Galamsey

Boniface Kalong

7/2/2022

[10pt]report

Contents

CHAPTER ONE

INTRODUCTION

All thing change, but how we respond to change is our responsibility, to fare it or embrasse 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 it's Environmental impact. Galamsey("gather them and sell"),(OwusuNimo2018) is the term given by local Ghanaian for illegal small-scale gold mining in Ghana (DavidYawDanquah2019). The major cause of Galamsey is unemployment among the youth in Ghana(Gracia2018). 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(News2009).

Illegal mining causes damage to the land and water supply(Ansah2017). 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(Allotey2017). 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(Gyekye2021).

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(Burrows2017). 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(Barenblitt2021). As a group, they are economically disad vantaged. 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.

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.

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.

Research Question

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

Significance Of The Study

Scope of The Study

Limitation Of The Study

Time series modeling aims to build an explanatory model of the data without over fitting the problem set, to use as simple a model as possible while accounting for as much of the data as possible. When breaking down time series data into component parts, remote sensing data has additional limitations that make this more challenging. It is almost inevitable that you will not get this same level of precision from remote sensing data. Additionally, atmospheric conditions can skew the visual results, where the hue of the vegetation changes drastically from image to image due to atmospheric conditions (fog,ground moisture, cloud cover).

Organization of The Study

CHAPTER TWO

LITERATURE REVIEW

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.

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.

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)* [6, 12, 23] and *Moving Average (MA)* [6, 23] models. Combining these two, the *Autoregressive Moving Average (ARMA)* [6, 12, 21, 23] and *Autoregressive Integrated Moving Average (ARIMA)* [6, 21, 23] models have been proposed in literature. The Autoregressive Fractionally Integrated Moving Average (ARFIMA) [9, 17] model generalizes **ARMA** and **ARIMA** models. For seasonal time series forecasting, a variation of **ARIMA**. The *Seasonal Autoregressive Integrated Moving Average (SARIMA)* [3, 6, 23] 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.
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.

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..

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).

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 $|\phi_1| < 1$, 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).

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:

$y_t = \omega + \theta e_{t-1} + e_t$

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).

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.

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. $\mu = E\left[x_t\right]$. 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. **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

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
- 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.

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

Time Series Forecasting Using Support Vector Machines

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].

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.

In this chapter we shall describe few important performance measures which are frequently used by researchers, with their salient features.

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, $\bar{y} = \dfrac{1}{n}\sum_{t=1}^{n} y_t$ is the test mean and

$\sigma^2 = \dfrac{1}{n-1}\sum_{t=1}^{n}(y_t - \bar{y})^2$is the test variance.

The Mean Forecast Error (MFE)

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

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 defferent 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 rate/ 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.

Also, in [16] studied the 20 years spatiotemporal **LULC** in Hawalbagh block India, the supervised classification using, Maximum Likelihood Classification was used. The result shows 5 LULC classes namely agriculture, barren, built-up, vegetation, and water body, where 3.51% and 3.55% increase in vegetation and built-up areas, while a decrease of 5.46%, 1.52% and 0.08% of barren land agriculture, and water body respectively was noticed.

Furthermore, in [17] study the **LULC** change of watershed in Pakistan from 1992 to 2012 using the supervised classification of maximum likelihood algorithm in Erdas Imagine. The finding shows 5 **LULC** classes agriculture, bare soil/rocks, settlements, vegetation and water. Also, the water body and vegetation are decreasing in favour of settlements, agriculture and bare soil rapidly from 38.2% and 74.3% respectively.

Also, in [18] study, both unsupervised (ISODATA) and supervised (MLA) methods were used for LULC classification. Change detection and Markov change analysis methods used to measure the LULC changes and generate future LULC map respectively in Mansoura and Talkha of Egypt from 1985 to 2010. The finding shows four LULC classes viz agriculture, barren land, built-up area and water body. Also, a significant change was noticed in agricultural land and built-up area to tune of 33% decrease and 30% increase respectively, while barren land and water bodies changes were minimal.

Similarly, in [19] studied the LULC classification of Sawantwadi taluka, in India. The hybrid, parametric (MLA and ISODATA), and nonparametric (DT) methods were used. The finding shows the classified LULC of the forest, water, built-up, agriculture, plantation, fallow land, open and dense shrubland, stone quarry,and grassland with an accuracy assessment of 93% and koppa of 0.92.

Also, in [20] measured the LULC change in Seramban. In the study, Natural Breaks (Jenks) and Normalized Difference Vegetation Index (NDVI) methods were used for classification and difference from 1990 to 2000. The result shows four classes of LULC viz barren land, built-up area, vegetation and water body. A 13% decrease in vegetation cover was noticed while other land use/ land cover increase by 3.7% accordingly with an accuracy assessment of 87% and 88% respectively.

Likewise, in [21] studied twenty-five years the spatiotemporal urban growth of Kuala Lumpur, using the Maximum Likelihood Classification (MLC) method for years 1989, 2001 and 2014. The result shows 4 LULC classes agriculture, urban/built-up, forest, and water body. Also, a rapid increase of the built-up areas and agricultural land was noticed while other land covers decrease very significantly.

Also, in [22] used NDVI method to study the LULC change of Sambas watershed, in Malaysia for the years 1990, 2002 and 2013. The results show 5 LULC classes viz barren land, forest, grassland, shrub and water body. A significant decrease in the forest cover was noticed, while barren land and grassland was increasing accordingly throughout the period.

Also, in [23] studied the ten years LULC changes of Aluva taluk, in India from 2000 to 2010. Supervised classification (MLA) and Change Detection Analysis were used for LULC classification and mapping. The result shows 8 LULC classes viz Agriculture, Built up, Cropland, Fallow land, Forest deciduous, Forest evergreen, Plantation, and Waterbody. A significant change in some LULC was noticed. Furthermore, in [24] used change detection matrix in the study of LULC change of Kolong River basin of India in the years 1967-68 and 2014. The finding shows six LULC classes viz agricultural land, built- up, forest, open space, shrub and wetland. A significant change in two primary land use, agricultural land and built-up area with the former decreasing in the year 1967-68 and the later increasing much in the year 2014 respectively.

Similarly, in [25] used supervised *Maximum likelihood classification (MLC)*, and *multi-layer Perceptron-Markov chain analysis(MLP-MCA)* to monitor the LULC changes as well predict future **LULC** changes in Patna India. The result shows seven classes of LULC viz agriculture, built-up, Fallow land, Riverbed, Shrubs, Vegetation, Water bodies. A slight change in the **LULC** was noticed across the classes, in a decrease and increases pattern. Also, the prediction shows significant changes in the built-up area.

Moreover, in [26] studied 31 years **LULC** change in Beressa Watershed Ethiopia from 1984 to 2015. Unsupervised ISODATA using Erdas imagine and Change Detection methods were used in **LULC** classification and change magnitude respectively. The finding shows six classes of **LULC** viz barren land, farmland, forest/plantation, grazing land, settlement, and water body. A continuous increase in settlement and farmland, while grazing land and barren were decreasing over the three decades.

Furthermore, in [27] studied **LULC** changes in Udhaim river basin in Iraq using Landsat TM image for 2006 and OLI 2015. Spectral indices **(NDVI, NDBI, NDWI, NDBaI, and CI)** methods were used to study for **LULC** classification and changes. The finding shows 5 **LULC** according to each index viz bare land, built-up, soil crust, vegetation and water body. Also, significant changes were noticed in the LULC with 3% increase soil crust, and 2.43%, 0.6%, 0.55% and 0.22% decrease in vegetation cover, built-up area, bare land and water body respectively.

Likewise, in [28] study the **LULC** of Kan basin from 2000 to 2016. Supervised classification of (MLA) method and Change Detection Analysis were used. The finding shows 5 **LULC** classes viz bareland, built up, garden, pasture, and water body. Also, found a slight increase of 0.3% and 0.2 % of pasture and build-up areas respectively, while bare land, garden, and water body decrease slightly over sixteen years by 0.4% and 0.01% respectively with an accuracy assessment of 86% and 89% for years 2000 and 2016 respectively. Moreover, in [29] study the **LULC** change of hotspot area in Pune region using Landsat images for 1972, 1992, and 2012. Change Detection and Statistical Cluster Analysis method for LULC change were used. The result shows 10 **LULC** classes viz cropland, fallow land, forests, industrial, rivers, rural, tree clads and Wastelands with an accuracy assessment range from 77% to 97%. A significant change was noticed, an increase in fallow land, industrial, and built-up areas around the hotspot region.

Also, in [30] studied ten years **LULC** change and transformations in Kanyakumari coast India. The study used supervised classification of (MLA) and Change Detection. The finding shows 8 **LULC** classes viz barren land, built-up, beach face, cultivable lands, fallow land, mining, vegetation, and water body. A significant change in the coastal **LULC** of Kanyakumari was noticed, some **LULC** changing to another over the ten-year period with accuracy assessment of 81.16% and 77.52% for image 2000 and 2011 respectively.

Also, in [31] study the **LULC** change in Tanguar Haor, Bangladesh. The study used supervised classification in (MLA) for classification and **CVA**, **NDVI** and **NDWI** analysis were used to for change detection analysis. The result shows 4 **LULC** classes deep water, vegetation, shallow water, and settlement. A significant change in the **LULC** with about 40% of the total area transformation, i.e. changing from one **LULC** to another.

Furthermore, in [32] used Google Earth and GIS Operation to study the LULC changes in Muar sub-district in Malaysia from the year 2010 to 2015. The results show 6 **LULC** classes viz agriculture, barren land, built-up, forest, open/ reaction space and roads. Also shows a significant change in the overall LULC across the district, i.e. some land covers a been converted into another form.

Also, in [33] studied the spatial-temporal **LULC** change in Astrakhan city, Russia, from the year 2000 to 2015. Supervised (MLA) and change detection analysis was used to classify the **LULC** and monitor the **LULC** change within the period. The result shows 5 **LULC** classes viz agriculture, bare-land, settlements, vegetation and water body. It further shows large vegetation dilapidation and water logging in different parts of the. Astrakhan city Also, in [34] used supervised (MLA) classification and Stochastic Markov (St Markov) method, to study the **LULC** change and predict future urban land use in Jodhpur City in India from 1990 to 2000. The finding shows 5 **LULC** classes viz built-up, mining area, other land, vegetation, and water body. It further shows the rates of changes from one LULC to the other.

Furthermore, in [35] studied the **LULC** change Khan-Kali watershed and Anas River from Gujarat, India between the year 2001 and 2011. Supervised classification (MLA) and **NDVI** and **NDWI** methods were used for LULC classification and change detection respectively. The results 7 **LULC** classes viz agriculture, barren land, built-up, forest, riverine sand, shrubland, and water body with an overall accuracy assessment of 91.8% and 95.5% for years 2001 and 2011 respectively. Also, a significant increase in the water body and shrub land while a decrease in the forest, barren land and riverine LULC.

Moreover, in [36] studied **LULC** classification in Okara, Pakistan. The supervised classification of **MLA** and *Synthetic Aperture Radar (SAR)* methods were used. The finding shows four classes of LULC, barren land, built-up, water body and vegetation, with an overall accuracy of 80% and 0.69 Kappa coefficients. Also, in [37] in their study, *Google Earth Engine (GEE)* and *Normalized Difference Vegetation Index (NDVI)* method were used for **LULC** classification and detect major **LULC** changes from 1985 to 2014 in Beijing respectively. The finding shows seven classes of **LULC** viz cropland, grassland, forest, shrub, water body, built-up, and barren land, with an overall accuracy of 86.61%.

“‘r if(!require("pacman"))install.packages("pacman") pacman::p$_{l}oad(char =$
$c('rgee','reticulate','raster','tidyverse','dplyr','sf','mapview','mapeddit','caret','forcats','reticulate','rgee','remotes','$
$F, update = F, character.only = T)$“‘

“‘ Warning in pacman::p$_{l}oad(char = c("rgee", "reticulate", "raster", "tidyverse", : Failedtoinstall/load :$
$tidyverse, mapeddit$“‘

“‘r library(rgee)

library(reticulate) ee$_{i}nstall()ee_{c}heck()$“‘

“‘r ee$_{I}nitialize("kalong", drive = TRUE)initializeGEE,$“‘

“‘ – rgee 1.1.4 ——————————————— earthengine-api 0.1.317 – v user: kalong v Google Drive credentials: “‘

“‘ Auto-refreshing stale OAuth token. “‘

“‘ v Google Drive credentials: FOUND v Initializing Google Earth Engine: v Initializing Google Earth Engine: DONE! v Earth Engine account: users/Earth$_{S}cience -$“‘

“‘r this will have you log in to Google Drive “‘

“‘r library(’sf’) Load shape file

setwd("C:/Users/Guy/Documents/GitHub/Artisanal-Mining-In-Ghana-Galamsey/New Regions") aoi <- read$_{s}f('Ghanashpfile/GHA/gadm41_{G}HA_{1}.shp')aoi < -st_{t}ransform(aoi, st_{c}rs(4326))aoi.ee < -st_{b}box(aoi)st_{a}s_{s}fc()sf_{a}s_{e}e()ConvertsittoanEarthEngineObject$“‘

These functions return the QA value from MODIS imagery and apply a quality Mask, returning quality masked EVI values, this technique was adapted from one presented by Cesar Aybar (one of the rgee authors) [here](https://csaybar.github.io/blog/2020/06/15/rgee$_{0}2_{i}o/$).

“‘r getQABits <- function(image, qa) Convert binary (character) to decimal (little endian) qa <- sum(2$^{(}which(rev(unlist(strsplit(as.character(qa), ""))) == 1)) - 1))Returnamaskbandimage, givingtheqavalue.image$bitwiseAnd(qa)$lt(1)$

mod.clean <- function(img) Extract the NDVI band ndvi$_{v}alues < -img$select("EVI") Extract the quality band ndvi$_{q}a < -img$select("SummaryQA") Select pixels to mask quality$_{m}ask < -getQABits(ndvi_{q}a, "11")Maskpixelswithvaluezero.ndvi_{v}alues$updateMask(quality$_{m}ask$)divide(ee$Image$constant(10000)) 0.0001 is the MODIS Scale Factor

modis.evi <- ee$ImageCollection("MODIS/006/MOD13Q1"$)filter(ee$Filter$date(’2000-01-01’,’2022-01-01’))$map(mod.clean)$“‘$Nowwewillcreateahexagonalgridoverthestudyarea$

“‘r library(tibble) aoi.proj <- st$_{t}ransform(aoi, st_{c}rs(2392))hex < -st_{m}ake_{g}rid(x = aoi.proj, cellsize = 17280, square = FALSE)st_{s}f()rowid_{t}o_{c}olumn('hex_{i}d')hex < -hex[aoi.proj,]plot(hex)$“‘

![](Galamsey$_{f}iles/figure - latex/unnamed - chunk - 6 - 1.pdf$) <! - - - - - > $Nowwewillusethegridcreatedabovetoextractthemean EVIvalueswithineachcellfortheyears 2000 - 2020.$

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.

“‘r converting the data to a transposed data frame tsv <- data.frame(evi = t(evi.df[i, 2:ncol(evi.df)])) colnames(tsv) <- c("evi") write.csv(tsv,"Data/tsv.csv") head(tsv) let’s take a look “‘

“‘ evi 2001-01-17 0.3103816 2001-03-22 0.6017811 2001-04-23 0.5585050 2002-01-17 0.3728227 2002-02-02 0.4369971 2002-04-07 0.5701539 “‘ CHAPTER THREE

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.

Research Design

Specification of the Model

Data Representation

The Analysis Of Variance (ANOVA) Method

The Empirical * Theory model

Assumptions Underlying EBCT Model 1

Parameter Estimation

“‘r We want to get an idea of the number of entries with no EVI value na.cnt <- length(tsv[is.na(tsv)]) evi.trend$na.cnt[i] < -na.cnttd < -tsvmutate(month = month(as.Date(rownames(tsv))), year = year(as.Date(rownames(tsv))))group_by(year, month)summarise(mean_evi = mean(evi, na.rm = T), .groups = "keep")as.data.frame()head(td)$“‘

“‘ year month mean$_evi1200110.31038162200130.60178113200140.55850504200210.37282275200220.43699716200240.6160278$“‘

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.

“‘r dx$mean_evi < -NAtdx < -rbind(td, dx)arrange(date)write.csv(tdx, "Data/tdx.csv")tdx < -read.csv("Data/tdx.csv")head(tdx)$“‘

“‘ X year month mean$_evidate11200110.31038162001 - 01 - 01221620012NA2001 - 02 - 0132200130.60178112001 - 03 - 0143200140.55850502001 - 04 - 01551020015NA2001 - 05 - 01661020016NA2001 - 06 - 01$“‘

“‘r na.cnt <- length(tdx[is.na(tdx)]) Convert data to time series. tdx <- ts(data = tdx$mean_evi, start = c(2001, 1), end = c(2019, 11), frequency = 12)plot(tdx)$“‘

![](Galamsey$_files/figure - latex/unnamed - chunk - 13 - 1.pdf$) <! - - - - >

“‘r library(imputeTS) tdx <- if(na.cnt > 0)imputeTS::na$_kalman(tdx, model = "auto.arima", smooth = T)elsetdxplot(tdx)$“‘

![](Galamsey$_files/figure - latex/unnamed - chunk - 14 - 1.pdf$) <! - - - - >

“‘r tdx.dcp <- stl(tdx, s.window = ’periodic’) plot(tdx.dcp) “‘

![](Galamsey$_files/figure - latex/unnamed - chunk - 15 - 1.pdf$) <! - - - - >

“‘r library(forecast) “‘

“‘ Warning: package ’forecast’ was built under R version 4.1.3 “‘

“‘r Tt <- trendcycle(tdx.dcp) St <- seasonal(tdx.dcp) Rt <- remainder(tdx.dcp) plot(Tt) “‘

![](Galamsey$_files/figure - latex/unnamed - chunk - 16 - 1.pdf$) <! - - - - >

“‘r plot(St) “‘

![](Galamsey$_files/figure - latex/unnamed - chunk - 16 - 2.pdf$) <! - - - - >

“‘r plot(Rt) “‘

![](Galamsey$_{f}iles/figure-latex/unnamed-chunk-16-3.pdf$) <! - - - - >
$StationarityWheninvestigatingatimeseries, oneofthefirstthingstocheckbeforebuildinganARIMAmodelistocheckthatthes$

Here, we will look at a couple methods for checking stationarity. If the time series is provided with seasonality, 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 for unit root Another test we can conduct is the 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).

"'r library(tseries) "'

"' Attaching package: 'tseries' "'

"' The following object is masked from 'package:imputeTS': na.remove "'

"'r adf.test(Rt) "'

"' Augmented Dickey-Fuller Test data: Rt Dickey-Fuller = -8.639, Lag order = 6, p-value = 0.01 alternative hypothesis: stationary "'

"'r adf.test(Tt) "'

"' Augmented Dickey-Fuller Test data: Tt Dickey-Fuller = -3.4545, Lag order = 6, p-value = 0.04798 alternative hypothesis: stationary "'

"'r adf.test(tdx) "'

"' 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

"'r plot.new() frame() The Stationary Signal and ACF plot(Rt,col= "red", main = "Stationary Signal") "'

![](Galamsey$_{f}iles/figure-latex/unnamed-chunk-18-1.pdf$) <! - - - - >

"'r acf(Rt, lag.max = length(Rt), xlab = "lag ", ylab = 'ACF', main = ") "'

![](Galamsey$_{f}iles/figure-latex/unnamed-chunk-18-2.pdf$) <! - - - - >

"'r The Trend Signal anf ACF

plot(Tt,col= "red",main = "Trend Signal") "'

![](Galamsey$_{f}iles/figure-latex/unnamed-chunk-18-3.pdf$) <! - - - - >

"'r acf(Tt, lag.max = length(Tt), xlab = "lag", ylab = "ACF", main = ") "'

![](Galamsey$_{f}iles/figure-latex/unnamed-chunk-18-4.pdf$) <! - - - - >
$Notably, thestationarysignal(topleft)resultsinfewsignificantlagsthatexceedtheconfidenceintervaloftheACF(bluedashed$

"'r tdx <- data.frame(time = c(1:length(tdx)), trend = tdx - tdx.dcp$time.series[,1])trend.summ <$ $-summary(lm(formula = trend\ time, data = tdx))$"'$CHAPTERFOUR$

ANALYSES AND FINDINGS

Summary Statistics

Distribution of Time Series Trend.

"'r ggdensity(tdx, x = "tmie",y = "trend",fill = "0073C2FF",color ="0073C2FF",add = "mean",rug = TRUE) "'

Time Series Trend Of Insurance Claim In Ghana Data

The Analysis Of Variance (ANOVA) Model Estimates.

The Empirical * Model Approach

Comparing ANOVA model and * model Premium E

Expected Claims versus Actual Claims

Goodness of Fit Test

“‘r plot(tdx) abline(a = trend.summ$coefficients[1,1], b = trend.summ$coefficients[2,1], col = ’blue’) “‘

![](Galamsey$_files/figure-latex/unnamed-chunk-21-1.pdf) <!----->$

“‘r Count of na values to dataframe Calculating Trend and Seasonal Strength evi.trend$NA_Values[i] <-na.cntevi.trend$Trend[i] <- trend.summ$coefficients[2,1]evi.trend$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$_value[i] <-trend.summ$coefficients[2,4] evi.trend$R_Squared[i] <-trend.summ$r.squared evi.trend$Standard_Error[i] <-trend.summ$sigma evi.trend[i,] “‘

“‘ hex$_idna.cntNA_ValuesTrendP_valueR_SquaredStandard_Error1340152 -$ $0.000111840.0059281960.033165540.03974499Trend_StrengthSeasonal_Strength$10.21“‘

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

CHAPTER FIVE

CONCLUSIONS AND RECOMMENDATIONS

Summary

Broadly speaking, in this introductory 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.

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.

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