Time Series Classification and Survival Analysis for Forecasting

Luckyn, Boma Josiah , Enoch, Joseph Diema

ABSTRACT: The research work is focused on the application of time series and survival analysis to forecast future values from previous values to assist government or organization, plan ahead with precise data using the Kaplan Meier survival method of different classification of time series data to obtain the desired objectives. The understanding of the analysis was gained through the health statistical data that have continuous regular time study. The decision to use time series data is to ascertain the regular time series which is the major variable to determine a survival rate at any given time. A customized software application was developed in Java (NetBeans) programming language for its implementation. The research extensively carried out testing using synthesized data to execute Time Series Classification and Survival Analysis with possible outcome for user interpretation on events. The results obtained from the implemented standalone software application, were validated with other proven data and results.

Index terms: Time series, survival analysis, Kaplan Meier, econometric analysis, biology analysis, health statistics, data mining.

IJSER

1 INTRODUCTION

Time series is a succession of discrete-time data points drawn in timely order and its sequence taken at equal times. Examples include Sunspot counts [10] and ocean tide heights, etc. Time series is plotted with line charts and used in processing signals, recognizing patterns, statistics, and forecasting weather. Moreover, it is used in predicting earthquake, communications, and engineering fields [2]. Time series has features such as means, standard deviations, maximum, and minimum values, skewness, and Kurtosis. Other related features include cross-correlations, Auto correlations, orders, parameters of AR (Auto Regression) part, and parameters of MA (Moving Average) part. Features classifying segments of time series are specific on domains but Fourier analysis tools allow examination of signals in time frequency domains. Time series is used in forecasting future values from previous values. Time series data is natural and temporal in its ordering, which makes it distinct from cross-sectional studies that do not have natural ordering.

Time series classification (TSC) of problems trains classifiers on sets of cases. Each case has ordered set of attributes that are real valued and labelled per class.

Time series classification involves analyzing data obtained in time series to extract meaningful statistics and other attributes of data.

Time series analysis is used to accomplish goals such as descriptive analysis, spectral analysis, forecasting, intervention, and explanative analysis. Descriptive analysis determines the trend and patterns of a time series through plotting and using other complex techniques. Analysts look at overall trends whether it is increasing or decreasing, cyclic patterns whether they are seasonal, outliers (erroneous data points), and turning points with different trends in data series [13]. Spectral analysis describes time series variation is accounted by cyclic components. Forecasting is aimed at identifying a future behaviour through a past behaviour by predicting confidence limits. Intervention analysis explains presence of events, which change time series. Explanative analysis (cross correlation) involves usage of one or many time series, which results in estimating dependent time series [13]. Regression analysis focuses on questions of statistical inference.

Continuous time series data involves continuous observations made, such as output of carbon dioxide from an engine. Discrete time series data involves observations made at certain times, such as composition of animal species every month [6]. Stationary data fluctuates around constant value while non-stationary data is a series of parameters within a cycle and one that changes over time. Deterministic data is exactly predicted while stochastic time series data is partly determined by past and future values and described within probability distribution.

Time series data transformed to stabilize variance using logarithmic transformation make seasonal effect additive, which makes constant effect year-to-year using logarithmic transformation, and make normal distributions of data to reduce skewness of data applied in appropriate statistics [1]. Frequency analysis decomposes time series into sine and cosine functions plotted by wavelengths. The wavelengths are analyzed to recognize the most relevant.

Survival analysis focuses on time to the event data. It is the use of statistical methods to analyze, estimate and interpret the survivor rate from survival data derived from laboratory studies, epidemiological studies, animal studies and economic studies. Survival data could be the outcomes of studying chronic diseases. Survival time is variable measuring time from set

starting time to a particular interest end. However, it is difficult to obtain complete survival time because of censoring. Censored data is found when some patients lose to the follow up when the period of study ends. Secondly, when certain subjects do not experience event before end of study or some individuals withdraw from the study[3].

The survival function is

S(t) = P (T > t) (1)

S(t) = Survival Function

P= Probability

T= Survival Time

Cumulative survival Function =

$$\frac{\text{\# Surviving past}}{\text{Total No. of subject at risk per Time}} \quad (2)$$

Related Works:

Survival analysis refers to a statistical analysis that delves into the assessment of the time duration interval until one or more events happens [28]. In business, finance or economic analysis, survival analysis is also referred to as duration modeling, most often dealing with answering financial or economic questions, for example how many employees will survive in a given industry during a serious recession causing massive layoffs. Survival analysis in these fields can also assess how certain factors play to increase or decrease the rate of survival or persistence of certain economic or financial events. Under the survival analysis branch of statistics, the time duration until a certain event occurs is normally the outcome variable.

The major factor in time analysis is measuring when an event will occur, or the hazard of an event occurring [14]. While normal statistical analysis such as regression analysis is done when there is both completeness and certainty of data, survival analysis is applied under circumstances where the data available lacks both. Therefore, the factors necessitating the application of survival analysis are the time-dependent covariates, and under circumstances where the data available is censored [15]. However, **Lee & Wang [16]** offers that it is not a necessary condition that data is censored so that survival analysis can be applied, but if the statistician or researcher only has access to censored data, there lacks another option to the application of survival analysis **p.12**.

Time-dependent covariates are the kind of variables that will change over time [17]. Therefore, when the data available for analysis is concerning the time-dependent variables, then, survival analysis can be applied to determine the outcome, defined in terms of time. On the other hand, the censored data can arise out of different circumstances. For example, the right data censoring is associated with an outcome that has not occurred for all participants, another example were some people losing their jobs while others retain them during a serious inflation or company reorganization [30].

The survival analysis models that have been developed offers varied means to analyzing censored or time-covariate data. The first model that tracks the survival rate of study subjects over time is the Kaplan-Meier Survival Curves, which applies only time as the function [29].

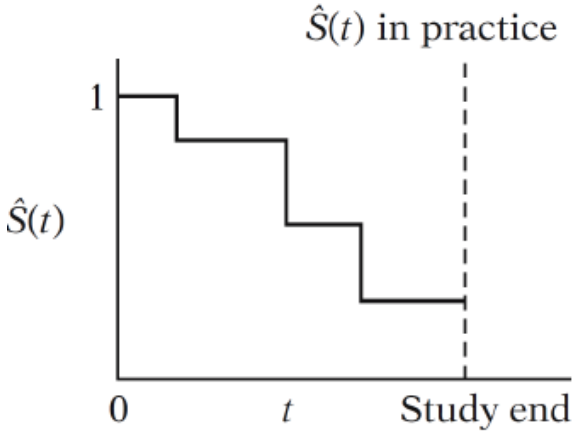

Figure 1 The Kaplan-Meier Survival Curves

S(t)- survival function, t- time data

Methods

The analysis of the proposed time series classification and survival analysis using Kaplan Meier survival estimation process in achieving the desired result is done by using a use case diagram and sequence diagrams

This aspect of the research specifies the requirement analysis of the Time series classification and survival analysis using Kaplan Meier survival estimation process in achieving the desired result. In trying to find out the required parameters to determine the variable to be used, several modelling techniques have been put into consideration like specifying the requirement model which lead to getting the functional and non-functional requirements. Subsequently, use case diagram, Use case Descriptions and User Interface prototypes were discussed.

PROBLEM ANALYSIS

The system is designed to facilitate users to engage in entering time specify data and process to plot graph that can be interpreted. The system should be able to manipulate the different data whether censored and

uncensored. Upon specifying the time or duration of the study of Unemployed graduates by the user. The user can exit from the study if not prepared to continue. Then the user can create new event of entering the collated data into the interface either manual or import into the system. When creating new event, the user is given permission to save, edit, delete, end or/and proceed to the next stage. The user from that point can start with the activity in the software that maps the different table that are necessary to achieve the graph using Kaplan Meier estimation method. That will enable the plotting of graph for interpretation by the user to management about the rate to which the unemployed graduates can be duly employed if they did not drop out of the study or before the study end. The system enables the user to view activity of the different table created. The statistical view as View activity should be properly displayed to cross check previous input data. This detail is sufficient to develop a standalone Java application to demonstrate the design and simulate the required functionalities by using a single computer for user to interpret.

ACTOR

USE CASE DESCRIPTION

To further inform the design, use case descriptions were used to inform design.

TABLE 2 USE CASE DETAILS

Use Case	Description
	The software opening
Specify Time range	A user is required to specify time range before event or not.
Add Event (Create New Event/Load Existing Event)	Upon opening, a user is required to import data or manually enter data set data as text file.
View Event	Upon Importing, a user would check the content imported and proceed into the next stage
Edit Event	Edit will enable for user to either cancel the process or correct error input
Delete Event	This process requires to cancel or close the package
Add Activity	Upon editing, a user is required to check the occurrence of Time Data as the system simulate the edited event data into different table
System function	This process enables the system when initiated to sort data and arrange them accordingly into their different tables in ascending order for analysis
Error handling	This process ensure that system try and catch error entries
View Activity	This process requires the system to enable the user to view numbers simulated by the system for analysis
Map with Activity	This process, the system checks the entered time data, manipulate them to get the frequency of occurrence of each time then checks the censorship or non-censorship by the time data(TD) and also checks the total number of time Data(TN) and counts down from the occurrence. Then Matches the horizontal(TD) and vertical (Survival function(Sf)) axis by using to enable the plotting of the graph
Graph	This process, enables the system to plot the graph of the values of the Survival Function(Sf) marched by the system with the Time data(TD)

FUNCTIONAL REQUIREMENTS

The specific GUI requirements are presented in the below table:

Table 3 **FUNCTIONAL REQUIREMENTS**

Req.ID	Description	Importance
FR1	The diagram must be able to display the visual representation to interface between the users and the system.	High
FR2	There must be at least some direct interaction with the diagram, such as clicking on an event or activity to see a popup of other activities specific events	High
FR3	Present errors check and warnings where necessary in order to prevent the user from manipulating error values.	High

IJSER

FR4	The typed specified time range must be within the speculated range meant to be examined.	High
FR5	When a specified time data range must be defined (Year, month, weekly…..), to ensure accurate presentations	High
FR6	When a censored Time Data is entered with plus sign manually the values must be consistent to avoid error else the system will assume other inconsistent values as non-censored Time Data.	High
FR7	The user should be able to progress with the program to the plotting of graph	High
FR8	User can edit/delete any event entered manually or loaded from existing event.	High
FR9	User should be able to view the Time data entered manually or loaded from existing event. And also check the added activity properly and even save on system before proceeding to plot the graph.	High
FR10	User must ensure that the graph should represent the plotting of the values of the Time Data(TD) against the Survival function(Sf).	High

TABLE 5 **NON-FUNCTIONAL REQUIREMENTS**

Req. ID	Description	Importance
NFR1	The GUI must implement required components which help with user interaction and restrict the Users selections by displaying warning and error dialogs when required.	High
NFR2	The application must be able to function on any computer system that supports java.	Medium
NFR3	Application program must be completed by the submission date of the dissertation.	High
NFR4	The application interface must include all required functionality and present it in a logical structure so that a user can learn how to use it with minimal effort.	High
NFR5	The GUI must be easy to Use and presentable; a prototype is required in order to agree with the user on the layout of the interface.	High

USE CASE MODEL

The use case model of this time series classification and survival analysis contains the use case diagrams, use case description expressed above and the user interface prototype. The use case diagram describes the collection of actions, use cases and their Communications. It can also be likened to describe the interaction between the User and system standpoint:

IJSER

functional description and its major processes (SW Chapter 1b n.d), for the case of this dissertation only one actor is considered.

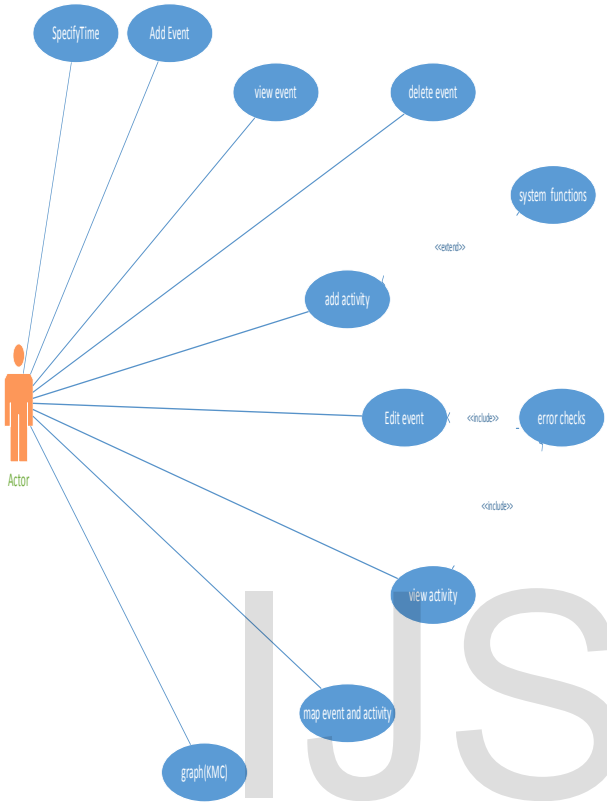

FIG. 2 USE CASE DIAGRAM

Actor could be human or other system or devices.

Oval Shape or ellipses represents the use cases.

Straight lines represent the communication association.

The dotted lines represent the use case relations i.e. (include or extend relation)

Further description of the use case is shown in Appendix A.

SEQUENCE DIAGRAM

Sequence diagram is an interaction diagram used to describe interactions between
objects and the lifelines when each process occurs. Sequence diagram contain the following elements: class roles (roles of object within interaction), lifelines (representation
of time an object exists), activations (time during an object is performing an operation) and
messages (communication between objects).

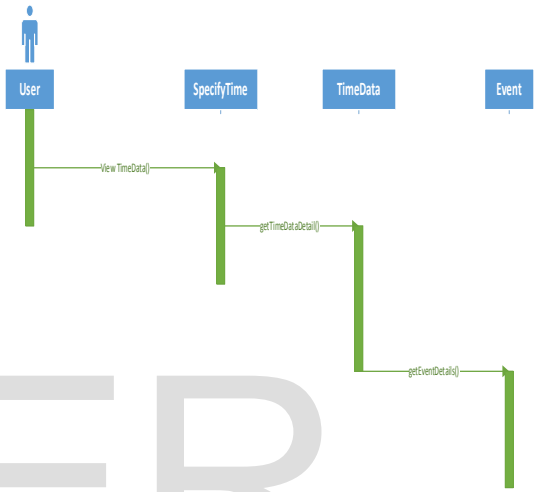

FIG. 3 **SEQUENCE DIAGRAM: VIEW EVENT**

The User open the software specifies on view the TimeData and then Presses Specify Time Button. The system gets the TimeData for the User to enter the desired input. When completed with the entry, the user presses the TimeData for the system proceeds to get the Event details.

PROCESSING STEP INCLUDE:

FIRST STAGE DATA ENTRY:

1. User may define the duration of the study by specifying whether daily, weekly, Monthly or yearly.
2. User must state the study period as defined though the input space can allow any text whether a character, number, or special character.

IJSER

3. Then User must proceed by clicking **TimeData** else **Cancel** the process to continues.

SECOND STAGE DATA ENTRY

1. User must click on **Create New event** for Input interface to display and accept necessary data from the study though the input interface can accept any text format like character, numbers or special character.
2. Upon entry of the text format, User can **remove** unnecessary input by **clicking** on the **clear button** or clicking on **clear error button** else unremoved error can be removed by user **manually** when noticed before proceeding but if not notice user still can proceed.
3. Upon proceeding, User must click on the **activity button** to proceed but if there are error not still removed then the system will generate a message to the user to **re-click** the **Create New Event** to correct necessary error still left before proceeding again to **Activity**.
4. Upon correcting the errors and proceeding to the activity button, the system will convert every proper numbers entered into floating numbers for final check of the user if all the desired numbers where properly capture.
5. User will then click on the **start button** on the interface for the **system** to plot the graph after manipulating the number into the **Kaplan Meier estimation method** and **mapping the tables** required for the **graph**.
6. User will present the graph to the management or for due **interpret** of the study.

APPENDIX B

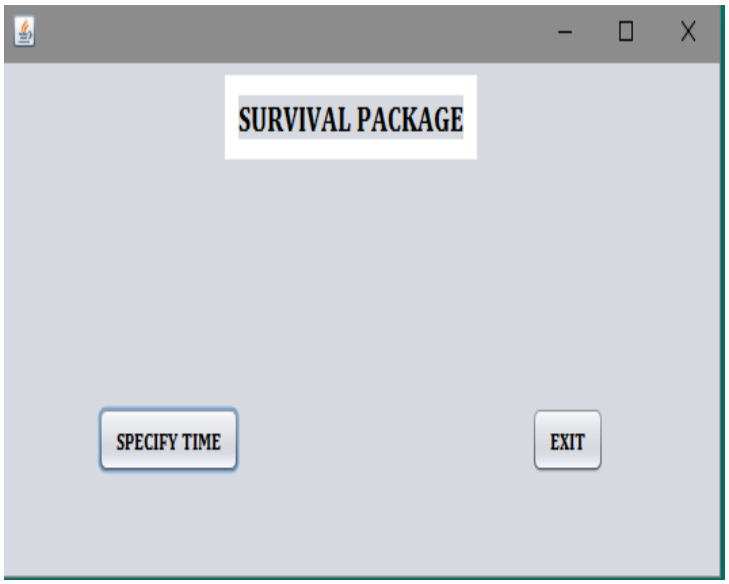

FIG. B1 OPENING PAGE OF THE SURVIVAL PACKAGE

When the Jar file is initiated /clicked, the opening page shows up for the user to start the Survival Package. The Package indicates the Name of the package and the exit button when a user is not ready to do any work. If he/she is willing to do a work then clicking the specify Time Button will enable to package to proceed.

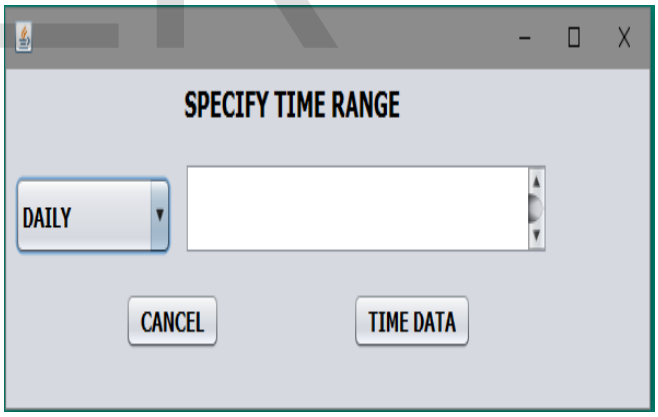

FIG. B2 ENTRY PAGE OF TIME DATA

On Clicking the Specify Time, the package opens up the selection of the Time range whether daily, weekly, monthly or yearly.

IJSER

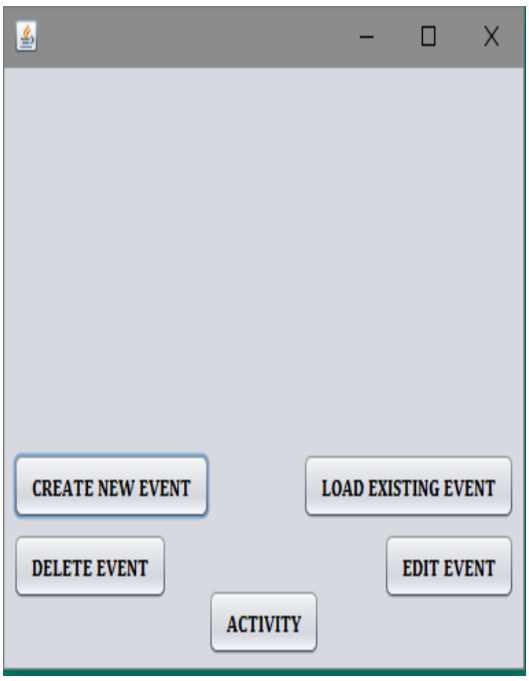

FIG. B3 SPECIFIES THE COLLECTION AND ANALYSIS OF THE TIME DATA.

On the Clicking Time Data Button, after entering the range under analysis. The Survival Package opens the interface to manually enter the desired Time Data by clicking on Create New Event. If the User does not want to proceed, she/he clicks on the Delete Event Button and the entire page opens for the user to start again. However, if the user has text data from any consultancy or research bodies and loads the file by clicking on the Load Existing Event. That will upload the existing files into the windows created by the system for editing of the file. The Activity on the Package page enable the system to analysis the mathematical function of the survival Analysis using Kaplan Meier Estimation approach.

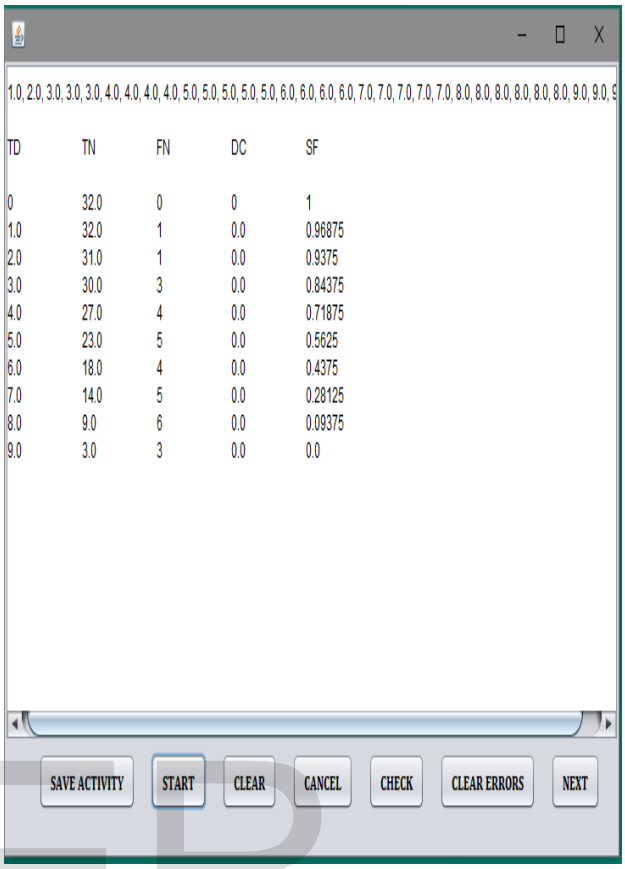

FIG. B4 WHEN THE START BUTTON IS CLICKED.

Upon clicking the start button, the entered data will than be expanded into the table to description the event into the graph.

IJSER

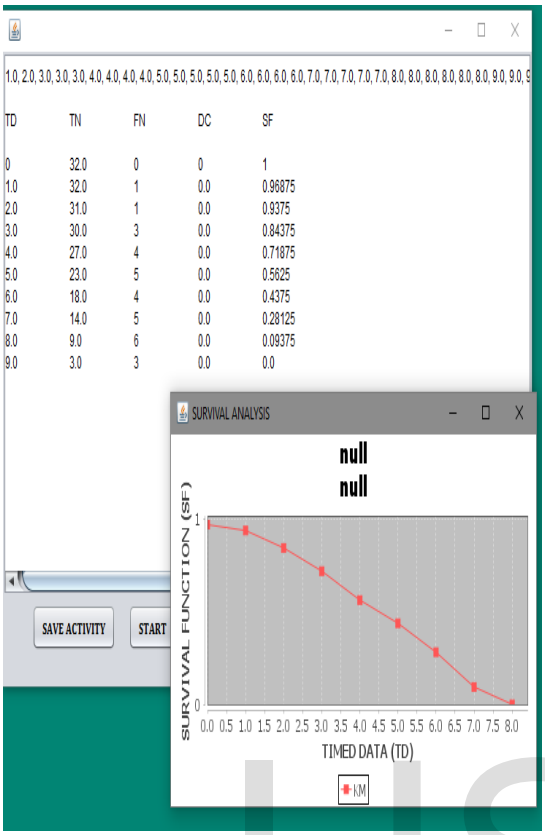

FIG. B5 WHEN THE NEXT BUTTON CLICK IS CLICKED

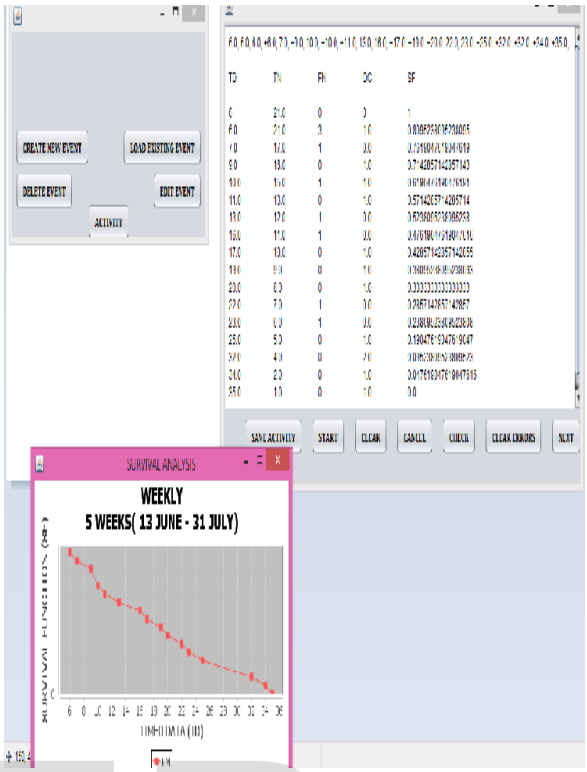

FIG. B6 RESULT

This explain the results obtained from the analysis of the event into diagrams. The data where obtained from censored data to obtain this graph.

CONCLUSION

The study which examined the Time Series classification and survival analysis has revealed that the development cycle of an application which supports time series data into the actualization of the Kaplan Meier Survival curve was successful. The related works gave an insight on the different time series and statistical complexities in time to an event research carried out previously. All the statistical errors and limitations observed were taken into consideration for the achievement of required results.

However, this study approach is new in econometric analysis where synthesized or collated time series data were being analyzed with the use of a Java programming language. The understanding of the analysis was gained through the health statistical data that has continuous regular time study. The decision to use time series data is to ascertain the regular time series which is the major variable to determine a survival rate at any given time. This study when applied with organizations (governmental or non-governmental) can be used to forecasting future event of survival at particular time interval.

IJSER

REFERENCE

[1.] Avent, S 2016, Time series analysis: an introduction, accessed 15 July 2016, http://userwww.sfsu.edu/efc/classes/biol710/timeseries/TimeSeriesAnalysis.html

[2.] Bartlett, P 2010, Introduction to time series analysis, accessed 15 July 2016, http://www.stat.berkeley.edu/~bartlett/courses/153-fall2010/lectures/1.pdf

[3.] BIOST 515, Lecture 15: Introduction to survival analysis, accessed 15 July 2016, http://www.stat.columbia.edu/~madigan/W2025/notes/survival.pdf

[4] Cox, D. R. & D. Oakes. (1984). Analysis of Survival Data, Chapman and Hall, New York.

[5] Chapter 10, n.d, Introduction to time series analysis, accessed 15 July 2016, http://www.biostat.jhsph.edu/~ririzarr/Teaching/754/section-10.pdf

[6] Davis, R n.d, Introduction to statistical analysis of time series, accessed 15 July 2016, http://www.stat.columbia.edu/~rdavis/lectures/Session6.pdf

[7] Deitel . I, Paul J. II. Deitel, Harvey M. Java (Computer program language) Title III, Prentice Hall 2011.

[8] Fox, J 2014, Lecture notes: Introduction to survival analysis, accessed 15 July 2016, http://socserv.mcmaster.ca/jfox/Courses/soc761/survival-analysis.pdf

[9] Guo, Shenyang. (2010) Survival analysis (Pocket guides to social work research methods).Oxford University Press, New York

[10] Hathaway, D. H. 2010, Living Rev. Solar Phys. 7,1

[11] Hosmer, D. W., & Lemeshow, S. (1999). Applied survival analysis: Regression modeling of time to event data. New York: John Wiley.

[12] Hougaard, P. (2000). Analysis of Multivariate Survival Data, Springer-Verlag, New York.

[13] Hyndman, R 2016, Time series analysis, accessed 15 July 2016, https://cran.r-project.org/web/views/TimeSeries.html

[14] Kalbfleish, J. D. & Prentice, R. L. (2002). The Statistical Analysis of Failure Data. Wiley.

[15] Klein, J & Moeschberger, M.L. (2003). Survival Analysis, Technique for Censored nd Truncated Data, 2nd ed., Springer-Verlag, New York

[16] Lee, T. & Wang, J. (2003). Statistical Methods for Survival Data Analysis, 3rd Edition, John Wiley and Son, New York.

[17] Louzada-Neto, F. (1997). Extended hazard regression model for rehability and survival analysis. Lifetime Data Analysis, 3, 367-381.

[18] Pereira, B. & Rao C. R. (2005). Survival Analysis Neural Networks. SBRN, 3(2) pp. 50-60.

[19] Pickup, M 2014, Introduction to time series analysis (Quantitative applications in the social sciences, Sage publication, New York.

[20] Rahbar, MH n.d, Introduction to survival analysis, Michigan State University Press, Michigan.

[21] Rupert G. & Miller, J. (1975). Survival Analysis, John Wiley and Son, New York.

[22] S.S. Alhir, "UML In a Nutshell," United States: O'Reilly , 1998.

[23] SW Chapter 14, n.d, Introduction to Time Series Regression and Forecasting, accessed 15 July 2016, http://www.ssc.upenn.edu/~fdiebold/Teaching104/Ch14_slides.pdf

[24] T. Yuan, "Lecture 3: Requirements Analysis,"presented for the Software Engineering,Unpublished (unpublished manuscript)

[25] T. Yuan, "Lecture 5", Computer Science Department, University of York, Unpublished (unpublished manuscript)

[26] T. Yuan, "Lecture 6: High level Design,"presented for the Software Engineering, module, Computer Science Department, University of York, Unpublished (unpublished manuscript)

[27] T. Yuan, "Lecture 2: Software process,"presented for the Software Engineering module, Computer Science Department, University of York, Unpublished (unpublished manuscript)

[28] Tableman, M. & Kim, J. (2003). Survival Analysis Using S: Analysis of Time-to-Event Data, Chapman and Hall, New York.

[29] Therneau, T. & Grambsch, P. (2000). Modeling Survival Data, Springer-Verlag, New York.

[30] Torben M. (2006). Dynamic Regression Models for Survival Data, Springer-Verlag, NewYork.

IJSER