAPTER FIVE