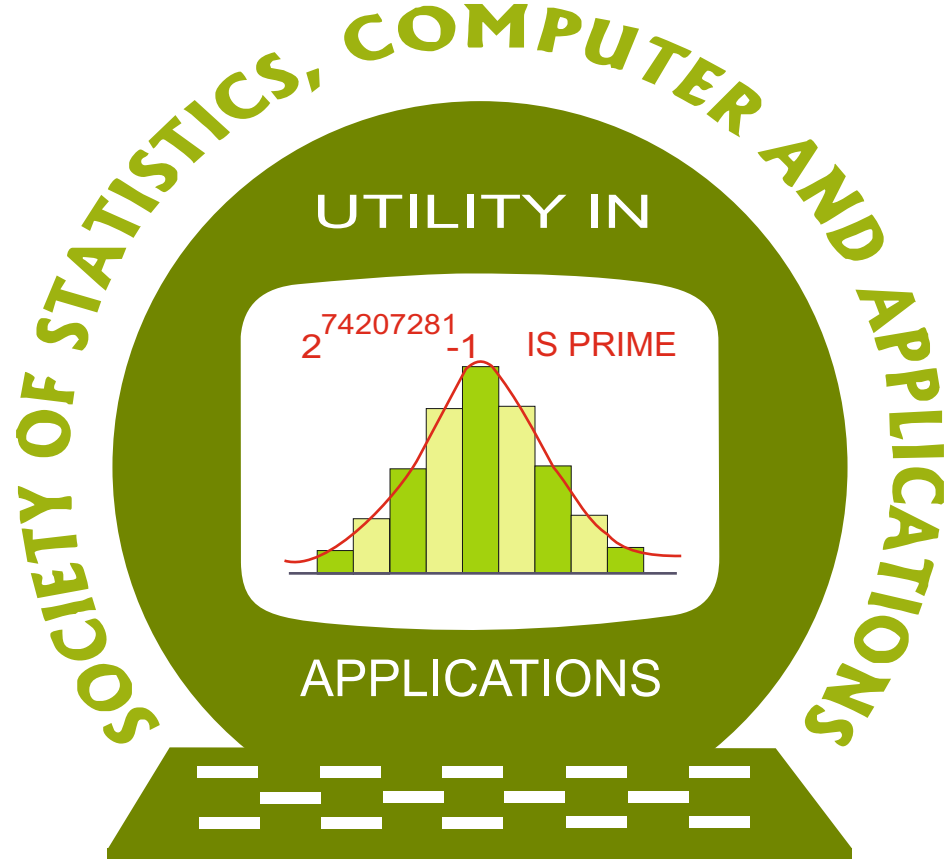
ISSN 2454-7395(online)

STATISTICS AND APPLICATIONS



FOUNDED 1998

Journal of the Society of Statistics, Computer and Applications

https://ssca.org.in/journal.html
Volume 21, N0. 2, 2023 (New Series)

Society of Statistics, Computer and Applications

Council and Office Bearers

Founder President
Late M.N. Das

President
V.K. Gupta

Executive President
Rajender Parsad

Patrons

A.C. Kulshreshtha, A.K. Nigam, Bikas Kumar Sinha, D.K. Ghosh
G.P. Samanta, K.J.S. Satyasai, P.P. Yadav, Pankaj Mittal
R.B. Barman, R.C. Agrawal, Rahul Mukerjee, Rajpal Singh

Vice Presidents

A. Dhandapani, Manish Sharma, P. Venkatesan, Praggya Das
Ramana V. Davuluri, S.D. Sharma, V.K. Bhatia

Secretary
D. Roy Choudhury

Foreign Secretary
Abhyuday Mandal

Treasurer
Ashish Das

Joint Secretaries

Aloke Lahiri, Shibani Roy Choudhury, Vishal Deo

Council Members

B. Re. Victor Babu, Manisha Pal, Mukesh Kumar, Parmil Kumar, Piyush Kant Rai
Rajni Jain, Rakhi Singh, Ranjit Kumar Paul, Raosaheb V. Latpate, Renu Kaul
S.A. Mir, Sapam Sobita Devi, V. Srinivasa Rao, V.M. Chacko, Vishnu Vardhan R.

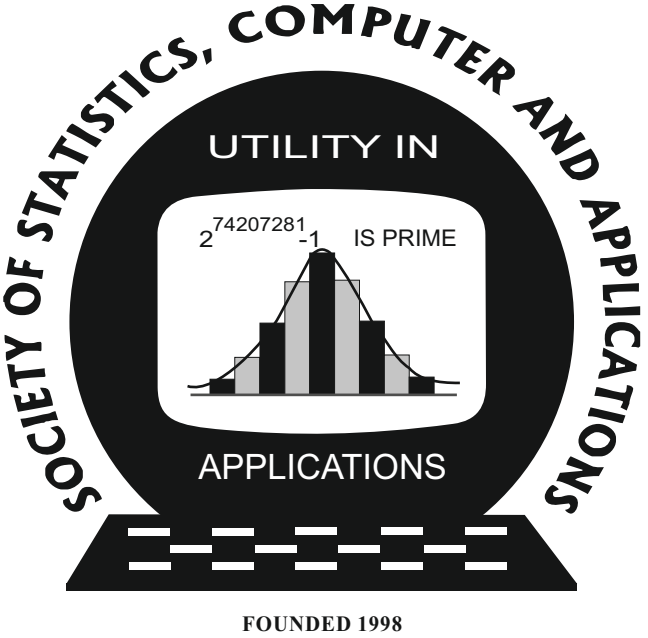
Ex-Officio Members (By Designation)

Director General, Central Statistics Office, Government of India, New Delhi
Director, ICAR-Indian Agricultural Statistics Research Institute, New Delhi
Chair Editor, Statistics and Applications
Executive Editor, Statistics and Applications

Society of Statistics, Computer and Applications
Registered Office: I-1703, Chittaranjan Park, New Delhi- 110019, INDIA
Mailing Address: B-133, Ground Floor, Chittaranjan Park, New Delhi-110019, INDIA

Statistics and Applications

ISSN 2454-7395(online)



Journal of the Society of Statistics, Computer and Applications

https://ssca.org.in/journal.html

Volume 21, No. 2, 2023 (New Series)

Statistics and Applications

Volume 21, No. 2, 2023 (New Series)

Murari Singh, Formerly at International Centre for Agricultural Research in the Dry Areas, Amman, Jordan; mandrsingh2010@gmail.com
Nripes Kumar Mandal, Flat No. 5, 141/2B, South Sinthee Road, Kolkata-700050; mandalnk2001@yahoo.co.in
P. Venkatesan, Professor Computational Biology SRIHER, Chennai, Adviser, CMRF, Chennai; venkaticmr@gmail.com
Pritam Ranjan, Indian Institute of Management, Indore - 453556; MP, India; pritam.ranjan@gmail.com
Ramana V. Davuluri, Department of Biomedical Informatics, Stony Brook University School of Medicine, Health Science Center Level 3, Room 043 Stony Brook, NY 11794-8322, USA; ramana.davuluri@stonybrookmedicine.edu; ramana.davuluri@gmail.com
S. Ejaz Ahmed, Faculty of Mathematics and Science, Mathematics and Statistics, Brock University, ON L2S 3A1, Canada; sahmed5@brocku.ca
Sanjay Chaudhuri, Department of Statistics and Applied Probability, National University of Singapore, Singapore -117546; stasc@nus.edu.sg
Sat N. Gupta, Department of Mathematics and Statistics, 126 Petty Building, The University of North Carolina at Greensboro, Greensboro, NC -27412, USA; sngupta@uncg.edu
Saumyadipta Pyne, Health Analytics Network, and Department of Statistics and Applied Probability, University of California Santa Barbara, USA; spyne@ucsb.edu, SPYNE@pitt.edu
Snigdhansu Chatterjee, School of Statistics, University of Minnesota, Minneapolis, MN -55455, USA; chatt019@umn.edu
T.V. Ramanathan; Department of Statistics; Savitribai Phule Pune University, Pune; madhavramanathan@gmail.com
Tapio Nummi, Faculty of Natural Sciences, Tampere University, Tampere Area, Finland; tapio.nummi@tuni.fi
Tathagata Bandyopadhyay, Indian Institute of Management Ahmedabad, Gujarat; tathagata.bandyopadhyay@gmail.com, tathagata@iima.ac.in
Tirupati Rao Padi, Department of Statistics, Ramanujan School of Mathematical Sciences, Pondicherry University, Puducherry; drtrpadi@gmail.com
V. Ramasubramanian, Division of Research Systems Management, ICAR - National Academy of Agricultural Research Management (NAARM), Rajendranagar, Hyderabad - 500 030, Telangana; ram.vaidhyanathan@gmail.com

CONTENTS

Statistics and Applications ISSN 2454-7395 (online)
Volume 21, No. 2, 2023 (New Series)

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 21, No. 2, 2023 (New Series), pp 1–15
http://www.ssca.org.in/journal

Analysis of River Water Quality Using Geo-Spatial and Temporal Data: A Case study

K. Muralidharan, Shrey Pandya, Aiman Shaikh, Parth Patel and Jayshree Vanzara
Department of Statistics
The Maharaja Sayajirao University of Baroda, Vadodara, 390002, India

Received: 10 April 2022; Revised: 27 June 2022; Accepted: 28 July 2022

Abstract

The conventional approach of water quality assessment via sampling followed by laboratory measurement methods comprises the analysis of different properties such as chemical, physical. The main idea behind detection of water quality parameters using imaging is based on the presence of pollutants in water and absorption of the incoming solar radiation. In this study, by considering data of the conventional water quality testing,an attempted is made to identify the association between the laboratory results and the indices and bands values obtained from spatial data, to determine their applicability in water quality estimation and prediction. The study makes use of two types of data, visually, spatial and non-spatial data. The spatial data used was Landsat-8 OLI from which the water index was calculated. While under non-spatial data ancillary information and water parameters were considered. Based on the analysis an approach was made to find the relation between Water Quality Index and spatial parameters. Further, a model was established to estimate WQI from spatial data.

Key words: Water pollutants; Spatial estimation; Regression analysis; Water quality index; Anthropogenic waste.

AMS Subject Classifications: 62K05, 05B05

1. Introduction

Water is an elixir of life. It is a precious natural resource and an important component of human survival and to maintain life cycle on our blue plant. Out of the total water reserves of the world, about 97% is salty water (marine) and only 3% is freshwater. Even this small fraction of freshwater is not available to us as most of it is locked up in polar ice caps and just 0.003% is readily available to us in the form of groundwater and surface water Pawan and Pradeep (2015). Due to its unique properties it is an essential part of all living organisms on the planet. Human beings depend on water for almost every development activities like drinking, irrigation and transportation, washing and waste disposal for industries and used as a coolant for thermal power plants. Water shapes the earth's surface

Corresponding Author: K. Muralidharan
Email: muralikustat@gmail.com

and regulates our climate. With increasing human population and rapid development, the world water withdrawal demands have increased many folds and a large proportion of the water withdrawal is polluted due to atmospheric activities. Rivers are the most important water resources. It has long been used for discharging wastes. Unfortunately, the rivers are being polluted by indiscriminate disposal of sewage, industrial wastes and by human activities Pawan and Pradeep (2015).

The conventional approach of water quality assessment *via* sampling followed by laboratory measurement methods comprises the analysis of different properties such as chemical, physical, biological and other indicators Ouma *et al.* (2018). However, water sampling and the subsequent measurements of water quality parameters (WQP) are helpful in representing point-based estimates of the quality of water conditions in terms of time and space both, while, obtaining spatial-temporal variations of water quality indices for large water bodies is very challenging Ritchie *et al.* (2015), ouma *et al.* (2018). Apart from the factors like tedious, work serious and exorbitant, some of the other significant limitations associated with the conventional method for water quality assessment are inability to monitor, forecast and manage the entire water body due to the water surface extent and its topographic characteristics and the lack of spatial-temporal data. To overcome these limitations, there is a need for technology which is fast, inexpensive, simple, automated and non-invasive in operational and productive aquatic environmental monitoring. Measurements and observations taken with such tools should provide essential information with respect to bio-geophysical water quality aspects Garaba *et al.* (2015), which is economically efficient, along with adequate spatial coverage, resolution and most important available on regular time intervals as well.

By utilizing remote detecting, the optically dynamic water constituents can be identified depending on their cooperation with light and the resulting change in the energy of the occurrence radiation as reflected from the water body Ritchie *et al.* (2015). The main idea behind detection of water quality parameters using imaginary data is based on pollutants present in water and absorption of the incoming solar radiation and the water quality can be correlated with the characteristics of the water segments, such as colour and transparency Dor and Ben-Yosef (1996). This implies that optical information can give an elective means to getting generally minimal expense and synchronous data on surface water quality conditions Dor and Ben-Yosef (1996), Dekker *et al.* (1993). Regardless of the capacity of remote detecting to be utilized for the appraisal of water quality with the ideal benefits of being convenient and practical, the procedure may not be adequately exact and should be benchmarked with the conventional testing techniques and field studies. That is, for better understanding, incorporated utilization of remote detecting, in-situ estimations and PC water quality displaying is probably going to bring about a more fiery information on the water quality in each surface water framework Gholizadeh *et al.* (2016).

Sampling and field measurements, have been the standard techniques that are been practiced since long in the determination of water quality with help of certain variables, at the same time various tests and approaches have been carried out for the estimation of different water parameters, in different case studies using novel methods and procedures. Despite being the traditional approach for water quality testing, the laboratory methods are unable to present the real-time spatial overview which is necessary for the monitoring of water quality at certain regular interval and make decisions Brivio *et al.* (2001). In this study, by considering data of the conventional water quality testing, an attempted is made

to identify the association between the laboratory results and the indices and bands values obtained from spatial data, to determine their applicability in water quality estimation and prediction. For further analysis, a correlation of the distribution of the measured WQP using laboratory measurements and the remote sensing models are spatially analysed. Retrieval of water quality characteristics from remote sensing was made possible using Landsat sensors, namely, operational land imager (OLI). This was used to establish relationship between water quality parameters, such as power of hydrogen (pH), temperature, dissolved oxygen (DO), biochemical oxygen demand (BOD) and chemical oxygen demand (COD). Nonetheless, dispect of many advantage that is possesd by Landsat the use of Landsat data has the few limitations: (1) the repeat cycle of sensor is of 16 days and that imposes major limitations on intra-seasonal monitoring more accurately, especially in areas characterized by frequent cloud cover and (2) the water quality parameter characteristics must be related to the inherent optical property (IOP) that can be measured by the satellite sensor Brezonik *et al.* (2005).

Remote detecting based models have broader uses in vast sea waters. While, investigations on inland freshwater bodies by that of remote detecting estimations are bit complex. Making it hard to foster functional freshwater remote detecting calculations. Besides, it is unimaginable to expect to utilize existing algorithmic models for exact water quality assessment. Not withstanding the calculations having been approved in explicit contextual investigations, the confined attribute of every space makes it important to rethink and revalidate the current calculations for their potential applications in other WQP forecast contextual analyses Ouma *et al.* (2018).

Remote sensing estimation of surface water quality is based on mapping the relationship between remote sensing multispectral signatures and measurements of ground truth data (*i.e.*, concentrations of SWQPs). Additionally, a remote sensing study of surface water quality requires multispectral data for the surface features, as they would be measured at ground level. Surface Water Quality Parameters (SWQPs) can be broadly classified into two main classes: optical and non-optical SWQPs. Optical parameters are optically sensitive parameters that can be sensed by remote sensing and hence can be approximated. A significant number of studies have been conducted for assessing optical parameters KC *et al.* (2019). A challenge is to approximate underlying relationship between both optical and non-optical parameters. Optical SWQPs, such as turbidity and total suspended solid (TSS) are most likely to affect the watercolour, the reflected signals and consequently can be detected by satellite sensors. On the other hand, non-optical SWQPs, such as COD, BOD, DO, total dissolved solid (TDS), pH and surface water temperature are less likely to affect the reflected radiation (Din). Mapping the relationship between Satellite Data and the Concentrations of SWQPs is achievable via regression techniques. Theoretically, the relationship between satellite multi-spectral signatures and the concentrations of SWQPs is too complex, especially in the presence of various pollutants at the same time. Moreover, it is very challenging for regression techniques to model such a complex relationship. The proposed solution aims at developing a novel artificial intelligence *i.e.*, learning-based) modelling method for mapping concentrations of both optical and non-optical SWQPs by using remotely sensed multispectral data Ouma *et al.* (2018).

The organization of the paper is as follows: The details about the methodology and study area are presented in Section 2. The data capture methods along with their specifica-

tions are discussed in Section 3. Section 4 presents the detailed analysis and interpretations. Some recommendations on the quality index are also studied in this section. The paper ends with a conclusion in the last section.

2. Materials and methods

2.1. Methodology

In this study two different types of data, visually, spatial and non-spatial data were used. The spatial data used was Landsat-8 OLI from which water index was calculated and water body area was extracted from the raw data using shapefile. While under non-spatial data ancillary information and water parameters were considered. Based on the analysis an approach was made to find the relation between Water Quality Index and spatial parameters. Further, a model was established to estimate WQI from spatial data.

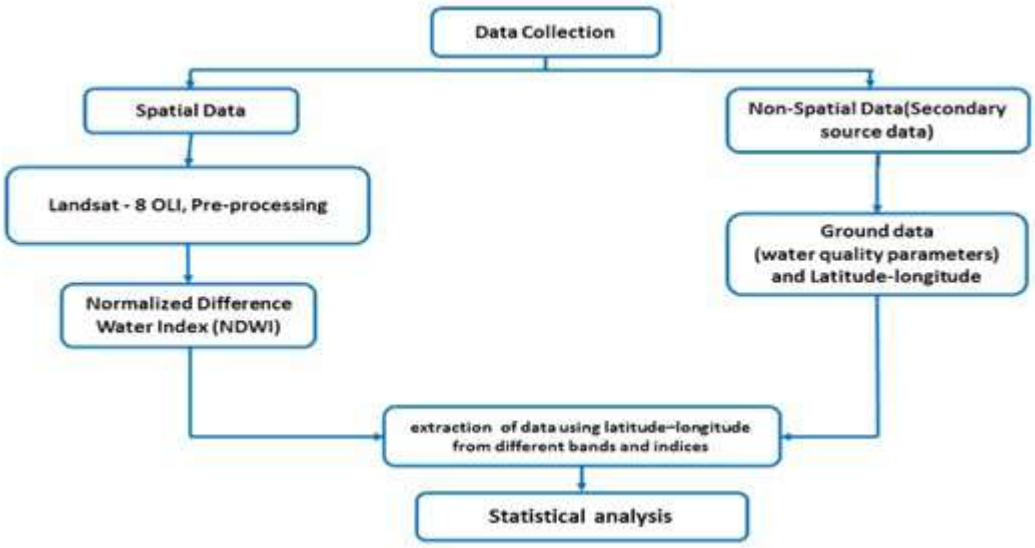


Figure 1: Methodology of the study

2.2. Study area

This study was carried out with respect to the location Dhuvaran situated in the Mahi basin. Dhuvaran (22.539188° N latitude and 72.412128° E longitude) is a remote village that comes under Khambhat taluka of Anand district, located at the point where Mahi river ends and the gulf of Khambhat starts. This village has a population size of 8043 of which 4168 are male and 3875 are female as per the population census 2011. The climate is semiarid with a temperature range of 15°C in winter and 34°C in summer. Significant rainfall occurs during the Southwest monsoon winds, from June to September and receive annual rainfall ranging from 20 inches to 30 inches. The location is very close to a nuclear power plant which is infamous for industrial pollution and anthropogenic waste accumulation. This effect the quality and consumption of water for everyday usages. The site is also famous for unusual climatic and temporal variations towards water scarcity and quality problems, which attracts environmentalists and statisticians to carry out studies to help policymakers to have interventions and strategies. The study area is shown in Figure 2.

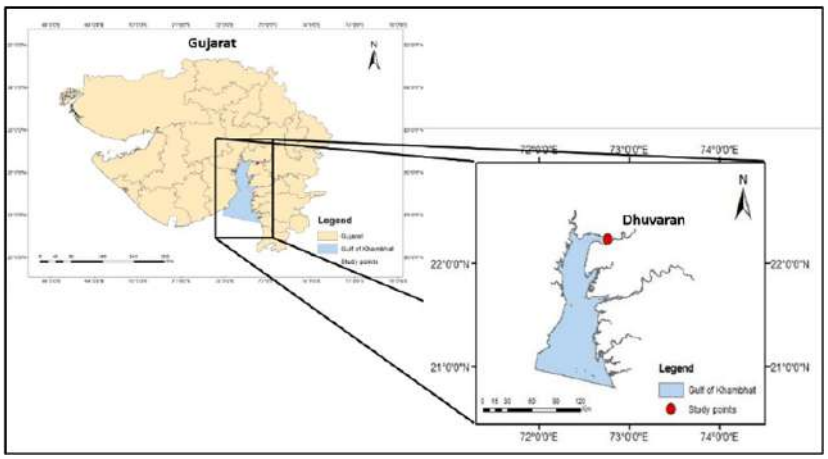


Figure 2: Study area map

3. Data presentation

3.1. Remote sensing data acquisition

Remote sensing data used for this context was based on the Landsat 8 optical land imager (OLI) level-2 imaginary (path:148 and row:45), acquired for free through united states geological survey (USGS) earth resources observation and science center (EROS) from January 2015 to March 2021. For this study bi-monthly data was considered with no or less cloud coverage. Overall, we had 51 spatiotemporal data scenes. The acquired data of the study area was already geometrically corrected and further, radiometric correction of multispectral imagery was done of acquired data by converting digital numbers (DNs) to the spectral radiance. Landsat 8 OLI level-2 processed data consists of 10 bands ranging from 435 nm-11190nm, which comprises of visible bands, NIR, SWIR and TRIS bands with a resolution of 30 meters for visible, NIR and SWIR; and 100 meters for TRIS. Data for the same can be acquired after every 15 days. Table 1 presents the specifications of these bands.

Table 1: L-8 OLI level-2 band description

Band Number	Band Description	Band Range (nm)
B1	Costal Aerosol	435 - 451
B2	Blue	452 -512
B3	Green	533 -590
B4	Red	636 -673
B5	Near Infrared (NIR)	851 -879
B6	Short Wave Infrared (SWIR-1)	1566 -1651
B7	SWIR-2	2107 -2294
B10	Thermal Infrared Sensor	10600 - 11190

3.2. Ground data

A sampling of surface water was collected from January 2010 to December 2020 from a predefined site. Parameters like pH and temperature were taken on the ground while, parameters like TDS, BOD, DO samples were brought to the lab for physicochemical experiments. Standard methods were carried out for capturing data related to these parameters Singh and Jayakumar (2016), APHA (2005). They are further considered for calculating the water quality index and were compared with the standards of WQI, as shown in Table 2.

Table 2: Water quality index scale

WQI	Rating
0-25	Excellent
26-50	Good
51-75	Poor
76-100	Very poor
Above 100	Unsuitable

4. Analysis and findings

4.1. Modified normalized difference water index (MNDWI)

MNDWI was proposed by Xu *et al.* (2006) who noticed a limitation about NDWI of not being able to suppress the signal reflected from the land and the build-up efficiently Yun Du *et al.* (2016), Xu. *et al.* (2006). Based on the finding, the proposed formula of MNDWI is shown as:

$$MNDWI = \frac{\rho_{\text{Green}} - \rho_{\text{SWIR}}}{\rho_{\text{Green}} + \rho_{\text{SWIR}}} \tag{1}$$

where ρ_{Green} is the top of atmosphere (TOA) reflectance of the green band and ρ_{SWIR} is the TOA reflectance of the SWIR band. In Landsat-8 OLI band 3 is mapped as a green band that has a spatial resolution of 30 m while band 6 is SWIR-1 and has the same spatial resolution of 30 m. So, with respect to resolution for Landsat 8 OLI formula can be rewritten as:

$$MNDWI_{30\text{ m}} = \frac{\rho_3 - \rho_6}{\rho_3 + \rho_6} \tag{2}$$

The normalized values were obtained by subtracting and adding the same bands in numerator and denominator and the values will range between -1 to +1.

4.2. Water parameters

The visualisation of ground and spatial data was done in the exploratory data analysis to understand the trend and distribution of the data. As illustrated in Figure 3, a line chart was created using the lower and higher limits set by the water pollution control board (WPCB/PCB) for each parameter. From 2013 to 2017, there was a notable shift in all Water parameters for the area, as shown in Figure 3. To capture the information's regarding the water quality, it was decided to have continuous monitoring and assessment of all variables as described below.

Dissolve oxygen (DO)

This parameter measures the amount of oxygen present in the water in dissolved formed. which is an important factor for the survival of the biotic components present under the water bodies. It depends on several factors like temperature, water agitation, type and number of aquatic plants and light penetration amount of dissolved suspended solids Sudarshan *et al.* (2018). The optimum range for good water quality ranges from 4-6 mg/l, which ensures healthy aquatic life in a water body Sawyer *et al.* (1994), Leo and Dekkar (2000), Burden *et al.* (2002), De (2003). Figure 3(a) indicates that the DO level dropped dramatically between 2013 and 2014, eventually reaching its lowest point in 2015. As a result, the survival rate of all biotic components presents in the waterbody in that area may have reduced leading to unsuitable for everyone's survival. The overall average of the DO data is 6.5 which falls under the range of good water quality which reflects a good aquatic life.

Biological oxygen demand (BOD)

BOD determines the strength in terms of oxygen required to stabilise the domestic and industrial wastes Shah *et al.* (2016). From 2012 to 2016 the demand for oxygen was high in the study area which resulted in a fall in the level of DO. On basis of which our assumption is, wastes released in the water body was not treated as per the standards and hence to stabilise more oxygen was required. The data on BOD showed a declining trend after 2016 with a severe fall till 2018 and after that, the value remained below constantly as observed in Figure 3(b).

pH

pH is one of the most common parameters affecting quality and hence is given due consideration in this study. Data related to this variable is directly captured from the field and no laboratory testing is performed on this. This parameter reflects the acidic or basic property of water (Figure 3(c)). The value of pH below 6.5 causes discontinuation in the making of vitamins in the human body. When pH becomes more than 8.5, the taste becomes saltier and causes eye irritation and skin disorders Gupta *et al.* (2017). The average pH value in our study is for the defined time frame was 7.9 which is close to 8 means salt contamination is more.

Temperature

Temperature is the easiest and common parameter but has a significant role to play for other parameters. Many parameters have a direct relationship to temperature. the temperature data chart showed an increasing trend, crossing the upper specification may limit during the 2017-2018 period (Figure 3(d)).

4.3. Measuring unit of variables

The source and nature of data considered in this study are different as a result the measuring units are also varying. Water parameters like DO, BOD, TDS have measuring unit milligram per liter, pH has an ordinal scale ranging from 0 to 14, temp. is measured

in degree celsius and bands of geospatial data has unit nanometer (nm) while WQI and MNDWI are unit free.

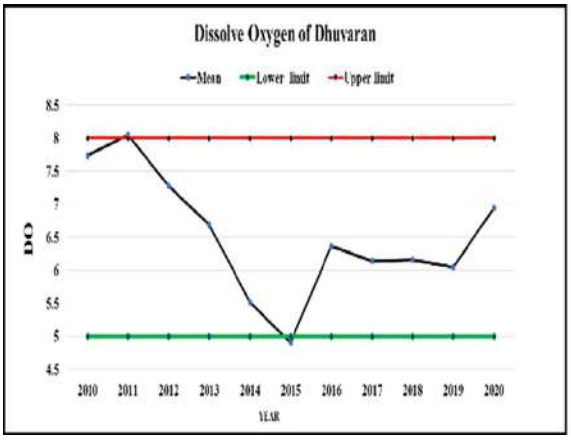


Figure 3(a): Control chart of DO

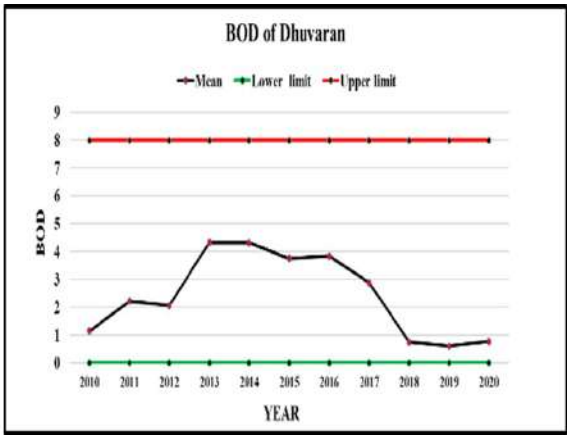


Figure 3(b): Control chart of BOD

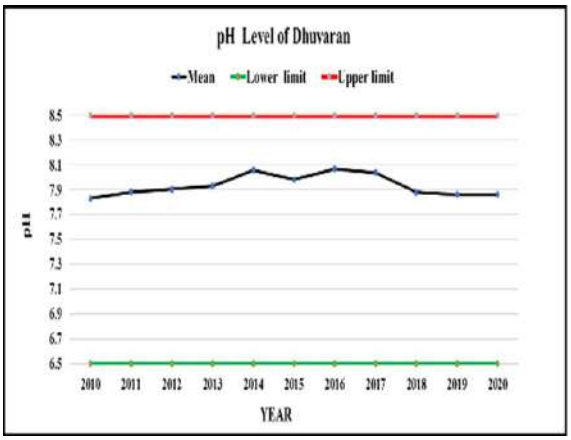


Figure 3(c): Control chart of pH

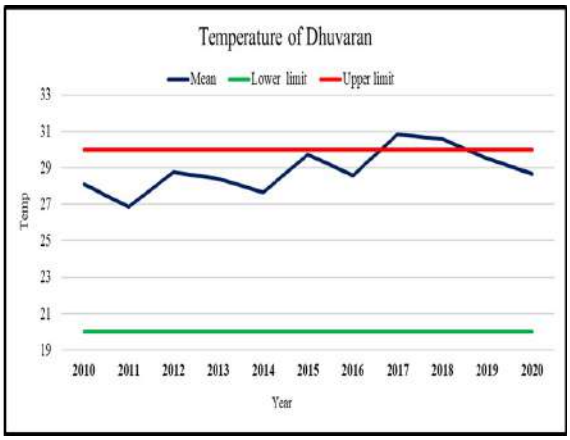


Figure 3(d): Control chart of temp.

4.4. Water quality index (WQI)

Despite monitoring individual water parameters, it is a bit difficult to assure the quality of water at a given point of time and location by looking at these parameters. Water quality plays an important role in such cases while making decisions with respect to water management and interventions for improving the quality. It defines the whole status of the water body by a single number and informs the public about its state APHA (2005), Sudarshan *et al.* (2018), Ashok *et al.* (2011). This single value gives information about the quality state of water at a given point of time for any space Alobaidy *et al.* (2010). In this study, weighted arithmetic mean WQI (WAWQI) is used to calculate the quality index Horton (1965), Sudarshan *et al.* (2018). Four parameters namely pH, DO, BOD, temperature were considered for calculating the WQI. Standards for drinking water was recommended by BSI (Indian standard specification for drinking water, 2012). The WAWQI is calculated as:

$$WAWQI = \frac{\sum_{i=0}^{i=n} W_i Q_i}{\sum_{i=0}^{i=n} W_i} \tag{3}$$

where unit weights (W_{i}) were calculated for each parameter by following the formula as given in Tiwari and Mishra (1985)

$$W_i = K_j \in \left(\frac{1}{s_n}\right), \tag{4}$$

where $K_j = \frac{1}{\frac{1}{S_1}+\frac{1}{S_2}+\ldots+\frac{1}{S_n}}$.

Now Quality rating scale ($\mathrm{Q_i}$) was calculated using the formula for all parameters except two variables DO and pH of pure water.

$$Q_i = \frac{Q_{(act)} - Q_{(ideal)}}{S_{(std)} - Q_{(ideal)}} \times 100 \tag{5}$$

where, W_i = unit weight of each water quality parameter, K = Proportionality constant, Q_{i} = Quality rating scale for each parameter, Q_{act}= Estimated concentration of $\mathrm{i^{th}}$ parameter in the analyzed water, Q_{ideal} = Value of the parameter in pure water, S_{std} = Standard value of ith parameter and n = No of water quality parameters.

For pH, value of $\mathrm{Q_i}$ is 7.0 and for DO, it is 14.0. The water quality index finally obtained is visualized in Figure 4. After calculating the weight of the parameter and quality rating scale, values were substituted in the final formula of WAWQI and then the index value were compared.

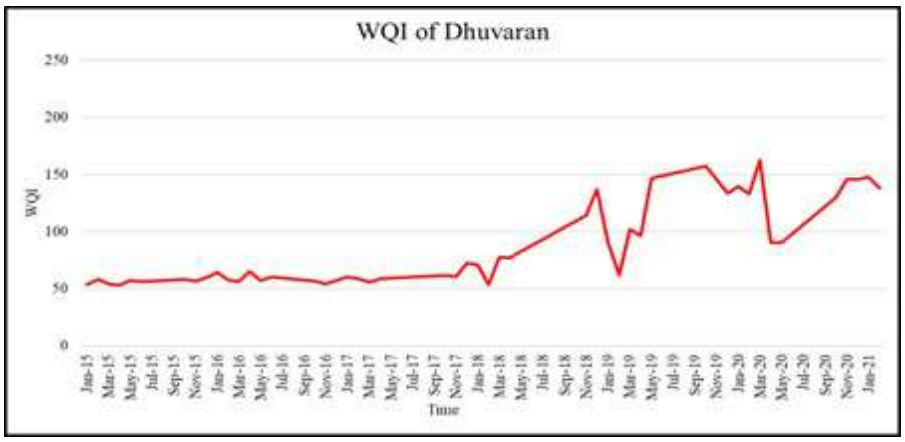


Figure 4: Control chart of Water Quality Index calculated using defined formula

However, looking at the WQI chart (Figure 4), it is seen that the variation was not suitable for an initial period, but with the passage of time, further, fluctuated significantly. This is attributed to an unknown lag effect and hence, the quality of water may not be suitable for consumption and domestic usages. As per our understating some of the factors that affect the quality can either be due to industrial disposal, human activities in the nearby areas, or can be underwater disturbances in aquatic life. To predict WQI and water quality parameters from non-conventional data, it's crucial to see if there's a relationship between the two types of data. To do so, a correlation matrix and a heatmap were created, as shown in Figure 5. It was discovered that DO and WQI had a very high correlation, whereas there was a moderate correlation between spatial and non-spatial data source variables.

Figure 5: Correlation matrix and heatmap chart

4.5. Multiple linear regression analysis

To understand the influence of reflectance data on the WQI and its parameters, a multiple linear regression was carried out to understand the significance of each variable. Regression analysis is the Statistical technique that is used for predicting dependent variables with the help of a single exploratory variable or multiple exploratory variables after expressing the linear relationships between them. This relationship can be written as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p + \epsilon \tag{6}$$

where,

Y: Dependent variable,

β_0: Intercept,

β_p: Slope,

X_p: Exploratory variables; p:*1,2,3,...,8*,

ϵ: Residual term.

Table 3: Regression models and R^2 values of respective models

Dependent variable	**Equation**	$\mathbf{R^2}$
DO	13.504 + 0.0032*$\mathbf{B1}$ - 0.0043*$\mathbf{B2}$ + 0.0005*$\mathbf{B3}$ + 0.0003*$\mathbf{B4}$ -0.0001*$\mathbf{B5}$ - 0.0005*$\mathbf{B6}$ + 0.0012* $\mathbf{B7}$ - 0.0002* $\mathbf{B10}$	0.12
BOD	37.4192 + 0.0094*$\mathbf{B1}$ - 0.0145*$\mathbf{B2}$ + 0.0032*$\mathbf{B3}$ + 0.0006*$\mathbf{B4}$ + 0.0001*$\mathbf{B5}$ + 0.0061*$\mathbf{B6}$ - 0.0076*$\mathbf{B7}$ - 0.0003*$\mathbf{B10}$	0.18
pH	8.0053 - 0.0017*$\mathbf{B1}$ + 0.0025*$\mathbf{B2}$ - 0.001*$\mathbf{B3}$ + 0.0002*$\mathbf{B4}$ - 0.0001*$\mathbf{B5}$ + 0.0003*$\mathbf{B6}$ - 0.0003*$\mathbf{B7}$	0.11
Temp.	- 49.7692 - 0.0072*$\mathbf{B1}$ + 0.0127*$\mathbf{B2}$ + 0.0031*$\mathbf{B3}$ - 0.0072*$\mathbf{B4}$ + 0.0008*$\mathbf{B5}$ - 0.0125*$\mathbf{B6}$ + 0.0138*$\mathbf{B7}$ + 0.0013*$\mathbf{B10}$	0.35
WQI	1033.5988 + 0.1408*$\mathbf{B1}$ - 0.2361*$\mathbf{B2}$ + 0.213*$\mathbf{B3}$ + 0.0105*$\mathbf{B4}$ + 0.0014*$\mathbf{B5}$ - 0.1124*$\mathbf{B6}$ - 0.0985*$\mathbf{B7}$ - 0.0054*$\mathbf{B10}$ - 3318.431*$\mathbf{MNDWI}$	0.21

***Note: The location considered for the study is geographically located at a point where the dispersion of soil in water is observed more often due to low and high tides. In geospatial data, the same is reflected and this makes it a bit difficult to establish the accurate relationship between DN values and laboratory data.**

Many studies are conducted where linear models Frenanda *et al.* (2020) and nonlinear KC *et al.* (2019) models are developed to predict WQI. In this study, we constructed various linear models for WQI and its associated parameters. See Table 3 for details along with the values of R^2. It is evident from the Table, that the regression coefficients are positively and inversely related in many cases. These coefficients helps us to understand the influence of the unit percentage change of independent variables on the dependent variable. Since all the abiotic factors are not included in the model, it is difficult to explain the amount of variation present in each model. However, the value of R^2 can explain the amount of variation to a useful extent. As per the analysis, the temperature model yields a high R^2 value as compared to any other model. This can be due to less time difference between spatio-temporal and ground data. The WQI model is also capable to explain around 21% of the total variation present in the data.

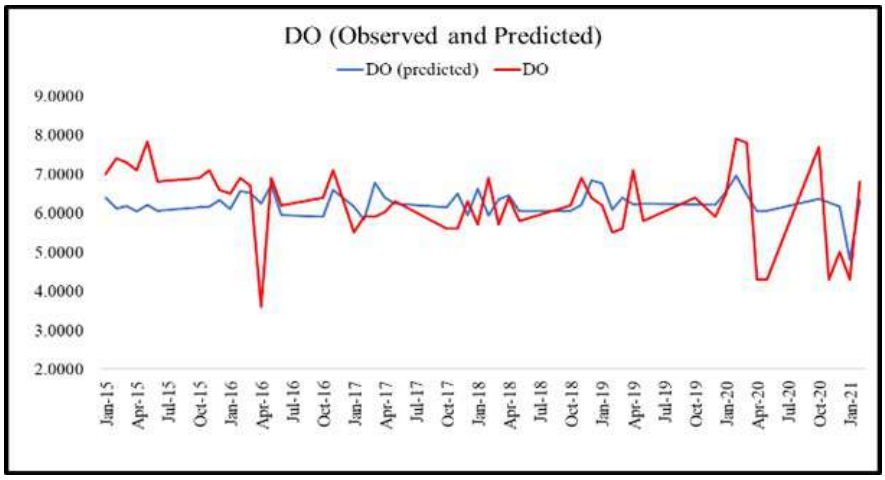


Figure 6(a): Chart of actual values and predicted values of DO

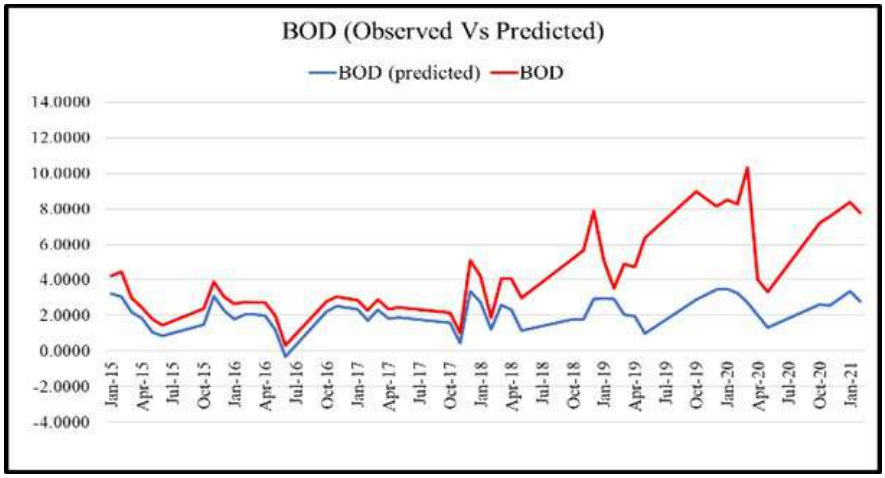


Figure 6(b): Chart of actual values and predicted values of BOD

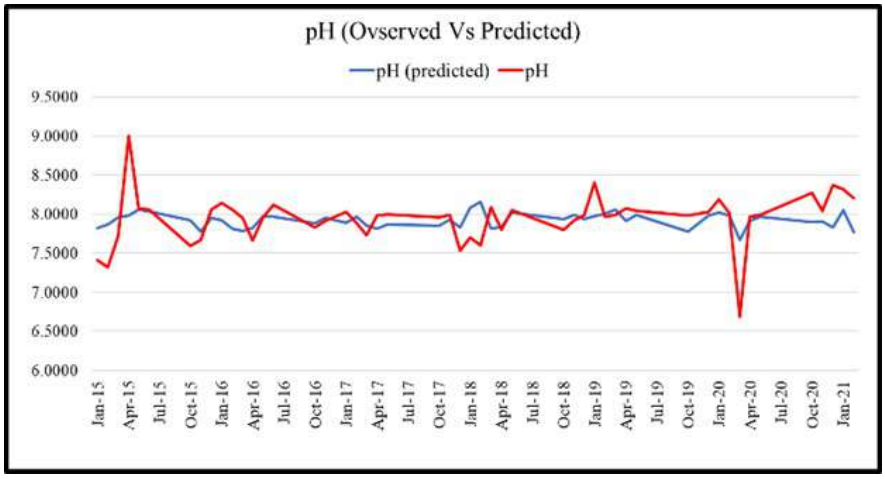


Figure 6(c): Chart of actual values and predicted values of pH

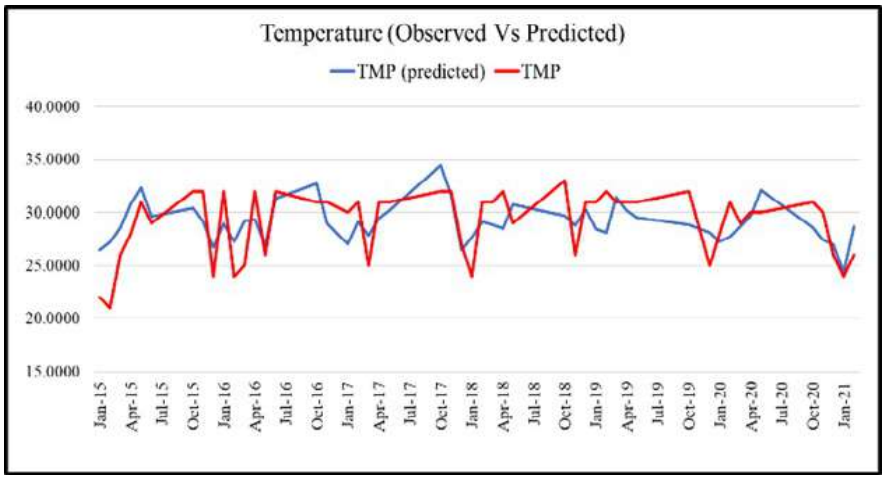


Figure 6(d): Chart of actual values and predicted values of temperature

To understand the prediction capability of each model, we present in Figures 6(a-d), the comparison of actual and predicted values of each variable. It appears that the line charts of projected values appeared to be in the same range as those of real values, however, the model was unable to suit the actual line chart partially or completely due to some elements not included in the model. Figure 6(a-d) shows how a model failed to anticipate a significant spike or dip in the data at some time. It's possible that this phenomenon is related to some climatic changes and fluctuations. The model for BOD was able to follow the trend of real values, but other climatic, physical and supporting variables were not included as factors, thus the projected values did not go together with the actual values. On other hand, the model was not able to predict random spike, which is evident from the decreasing trend as seen in Figure 6(c) of pH. The same thing happened with the temperature chart as well.

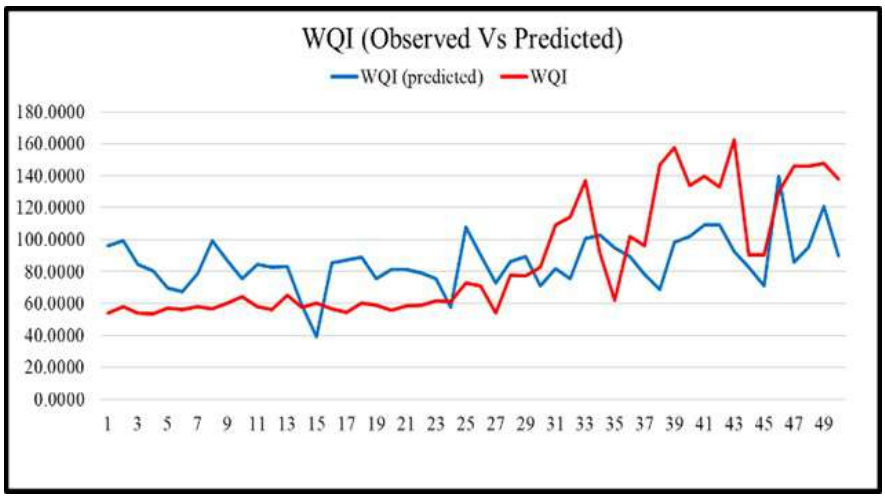


Figure 7: Chart of actual values and predicted values of WQI

It is observed that about 21% of the variation was explained by the WIQ model as per Table 3. However, on plotting the actual and predicted values it can be observed that the model was able to capture the trend in a satisfactory way as shown in Figure 7.

5. Discussion and limitations

The goal of this research is to model parameters using spatial data. In which it was discovered that if we have a region where two separate water bodies are linked, such as our research area, spatial data alone is insufficient to predict water quality properties. Other environmental factors, as well as the lunar cycle, will play a key influence here. We believe that more abiotic elements can be added to every model to make it more useful and trustworthy.

Acknowledgements

The authors thank Dr. Kauresh Vachhrajani, of department of Environmental Science, The Maharaja Sayajirao University of Baroda for his valuable suggestions and providing some assistance in acquiring the ground data. Also, the authors are grateful to the editor of the journal and the anonymous reviewers for their valuable suggestions comments and suggestions of generously listing many useful references.

References

Alobaidy, A. H. A. J., Abid, H. S., and Maulood, B. K. (2010). Application of water quality index for assessment of Dokan lake ecosystem, Kurdistan region. *Journal of Water Resource and Protection*, **2**, 792-798.

Ashok, L., Sharma, T. C., and Jean-Francois, B. (2011). A review of genesis and evolution of water quality index (WQI) and some future directions. *Water Quality Exposure and Health*, **3**, 11-24.

APHA (American Public Health Association) (2005). *Standard Methods for the Examination of Water and Wastewate.* Washington D. C., USA.

Brezonik, P., Menken, K. D., and Bauer, M. (2005). Landsat-based remote sensing of lake water quality characteristics, including chlorophyll and colored dissolved organic matter (CDOM). *Lake and Reservoir Management*, **21**, 373–382.

Brivio, P. A., Giardino, C., and Zilioli, E. (2001). Validation of satellite data for quality assurance in lake monitoring applications. *Science of the Total Environment*, **268**, 3–18.

Burden, F. R., Mc. Kelvin, I., Forstuner, U., and Guenther, A. (2002). *Environmental Monitoring Handbook.* Mc Graw-Hill handbooks, New York, 3.1-3.21.

De, A. K. (2003). *Environmental Chemistry*, 5th Edition. New Age International Publisher, New- Delhi, 190,215,242-244.

Dekker, A. G. and Peters, S. W. M. (1993). The use of the thematic mapper for the analysis of eutrophic lakes: a case study in the Netherlands. *International Journal of Remote Sensing*, **14**, 799–821.

Dor, I. and Ben-Yosef, N. (1996). Monitoring effluent quality in hypertrophic wastewater reservoirs using remote sensing. *Water Science and Technology*, **33**, 23–29.

Garaba, S. P., Friedrichs, A., Voß, D., and Zielinski, O. (2015). Classifying natural waters with the forel-ule colour index system: results, applications, correlations and crowdsourcing. *International Journal of Environmental Research and Public Health*, **12**, 16096–16109.

Gholizadeh, M., Melesse, A., and Reddi, L. (2016). A comprehensive review on water quality parameters estimation using remote sensing techniques. *Sensors*, **16**, 1298–1306

Gupta, N., Pande, P., and Hussain, J. (2017). Effect of physciochemical and biological parameters on the quality of river water of Narmada, Madhya Pradesh, India. *Water Science*, **31**, 11-33

Horton, R. K. (1965). An index number system for rating quality. *Journal of Water Pollution Control Federation*, **37**, 300-305.

KC, A., Chalise, A., Parajuli, D., Dhital, N., Shrestha, S., and Kandel, T. (2019). Surface water quality assessment using remote sensing, gis and artificial intelligence. *Nepal Engineers' Association*, Gandaki Province ISSN: 2676-1416, **1**, 113-122.

LANDSAT 8 (L8) data users handbook (2015), Department of the Interior U.S. Geological Survey. Version 1. Available from: http://landsat.usgs.gov/documents/Landsat8DataUsersHandbook.pdf

Leo, M. L. and Dekkar, M. (2000). *Hand Book of Water Analysis.* Marcel Dekker, New York, 1-25,115-117,143,175,223-226,261,273,767

Lu, D., Mausel, P., Brondizio, E., and Moran, E. (2002). Assessment of atmospheric correction methods for Landsat TM data applicable to Amazon basin LBA research. *International Journal of Remote Sensing*, **23**, 2651-2671

Ouma, Y. O., Waga, J., Okech, M., Lavisa, O., and Mbuthia, D. (2018). Estimation of reservoir bio-optical water quality parameters using smartphone sensor apps and landsat ETM+: review and comparative experimental results. *Journal of Sensors*, **2018**.

Pawan, K. S. and Pradeep, S. (2015). Analysis of water quality of river Narmada. *International Journal of Current Research*, **7**, 24073-24076.

Pizani, F. M. C., Maillard, P., Ferreira, A. F. F., and Amorim, C. C. (2020). Estimation of water quality in a reservoir from Sentinel-2 and landsat-8 OLI sensors. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Science*, **3**, 401-408.

Ritchie, J. C., Zimba, P. V., and Everitt, J. H. (2003). Remote sensing techniques to assess water quality. *Photogrammetric Engineering and Remote Sensing*, **69**, 695–704

Sawyer, C. N., Mc Carthy, P. L., and Parkin G. F. (1994). *Chemistry for Environmental Engineering*, 4th Edition. Mc Graw-Hill International Edition, New York, 365-577.

Shah, K. A. and Joshi, G. S. (2016). Development of water pollution model: A case study of Mahi river basin, Gujarat, India. *Journal of Environmental Science, Toxicology and Food Technology*, **10**, 59-64.

Simoes, F. S., Moreira, A. B., Bisinoti, M. C., Gimenez, S. M. N., and Yabe, M.J.S. (2008). Water quality index as simple indicator of aquaculture effects on aquatic bodies. *Ecological Indicators*, **8**, 476-484.

Singh, A. K. and Jayakumar, S. (2016). Water quality assessment of Kanwar lake, Begusarai, Bihar, India. *Imperial Journal of Interdisciplinary Research*, **2**, 793-803.

Sudarshan, P., Mahesh, M. K., and Ramachandra, T. V. (2018). Assessment of seasonal variation in water quality and Water Quality Index (WQI) of Hebbal Lake, Bangalore, India. *Journal of Environment and Ecology*, **37**, 309-317.

Trivedy, R. K. and Goel, P. K. (1986). *Chemical and Biological Methods for Water Pollution Studies.* Environmental Publication, India.

Xu, H. Q. (2006). Modification of normalised difference water index (NDWI) to enhance open water features in remotely sensed imagery. *International Journal Remote Sensing*, **27**, 3025–3033.

Statistics and Application {ISSN 2454-7395 (online)}
Volume 21, No. 2, 2023 (New Series), pp 17-36
http://www.ssca.org.in/journal

Estimating Finite Population Mean using Multiple Parameters of an Ancillary Variable

Deepak Singh[1], Rohini Yadav[2] and Rajesh Singh[3]

[1]*ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India.*
[2]*Department of Statistics, University of Lucknow, Lucknow, India.*
[3]*Department of Statistics, Institute of Science, Banaras Hindu University, Varanasi, India*

Received: 20 June 2022; Revised: 25 August 2022; Accepted: 09 September 2022

Abstract

This study deals with an improved class of estimators for estimating the unknown finite population mean of the study variable using auxiliary information. It has been developed by using the power transformation in Singh and Yadav (2017) family of estimators. The expression for bias and mean squared error of the proposed estimator is derived under large sample approximation. The conditions have been derived for the suggested class of estimators under which it performs better than the estimators considered in this study. The theoretical results are supported by numerical illustration. Two phase sampling version of the proposed family of estimators is suggested and its properties are also studied.

Keywords: Study variable; Auxiliary variable; Bias; Mean squared error; Ratio-product-ratio type estimator; Simple random sampling; Double sampling.

Mathematics Subject Classification Code: 62D05.

1. Introduction and notations used

In survey sampling, it is well recognized that the use of auxiliary information results in substantial gain in efficiency over the estimators which do not utilize such information. When the auxiliary variable is available, the ratio, product and regression methods of estimation are the classical examples, which uses auxiliary information and are better than usual mean estimator.

Let there be a finite population $U=\{U_1,U_2,...U_N\}$ of N units and $(y,\ x)$ be the study and auxiliary variables assuming real non-negative values of the finite population U. The population means of the study and auxiliary variables are denoted by $\bar{Y}=\sum_{i=1}^{N}Y_i/N$, $\bar{X}=\sum_{i=1}^{N}X_i/N$; respectively, and sample means by $\bar{y}=\sum_{i=1}^{n}y_i/n$, $\bar{x}=\sum_{i=1}^{n}x_i/n$ respectively.

Corresponding Author: Rohini Yadav
Email: rohiniyadav.ism@gmail.com

Some common notations used in this paper are-

The population variance of the study variable y: $S_y^2 = \left\{1/\left(N-1\right)\right\}\sum_{i=1}^{N}\left(y_i - \bar{Y}\right)^2$

The population variance of the auxiliary variable x: $S_x^2 = \left\{1/\left(N-1\right)\right\}\sum_{i=1}^{N}\left(x_i - \bar{X}\right)^2$

The population covariance: $S_{xy} = \left\{1/\left(N-1\right)\right\}\sum_{i=1}^{N}\left(x_i - \bar{X}\right)\left(y_i - \bar{Y}\right)$

The coefficient of variation of x: $C_x = S_x / \bar{X}$,

The coefficient of variation of y: $C_y = S_y / \bar{Y}$

The population correlation coefficient of x and y: $\rho = S_{xy} / S_x S_y$

and $C = \rho C_y / C_x$.

To estimate the unknown population mean of the study variable $\bar{Y}$, let n pairs of sample observations $(y_i, x_i), i = 1,2,...n$ are drawn using simple sampling without replacement (SRSWOR) from the population U for the study and auxiliary variables respectively. In case no auxiliary information is available, the mean squared error of usual unbiased estimator for population mean under SRSWOR is given by

$$\mathrm{MSE}\left(\bar{y}\right) = \eta \bar{Y}^2 C_y^2 \tag{1}$$

where, $\eta = n^{-1}\left(1-f\right)$, $f = n/N$ (sample fraction).

It is assumed that the population mean of the auxiliary variable $\bar{X}$ is known. The classical ratio estimator $\bar{y}_R = \bar{y}\left(\bar{X}/\bar{x}\right)$ suggested by Cochran (1940) is useful when the study variable and auxiliary variable are positively correlated but when study variable and auxiliary variable are negatively correlated, product estimator $\bar{y}_P = \bar{y}\left(\bar{x}/\bar{X}\right)$ given by Murthy (1964) is more appropriate. The expression for biases and mean squared errors for ratio and product estimators are respectively given by-

$$\mathrm{Bias}\left(\bar{y}_R\right) = \eta \bar{Y} C_x^2 \left(1 - C\right) \tag{2}$$

$$\mathrm{MSE}\left(\bar{y}_R\right) = \eta \bar{Y}^2 \left[C_y^2 + C_x^2\left(1 - 2C\right)\right] \tag{3}$$

$$\mathrm{Bias}\left(\bar{y}_P\right) = \eta \bar{Y} C_x^2 C \tag{4}$$

$$\mathrm{MSE}\left(\bar{y}_P\right) = \eta \bar{Y}^2 \left[C_y^2 + C_x^2\left(1 + 2C\right)\right] \tag{5}$$

Improved estimators for estimating unknown population mean of the study variable utilizing auxiliary are studied by various authors viz. Searls (1964), Upadhyaya *et al.* (1985), Upadhyaya and Singh (1999), Singh and Ruiz Espejo (2003), Upadhyaya *et al.* (2011), Yadav *et al.* (2012), Yadav *et al.* (2013) etc. and the references cited therein.

Chami *et al*. (2012) proposed two-parameter ratio-product-ratio estimator for estimating unknown population mean of the study variable is given by

$$T_{\alpha,1-\alpha,\beta}^{1,1}=\bar{y}\left[\alpha\left\{\frac{(1-\beta)\bar{x}+\beta\bar{X})}{\beta\bar{x}+(1-\beta)\bar{X}}\right\}+(1-\alpha)\left\{\frac{\beta\bar{x}+(1-\beta)\bar{X}}{(1-\beta)\bar{x}+\beta\bar{X}}\right\}\right] \tag{6}$$

Following Chami *et al*. (2012), Singh and Yadav (2017) proposed a ratio-product-ratio family of estimators given by

$$T_{\alpha_1,\alpha_2,\beta}^{\delta,1}=\bar{y}\left[\alpha_1\left\{\frac{(1-\beta)\bar{x}+\beta\bar{X})}{\beta\bar{x}+(1-\beta)\bar{X}}\right\}^{\delta}+\alpha_2\left\{\frac{\beta\bar{x}+(1-\beta)\bar{X}}{(1-\beta)\bar{x}+\beta\bar{X}}\right\}\right] \tag{7}$$

In this paper, a generalized family of ratio-product-ratio type estimators for estimating the population mean of study variable y is proposed which generalizes the earlier works of Chami *et al*. (2012) and Singh and Yadav (2017). It is assumed throughout the paper that the population size N is very large so that the finite population correction term is ignored and $(N-1)\cong N$.

2. The proposed family of estimators

Motivated by Singh and Yadav (2017), we have proposed the following five-parameter ratio-product-ratio type estimator for estimating the population mean $\bar{Y}$ as follows

$$T_{\alpha_1,\alpha_2,\beta}^{\delta,\gamma}=\bar{y}\left[\alpha_1\left\{\frac{(1-\beta)\bar{x}+\beta\bar{X})}{\beta\bar{x}+(1-\beta)\bar{X}}\right\}^{\delta}+\alpha_2\left\{\frac{\beta\bar{x}+(1-\beta)\bar{X}}{(1-\beta)\bar{x}+\beta\bar{X}}\right\}^{\gamma}\right] \tag{8}$$

where α_1, α_2 are constants to be determined such that MSE of the generalized class is minimum, and δ, γ are constants which take finite values for designing the different estimators and β can take any values of the known parameters like coefficient of variation, coefficient of skewness, coefficient of kurtosis and the correlation coefficient (see Singh and Kumar (2011) and Singh and Solanki (2012)). Introducing power transformation in the product type part of the Singh and Yadav (2017) family of estimators $T_{\alpha_1,\alpha_2,\beta}^{\delta,1}$ in the form of γ substantially improves the efficiency of the Singh and Yadav (2017) estimator.

2.1. First-degree approximation to the bias and mean squared error

To obtain the bias and mean squared error (MSE) up to first-degree approximation, we define the following relative error terms

$$\bar{y}=\bar{Y}(1+e_o) \text{ and } \bar{x}=\bar{X}(1+e_1)$$

such that

$E(e_o)=E(e_1)=0,\ E(e_o^2)=\eta C_y^2,\ E(e_1^2)=\eta C_x^2$ and $E(e_o e_1)=\eta\rho C_y C_x=\eta C C_x^2$

We assume that the sample size n is large enough such that contributions from $E(e_o^i)$, $E(e_1^i)$ when $i>2$ and $E(e_o^i e_1^j)$ when $(i+j)>2$ are negligible. Expressing the equation (8) in error terms $(e_i's)$, we get

$$T_{\alpha_1,\alpha_2,\beta}^{\delta,\gamma}=\bar{y}(1+e_o)\left[\alpha_1\{1+(1-\beta)e_1\}^{\delta}(1+\beta e_1)^{-\delta}+\alpha_2(1+\beta e_1)^{\gamma}\{1+(1-\beta)e_1\}^{-\gamma}\right] \tag{9}$$

Expanding $\{1+(1-\beta)e_1\}^{\delta}$, $(1+\beta e_1)^{-\delta}$, $(1+\beta e_1)^{\gamma}$ and $\{1+(1-\beta)e_1\}^{-\gamma}$ as a series in powers of e_1, and assuming $|e_1|<\min.\{1/|\beta|,1/|(1-\beta)|\}$, keeping series up to $O(e_1^2)$ and neglecting higher orders, the bias of $T_{\alpha_1,\alpha_2,\beta}^{\delta,\gamma}$ to order $O(n^{-1})$ is obtained as

$$B(T_{\alpha_1,\alpha_2,\beta}^{\delta,\gamma})=\bar{Y}\begin{bmatrix}\alpha_1\left\{1+\dfrac{\delta(1-2\beta)}{2}\eta C_X^2\left(\delta(1-2\beta)+2C-1\right)\right\}\\ +\alpha_2\left\{1+\dfrac{\gamma(1-2\beta)}{2}\eta C_X^2\left(\gamma(1-2\beta)-2C+1\right)\right\}-1\end{bmatrix} \tag{10}$$

The bias tends to zero when n tends to N and $\alpha_1+\alpha_2=1$. The $MSE(T_{\alpha_1,\alpha_2,\beta}^{\delta,\gamma})$ of the suggested family of estimators to the first degree of approximation is given by

$$MSE(T_{\alpha_1,\alpha_2,\beta}^{\delta,\gamma})=\bar{Y}^2\left(1+\alpha_1^2Z_1+\alpha_2^2Z_2+2\alpha_1\alpha_2Z_3-2\alpha_1Z_4-2\alpha_2Z_5\right) \tag{11}$$

where,

$$Z_1=1+\eta\left[C_y^2+C_x^2\delta(1-2\beta)\{2\delta(1-2\beta)+4C-1\}\right]$$
$$Z_2=1+\eta\left[C_y^2+C_x^2\gamma(1-2\beta)\{2\gamma(1-2\beta)-4C+1\}\right]$$
$$Z_3=1+\eta\left[C_y^2+C_x^2\frac{(1-2\beta)(\delta-\gamma)}{2}\{(1-2\beta)(\delta-\gamma)+4C-1\}\right]$$
$$Z_4=1+\eta C_x^2\frac{\delta(1-2\beta)}{2}\{\delta(1-2\beta)+2C-1\}$$
$$Z_5=1+\eta C_x^2\frac{\gamma(1-2\beta)}{2}\{\gamma(1-2\beta)-2C+1\}$$

Differentiating the $MSE(T_{\alpha_1,\alpha_2,\beta}^{\delta,\gamma})$ at equation (11) with respect to α_1 and α_2 and equating them to zero, we get

$$\begin{bmatrix}Z_1 & Z_3\\ Z_3 & Z_2\end{bmatrix}\begin{bmatrix}\alpha_1\\ \alpha_2\end{bmatrix}=\begin{bmatrix}Z_4\\ Z_5\end{bmatrix} \tag{12}$$

Solving equation (12), we get the optimum values of α_1 and α_2 respectively as

$$\left.\begin{aligned}\alpha_1 &= \left(\frac{Z_2Z_4 - Z_3Z_5}{Z_1Z_2 - Z_3^2}\right) = \alpha_1' \\ \alpha_2 &= \left(\frac{Z_1Z_5 - Z_3Z_4}{Z_1Z_2 - Z_3^2}\right) = \alpha_2'\end{aligned}\right\} \tag{13}$$

Putting the optimum values α_1' and α_2' in place of α_1 and α_2 in equation (11), the minimum MSE of the suggested estimator is given by

$$MSE_{\min}\left(T_{\alpha_1,\alpha_2,\beta}^{\delta,\gamma}\right) = \bar{Y}^2\left\{1 - \frac{\left(Z_2Z_4^2 + Z_1Z_5^2 - 2Z_3Z_4Z_5\right)}{Z_1Z_2 - Z_3^2}\right\} \tag{14}$$

The equation (14) provides the minimum value of the MSE of the proposed family of estimator $T_{\alpha_1,\alpha_2,\beta}^{\delta,\gamma}$.

3. A four-parameter ratio-product-ratio estimator

Putting $\alpha_1 = \alpha$ and $\alpha_2 = 1-\alpha$, the four parameters ratio-product-ratio estimator $T_{\alpha,1-\alpha,\beta}^{\delta,\gamma}$ is given by

$$T_{\alpha,1-\alpha,\beta}^{\delta,\gamma} = \bar{y}\left[\alpha\left(\frac{(1-\beta)\bar{x} + \beta\bar{X})}{\beta\bar{x} + (1-\beta)\bar{X}}\right)^{\delta} + (1-\alpha)\left(\frac{\beta\bar{x} + (1-\beta)\bar{X}}{(1-\beta)\bar{x} + \beta\bar{X}}\right)^{\gamma}\right] \tag{15}$$

where $\alpha_1 + \alpha_2 = 1$. The bias and MSE of the estimator upto the first degree of approximation are respectively given by

$$B\left(T_{\alpha,1-\alpha,\beta}^{\delta,\gamma}\right) = \bar{Y}\frac{(1-2\beta)}{2}\eta C_X^2\left[(1-2\beta)\left\{\alpha\left(\delta^2 - \gamma^2\right) + \gamma^2\right\} + (2C-1)\left\{\alpha\left(\delta+\gamma\right) - \gamma\right\}\right] \tag{16}$$

$$\left.\begin{aligned}&MSE\left(T_{\alpha,1-\alpha,\beta}^{\delta,\gamma}\right) = \bar{Y}^2\eta\left[C_y^2 + C_x^2(1-2\beta)\left\{\gamma - \alpha\left(\delta+\gamma\right)\right\}\left[(1-2\beta)\left\{\gamma - \alpha\left(\delta+\gamma\right)\right\} - 2C\right]\right] \\ &\text{or} \\ &MSE\left(T_{\alpha,1-\alpha,\beta}^{\delta,\gamma}\right) = \bar{Y}^2\left[(1 + Z_2 - 2Z_5) + \alpha^2\left(Z_1 + Z_2 - 2Z_3\right) - 2\alpha\left(Z_2 - Z_3 + Z_4 - Z_5\right)\right]\end{aligned}\right\} \tag{17}$$

$MSE\left(T_{\alpha,1-\alpha,\beta}^{\delta,\gamma}\right)$ is minimum when $\alpha = \dfrac{\left(Z_2 - Z_3 + Z_4 - Z_5\right)}{\left(Z_1 + Z_2 - 2Z_3\right)} = \alpha_{opt}$

and is given by

$$MSE_{\min}\left(T_{\alpha,1-\alpha,\beta}^{\delta,\gamma}\right) = \bar{Y}^2\left[1 + Z_2 - 2Z_5 - \frac{2\left(Z_2 - Z_3 + Z_4 - Z_5\right)^2}{Z_1 + Z_2 - 2Z_3}\right] = \eta S_Y^2\left(1-\rho^2\right) = MSE\left(\bar{y}_{lr}\right) \tag{18}$$

In equation (18), $MSE\left(\bar{y}_{lr}\right)$ indicates the mean square error of the linear regression estimator $\bar{y}_{lr} = \bar{y} + \beta\left(\bar{x} - \bar{X}\right)$. So, $T_{\alpha,1-\alpha,\beta}^{\delta,\gamma}$ is equally efficient to the regression estimator.

4. Efficiency comparison

From equations (1), (3), (5) and (18), we get

$$MSE(\bar{y})-MSE_{\min}(T^{\delta,\gamma}_{\alpha,1-\alpha,\beta})=\eta S_Y^2\rho^2>0 \tag{19}$$

$$MSE(\bar{y}_R)-MSE_{\min}(T^{\delta,\gamma}_{\alpha,1-\alpha,\beta})=\eta\bar{Y}^2C_x^2(1-C)^2>0 \tag{20}$$

$$MSE(\bar{y}_p)-MSE_{\min}(T^{\delta,\gamma}_{\alpha,1-\alpha,\beta})=\eta\bar{Y}^2C_x^2(1+C)^2>0 \tag{21}$$

Hence, the $T^{\delta,\gamma}_{\alpha,1-\alpha,\beta}$ is more efficient than sample mean $\bar{y}$, ratio $\bar{y}_R$ and product $\bar{y}_P$ estimator. The minimum MSE of proposed family of estimators $T^{\delta,\gamma}_{\alpha_1,\alpha_2,\beta}$ is compared with that of four-parameter sub-family of estimators $MSE\left(T^{\delta,\gamma}_{\alpha,1-\alpha,\beta}\right)$ as

$$MSE_{\min}\left(T^{\delta,\gamma}_{\alpha,1-\alpha,\beta}\right)-MSE_{\min}\left(T^{\delta,\gamma}_{\alpha_1,\alpha_2,\beta}\right)=\bar{Y}^2\frac{\left[Z_1\left(Z_2-Z_5\right)+Z_4\left(Z_3-Z_2\right)+Z_3\left(Z_5-Z_3\right)\right]^2}{\left(Z_1+Z_2-2Z_3\right)\left(Z_1Z_2-Z_3^2\right)}>0 \tag{22}$$

In case of $T^{\delta,\gamma}_{\alpha,1-\alpha,\beta}$ subfamily of the estimators i.e. $T^{\delta,\gamma}_{\alpha,1-\alpha,\beta}$, due to the restriction on α_1 and α_2 $\left(\alpha_1+\alpha_2=1\right)$, both the α and $1-\alpha$ coefficients of ratio and product type part of family of estimators are interdependent to each other that leads to obtain the minimum mean square error of $T^{\delta,\gamma}_{\alpha,1-\alpha,\beta}$ under $\alpha_1+\alpha_2=1$ restriction at the optimum value of α i.e. $\alpha_{opt}=\frac{\left(Z_2-Z_3+Z_4-Z_5\right)}{\left(Z_1+Z_2-2Z_3\right)}$. For the proposed family of estimators $T^{\delta,\gamma}_{\alpha_1,\alpha_2,\beta}$ there is no restriction on α_1 and α_2 constants, therefore, the ratio and product type part of family of estimators are independent to each other, which leads to obtain the optimum values of α_1 and α_2 separately i.e. $\alpha_1=\left(\frac{Z_2Z_4-Z_3Z_5}{Z_1Z_2-Z_3^2}\right)=\alpha_1'$ and $\alpha_2=\left(\frac{Z_1Z_5-Z_3Z_4}{Z_1Z_2-Z_3^2}\right)=\alpha_2'$. Since, the ratio and product part of the proposed family of estimators are optimized separately, the minimum mean square error of the proposed family of estimators $T^{\delta,\gamma}_{\alpha_1,\alpha_2,\beta}$ will be always lesser than its subfamily of estimators $T^{\delta,\gamma}_{\alpha,1-\alpha,\beta}$. From equation 22, it is inferred that proposed family of estimators is more efficient than its subfamily of estimators $\left(T^{\delta,\gamma}_{\alpha,1-\alpha,\beta}\right)$. Therefore, the suggested family of estimators $T^{\delta,\gamma}_{\alpha_1,\alpha_2,\beta}$ is more efficient in comparison to the sample mean, ratio, product, regression, and Chami *et al*. (2012) estimator.

Comparing MSE of the proposed subfamily of estimators $T^{\delta,\gamma}_{\alpha,1-\alpha,\beta}$ and Singh and Yadav (2017) subfamily of estimators $T^{\delta,1}_{\alpha,1-\alpha,\beta}$, we get

$$MSE\left(T^{\delta,1}_{\alpha,1-\alpha,\beta}\right)-MSE\left(T^{\delta,\gamma}_{\alpha,1-\alpha,\beta}\right)=\eta C_x^2\bar{Y}^2\left[\begin{matrix}(1-2\beta)(1-\gamma)\\ \left[(1-2\beta)(1+\gamma)-2\alpha\left\{(1+\delta)-2C\right\}\right]\end{matrix}\right]>0 \tag{23}$$

Therefore, we get the following conditions for efficiency for $T_{\alpha,1-\alpha,\beta}^{\delta,\gamma}$ to be better performer than $T_{\alpha,1-\alpha,\beta}^{\delta,1}$,

A. $\gamma < 1,\ \beta < 1/2,\ C > (1-2\beta)\left[(1+\gamma)-2\alpha(1+\delta)\right]/2$

B. $\gamma > 1,\ \beta < 1/2,\ C < (1-2\beta)\left[(1+\gamma)-2\alpha(1+\delta)\right]/2$

C. $\gamma < 1,\ \beta > 1/2,\ C < (1-2\beta)\left[(1+\gamma)-2\alpha(1+\delta)\right]/2$

D. $\gamma > 1,\ \beta > 1/2,\ C > (1-2\beta)\left[(1+\gamma)-2\alpha(1+\delta)\right]/2$

Some known members of the proposed family of estimators $\boldsymbol{T_{\alpha_1,\alpha_2,\beta}^{\delta,\gamma}}$ as well as sub-family of estimators $\boldsymbol{T_{\alpha,1-\alpha,\beta}^{\delta,\gamma}}$ and some new members of the proposed family of estimators along with their corresponding members of Singh and Yadav (2017) estimator are given in the Table 1, Table 2 and Table 3 (see appendix) respectively.

5. Empirical study

To illustrate the relative performance of the members of the proposed family of estimators with other estimators considered in this article, the three natural populations from literature are considered whose descriptions are given below:

Population I [Source: Chami *et. al.* (2012)]

y: Maximum daily values (in feet) of groundwater for the period of October 2009 to September 2010 collected at site number 02290829501 located in Florida.

x: Maximum daily values (in feet) of groundwater for the period of October 2008 to September 2009 collected at site number 02290829501 located in Florida

$N=365,\ n=112,\ \bar{Y}=0.5832,\ \bar{X}=0.6277,\ C_x=1.1504,\ C_y=0.7681,\ \rho=0.9125.$

Population II [Source: Steel and Torrie (1960), pp. 282]

y: Log of leaf burn in sack

x: Chlorine percentage

$N=30,\ n=6,\ \bar{Y}=0.6860,\ \bar{X}=0.8077,\ C_y=0.7001,\ C_x=0.7493,\ \rho=0.4996.$

Population-III [Source: Murthy (1967), pp. 399]

y: Area under wheat in 1964

x: Area under wheat in 1963

$N=34,\ \bar{Y}=199.4411,\ \bar{X}=208.8823,\ C_y=0.753193,\ C_x=0.720486,\ \rho=0.9801.$

The percent relative efficiencies (PREs) of the suggested members of family of estimators

$T^{\delta,\gamma}_{\alpha_1,\alpha_2,\beta}$ and $T^{\delta,1}_{\alpha_1,\alpha_2,\beta}$ are compared with the linear regression estimator $\bar{y}_{lr}$ by using the following formula:

$$PRE\left(t,\bar{y}_{lr}\right)=\frac{Var\left(\bar{y}_{lr}\right)}{MSE_{\min}\left(t\right)}*100 \text{ ; where } t=T^{\delta,\gamma}_{\alpha_1,\alpha_2,\beta} \text{ and } T^{\delta,1}_{\alpha_1,\alpha_2,\beta}$$

The percentage gain in PRE due to the effect of optimized γ in proposed estimator (power transformation in Singh and Yadav (2017) estimator) is given by

$$\%\ gain\ in\ PRE\ =\left\{\left(B-A\right)/A\right\}*100$$

where A is Singh and Yadav (2017) estimator and B is our proposed estimator. The average percentage gain in PRE due to the effect of optimized γ (proposed estimator) in the Singh and Yadav (2017) estimator over three populations considered for study is given by

$$Avg\ \%\ \text{gain}\ in\ PRE=\frac{1}{3}\sum_{i=1}^{3}\left\{\frac{\left(B_i-A_i\right)}{A_i}\right\}*100,\ \left(i=1\text{ to }3\right)$$

In Table 4 (see appendix), the PRE of proposed $T^{\delta,\gamma}_{\alpha_1,\alpha_2,\beta}$ is more to corresponding $T^{\delta,1}_{\alpha_1,\alpha_2,\beta}$ estimator except for four estimators *w.r.t.* population III where the efficiencies are equal. Therefore, it may be concluded that the $T^{\delta,\gamma}_{\alpha_1,\alpha_2,\beta}$ are either more or equally efficient to $T^{\delta,1}_{\alpha_1,\alpha_2,\beta}$ with respect to almost all the estimators when $\alpha_1=\alpha_1'$ and $\alpha_2=\alpha_2'$ which is indicated by average % gain in PRE. Comparing Table 4 (see appendix), it is concluded that all the $T^{\delta,\gamma}_{\alpha_1,\alpha_2,\beta}$ estimators are more efficient than linear regression estimator.

To compare the performance of $T^{\delta,\gamma}_{\alpha_1,\alpha_2,\beta}$ with $T^{\delta,1}_{\alpha_1,\alpha_2,\beta}$, different combinations have been developed for $T^{\delta,\gamma}_{\alpha_1,\alpha_2,\beta}$ (5β x 4δ x $4\gamma=80$ combinations) and $T^{\delta,1}_{\alpha_1,\alpha_2,\beta}$ (5β x $4\delta=20$ combinations) for δ and γ taking values = 1, 1.5, 2, 2.5 for $\beta=1,\rho,C_y,C_x$ and C. In table 5 (see appendix), out of 80 combinations of $T^{\delta,\gamma}_{\alpha_1,\alpha_2,\beta}$ and 20 combinations of $T^{\delta,1}_{\alpha_1,\alpha_2,\beta}$ the best performing estimator for same value of β of $T^{\delta,\gamma}_{\alpha_1,\alpha_2,\beta}$ and $T^{\delta,1}_{\alpha_1,\alpha_2,\beta}$ with respect to population I has been retained and presented.

It is clear from the Table 5 (see appendix) that at same value of β the % gain in efficiency ranges from 0 to 168.07% which concludes that at same value of β, the $T^{\delta,\gamma}_{\alpha_1,\alpha_2,\beta}$ are either equally or more efficient to $T^{\delta,1}_{\alpha_1,\alpha_2,\beta}$ at $\alpha_1=\alpha_1'$ and $\alpha_2=\alpha_2'$.

In Table 6 (see appendix), out of 80 combinations of $T^{\delta,\gamma}_{\alpha_1,\alpha_2,\beta}$ and 20 combinations of $T^{\delta,1}_{\alpha_1,\alpha_2,\beta}$ the top performing estimator with respect to each population has been retained and presented.

From Table 6 (see appendix) , it is clear that the proposed family of estimators is more efficient than Singh and Yadav (2017) family of estimators when dealing with practical and real-world problems where the % gain in efficiency ranges from 20.26 to 564.64 with

respect to three populations.

From the empirical study, it is concluded that $T_{\alpha_1,\alpha_2,\beta}^{\delta,\gamma}$ should be preferred over $T_{\alpha_1,\alpha_2,\beta}^{\delta,1}$ Singh and Yadav (2017) family of estimators. Thus, we recommend the use of proposed family of estimators in practice.

6. The proposed family of estimators in double (two-phase) sampling

In some practical situations, the value of the population mean of the auxiliary variable is unavailable. In such situations, double sampling (two-phase sampling) is used to estimate the population mean $\bar{X}$, from a large sample of size n' drawn from population. A second sample of size $n\ (<n')$ is drawn from this preliminary large sample to observes the study variable y.

Let $\bar{x}' = \sum_{i=1}^{n} x_i / n'$, $\bar{y} = \sum_{i=1}^{n} y_i / n$, and $\bar{x} = \sum_{i=1}^{n} x_i / n$. The usual ratio estimator, product estimator and regression estimators of population mean of study variable y in double sampling are respectively defined as

$$\bar{y}_{Rd} = \bar{y}\left(\frac{\bar{x}'}{\bar{x}}\right) \tag{24}$$

$$\bar{y}_{Pd} = \bar{y}\left(\frac{\bar{x}}{\bar{x}'}\right) \tag{25}$$

$$\bar{y}_{lrd} = \bar{y} + \hat{\beta}\left(\bar{x}' - \bar{x}\right) \tag{26}$$

where $\hat{\beta} = \left(s_{xy} / s_{x^2}\right)$ is the sample regression coefficient.

The double sampling version of suggested generalized family of estimators $D_{\alpha_1,\alpha_2,\beta}^{\delta,\gamma}$ is defined as

$$D_{\alpha_1,\alpha_2,\beta}^{\delta,\gamma} = \bar{y}\left[\alpha_1\left\{\frac{(1-\beta)\bar{x}+\beta\bar{x}')}{\beta\bar{x}'+(1-\beta)\bar{x}}\right\}^{\delta} + \alpha_2\left\{\frac{\beta\bar{x}+(1-\beta)\bar{x}'}{(1-\beta)\bar{x}+\beta\bar{x}'}\right\}^{\gamma}\right] \tag{27}$$

where δ, γ are real constants and β can take values of known parameters like coefficient of variation, coefficient of skewness, coefficient of kurtosis and the correlation coefficient along with real constants and $(\alpha_1,\ \alpha_2)$ are suitably chosen constants such that the mean squared error (MSE) of the developed estimator is minimal. To obtain the bias and mean squared error (MSE) up to first-degree approximation, we define the following relative error terms

$\bar{y} = \bar{Y}(1+e_o)$, $\bar{x} = \bar{X}(1+e_1)$ and $\bar{x}' = \bar{X}(1+e_1')$

such that $E(e_o) = E(e_1) = E(e_1') = 0$, $E(e_o^2) = \eta C_y^2 = V_{(0,2)}$, $E(e_1^2) = \eta C_x^2 = V_{(2,0)}$, $E(e_1'^2) = \eta' C_x^2 = V'_{(2,0)}$, $E(e_o e_1) = \eta\rho C_y C_x = \eta C C_x^2 = V_{(0,1)}$, $E(e_1 e_1') = \eta' C_x^2 = V'_{(1,1)}$ and $E(e_o e_1') = \eta'\rho C_y C_x = \eta' C C_x^2 = V'_{(0,1)}$

We consider the following notations for getting the expression of bias and MSE of the proposed estimator

$$\alpha_1\delta-\alpha_2\gamma=K$$

$$\alpha_1\delta^2-\alpha_2\gamma^2=L$$

$$\alpha_1\delta\left\{\delta\left(1-2\beta\right)-1\right\}+\alpha_2\gamma\left\{\gamma\left(1-2\beta\right)+1\right\}=S$$

$$\alpha_1\delta\left\{\delta\left(1-2\beta\right)+1\right\}+\alpha_2\gamma\left\{\gamma\left(1-2\beta\right)-1\right\}=T$$

$$\psi=V_{(2,0)}+V'_{(2,0)}-2V'_{(1,1)}$$

$$\omega=\gamma\left(1-2\beta\right)\left[V_{(2,0)}\left\{\gamma\left(1-2\beta\right)+1\right\}+V'_{(2,0)}\left\{\gamma\left(1-2\beta\right)-1\right\}-2V_{(0,1)}+2V'_{(0,1)}-2V'_{(1,1)}\gamma\left(1-2\beta\right)\right]$$

$$\mu=\delta\left(1-2\beta\right)\left[V_{(2,0)}\left\{\delta\left(1-2\beta\right)-1\right\}+V'_{(2,0)}\left\{\delta\left(1-2\beta\right)+1\right\}+2V_{(0,1)}-2V'_{(0,1)}-2V'_{(1,1)}\delta\left(1-2\beta\right)\right]$$

$$\tau=\delta\left(1-2\beta\right)\left[\frac{V_{(2,0)}}{2}\left\{\delta\left(1-2\beta\right)-1\right\}+\frac{V'_{(2,0)}}{2}\left\{\delta\left(1-2\beta\right)+1\right\}+2V_{(0,1)}-2V'_{(0,1)}-V'_{(1,1)}\delta\left(1-2\beta\right)\right]$$

$$\phi=\gamma\left(1-2\beta\right)\left[\frac{V_{(2,0)}}{2}\left\{\gamma\left(1-2\beta\right)+1\right\}+\frac{V'_{(2,0)}}{2}\left\{\gamma\left(1-2\beta\right)-1\right\}-2V_{(0,1)}+2V'_{(0,1)}-V'_{(1,1)}\gamma\left(1-2\beta\right)\right]$$

$$\xi=1+V_{(0,2)}$$

We assume that the sample size n is large enough such that contributions from $E\left(e_o^i\right)$, $E\left(e_1'^i\right)$ when $i>2$ and $E\left(e_o^i e_1^j\right)$ when $(i+j)>2$ are negligible. Expressing the equation (27) in error terms $(e_i's)$, we get

$$D^{\delta,\gamma}_{\alpha_1,\alpha_2,\beta}=\bar{Y}\left(1+e_o\right)=\bar{y}\left[\alpha_1\left\{\frac{1+\beta e_1'+\left(1-\beta\right)e_1}{1+\beta e_1+\left(1-\beta\right)e_1'}\right\}^{\delta}+\alpha_2\left\{\frac{1+\beta e_1+\left(1-\beta\right)e'}{1+\beta e_1'+\left(1-\beta\right)e_1}\right\}^{\gamma}\right] \tag{28}$$

Expanding $\left\{1+\beta e_1'+\left(1-\beta\right)e_1\right\}^{\delta}$, $\left\{1+\beta e_1+\left(1-\beta\right)e'\right\}^{-\delta}$, $\left\{1+\beta e_1+\left(1-\beta\right)e'\right\}^{\gamma}$ and $\left\{1+\beta e_1'+\left(1-\beta\right)e_1\right\}^{-\gamma}$ as a series in powers of e_1 and e_1', it is assumed that $|e_1|<\min.\left\{\frac{1}{|\beta|},\frac{1}{|(1-\beta)|}\right\}$. Keeping series up to $O\left(e_1^2\right)$ and neglecting higher orders, the bias of $D^{\delta,\gamma}_{\alpha_1,\alpha_2,\beta}$ to order $O(n^{-1})$ is obtained as

$$B\left(D^{\delta,\gamma}_{\alpha_1,\alpha_2,\beta}\right)=\bar{Y}\left[\alpha_1\left(1+\frac{\mu}{2}\right)+\alpha_2\left(1+\frac{\omega}{2}\right)-1\right] \tag{29}$$

The MSE of the proposed estimator is given by

$$MSE\left(D^{\delta,\gamma}_{\alpha_1,\alpha_2,\beta}\right)=\bar{Y}^2\left(1+\alpha_1^2Z_1+\alpha_2^2Z_2+2\alpha_1\alpha_2Z_3-2\alpha_1Z_4-2\alpha_2Z_5\right) \tag{30}$$

where,

$$Z_1 = \xi + 2\tau + \delta^2\psi\left(1-2\beta\right)^2$$

$$Z_2 = \xi + 2\phi + \gamma^2\psi\left(1-2\beta\right)^2$$

$$Z_3 = \xi + \phi + \tau - \delta\gamma\psi\left(1-2\beta\right)^2$$

$$Z_4 = \frac{\mu+2}{2}$$

$$Z_5 = \frac{\omega+2}{2}$$

Differentiating the $MSE\left(D^{\delta,\gamma}_{\alpha_1,\alpha_2,\beta}\right)$ with respect to α_1 and α_2 and equating them to zero, we have

$$\begin{bmatrix} Z_1 & Z_3 \\ Z_3 & Z_2 \end{bmatrix}\begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} = \begin{bmatrix} Z_4 \\ Z_5 \end{bmatrix} \tag{31}$$

Solving equation (31), we get the optimum values of α_1 and α_2 as

$$\begin{aligned} \alpha_1 &= \left(\frac{Z_2Z_4 - Z_3Z_5}{Z_1Z_2 - Z_3^2}\right) = \alpha_1' \\ \alpha_2 &= \left(\frac{Z_1Z_5 - Z_3Z_4}{Z_1Z_2 - Z_3^2}\right) = \alpha_2' \end{aligned} \tag{32}$$

Putting the optimum values α_1' and α_2' in place of α_1 and α_2 in equation (29) and (30), the optimum bias and the minimum mean square error of $D^{\delta,\gamma}_{\alpha_1,\alpha_2,\beta}$ is obtained as

$$B_{opt}\left(D^{\delta,\gamma}_{\alpha_1,\alpha_2,\beta}\right) = -\bar{Y}\left\{1 - \frac{\left(Z_2Z_4^2 + Z_1Z_5^2 - 2Z_3Z_4Z_5\right)}{Z_1Z_2 - Z_3^2}\right\} \tag{33}$$

$$MSE_{\min}\left(D^{\delta,\gamma}_{\alpha_1,\alpha_2,\beta}\right) = \bar{Y}^2\left\{1 - \frac{\left(Z_2Z_4^2 + Z_1Z_5^2 - 2Z_3Z_4Z_5\right)}{Z_1Z_2 - Z_3^2}\right\} \tag{34}$$

The equation (34) provides the minimum value of the MSE of the proposed family of estimator $D^{\delta,\gamma}_{\alpha_1,\alpha_2,\beta}$.

6.1. Particular case in two-phase sampling

For $\alpha_1 + \alpha_2 = 1$ the suggested family reduces to the following family of estimators

$$D^{\delta,\gamma}_{\alpha,1-\alpha,\beta} = \bar{y}\left[\alpha\left(\frac{\left(1-\beta\right)\bar{x} + \beta\bar{X})}{\beta\bar{x} + \left(1-\beta\right)\bar{X}}\right)^{\delta} + (1-\alpha)\left(\frac{\beta\bar{x} + \left(1-\beta\right)\bar{X}}{\left(1-\beta\right)\bar{x} + \beta\bar{X}}\right)^{\gamma}\right] \tag{35}$$

The bias and MSE of the estimator to the first degree of approximation are derived as

$$B\left(D^{\delta,\gamma}_{\alpha_1,\alpha_2,\beta}\right)=\frac{\overline{Y}}{2}\left[\omega+\alpha\left(\mu-\omega\right)\right] \tag{36}$$

$$MSE\left(D^{\delta,\gamma}_{\alpha,1-\alpha,\beta}\right)=\overline{Y}^2\left[\left(1+Z_2-2Z_5\right)+\alpha^2\left(Z_1+Z_2-2Z_3\right)-2\alpha\left(Z_2-Z_3+Z_4-Z_5\right)\right] \tag{37}$$

For minima, we took gradient $\overline{V}=\left(\frac{\partial}{\partial\alpha}\right)$ of $MSE\left(D^{\delta,\gamma}_{\alpha,1-\alpha,\beta}\right)$ and equating it to zero, we get $\beta=\left(1/2\right)$and $C=\left(1-2\beta\right)\left\{\gamma-\alpha\left(\gamma+\delta\right)\right\}$.

When $\beta=\left(1/2\right)$is used in $D^{\delta,\gamma}_{\alpha,1-\alpha,\beta}$, we get the usual unbiased estimator of $\overline{y}$ and the MSE expression becomes

$$MSE\left(D^{\delta,\gamma}_{\frac{\gamma}{\gamma+\delta},1-\frac{\gamma}{\gamma+\delta},\frac{1}{2}}\right)=\eta S_y^2 \tag{38}$$

and when $C=\left(1-2\beta\right)\left\{\gamma-\alpha\left(\gamma+\delta\right)\right\}$ is used in $D^{\delta,\gamma}_{\alpha,1-\alpha,\beta}$, we get the asymptotically optimum estimator (AOE) $D^{\delta,\gamma}_{\alpha,1-\alpha,\beta}$and the MSE expression transforms into

$$MSE\left(D^{\delta,\gamma}_{\alpha,1-\alpha,\beta}\right)=\eta^* S_y^2\left(1-\rho^2\right)+\eta' S_y^2 \tag{39}$$

Also, $MSE\left(D^{\delta,\gamma}_{\alpha,1-\alpha,\beta}\right)$ is minimum when $\alpha=\frac{\left(Z_2-Z_3+Z_4-Z_5\right)}{\left(Z_1+Z_2-2Z_3\right)}=\alpha_{opt}$ *i.e.*

$$\begin{aligned} MSE_{\min}\left(D^{\delta,\gamma}_{\alpha,1-\alpha,\beta}\right)&=\overline{Y}^2\left[1+Z_2-2Z_5-\frac{2\left(Z_2-Z_3+Z_4-Z_5\right)^2}{Z_1+Z_2-2Z_3}\right] \\ &=\eta^* S_y^2\left(1-\rho^2\right)+\eta' S_y^2=MSE\left(\overline{y}_{lrd}\right)\end{aligned} \tag{40}$$

where $\eta'=\left(N-n'\right)/\left(Nn'\right)^{-1}$ and $\eta^*=\eta-\eta'=\left(n'-n\right)\left(nn'\right)^{-1}$.

Here, $MSE\left(\overline{y}_{lrd}\right)$ is the MSE of the double sampling version of linear regression estimator. Therefore, the estimator $D^{\delta,\gamma}_{\alpha,1-\alpha,\beta}$ is equally efficient to double sampling version of linear regression estimator.

7. Efficiency comparison

The suggested class of estimators is compared with $\overline{y}_d$, $\overline{y}_{Rd}$ and $\overline{y}_{Pd}$ in terms of MSE's.

$$MSE(\overline{y}_d)=\eta' S_y^2 \tag{41}$$

$$MSE(\overline{y}_{Rd})=\eta\overline{Y}^2\left[\eta C_y^2+\eta^* C_x^2(1-2C)\right] \tag{42}$$

$$MSE(\bar{y}_{Pd}) = \eta \bar{Y}^2 \left[\eta C_y^2 + \eta^* C_x^2 (1 + 2C) \right] \tag{43}$$

From equations (34), (40), (41), (42) and (43) we get

$$MSE\left(\bar{y}_d\right) - MSE_{\min}\left(D^{\delta,\gamma}_{\alpha,1-\alpha,\beta}\right) = \eta^* \bar{Y}^2 C_y^2 \rho^2 \tag{44}$$

$$MSE\left(\bar{y}_{Rd}\right) - MSE_{\min}\left(D^{\delta,\gamma}_{\alpha,1-\alpha,\beta}\right) = \eta^* \bar{Y}^2 C_x^2 \left(1 - C\right)^2 \tag{45}$$

$$MSE\left(\bar{y}_{Rd}\right) - MSE_{\min}\left(D^{\delta,\gamma}_{\alpha,1-\alpha,\beta}\right) = \eta^* \bar{Y}^2 C_x^2 \left(1 - C\right)^2 \tag{46}$$

$$MSE_{\min}\left(D^{\delta,\gamma}_{\alpha,1-\alpha,\beta}\right) - MSE_{\min}\left(D^{\delta,\gamma}_{\alpha_1,\alpha_2,\beta}\right) = \bar{Y}^2 \frac{\left[Z_1\left(Z_2 - Z_5\right) + Z_4\left(Z_3 - Z_2\right) + Z_3\left(Z_5 - Z_3\right)\right]^2}{\left(Z_1 + Z_2 - 2Z_3\right)\left(Z_1 Z_2 - Z_3^2\right)} > 0 \tag{47}$$

The minimum MSE of $D^{\delta,\gamma}_{\alpha,1-\alpha,\beta}$ subfamily of estimators will always be larger than the proposed family of estimators $D^{\delta,\gamma}_{\alpha_1,\alpha_2,\beta}$ as the coefficients of ratio and product type part of family of estimators in subfamily are interdependent in $D^{\delta,\gamma}_{\alpha,1-\alpha,\beta}$ while they are independent in $D^{\delta,\gamma}_{\alpha_1,\alpha_2,\beta}$.

From equations (44) to (47), it is clear that the proposed family of estimators $D^{\delta,\gamma}_{\alpha_1,\alpha_2,\beta}$ is more efficient than its subfamily of estimators $D^{\delta,\gamma}_{\alpha,1-\alpha,\beta}$, double sampling version of sample mean estimator $\bar{y}_d$, ratio estimator $\bar{y}_{Rd}$, and product estimator $\bar{y}_{Pd}$.

8. Empirical study

To illustrate the relative performance of the members of the proposed family of estimators with other estimators considered in this article, the two natural populations from literature are considered whose descriptions are given below:

Population I [Source: Koyunchu and Kadilar (2009)]

y: Number of teachers teaching in both primary and secondary schools.
x: Number of students studying in both primary and secondary schools.

$N = 923,\ n' = 270,\ n = 180,\ \bar{Y} = 436.43,\ \bar{X} = 11440.50,\ C_x = 1.86,\ C_y = 1.72,\ C = 0.88,$
$\rho = 0.95.$

Population II [Source: Cochran (1977, p.172)]

y: Production of peaches (I bushels).
x: Peach trees in an orchard.

$N = 256,\ n' = 150,\ n = 100,\ \bar{Y} = 56.47,\ \bar{X} = 44.45,\ C_x = 1.40,\ C_y = 1.42,\ C = 0.90,$
$\rho = 0.89.$

The percent relative efficiencies (*PREs)* of the suggested family of estimators $D^{\delta,\gamma}_{\alpha_1,\alpha_2,\beta}$ are compared with the double sampling version linear regression estimator $\bar{y}_{lrd}$ by using the following formulas:

$$PRE\left(D^{\delta,\gamma}_{\alpha_1,\alpha_2,\beta},\bar{y}_{lrd}\right)=\frac{\left(\eta * S_y^2\left(1-\rho^2\right)+\eta' S_y^2\right)}{\bar{Y}^2\left[1-\frac{\left(Z_2Z_4^2+Z_1Z_5^2-2Z_3Z_4Z_5\right)}{Z_1Z_2-Z_3^2}\right]}*100 \tag{48}$$

It is observed from the Table 7 (see appendix) that all the members of proposed family of estimators are performing better in comparison to double sampling linear regression estimator, therefore, the members are also efficient to double sampling version ratio and product estimator.

9. Conclusion

We have dealt with the problem of estimating the population mean $\bar{Y}$ of study variable using the auxiliary information in the form of different parameters of the variable x .The proposed family of estimators are very wide and many new estimators can be derived from the suggested class of estimators. It includes all the estimators recently proposed by Singh and Yadav (2017) along with the two parameters ratio-product-ratio estimator proposed by Chami *et al.* (2012).

To judge the performance of the proposed family of estimators with other estimators, an empirical study has been carried out. From the Table 4 (see appendix), it is observed that the suggested family of estimators is efficient to sample mean estimator $\bar{y}$, linear regression estimator $\bar{y}_{lr}$ and Singh and Yadav (2017) estimators. At same value of β, proposed family of estimators should be preferred over Singh and Yadav (2017) estimators (Table 5 (see appendix)) . For identifying the most efficient estimator, the proposed family of estimators should be preferred over Singh and Yadav (2017) family of estimators (Table 6 (see appendix)). All the members of double sampling version of the proposed family of estimators are efficient to double sampling version of ratio, product, and linear regression estimator. Thus, we recommend the use of proposed family of estimators in practice.

References

Cochran, W. G. (1977). *Sampling Techniques*. New-York: John Wiley and Sons.

Chami, S. P., Singh, B. and Thomas, D. (2012). A two-parameter ratio-product-ratio estimator using auxiliary information. *ISRN Probability and Statistics*, 1-15.

Cochran W. G. (1940). The estimation of the yields of the cereal experiments by sampling for the ratio of grain to total produce. *The Journal of Agricultural Science*, **59**, 1225-1226.

Gupta, P. C. and Kothwala, N. H. (1990). A study of second order approximation for some product type estimators. *Journal of the Indian Society of Agricultural Statistics*, **42**, 171-185.

Kadilar, C. and Cingi, H. (2006). An improvement in estimating the population mean by using the correlation coefficient. *Hacettepe Journal of Mathematics and Statistics,* **35**, 103-109.

Murthy, M. N. (1964). Product method of estimation. *Sankhya, The Indian Journal of Statistics -Series A,* **26**, 69-74.

Murthy, M. N. (1967). *Sampling Theory and Methods*. Calcutta: Statistical Publishing Society.

Pandey, G. S. (1980). Product-cum-power estimators. *Calcutta Statistical Association Bulletin,* **29**, 103-108.

Searls, D. T. (1964). The utilization of known coefficient of variation in the estimation procedure. *Journal of the American Statistical Association,* **59**, 1125-1126.

Singh, H. P. and Ruiz-Espejo, M. (2003). On linear regression and ratio-product estimation of a finite population mean. *Statistics,* **52**, 59-67.

Singh, H. P. and Solanki, R. S. (2012). A new procedure for variance estimation in SRS using auxiliary information. Statistical Papers, **54**, 479–497. DOI 10.1007/s00362-012-0445-2.

Singh, H. P. and Yadav, A. (2017). An improved family of estimators of finite population mean using information on an auxiliary variable in sample surveys. *Journal of Applied Mathematics, Statistics and Informatics,* **13**, 77-107.

Singh, R. and Kumar, M. (2011): A note on transformations on auxiliary variable in survey sampling. *Model Assisted Statistics and Applications*, **6**, 17-19.

Srivastava, S. K. (1967). An estimator using auxiliary information in sample surveys. *Calcutta Statistical Association Bulletin,* **16**, 121–132.

Steel, R. G. D. and Torrie, J. H. (1960). *Principles and Procedures of Statistics*. McGraw Hill Book Co.

Sukhatme, P. V. and Sukhatme, B. V. (1970). *Sampling Theory of Surveys with Applications.* Ames, IO: Iowa State University Press.

Swain, A. K. P. C. (2014). On an improved ratio type estimator of finite population mean in sample surveys. *Revista de Investigacion Operacional,* **35**, 49-57.

Upadhyaya, L. N., Singh, H. P., and Vos, J. W. E. (1985). On the estimation of population means and ratios using supplementary information. *Statistica Neerlandica,* **39**, 309-318.

Upadhyaya, L. N. and Singh, H. P. (1999). Use of transformed auxiliary variable in estimating the finite population means. *Biometrical Journal*, **41**, 627-636.

Upadhyaya, L. N., Singh, H. P., Chatterjee, S., and Yadav, R. (2011). A generalized family of transformed ratio-product estimators in sample surveys. *Model Assisted Statistics and Applications,* **6**, 137-150.

Yadav, R., Upadhyaya, L. N., Singh, H. P., and Chatterjee, S. (2012). Almost unbiased ratio and product type exponential estimators. *Statistics in Transition,* **13**, 537-550.

Yadav, R., Upadhyaya, L. N., Singh, H. P. and Chatterjee, S. (2013). A generalized family of transformed ratio-product estimators for variance in sample surveys. *Communications in Statistics - Theory and Methods,* **42**, 1839-1850.

Appendix

Table 1: Some known members of the proposed family of estimators $T^{\delta,\gamma}_{\alpha_1,\alpha_2,\beta}$

S.No.	Value of constants $(\alpha_1,\alpha_2,\beta,\delta,\gamma)$	Estimator	S.No.	Value of constants $(\alpha_1,\alpha_2,\beta,\delta,\gamma)$	Estimator
1	$(\alpha_1, \alpha_2, 1, 0, 1)$ Upadhyaya *et al.*(1985) estimator	$T^{0,1}_{\alpha_1,\alpha_2,1}$	2.	$(\alpha_1, \alpha_2, 1, 1, 0)$ Upadhyaya *et al.* (1985) estimator	$T^{1,0}_{\alpha_1,\alpha_2,1}$

Table 2: Some known members of the $T^{\delta,\gamma}_{\alpha,1-\alpha,\beta}$ sub-family of estimators

S.No.	Value of constants $(\alpha_1,\alpha_2,\beta,\delta,\gamma)$	Estimator	S.No.	Value of constants $(\alpha_1,\alpha_2,\beta,\delta,\gamma)$	Estimator
1.	$(\alpha, 1\text{-}\alpha, 1, 1, 1)$ Singh & Ruiz Espejo (2003)	$T^{1,1}_{\alpha,1-\alpha,1}$	4.	$(\alpha, 1\text{-}\alpha, 1, \delta, 1)$ Pandey (1980)	$T^{\delta,1}_{\alpha,1-\alpha,1}$
2.	$(\alpha, 1\text{-}\alpha, \beta, 1, 1)$ Chami *et al*. (2012)	$T^{1,1}_{\alpha,1-\alpha,\beta}$	5.	$(1, 0, 1, 1/2, *)$ Swain (2014)	$T^{1/2,*}_{1,0,1}$
3.	$(1, 0, 1, 2, *)$ Kadilar and Cingi (2006)	$T^{2,*}_{1,0.1}$	6.	$(1, 0, 1, \delta, *)$ Srivastava (1967)	$T^{\delta,*}_{1,0,1}$

Table 3: Some members of the proposed family of estimators along with their corresponding members of Singh and Yadav (2017) estimator

	Members of estimator $T^{\delta,\gamma}_{\alpha_1,\alpha_2,\beta}$		**Corresponding members of Singh and Yadav (2017) Estimator** $T^{\delta,1}_{\alpha_1,\alpha_2,\beta}$ **at** $\gamma=1$	
Sl. No.	**Value of constants** $(\alpha_1,\alpha_2,\beta,\delta,\gamma)$	**Estimator**	**Value of constants** $(\alpha_1,\alpha_2,\beta,\delta,\gamma)$	**Estimator**
1.	$(\alpha_1, \alpha_2, 1, 0, \gamma)$	$T^{0,\gamma}_{\alpha_1,\alpha_2,1}$	$(\alpha_1, \alpha_2, 1, 0, 1)$	$T^{0,1}_{\alpha_1,\alpha_2,1}$
2.	$(\alpha_1, \alpha_2, 0, 0, \gamma)$	$T^{0,\gamma}_{\alpha_1,\alpha_2,0}$	$(\alpha_1, \alpha_2, 0, 0, 1)$	$T^{0,1}_{\alpha_1,\alpha_2,0}$
3.	$(\alpha_1, \alpha_2, 1, 1/2, \gamma)$	$T^{1/2,\gamma}_{\alpha_1,\alpha_2,1}$	$(\alpha_1, \alpha_2, 1, 1/2, 1)$	$T^{1/2,1}_{\alpha_1,\alpha_2,1}$
4.	$(\alpha_1, \alpha_2, 1, 2, \gamma)$	$T^{2,\gamma}_{\alpha_1,\alpha_2,1}$	$(\alpha_1, \alpha_2, 1, 2, 1)$	$T^{2,1}_{\alpha_1,\alpha_2,1}$
5.	$(\alpha_1, \alpha_2, C_x, 1/2, \gamma)$	$T^{1/2,\gamma}_{\alpha_1,\alpha_2,C_x}$	$(\alpha_1, \alpha_2, C_x, 1/2, 1)$	$T^{1/2,1}_{\alpha_1,\alpha_2,C_x}$
6.	$(\alpha_1, \alpha_2, C_x, 1, \gamma)$	$T^{1,\gamma}_{\alpha_1,\alpha_2,C_x}$	$(\alpha_1, \alpha_2, C_x, 1, 1)$	$T^{1,1}_{\alpha_1,\alpha_2,C_x}$
7.	$(\alpha_1, \alpha_2, C_x, 2, \gamma)$	$T^{2,\gamma}_{\alpha_1,\alpha_2,C_x}$	$(\alpha_1, \alpha_2, C_x, 2, 1)$	$T^{2,1}_{\alpha_1,\alpha_2,C_x}$
8.	$(\alpha_1, \alpha_2, C_x, 0, \gamma)$	$T^{0,\gamma}_{\alpha_1,\alpha_2,C_x}$	$(\alpha_1, \alpha_2, C_x, 0, 1)$	$T^{0,1}_{\alpha_1,\alpha_2,C_x}$
9.	$(\alpha_1, \alpha_2, \rho, 1, \gamma)$	$T^{1,\gamma}_{\alpha_1,\alpha_2,\rho}$	$(\alpha_1, \alpha_2, \rho, 1, 1)$	$T^{1,1}_{\alpha_1,\alpha_2,\rho}$
10.	$(\alpha_1, \alpha_2, \rho, 2, \gamma)$	$T^{2,\gamma}_{\alpha_1,\alpha_2,\rho}$	$(\alpha_1, \alpha_2, \rho, 2, 1)$	$T^{2,1}_{\alpha_1,\alpha_2,\rho}$
11.	$(\alpha_1, \alpha_2, C_y, 1, \gamma)$	$T^{1,\gamma}_{\alpha_1,\alpha_2,C_y}$	$(\alpha_1, \alpha_2, C_y, 1, 1)$	$T^{1,1}_{\alpha_1,\alpha_2,C_y}$
12.	$(\alpha_1, \alpha_2, C_y, 2, \gamma)$	$T^{2,\gamma}_{\alpha_1,\alpha_2,C_y}$	$(\alpha_1, \alpha_2, C_y, 2, 1)$	$T^{2,1}_{\alpha_1,\alpha_2,C_y}$
13.	$(0.5, 0.5, 1, 2, \gamma)$	$T^{2,\gamma}_{0.5,0.5,1}$	$(0.5, 0.5, 1, 2, 1)$	$T^{2,1}_{0.5,0.5,1}$
14.	$(0.5, 0.5, C_x, 2, \gamma)$	$T^{2,\gamma}_{0.5,0.5,C_x}$	$(0.5, 0.5, C_x, 2, 1)$	$T^{2,1}_{0.5,0.5,C_x}$
15.	$(0.5, 0.5, \rho, 1, \gamma)$	$T^{2,\gamma}_{0.5,0.5,\rho}$	$(0.5, 0.5, \rho, 1, 1)$	$T^{2,1}_{0.5,0.5,\rho}$
16.	$(0.5, 0.5, \rho, 2, \gamma)$	$T^{2,\gamma}_{0.5,0.5,C_y}$	$(0.5, 0.5, \rho, 2, 1)$	$T^{2,1}_{0.5,0.5,C_y}$
17.	$(0.5, 0.5, C_y, 2, \gamma)$	$T^{2,\gamma}_{0.5,0.5,C_y}$	$(0.5, 0.5, C_y, 2, 1)$	$T^{2,1}_{0.5,0.5,C_y}$

Note: *Estimators from S. No. 1 to 12 corresponds to* $\alpha_1=\alpha_1,\ \alpha_2=\alpha_2$ *and estimators from S. No. 13 to 17 corresponds for* $\alpha_1=\alpha,\ \alpha_2=1-\alpha$

Table 4: PREs of several members of proposed family of estimators and Singh and Yadav (2017) estimator due to $\gamma = 1$

Estimator	Real values		$T^{\delta,1}_{\alpha_1,\alpha_2,\beta}$	$T^{\delta,\gamma}_{\alpha_1,\alpha_2,\beta}$
	β	δ	PRE (A)	PRE (B)
Population I				
1	1	0	$100.88\,(\gamma = 1)$	$470.61\,(\gamma = 8)$
2	0	0	$100.98\,(\gamma = 1)$	$1545.40\,(\gamma = 9)$
3	1	2	$122.64\,(\gamma = 1)$	$435.12\,(\gamma = 3)$
4	C_x	0.5	$100.50\,(\gamma = 1)$	$138.31\,(\gamma = 16)$
5	C_x	1	$108.47\,(\gamma = 1)$	$466.59\,(\gamma = 5)$
6	C_x	2	$187.59\,(\gamma = 1)$	$478.72\,(\gamma = 1.5)$
7	C_x	0	$101.60\,(\gamma = 1)$	$388.53\,(\gamma = 6)$
8	ρ	1	$101.60\,(\gamma = 1)$	$388.53\,(\gamma = 20)$
9	ρ	2	$101.24\,(\gamma = 1)$	$869.98\,(\gamma = 5)$
10	C_y	2	$102.25\,(\gamma = 1)$	$1254.19\,(\gamma = 19)$
Population II				
1	1	0	$102.23\,(\gamma = 1)$	$207.24\,(\gamma = -5)$
2	0	0	$109.61\,(\gamma = 1)$	$207.24\,(\gamma = 5)$
3	1	2	$151.75\,(\gamma = 1)$	$1461.06\,(\gamma = 3)$
4	C_x	0.5	$105.20\,(\gamma = 1)$	$113.62\,(\gamma = -5)$
5	C_x	1	$107.24\,(\gamma = 1)$	$108.42\,(\gamma = 5)$
6	C_x	2	$112.53\,(\gamma = 1)$	$132.25\,(\gamma = 5)$
7	C_x	0	$103.52\,(\gamma = 1)$	$121.42\,(\gamma = -5)$
8	ρ	1	$103.52\,(\gamma = 1)$	$121.42\,(\gamma = 5)$
9	ρ	2	$105.14\,(\gamma = 1)$	$105.15\,(\gamma = 5)$
10	C_y	2	$109.88\,(\gamma = 1)$	$117.19\,(\gamma = 5)$
Population III				
1	1	0	$3725.78\,(\gamma = 1)$	$3725.78\,(\gamma = 1)$
2	0	0	$100.02\,(\gamma = 1)$	$902.26\,(\gamma = 3)$

3	1	2	$560.56\,(\gamma=1)$	$560.56\,(\gamma=1)$
4	C_x	0.5	$139.24\,(\gamma=1)$	$496.05\,(\gamma=3)$
5	C_x	1	$116.27\,(\gamma=1)$	$567.12\,(\gamma=5)$
6	C_x	2	$100.47\,(\gamma=1)$	$116.15\,(\gamma=10)$
7	C_x	0	$191.26\,(\gamma=1)$	$672.73\,(\gamma=2)$
8	ρ	1	$191.26\,(\gamma=1)$	$672.73\,(\gamma=10)$
9	ρ	2	$100.09\,(\gamma=1)$	$353.49\,(\gamma=1)$
10	C_y	2	$100.03\,(\gamma=1)$	$100.03\,(\gamma=1)$

Note: (1) “A” indicates Singh and Yadav (2017) estimator; (2) “B” indicates proposed estimator.

Table 5: Top performing estimators of $T^{\delta,\gamma}_{\alpha_1,\alpha_2,\beta}$ and $T^{\delta,1}_{\alpha_1,\alpha_2,\beta}$ at constant β when $\alpha_1=\alpha_1'$ and $\alpha_2=\alpha_2'$ for Population-I

S.no.	β	$T^{\delta,1}_{\alpha_1,\alpha_2,\beta}$		$T^{\delta,\gamma}_{\alpha_1,\alpha_2,\beta}$			% gain in PRE
		δ	PRE (A)	δ	γ	PRE (B)	
1.	1	2.5	145.85	2.5	2.0	390.98	**168.07**
2.	1.1504	2.5	455.30	1.5	2.5	547.56	20.26
3.	0.9125	2.5	119.60	2.5	2.5	193.41	61.72
4.	0.7681	2.5	104.15	2.5	2.5	109.65	5.28
5.	0.6092	2.5	100.21	2.5	1	100.21	0.00

Table 6: Top performing estimators of $T^{\delta,\gamma}_{\alpha_1,\alpha_2,\beta}$ and $T^{\delta,1}_{\alpha_1,\alpha_2,\beta}$ when $\alpha_1=\alpha_1'$ and $\alpha_2=\alpha_2'$ for three populations

Pop.	$T^{\delta,1}_{\alpha_1,\alpha_2,\beta}$			$T^{\delta,\gamma}_{\alpha_1,\alpha_2,\beta}$				% gain in PRE
	β	δ	PRE (A)	β	δ	γ	PRE (B)	
i.	C_x	2.5	455.30	C_x	1.5	2.5	547.56	20.26
ii.	1	2.5	189.85	1	2.5	2	710.58	274.29
iii.	1	1	560.57	1	0	1	3725.78	**564.64**

Table 7: PREs of $D^{\delta,\gamma}_{\alpha_1,\alpha_2,\beta}$ family of estimators when $\alpha_1 = \alpha_1'$ and $\alpha_2 = \alpha_2'$

Estimator	Real values	Population I		Population II	
	β	(δ, γ)	PRE of $D^{\delta,\gamma}_{\alpha_1,\alpha_2,\beta}$ w.r.t. $MSE(\bar{y}_{lrd})$	(δ, γ)	PRE of $D^{\delta,\gamma}_{\alpha_1,\alpha_2,\beta}$ w.r.t. $MSE(\bar{y}_{lrd})$
1	1	$(\delta=1,\gamma=8)$	**141465.86**	$(\delta=5,\gamma=5)$	1212.20
2	0	$(\delta=2,\gamma=10)$	1977.39	$(\delta=4,\gamma=6)$	3448.88
3	ρ	$(\delta=1,\gamma=10)$	956.79	$(\delta=2,\gamma=12)$	6151.28
4	C_x	$(\delta=0.5,\gamma=1)$	596.20	$(\delta=0.5,\gamma=6)$	1317.25
5	C_y	$(\delta=0.5,\gamma=1)$	272.94	$(\delta=0.5,\gamma=6)$	**15756.31**
6	C	$(\delta=1,\gamma=20)$	16963.29	$(\delta=2,\gamma=11)$	12634.51
7	f	$(\delta=12,\gamma=6)$	5528.56	$(\delta=0.5,\gamma=2.5)$	445.12
8	$1-f$	$(\delta=2,\gamma=10)$	382.42	$(\delta=4,\gamma=0.5)$	547.72
9	$1/1+f$	$(\delta=2,\gamma=10)$	1360.60	$(\delta=10,\gamma=12)$	1943.52
10	$2f/1+f$	$(\delta=0.5,\gamma=1)$	302.70	$(\delta=5,\gamma=1.5)$	276.38
11	$f/1-f$	$(\delta=1,\gamma=2)$	255.01	$(\delta=1,\gamma=1.5)$	268.89
12	$1-f/1+f$	$(\delta=10,\gamma=15)$	813.12	$(\delta=1,\gamma=5)$	305.58

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 21, No. 2, 2023 (New Series), pp 37-58
http://www.ssca.org.in/journal

Prediction Intervals in ARCH Models Using Sieve Bootstrap Robust Against Outliers

Samir Barman, Ramasubramanian V., Mrinmoy Ray and Ranjit Kumar Paul
ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India

Received: 07 May 2022; Revised: 28 August 2022; Accepted: 30 September 2022

Abstract

One of the primary goals of time series (TS) modeling is to forecast future observations. Although point forecasts are the most common type of prediction, interval forecasts are more informative and are typically obtained as prediction intervals (PIs). For non-linear TS data, the ARCH model is one of the widely used models. The Sieve Bootstrap method is a popular method for constructing PIs in TS models. The TS data are not always free from outliers, whose presence may result in an increase in the length of PIs obtained also with poor coverage. In this study, two new robust Sieve Bootstrap approaches based on weighted least squares estimation have been proposed to deal with the presence of outliers for developing PIs for both returns and volatilities in the ARCH model setup. The performances of the proposed methods viz., Robust Unconditional Sieve Bootstrap (RUSB) and Robust Sieve Bootstrap (RSB) for constructing PIs using both simulated as well as real data sets have been found to be better when compared with their existing counterparts.

Key words: Coverage probability; Innovative outlier; Length of prediction interval; Return; Volatility; Weighted least squares.

1. Introduction

A time series (TS) is an ordered sequence of data points observed over time, typically at equally spaced time intervals. The analysis of TS is essential not only in agriculture but also in other diverse fields such as economics, finance pattern recognition, tourism etc. In all these areas, TS methodologies are used not only to model TS data, but also to forecast future values of such processes. TS predictions can be observed either as point or interval estimates. Point estimation is concerned with predicting a single value from a set of observations, whereas interval estimation provides prediction intervals (PIs), with some probability, within which forecasted future values will lie. There are many reasons for preferring PIs over point estimates. PIs help to assess the future uncertainty in a broad manner for better risk management decisions, plan different strategies for the range of possible outcomes, and explore scenarios based on different assumptions more carefully and so on. A good account

Corresponding Author: Ramasubramanian V.
Email: ram.vaidhyanathan@gmail.com

on PIs in TS can be found in many books, to cite a few, Politis *et al.* (1999), Chatfield (2000) and Lahiri (2003).

In the context of agricultural commodity price or any financial TS data, generally linear TS models with homoscedastic error variance are popularly used until it need to deal with volatile data. Volatility being the sudden unexpected rise or fall in TS, measuring it plays an important role in assigning risk and uncertainty. While modeling TS, a series is said to be volatile when a few error terms are larger than the others and are responsible for the unique behavior of the series, resulting in heteroscedasticity. To deal with volatilities and non-linear dynamics, the Auto-Regressive Conditional Heteroscedastic (ARCH) model proposed by Engle (1982) where the idea is to model volatilities as a linear function of previous returns, is popularly employed. By adding a moving average part, the ARCH model was generalized by Bollerslev (1986) in the form of the Generalized ARCH (GARCH) model for the parsimonious representation of ARCH. In the GARCH model, the conditional variance is also a linear function of its own lags. In this context, the GARCH model became the most popularly used for modeling volatility and obtaining dynamic PIs for returns and volatilities. Many recent studies are found on the non-linear TS processes in modeling volatilities (to cite a few, see, Bhardwaj *et al.*, 2014; Lama *et al.*, 2015; Bentes, 1015; Dyhrberg, 2016).

Existing literature mainly focused on point forecasts of volatilities and little attention has been given to constructing the PIs (Baillie and Bollerslev, 1992; Andersen and Bollerslev, 1998; Andersen *et al.*, 2001; Poon, 2005). However, the construction of PIs in TS models with finite parameters, requires knowledge of the distribution of the observed data, which is typically unknown in practice. Several studies have shown that when the underlying distributional assumptions are violated the resulting PIs can be adversely affected yielding poor results (Thombs and Schucany, 1990). The construction of PIs in TS models with finite parameters and with known innovative processes has been widely discussed in the literature and it has been found that these PIs are extremely sensitive to the presence of outliers (Tsay, 1988, 2010). Moreover, over time, several distribution-free methods, using resampling techniques using Bootstrap method, have been proposed as an alternative for the construction of PIs. One of the popular and effective Bootstrap procedures is residual-based resampling *i.e.* resampling the residuals from the fitted model on the TS (Bühlmann, 2002; Politis, 2003; Härdle *et al.*, 2003). Miguel and Olave (1999) first proposed a Bootstrap procedure for a non-linear ARCH model for the construction of PIs for return and volatilities by directly adding resampled residuals from the ARCH model to the respective point forecasts. This work was improved by Reeves (2005) by adding an additional step of re-estimating the ARCH parameters for each Bootstrap realization of the returns, which considered the variability of the estimated parameters of the ARCH model. Further, Pascual *et al.* (2006) extended these procedures for the GARCH model in different ways and obtained the PIs for both returns and volatilities which were found to be well-calibrated *i.e.*, the number of observed data falling within PIs coincided with the declared coverage. However, these procedures involve the estimation of ARCH/GARCH parameters by maximum likelihood (ML) estimation and are computationally expensive. Hence as an improvement over these, Chen *et al.* (2011) proposed a computationally efficient and distribution-free resampling technique for developing PIs for both returns and volatilities in ARCH and GARCH processes. Their method was based on the Sieve Bootstrap procedure used in the linear model AR/ARMA representation of the ARCH/GARCH process. In particular, the squared returns from the ARCH/GARCH model is a linear process that follows an AR/ARMA process (Tsay, 2010;

Box *et al.*, 2015). Bose and Mukherjee (2009) proposed a weighted linear estimator (WLE) to estimate the ARCH parameters, and a corresponding Bootstrap weighted linear estimator (BWLE). An alternative WLE method in the context of multivariate ARCH models was proposed by Iqbal (2011) and improved results were reported. Later, Iqbal and Chand (2013) constructed efficient PIs for returns and volatility for ARCH models using a particular version of residual Bootstrap. Further Pan and Politis (2016) proposed a Bootstrap algorithm for developing PIs for ARCH models based on BWLE. However, these above-mentioned approaches including the Sieve Bootstrap procedure are affected by the presence of innovative outliers, resulting in an undesirable increase in the length of the PIs. In recent times, Ulloa *et al.* (2014) and Allende *et al.* (2015) have proposed a residual-based resampling technique for developing robust PIs for returns and volatilities for GARCH models based on the winsorized residuals. Trucíos *et al.* (2017) constructed Bootstrap densities for returns and volatilities using a robust parameter estimator based on variance-targeting implemented together with an adequate modification of the volatility filter in analyzing the effect of additive outliers. Beyaztas and Shang (2020) proposed a robust Bootstrap technique for PI construction in AR models based on weighted likelihood estimates and weighted residuals. The presence of outliers can have an impact on TS analysis, leading to incorrect model identification and parameter estimation and TS forecasts obtained from such models could be erroneous. Hence, there is always a need to develop improved and computationally efficient Bootstrap methods in computing PIs for TS aimed at providing better forecasts. In this study, the focus is on developing models robust against the presence of outliers to get improved PIs. This approach of robust modeling has been applied using the Sieve Bootstrap procedure for developing PIs for both return and volatilities in the ARCH model setup. In addition, instead of applying least square estimation (see, Chen *et al.*, 2011), a weighted least squares (WLS) estimation has been applied. The details of the new WLS method and the proposed Bootstrap procedure have been described in subsequent sections.

Towards this end, two new Bootstrap approaches for constructing PIs have been proposed in this study. The remainder of the article is organized as follows. The next section discusses the two proposed methods by first describing about the ARCH models and the weighted least squares procedure employed. Thereafter Section 3 deals with the results of the simulation study conducted followed by Section 4 which contains a case study on a real data set. The paper is signed off with concluding remarks in Section 5.

2. Methodology

2.1. ARCH models

A non-linear TS model can be expressed as $y_t = f(\varepsilon_t, \varepsilon_{t-1}, \dots)$ where $f(\cdot)$ is the non-linear function of past and present random shocks. In such a setup, consider a TS $\{y_t\}_{t=1}^n$ following ARCH(p) process, $p \geq 1$ has the following representation:

$$y_t = \sigma_t \varepsilon_t \tag{1}$$

$$\sigma_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i y_{t-i}^2 \tag{2}$$

where $\{\varepsilon_t\}_{t=1}^n$ is a sequence of independently and identically distributed (*i.i.d.*) random variables with zero mean and unit variance and $E(\varepsilon_t^4) < \infty$; the volatility process $\{\sigma_t\}_{t=1}^n$ is

a stochastic process assumed to be independent of $\{\varepsilon_t\}_{t=1}^n$; α_0, α_i's are unknown parameters satisfying $\alpha_0, \alpha_i \geq 0$, for $i = 1, 2, \ldots, p$. The process is assumed to be weakly stationary (Tsay, 2010) *i.e.* $\sum_{i=1}^p \alpha_i < 1$ is satisfied. Further, it is assumed that the strict stationarity conditions of $\{y_t\}_{t=1}^n$ given in Bougerol and Picard (1992a, 1992b) hold.

Despite the non-linear nature of variance in ARCH models, they can be represented by means of the linear AR model (Tsay, 2010; Box *et al.*, 2015). In particular, the squared returns of an ARCH model is a linear process that can be written as an AR representation. From (1) and (2),

$$y_t^2 = \sigma_t^2 \varepsilon_t^2 \tag{3}$$

$$\alpha_0 + \sum_{i=1}^{p} \alpha_i y_{t-i}^2 = \sigma_t^2 \tag{4}$$

Subtracting equation (4) from equation (3),

$$y_t^2 - \left(\alpha_0 + \sum_{i=1}^{p} \alpha_i y_{t-i}^2\right) = \sigma_t^2 \varepsilon_t^2 - \sigma_t^2 \tag{5}$$

Let, $\nu_t = \sigma_t^2 \varepsilon_t^2 - \sigma_t^2 = y_t^2 - \sigma_t^2$, and by substituting $\sigma_t^2 = y_t^2 - \nu_t$ in (4) yielding,

$$y_t^2 - \nu_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i y_{t-i}^2$$

$$y_t^2 = \alpha_0 + \sum_{i=1}^{p} \alpha_i y_{t-i}^2 + \nu_t \tag{6}$$

where $\{y_t^2\}_{t=1}^n$is an AR(p) process and $\nu_t = y_t^2 - \sigma_t^2$ is white noise but not *i.i.d.*, in general. Under strict stationarity assumptions of $\{y_t\}_{t=1}^n$, innovations $\{\nu_t\}_{t=1}^n$ are identically distributed.
Let $p = 1$, then, $\{y_t\}_{t=1}^n$ follows ARCH(1):

$$y_t = \sigma_t \varepsilon_t \tag{7}$$

$$\sigma_t^2 = \alpha_0 + \alpha_1 y_{t-1}^2 \tag{8}$$

Then from equation (6), ARCH(1) can be expressed in AR(1) form:

$$y_t^2 = \alpha_0 + \alpha_1 y_{t-1}^2 + \nu_t \tag{9}$$

Similarly, suppose $\{y_t\}_{t=1}^n$ follows an ARCH(2), then it can be rewritten in AR(2) form as:

$$y_t^2 = \alpha_0 + \alpha_1 y_{t-1}^2 + \alpha_2 y_{t-2}^2 + \nu_t \tag{10}$$

2.2. Weighted least squares (WLS) estimation

In this study, following Chen *et al.* (2011), the AR parameterization of the ARCH model presented in equation (6) has been considered and estimated using WLS estimation for constructing the PIs. Let, $x_t = y_t^2$ and for an ARCH model equation (6) can be written as,

$$x_t = \alpha_0 + \sum_{i=1}^{p} \alpha_i x_{t-i} + \nu_t \tag{11}$$

The Least Squares (LS) estimators of an AR(p) model are obtained by fitting a linear regression of x_t onto $x_{t-1}, x_{t-2}, \ldots, x_{t-m}$. In matrix notation, let $\boldsymbol{z}$ and $\boldsymbol{X}$ as follows:

$$\boldsymbol{z} = \begin{bmatrix} x_{p+1} \\ \vdots \\ x_n \end{bmatrix} \text{ and } \boldsymbol{X} = \begin{bmatrix} 1 & x_p & x_{p-1} & \cdots & x_1 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n-1} & x_{n-2} & \cdots & x_{n-p} \end{bmatrix}$$

The LS estimate of parameters $\widehat{\Phi} = (\widehat{\alpha}_0, \widehat{\alpha}_1, \ldots, \widehat{\alpha}_p)'$ is obtained as

$$\widehat{\Phi} = \left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{z} \tag{12}$$

with $\boldsymbol{X}'\boldsymbol{X}$ is non-singular.

It is a known fact that when the TS data are contaminated with outliers, the LS estimates of model parameters are affected *i.e.* they produce biased estimates and the errors computed corresponding to outliers will be large. Thus the Bootstrap PIs based on LS estimates may not provide reliable results in the presence of outliers. Therefore it is proposed to construct robust Bootstrap PIs for the ARCH process based on WLS estimates of parameters, on similar lines to the weighted procedure employed in the case of likelihood estimation by Markatou *et al.* (1998) and Beyaztas and Shang (2020); also in partial least squares estimation by Beyaztas and Shang (2021) to improve the robustness of the estimates.

Now, from equation (11), let $\nu_t(\Phi) = \nu_t(\Phi|x_t) = x_t - \alpha_0 - \sum_{i=1}^{p} \alpha_i x_{t-i}$ for $t = p+1,\ p+2, \ldots, n$ be the model residuals, where the values of ν_t for $t \leq p$ are taken as zero. Let $f^*(\cdot)$ be the non-parametric kernel density estimator and $m^*(\cdot)$ be the smoothed model density, respectively, defined as follows:

$$f^*\left(\nu_t(\Phi), \hat{F}_\nu(\Phi)\right) = \int k\left(\nu_t(\Phi), r, d\right) d\hat{F}_\nu(r,\Phi) \quad \forall\ t = 1, 2, \ldots, n$$

$$m^*\left(\nu_t(\Phi), \sigma^2\right) = \int k\left(\nu_t(\Phi), r, d\right) dM\left(r, \sigma^2\right)$$

where $\hat{F}_\nu(\Phi)$ is the empirical cumulative distribution function based on $\nu_t(\Phi)$ and $M(\sigma^2)$ is actual assumed model distribution function with variance σ^2, such as general normal distribution with zero mean and variance σ^2. Function $k\left(\nu_t(\Phi), r, d\right)$ is the kernel density with bandwidth d. The weight function, say $w(\cdot)$, is defined according to the minimum discrepancy measure, as a measure of agreement between the parametric model of the error and the actual residuals. Following Beyaztas and Shang (2020, 2021), the Pearson residual δ_t is then defined as:

$$\delta_t = \delta\left(\nu_t(\Phi); M\left(\sigma^2\right), \hat{F}_\nu(\Phi)\right) = \frac{f^*\left(\nu_t(\Phi), \hat{F}_\nu(\Phi)\right) - m^*\left(\nu_t(\Phi), \sigma^2\right)}{m^*\left(\nu_t(\Phi), \sigma^2\right)} \quad \forall\ t = 1, 2, \ldots, n \tag{13}$$

and weight function $w(\delta_t)$ is then defined as:

$$w(\delta_t) = w\left(\nu_t(\Phi); M\left(\sigma^2\right), \hat{F}_\nu(\Phi)\right) = \min\left\{1, \frac{\left[A(\delta_t) + 1\right]^+}{\delta_t + 1}\right\} \tag{14}$$

where $[\cdot]^+$ indicates the positive part and $A(\cdot)$ denotes the residual adjustment function (RAF) of Lindsay (1994) (here in this study, Hellinger RAF $A(\delta) = 2\left[(\delta+1)^{1/2} - 1\right]$ have been used). Then the WLS estimate for Φ is obtained as:

$$\widehat{\Phi}^w = \left(\boldsymbol{X}'\boldsymbol{W}\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{W}\boldsymbol{z} \tag{15}$$

where $\boldsymbol{W} = diag\left(w\left(\delta_t\right)\right)$ and $\widehat{\Phi}^w = \left(\widehat{\alpha}_0^w, \widehat{\alpha}_1^w, \ldots, \widehat{\alpha}_p^w\right)'$. From equations (13) and (14), it can be seen that when the model assumptions are holding good and with no outliers present in the data, δ_t converges to zero and $w\left(\delta_t\right)$ converges to 1. Similarly, in the presence of outliers, δ_t will be larger and corresponding $w\left(\delta_t\right)$ will be smaller than 1 *i.e.* the outlier observations will get less weight.

2.3. Robust bootstrap procedures

Sieve Bootstrap was first proposed by Buhlmann (1997) as a variation in Bootstrap process where sieves of linear autoregressive processes are used to approximate the underlying process to estimate the distribution of a statistical quantity of the process. The idea of Sieve Bootstrap is that it involves the sampling of the residuals of a fitted autoregressive or AR(p_n) models of order p_n, where $p_n \to \infty$ as $n \to \infty$, and then new Bootstrap realizations are generated from the resampled residuals. In this study, two new Bootstrap methods robust against outliers have been proposed for constructing PIs for an ARCH model. The first one *i.e.* robust unconditional Sieve Bootstrap (RUSB) is an improvement of the unconditional Sieve Bootstrap (USB) method for the ARCH process proposed by Chen *et al.* (2011) and the second one *i.e.* robust Sieve Bootstrap (RSB) is a modification of the SB method described by Tresch (2015). In both the existing methods, the estimation of parameters was done by the ordinary least squares method. This estimation yields poor results in the presence of outliers. To handle such outliers, here the estimations of parameters have been done by the WLS procedure.

Let $\{y_t\}_{t=1}^n$ follows the realization of an ARCH(p) process and it has the model representation given in equation (1), equation (2) and its AR representation in equation (6). Further letting $x_t = y_t^2$ for $t = 1, 2, \ldots n$, it can be easily presented by equation (11).

2.3.1. Robust unconditional sieve bootstrap (RUSB) method

The steps involved in this proposed algorithm are as follows:

1. Considering the model representation of equation (11), estimate the ARMA coefficients $\widehat{\Phi}^w = \left(\widehat{\alpha}_0^w, \widehat{\alpha}_1^w, \ldots, \widehat{\alpha}_p^w\right)'$ using the WLS method as in equation (15).

2. Estimate the residuals $\{\widehat{\nu}_t\}_{t=p+1}^n$ as

$$\widetilde{\nu}_t = x_t - \widehat{\alpha}_0^w - \sum_{i=1}^{p} \widehat{\alpha}_i^w x_{t-i} \tag{16}$$

where $\widetilde{\nu}_t = 0,$ for $t = 1, 2, \ldots, p$.

3. Center the estimated residuals $\widehat{\nu}_t = \widetilde{\nu}_t - (n-p)^{-1}\sum_{t=p+1}^{n}\widetilde{\nu}_t$ and then calculate the empirical distribution of the centered residuals as

$$\hat{F}_{\widehat{\nu}_t}(x) = (n-p)^{-1}\sum_{t=p+1}^{n} I_{(-\infty,x]}(\widehat{\nu}_t) \tag{17}$$

4. Resample with replacement, Bootstrap innovations $\{\nu_t^*\}$ from $\hat{F}_{\widehat{\nu}_t}(x)$.

5. Generate the Bootstrap sample of squared return x_t^*, where $x_t^* = y_t^{2*}$, by the recursion

$$x_t^* = \widehat{\alpha}_0^w + \sum_{i=1}^{p}\widehat{\alpha}_i^w x_{t-i}^* + \nu_t^* \tag{18}$$

where $x_t^* = \widehat{\alpha}_0^w/\{1-\sum_{i=1}^{p}\widehat{\alpha}_i^w\}$ and $\nu_t^* = 0$ for $t \le p$. Generate $(n+200)$ values of x_t^* and then drop the first 200 "burn-in" observations to reduce the effect of the starting values as asymptotically negligible. (Kreiss and Franke, 1992).

6. Now given $\{x_t^*\}_{t=1}^n$ from Step 5, fit the model given by equation (11) then estimate the coefficients by the WLS method, and let the resultant estimated coefficients be $\widehat{\Phi}^{w*} = \left(\widehat{\alpha}_0^{w*}, \widehat{\alpha}_1^{w*}, \ldots, \widehat{\alpha}_p^{w*}\right)'$.

7. Then Bootstrap sample of volatility $\{\sigma_t^{2*}\}_{t=1}^n$ is obtained as

$$\sigma_t^{2*} = \widehat{\alpha}_0^{w*} + \sum_{i=1}^{p}\widehat{\alpha}_i^{w*} x_{t-i}^* \quad \text{for } t = p+1, p+2, \ldots, n. \tag{19}$$

where $\sigma_t^{2*} = \widehat{\alpha}_0^w/\{1-\sum_{i=1}^{p}\widehat{\alpha}_i^w\}$ for $t = 1, \ldots, p$.

8. Again sample with replacement, Bootstrap innovations $\left\{\nu_{n+h}^*\right\}_{h=1}^{s}$, $s > 0$, from $\hat{F}_{\widehat{\nu}_t}(x)$ to obtain future Bootstrap observations.

9. Compute the h-step ahead, $h = 1, 2, \ldots, s$, future Bootstrap observations for squared returns x_{n+h}^* and volatility σ_{n+h}^{2*} by the recursions

$$x_{n+h}^* = \widehat{\alpha}_0^{w*} + \sum_{i=1}^{p}\widehat{\alpha}_i^{w*} x_{n+h-i}^* + \nu_{n+h}^* \tag{20}$$

$$\sigma_{n+h}^{2*} = \widehat{\alpha}_0^{w*} + \sum_{i=1}^{p}\widehat{\alpha}_i^{w*} x_{n+h-i}^* \tag{21}$$

where $x_{n+h}^* = x_{n+h}$ for $h \le 0$.

10. Repeat Steps 4 to 9 B times to generate B Bootstrap replicates.

11. Obtain the empirical Bootstrap distribution function $\hat{F}_{x_{n+h}^*}^*$ of x_{n+h}^*, where $x_{n+h}^* = y_{n+h}^{2*}$, to approximate the unknown distribution of x_{n+k} given the observed sample and $\hat{F}_{\sigma_{n+h}^{2*}}^*$

of σ^{2*}_{n+h} to approximate the unknown distribution σ^2_{n+h}.
The $(1-\alpha)\,100\%$ PIs for future returns y_{n+h} is given by

$$\left[Q^*_{n+h}\left(\alpha/2\right),\ Q^*_{n+h}(1-\alpha/2)\right] \tag{22}$$

where $Q^*_{n+h}\left(\alpha/2\right) = -\sqrt{H^*_{n+h}\left(1-\alpha\right)}$ and $Q^*_{n+h}\left(1-\alpha/2\right) = \sqrt{H^*_{n+h}\left(1-\alpha\right)}$ where $H^*_{n+h}\left(1-\alpha\right)$ is the $(1-\alpha)$ quantile of $\hat{F}^*_{x^*_{n+h}}$.
Similarly, the $(1-\alpha)\,100\%$ PIs for σ^2_{n+h} is given by

$$\left[0,\ K^*_{n+h}(1-\alpha)\right] \tag{23}$$

where $K^*_{n+h}(1-\alpha)$ is the $(1-\alpha)$ quantile of $\hat{F}^*_{\sigma^{2*}_{n+h}}$.

2.3.2. Robust sieve bootstrap (RSB) method

It is possible to write that an AR process of $\{x_t\}_{t=1}^n$, as in equation (11), in the form of an infinite AR representation:

$$\sum_{j=0}^{\infty}\varphi_j\left(x_{t-j}-\mu_x\right)=\nu_t\ ,\ \varphi_0=1,\quad for\ \ t\in\mathbb{Z} \tag{24}$$

with coefficients satisfying the condition $\sum_{j=0}^{\infty}\varphi_j^2<\infty$. Let the parameter μ_x be estimated by its empirical mean $\overline{x}=\frac{1}{n}\sum_{t=1}^n x_t$, as has been done by Alonso *et al.* (2002, 2003, 2004). The steps involved in the proposed algorithm are as follows:

1. For the given realization of squared return series, $\{x_t\}_{t=1}^n$, select the maximum order $p_{\max}=p\left(n\right)$ of the AR approximation and using AICC criteria, obtain the optimum order. The optimum order has been considered as $\hat{p}=p_{AICC}+1$ for the order of the AR model to be fitted to the observed data. In the Monte Carlo simulation, $p_{\max}=p\left(n\right)$ was taken as $(n/10)$, as recommended by Bhansali (1983) where n is the sample size.

2. Estimate the coefficients of AR($\hat{p}$) process using the WLS method described in equation (15). Let the estimates be $\widehat{\varphi}^w_1,\widehat{\varphi}^w_2,\ldots,\widehat{\varphi}^w_{\hat{p}}$ in place of the Yule-Walker method used for coefficient estimation within Tresch (2015).

3. Compute the $(n-\hat{p})$ residuals as

$$\widetilde{\nu}_t=\sum_{j=0}^{\hat{p}}\widehat{\varphi}^w_j\left(x_{t-j}-\overline{x}\right);\quad \widehat{\varphi}^w_0=1,\ t\in(\hat{p}+1,\hat{p}+2,\ldots,n) \tag{25}$$

where $\overline{x}$ is the mean of $\{x_t\}_{t=1}^n$.

4. Center the residuals as $\widehat{\nu}_t=\widetilde{\nu}_t-\overline{\widetilde{\nu}}_t$, where $\overline{\widetilde{\nu}}_t=\left(n-\hat{p}\right)^{-1}\sum_{t=\hat{p}+1}^n\widetilde{\nu}_t$. Then compute the empirical distribution function of the centered residuals $\hat{F}_{\widehat{\nu}}(x)=\left(n-\hat{p}\right)^{-1}\sum_{t=\hat{p}+1}^n I_{(-\infty,x]}\left(\widehat{\nu}_t\right)$.

5. Resample with replacement, Bootstrap innovations ν_t^* from this distribution $\hat{F}_{\widehat{\nu}}(x)$ for $t = -199, -198, \ldots, 0, 1, \ldots, n$.

6. Generate the Bootstrap series x_t^*, $t = -199, -198, \ldots, 0, 1, \ldots, n$ by the recursion as:

$$\sum_{j=0}^{\hat{p}} \widehat{\varphi}_j^w \left(x_{t-j}^* - \overline{x} \right) = \nu_t^* \tag{26}$$

where the first $\hat{p}$ values are taken as $x_t^* = \overline{x}$. Then drop the first 200 "burn-in" observations to reduce the effect of the starting values as asymptotically negligible.

7. Fit an $AR\ (\hat{p})$ model to the pseudo-data $\{x_1^*, x_2^*, \ldots, x_n^*\}$, re-estimate the coefficients using the WLS method and let the estimated coefficients be $\widehat{\varphi}_1^{w*}, \widehat{\varphi}_2^{w*}, \ldots, \widehat{\varphi}_{\hat{p}}^{w*}$.

8. Using the new coefficients $\widehat{\varphi}_1^{w*}, \widehat{\varphi}_2^{w*}, \ldots, \widehat{\varphi}_{\hat{p}}^{w*}$, compute the h-step ahead future Bootstrap observations by the recursion as:

$$x_{n+h}^* - \overline{x} = -\sum_{j=1}^{\hat{p}} \widehat{\varphi}_j^{w*} \left(x_{n+h-j}^* - \overline{x} \right) + \nu_{n+h}^{**} \tag{27}$$

where $x_t^* = x_t$ when $t \leq n$ with ν_{n+h}^{**} for $h = 1, 2, \ldots, s$, resampled from $\hat{F}_{\widehat{\nu}}(x)$.
Also, instead of employing fixed $\overline{x}$, here the mean of the Bootstrap series $\overline{x}^*$ has been employed as an estimate of the mean μ_x at individual Bootstrap prediction, following Mukhopadhyay and Samaranayake (2010), since it includes sampling variability. So to account for the sampling variability due to the estimate of the mean μ_x of the TS, add $(\overline{x}^* - \overline{x})$ to predict future observations x_{n+h}^*. Thus the future Bootstrap squared return is then $\hat{x}_{n+h}^* = x_{n+h}^* + \overline{x}^* - \overline{x}$ for $h = 1, 2, \ldots, s$.

9. Using the future values x_{n+h}^* and the relationship for AR and ARCH/GARCH process, the future volatility can be calculated by the following recursion:

$$\sigma_{n+h}^{2*} = \overline{x}^* - \sum_{j=1}^{\hat{p}} \widehat{\varphi}_j^{w*} \left(x_{n+h-j}^* - \overline{x} \right) \tag{28}$$

where $x_{n+h-j}^* = x_{n+h}$for $h \leq 0$.

10. Repeat steps 4 to 9 B times to generate B Bootstrap replicates. Then obtain the empirical Bootstrap distribution function $\hat{F}_{x_{n+h}^*}^*$ of x_{n+h}^*, where $x_{n+h}^* = y_{n+h}^{2*}$, to approximate the unknown distribution of x_{n+k} given the observed sample and $\hat{F}_{\sigma_{n+h}^{2*}}^*$ of σ_{n+h}^{2*} to approximate the unknown distribution σ_{n+h}^2.

11. The $(1-\alpha)\,100\%$ PIs for future return y_{n+h} is given by:

$$\left[Q_{n+h}^*(\alpha/2),\ Q_{n+h}^*(1-\alpha/2) \right] \tag{29}$$

where $Q_{n+h}^*(\alpha/2) = -\sqrt{H_{n+h}^*(1-\alpha)}$ and $Q_{n+h}^*(1-\alpha/2) = \sqrt{H_{n+h}^*(1-\alpha)}$ where $H_{n+h}^*(1-\alpha)$ is the $(1-\alpha)$ quantile of $\hat{F}_{x_{n+h}^*}^*$.
Similarly, the $(1-\alpha)\,100\%$ PIs for σ_{n+h}^2 is given by:

$$\left[0,\ K_{n+h}^*(1-\alpha) \right] \tag{30}$$

where $K^*_{n+h}(1-\alpha)$ is the $(1-\alpha)$ quantile of $\hat{F}^*_{\sigma^{2*}_{n+h}}$.

In the SB method by Tresch (2015), the future volatilities have been calculated by the recursion of $\sigma^{2*}_{n+h} = x^*_{n+h} - \sum_{j=1}^{\hat{p}} \widehat{\varphi}^*_j x^*_{n+h-j}$ for $h = 1, 2, \ldots, s$. This has been changed in RSB and given in (28). It is also noted that the use of $\overline{x}^*$ in the second proposed method has been done which incorporates the advantage of the Bootstrap sampling variability on future volatilities.

A schematic diagram of the method in Section 2.3.1. is given in the Figure 1 below.

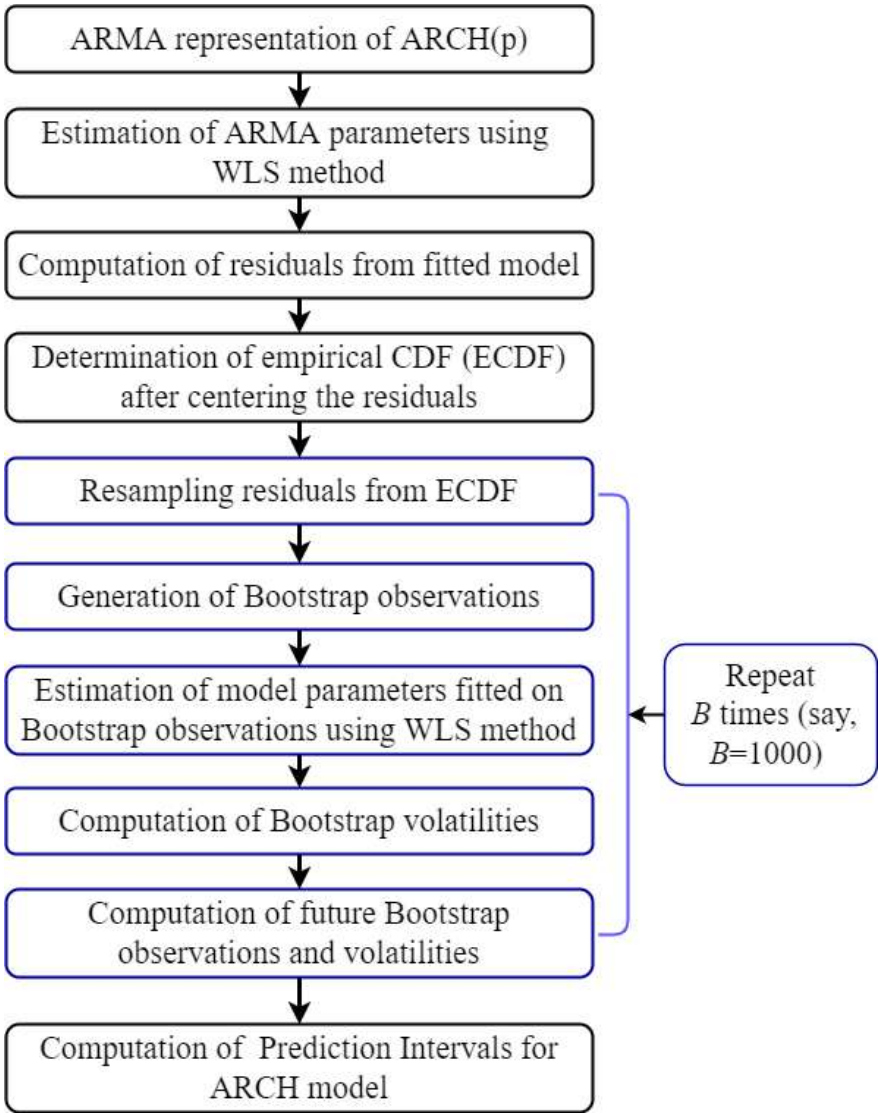


Figure 1: A Schematic diagram of the algorithm in 2.3.1.

3. Simulation results

To compare the finite sample performance of the proposed Bootstrap methods with the existing Bootstrap methods, a Monte-Carlo simulation study has been carried out on

an ARCH(2) model for varying sample sizes and with data having no contamination and also with contamination (read innovative outliers). Data were generated using the following ARCH(2) model for heteroscedastic errors:

$$y_t = \sigma_t \varepsilon_t \tag{31}$$

$$\sigma_t^2 = 0.1 + 0.2y_{t-1}^2 + 0.15y_{t-2}^2 \tag{32}$$

generated separately considering two different distributions for the innovation process $\{\varepsilon_t\}_{t\in\mathbb{Z}}$ given as (i) $N(0,1)$ and (ii) $(1-\zeta)N(0,1) + \zeta N(0,10)$. Here the level of contamination has been taken as $\zeta = 0.05$. The sample sizes considered were 300 and 1000. For each combination of error distribution and sample size, to start with, the simulated datasets, y_t and σ_t^2, from ARCH(2) process were generated and then $R = 1000$ future values, y_{n+h} and σ_{n+h}^2, for each future lead $h = 1,\ 2, \ldots,\ 20$ were generated from the underlying model using the true values of the parameter coefficients for each simulation. Furthermore, for each Bootstrap procedure (both existing and proposed), $B = 1000$ Bootstrap pseudo-series were generated to obtain Bootstrap PIs for nominal coverages of 95%. These procedures were repeated $N = 1000$ times to calculate the average values of the performance metrics described subsequently.

The empirical or theoretical length of the PIs of y_{t+h} for i^{th} simulation run, $i = 1, 2, \ldots N$, was calculated as $L_{T,y}(i) = \left[y_{n+h}^{(R)}(1-\alpha/2) - y_{n+h}^{(R)}(\alpha/2)\right]$, the difference between $(1-\alpha/2)\,100^{th}$ and $(\alpha/2)\,100^{th}$ percentile point of the empirical distribution of the R future returns. Then mean theoretical length of return is $\overline{L}_{T,y} = N^{-1}\sum_{i=1}^{N} L_{T,y}(i)$. Similarly the mean theoretical length of the PIs of σ_{n+h}^2 is calculated as: $\overline{L}_{T,\sigma^2} = N^{-1}\sum_{i=1}^{N} L_{T,\sigma^2}(i)$, where $L_{T,\sigma^2}(i) = \left[\sigma_{n+h}^{2,(R)}(1-\alpha/2) - \sigma_{n+h}^{2,(R)}(\alpha/2)\right]$, the difference between $(1-\alpha/2)\,100^{th}$ and $(\alpha/2)\,100^{th}$ percentile point of the empirical distribution of the R future volatilities.

The coverage probability (CP) of returns y_{t+h} for i^{th} simulation run is then calculated as the $C_y(i) = R^{-1}\sum_{r=1}^{R} I_{\left[Q^*(\alpha/2)\le y_{n+h}^{(r)}(i)\le Q^*(1-\alpha/2)\right]}$, where $Q^*(\alpha/2)$ is the $(\alpha/2)^{th}$ quantile of the estimated Bootstrap distribution and $y_{n+h}^{(r)}(i)$ is r^{th} future return value, $r = 1,\ 2, \ldots, R$, generated at i^{th} simulation, $i = 1, 2, \ldots N$. Similarly, the CP of volatility σ_{n+h}^2 for i^{th} simulation $\sigma_{n+h}^{2,(r)}(i)$ is then calculated as the $C_{\sigma^2}(i) = R^{-1}\sum_{r=1}^{R} I_{\left[0\le\sigma_{n+h}^{2,(r)}(i)\le K_{n+h}^*(1-\alpha)\right]}$, where $K^*(\alpha)$ is the α^{th} quantile of the estimated Bootstrap distribution and $\sigma_{n+h}^{2,(r)}(i)$ is r^{th} future volatility value, $r = 1,\ 2, \ldots,\ R$, generated at i^{th} simulation.

The Bootstrap length of returns y_{t+h} and volatility σ_{n+h}^2 for i^{th} simulation run is calculated as $L_{B,y}(i) = [Q^*(1-\alpha/2) - Q^*(\alpha/2)]$ and $L_{B,\sigma^2}(i) = K^*(1-\alpha)$, respectively. Finally, the following performance evaluation measures were calculated:

- Mean Return Coverage (CVR_{ret}): $\overline{C}_y = N^{-1}\sum_{i=1}^{N} C_y(i)$
- Mean Volatility Coverage (CVR_{vol}): $\overline{C}_{\sigma^2} = N^{-1}\sum_{i=1}^{N} C_{\sigma^2}(i)$
- Standard Error of CVR_{ret}: $se\left(\overline{C}_y\right) = \left\{[N(N-1)]^{-1}\sum_{i=1}^{N}\left[C_y(i) - \overline{C}_y\right]^2\right\}^{1/2}$

- Standard Error of CVR_{vol}: $se\left(\overline{C}_{\sigma^2}\right) = \left\{[N(N-1)]^{-1}\sum_{i=1}^{N}\left[C_{\sigma^2}(i) - \overline{C}_{\sigma^2}\right]^2\right\}^{1/2}$
- Mean length of Return (LEN_{ret}): $\overline{L}_{B,y} = N^{-1}\sum_{i=1}^{N} L_{B,y}(i)$
- Mean length of Volatility (LEN_{vol}): $\overline{L}_{B,\sigma^2} = N^{-1}\sum_{i=1}^{N} L_{B,\sigma^2}(i)$
- Standard Error of LEN_{ret}: $se\left(\overline{L}_{B,y}\right) = \left\{[N(N-1)]^{-1}\sum_{i=1}^{N}\left[L_{B,y}(i) - \overline{L}_{B,y}\right]^2\right\}^{1/2}$
- Standard Error of LEN_{vol}: $se\left(\overline{L}_{B,\sigma^2}\right) = \left\{[N(N-1)]^{-1}\sum_{i=1}^{N}\left[L_{B,\sigma^2}(i) - \overline{L}_{B,\sigma^2}\right]^2\right\}^{1/2}$

$CQ_{ret} = \left|1 - \left(\overline{L}_{B,y}/\overline{L}_{T,y}\right)\right| + |1 - (CVR_{ret}/CVR_{T,y})|$

$CQ_{vol} = \left|1 - \left(\overline{L}_{B,\sigma^2}/\overline{L}_{T,\sigma^2}\right)\right| + |1 - (CVR_{vol}/CVR_{T,vol})|$

where $CVR_{T,(.)}$ is the $(1-\alpha)\,\%$ nominal coverage. Here, CQ is an index of coverage quality. Therefore the simulation results have been summarized in different tables that contain the mean coverage (CVR), mean length of the intervals (LEN), standard error of mean coverage (SE), and standard error of mean length of the intervals (SE) for different combinations. The performances of the proposed methods were compared to the existing unconditional Sieve Bootstrap (USB) proposed by Chen *et al.* (2011) method and Sieve Bootstrap (SB) by Tresch (2015) for constructing PIs. The proposed approaches are given in Sections 2.3.1 and 2.3.2 respectively.

It is noted that, for the case of $h = 1$, equations (21) and (28) both will have their Bootstrap volatilities as constant and hence the computation of PIs of their one-step-ahead forecast volatilities are not appropriate and hence not given in the following tables.

In Tables 1 through 4, results of the comparisons of PIs for h= 1, 5, 10, 15 and 20 steps ahead of the described methods have been presented for comparison purposes.

Tables 1 and 2 provide the results pertaining to the ARCH(2) model without contaminated innovations. From Tables 1 and 2, it can be seen that all methods have almost similar results in terms of coverage and length of intervals. It can also be seen that the proposed method RSB is performing almost at par with SB when coverage probabilities are compared while lengths of PIs of RUSB are always found to be less than the existing method *i.e.* USB. The same conclusion can be drawn when we compare the proposed method RUSB with the existing method USB. When the lengths of PIs of two proposed methods RSB and RUSB are compared, by and large, RUSB is always better than RSB both for returns and volatilities.

From Tables 3 and 4, a striking feature of the proposed method RUSB which can be seen is that the length of PIs across all forecast horizons for both returns and volatilities have been found to be less as compared to those of the existing methods SB and USB and also of the proposed method RSB when the data is contaminated. The feature of obtaining the order of model by Sieve approximation rather than assumed to be fixed beforehand has yielded better coverage in the case of the proposed method RSB and the existing method SB (in which such a feature is there) as compared to the other two methods *viz.* proposed method RUSB and existing method USB. It can also be seen from Tables 3 and 4 that

the proposed methods were able to tackle the inflation of variances and at the same time maintains the length of PIs.

Another inference that can be drawn when the coverage of volatility are considered is that the proposed methods performed well in case of contaminated data. It can also be seen that the length of PIs for the proposed method RSB is always less than those of the existing methods USB and SB. Even though the lengths of PIs of the proposed method RSB are larger than the RUSB, it can be seen that the coverages obtained from RSB are always better than RUSB for both returns and volatilities in the case of contaminated data. When both coverages and lengths of PIs are considered together, as per the combined measure CQ_{ret} and CQ_{vol}, the RUSB has been found far better than others.

4. Case Study

In this section, the performance of the proposed methods RSB and RUSB in comparison with the existing methods USB and SB have been presented with real-time series data. Monthly onion price (Rs/quintal) data at Delhi market has been used for validating the methods. It pertains to the period January 2003 to February 2022, with a total of 230 observations. Data were collected from the secondary source available at National Horticultural Research and Development Foundation, New Delhi, India (NHRDF, 2003-2022). The methods were applied to the return of monthly Onion price at Delhi market data. The returns are more frequently used than the price time series, because returns do not depend on units, making the comparison easier. The return series is obtained as follows:

$$y_t = \frac{P_t - P_{t-1}}{P_{t-1}} \tag{33}$$

where P_t is the monthly onion price at time t. The price series is shown in Figure 2 and the return series y_t is shown in Figure 3. ADF test has been employed on the return series y_t which revealed that it is stationary. From Figure 3 and Figure 4, the data reveals the presence of outliers. Table 5 presents the summary statistics of the return data series. As the estimated kurtosis is higher than 3, indicates that the return series is leptokurtic.

Now Lagrange-Multiplier (LM) test confirmed the presence of the ARCH effect on this return series. It was found that ARCH(1) is a suitable model for return series y_t. The data set has been partitioned into an in-sample estimation set from January 2003 to December 2020 and an out-sample set from January 2021 to February 2022 for validation. That is out of 230 sample observations 14 observations have been set aside for predictions purpose. From equation (9), by fitting AR(1) model on y_t^2 using LS estimation, the resulting estimated model is

$$y_t^2 = \ 0.0670 + 0.1026 y_{t-1}^2 \tag{34}$$

i.e. $\widehat{\alpha}_0 = 0.0670$ and $\widehat{\alpha}_1 = 0.1026$. It can also be seen that $\widehat{\alpha}_1^2 \leq 1/3$, and hence indicates strictly stationarity (and hence weakly stationarity also) of return series (Tsay, 2010; Box *et al.*, 2015).

Figures 5 and 6 pertain to the PIs for the returns and volatilities from thevarious methods. In case of PIs of returns, lower and upper boundaries of the PIs were obtained,

Table 1: Simulated results of ARCH(2) model for sample size 300 and standard normal innovation and no contamination

h	Method	$\boldsymbol{CVR_{ret}}$ **(SE)**	$\boldsymbol{LEN_{ret}}$ (SE)	$\boldsymbol{CQ_{ret}}$	$\boldsymbol{CVR_{vol}}$ (SE)	$\boldsymbol{LEN_{vol}}$ (SE)	$\boldsymbol{CQ_{vol}}$
1	-	95%	1.501	-	95%	-	-
	USB	0.9481 (0.0014)	1.514 (0.0065)	0.0098	-	-	-
	SB	0.9461 (0.0012)	1.511 (0.0051)	0.0098	-	-	-
	RSB	0.9462 (0.0012)	1.512 (0.0049)	0.0101	-	-	-
	RUSB	0.9480 (0.0014)	1.514 (0.0065)	0.0096	-	-	-
5	-	95%	1.535	-	95%	0.274	-
	USB	0.9465 (0.0009)	1.542 (0.0063)	0.0082	0.9162 (0.0127)	0.273 (0.0040)	0.0418
	SB	0.9468 (0.0006)	1.543 (0.0046)	0.0086	0.9021 (0.0142)	0.273 (0.0032)	0.0563
	RSB	0.9467 (0.0006)	1.541 (0.0043)	0.0073	0.9083 (0.0098)	0.272 (0.0027)	0.0519
	RUSB	0.9462 (0.0009)	1.539 (0.0061)	0.0069	0.9153 (0.0127)	0.270 (0.0036)	0.0529
10	-	95%	1.539	-	95%	0.273	-
	USB	0.9465 (0.0008)	1.542 (0.0063)	0.0057	0.9156 (0.0127)	0.273 (0.0040)	0.0373
	SB	0.9462 (0.0006)	1.542 (0.0047)	0.0058	0.9015 (0.0142)	0.275 (0.0034)	0.0576
	RSB	0.9463 (0.0006)	1.540 (0.0044)	0.0046	0.9078 (0.0098)	0.274 (0.0028)	0.0463
	RUSB	0.9463 (0.0008)	1.540 (0.0061)	0.0044	0.9145 (0.0127)	0.270 (0.0036)	0.0498
15	-	95%	1.535	-	95%	0.274	-
	USB	0.9458 (0.0008)	1.541 (0.0062)	0.0082	0.9157 (0.0127)	0.274 (0.0041)	0.0368
	SB	0.9468 (0.0006)	1.540 (0.0045)	0.0065	0.9023 (0.0142)	0.275 (0.0033)	0.0520
	RSB	0.9468 (0.0006)	1.538 (0.0042)	0.0053	0.9082 (0.0098)	0.273 (0.0021)	0.0473
	RUSB	0.9456 (0.0008)	1.539 (0.0059)	0.0068	0.9147 (0.0127)	0.270 (0.0036)	0.0521
20	-	95%	1.535	-	95%	0.2740	-
	USB	0.9471 (0.0009)	1.543 (0.0062)	0.0089	0.9163 (0.0127)	0.274 (0.0040)	0.0358
	SB	0.9472 (0.0006)	1.546 (0.0047)	0.0106	0.9021 (0.0142)	0.275 (0.0035)	0.0537
	RSB	0.9472 (0.0006)	1.544 (0.0044)	0.0093	0.9080 (0.0098)	0.273 (0.0027)	0.0475
	RUSB	0.9468 (0.0009)	1.541 (0.0059)	0.0075	0.9153 (0.0127)	0.270 (0.0035)	0.0497

but since the volatility is non-negative, only the upper boundary has been obtained and the lower boundary has been assumed to be zero. It can be found that the PIs for returns developed by all methods contained all the future returns. At some points, it is clearly visible that the proposed methods have smaller lengths as compared to existing methods. As volatilities are not directly observable, once the parameters were estimated, volatilities have been estimated using the following equation (Ullao *et al.*, 2014):

$$\sigma_t^2 = \widehat{\alpha}_0 + \widehat{\alpha}_1 y_{t-1}^2 \tag{35}$$

where y_{t-1} corresponds to the observed past return series. It can be clearly seen that the

Table 2: Simulated results of ARCH(2) model for sample size 1000 and standard normal innovation and no contamination

h	Method	$\boldsymbol{CVR_{ret}}$ **(SE)**	$\boldsymbol{LEN_{ret}}$ (SE)	$\boldsymbol{CQ_{ret}}$	$\boldsymbol{CVR_{vol}}$ (SE)	$\boldsymbol{LEN_{vol}}$ (SE)	$\boldsymbol{CQ_{vol}}$
1	-	95%	1.513	-	95%	-	-
	USB	0.9479 (0.0015)	1.522 (0.0053)	0.0080	-	-	-
	SB	0.9471 (0.0011)	1.519 (0.0038)	0.0068	-	-	-
	RSB	0.9473 (0.0011)	1.519 (0.0038)	0.0070	-	-	-
	RUSB	0.9478 (0.0015)	1.522 (0.0052)	0.0081	-	-	-
5	-	95%	1.538	-	95%	0.275	-
	USB	0.9475 ()0.0006	1.539 (0.0038)	0.0031	0.9381 (0.0144)	0.275 (0.0023)	0.0125
	SB	0.9482 (0.0004)	1.540 (0.0027)	0.0031	0.9374 (0.0136)	0.274 (0.0017)	0.0144
	RSB	0.9481 (0.0004)	1.539 (0.0027)	0.0025	0.9371 (0.0086)	0.272 (0.0015)	0.0216
	RUSB	0.9472 (0.0006)	1.537 (0.0037)	0.0038	0.9369 (0.0144)	0.271 (0.0021)	0.0251
10	-	95%	1.537	-	95%	0.274	-
	USB	0.9471 (0.0006)	1.537 (0.0039)	0.0031	0.9385 (0.0144)	0.275 (0.0023)	0.0183
	SB	0.9483 (0.0004)	1.539 (0.0027)	0.0030	0.9375 (0.0136)	0.275 (0.0017)	0.0183
	RSB	0.9482 (0.0004)	1.538 (0.0026)	0.0024	0.9373 (0.0086)	0.273 (0.0015)	0.0170
	RUSB	0.9469 (0.0006)	1.535 (0.0038)	0.0046	0.9374 (0.0144)	0.272 (0.0020)	0.0184
15	-	95%	1.537	-	95%	0.274	-
	USB	0.9483 (0.0006)	1.538 (0.0038)	0.0024	0.9387 (0.0144)	0.275 (0.0023)	0.0166
	SB	0.9482 (0.0004)	1.540 (0.0026)	0.0040	0.9374 (0.0136)	0.274 (0.0016)	0.0155
	RSB	0.9480 (0.0004)	1.538 (0.0026)	0.0031	0.9370 (0.0086)	0.272 (0.0015)	0.0195
	RUSB	0.9481 (0.0006)	1.536 (0.0037)	0.0027	0.9374 (0.0144)	0.272 (0.0020)	0.0202
20	-	95%	1.539	-	95%	0.274	-
	USB	0.9473 (0.0006)	1.535 (0.0039)	0.0054	0.9384 (0.0144)	0.275 (0.0022)	0.0162
	SB	0.9475 (0.0004)	1.536 (0.0026)	0.0049	0.9369 (0.0136)	0.274 (0.0016)	0.0167
	RSB	0.9474 (0.0004)	1.535 (0.0025)	0.0054	0.9367 (0.0086)	0.272 (0.0015)	0.0191
	RUSB	0.9471 (0.0006)	1.534 (0.0038)	0.0068	0.9372 (0.0144)	0.272 (0.0020)	0.0190

proposed method RSB and SB are almost close to each other and cover all future volatilities. RUSB has a very small length for PIs.

5. Conclusion

In this study, two new robust Sieve Bootstrap approaches based on weighted least squares estimation have been proposed to deal with the presence of outliers for developing PIs for both returns and volatilities in the ARCH model setup. The performances of the proposed methods *viz.*, Robust Unconditional Sieve Bootstrap (RUSB) and Robust Sieve Bootstrap (RSB) for constructing PIs using both simulated as well as real data sets have been found

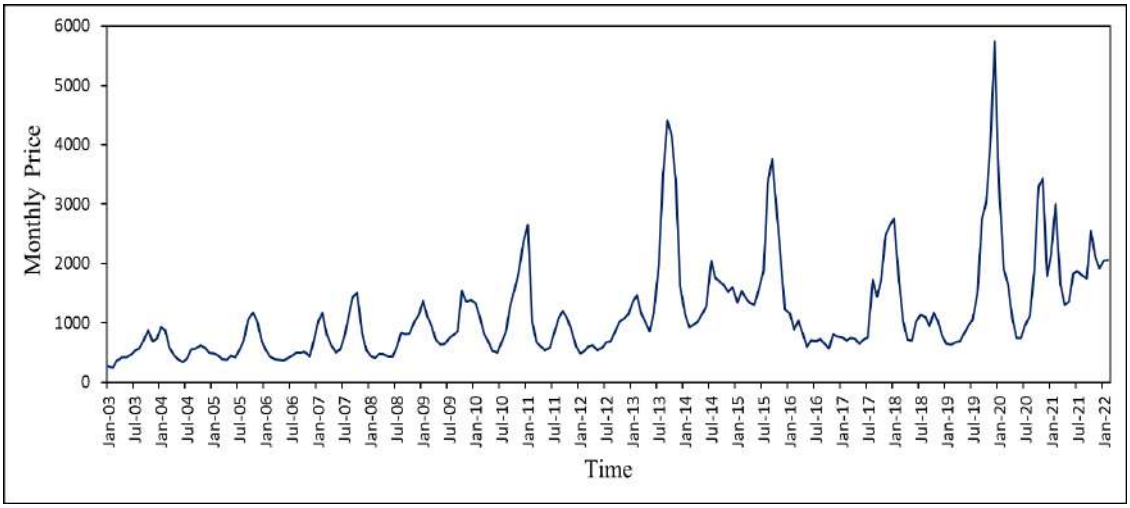


Figure 2: Monthly Onion price at Delhi market data from January 2003 to February 2022, with a total of 230 observations

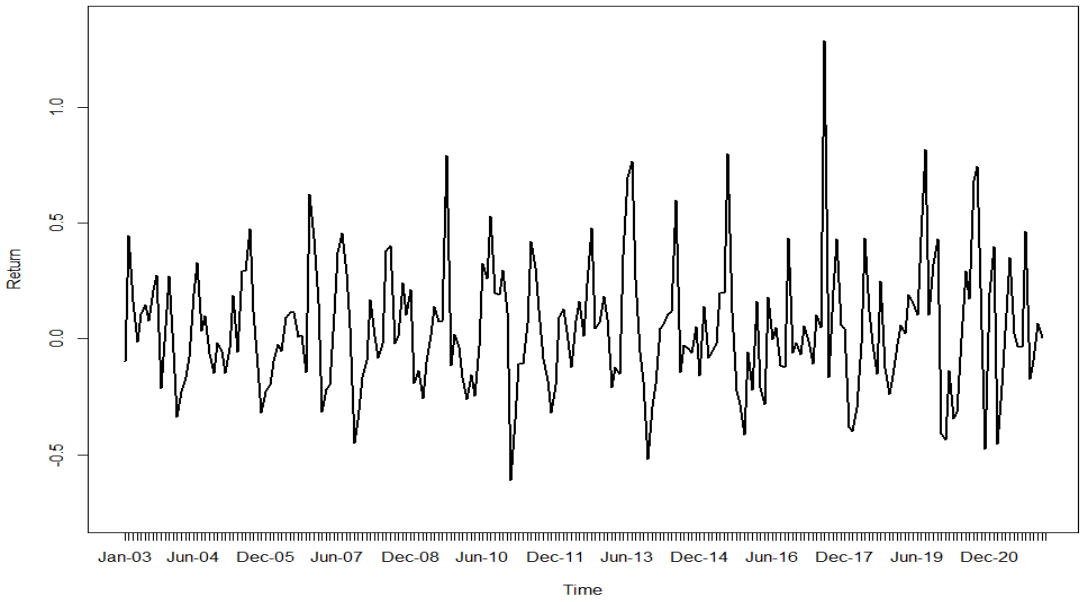


Figure 3: Time plot of returns of monthly onion price data of Delhi market

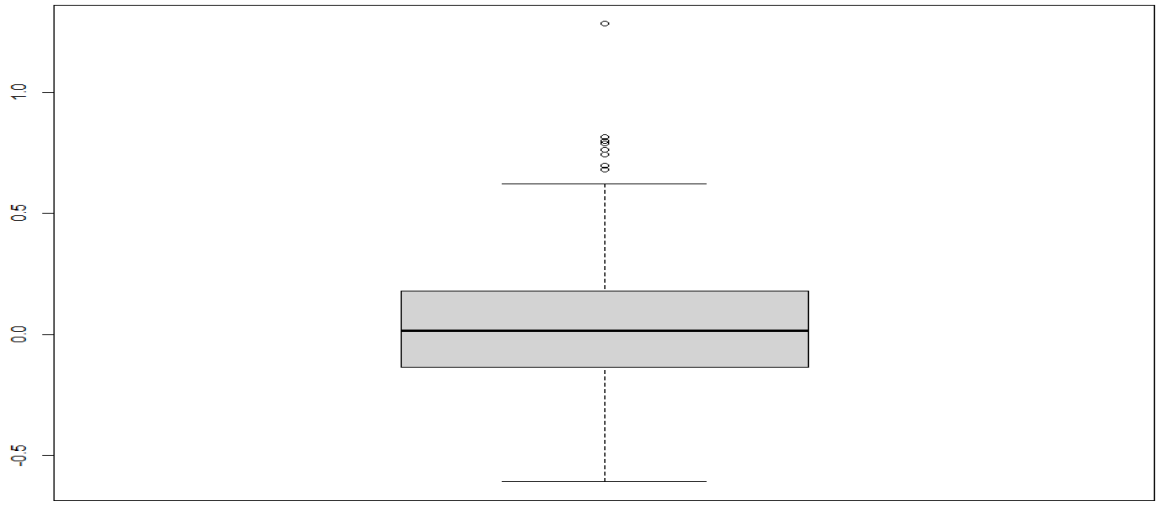


Figure 4: Box plot of return series of monthly onion price of Delhi market

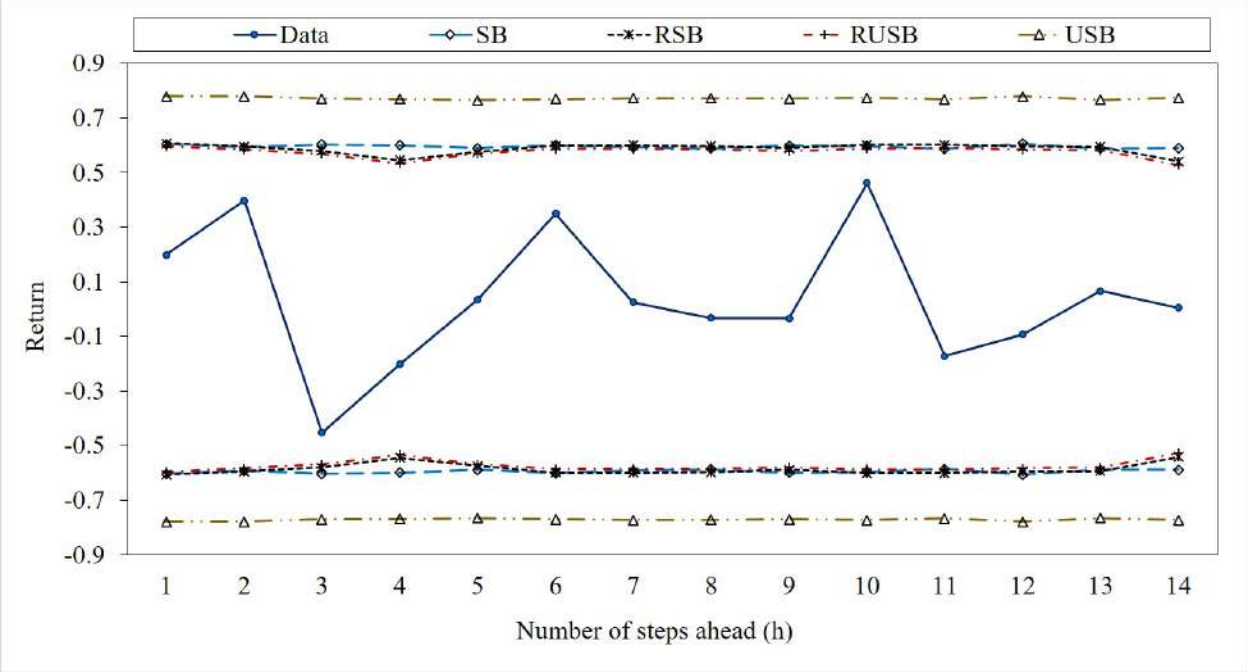


Figure 5: Prediction intervals for returns of monthly onion price of Delhi market for forecast horizons $h = 1, 2, \cdots, 14$

Table 3: Simulated results of ARCH(2) model for sample size 300 with 5% contaminated normal innovation

h	Method	$\boldsymbol{CVR_{ret}}$ **(SE)**	$\boldsymbol{LEN_{ret}}$ (SE)	$\boldsymbol{CQ_{ret}}$	$\boldsymbol{CVR_{vol}}$ (SE)	$\boldsymbol{LEN_{vol}}$ (SE)	$\boldsymbol{CQ_{vol}}$
1	-	95%	1.820	-	95%	-	-
	USB	0.9607 (0.0017)	2.151 (0.0198)	0.1935	-	-	-
	SB	0.9480 (0.0015)	1.906 (0.0102)	0.0493	-	-	-
	RSB	0.9491 (0.0016)	1.941 (0.0135)	0.0675	-	-	-
	RUSB	0.9423 (0.0022)	1.839 (0.0126)	0.0186	-	-	-
5	-	95%	1.955	-	95%	0.433	-
	USB	0.9626 (0.0009)	2.446 (0.0346)	0.2646	0.9749 (0.0030)	1.138 (0.0641)	1.6523
	SB	0.9518 (0.0007)	2.117 (0.0151)	0.0846	0.9441 (0.0069)	0.581 (0.0134)	0.3459
	RSB	0.9500 (0.0007)	2.072 (0.0162)	0.0597	0.9465 (0.0056)	0.576 (0.0291)	0.3335
	RUSB	0.9440 (0.0009)	1.949 (0.0132)	0.0096	0.9363 (0.0078)	0.436 (0.0084)	0.0200
10	-	95%	1.960	-	95%	0.429	-
	USB	0.9636 (0.0009)	2.490 (0.0380)	0.2849	0.9741 (0.0030)	1.216 (0.0788)	1.8600
	SB	0.9514 (0.0007)	2.126 (0.0163)	0.0864	0.9433 (0.0069)	0.601 (0.0162)	0.4077
	RSB	0.9496 (0.0006)	2.081 (0.0188)	0.0622	0.9455 (0.0056)	0.601 (0.0399)	0.4066
	RUSB	0.9447 (0.0009)	1.970 (0.0150)	0.0106	0.9348 (0.0078)	0.443 (0.0094)	0.0479
15	-	95%	1.965	-	95%	0.432	-
	USB	0.9626 (0.0009)	2.484 (0.0392)	0.2776	0.9746 (0.0030)	1.234 (0.0861)	1.8824
	SB	0.9512 (0.0007)	2.126 (0.0159)	0.0832	0.9430 (0.0069)	0.605 (0.0164)	0.4078
	RSB	0.9491 (0.0007)	2.083 (0.0210)	0.0614	0.9451 (0.0056)	0.605 (0.0496)	0.4045
	RUSB	0.9438 (0.0009)	1.953 (0.0140)	0.0125	0.9354 (0.0078)	0.440 (0.0089)	0.0341
20	-	95%	1.966	-	95%	0.432	-
	USB	0.9628 (0.0009)	2.488 (0.0395)	0.2793	0.9738 (0.0030)	1.241 (0.0885)	1.9020
	SB	0.9511 (0.0007)	2.132 (0.0164)	0.0859	0.9431 (0.0069)	0.609 (0.0179)	0.4193
	RSB	0.9491 (0.0007)	2.090 (0.0216)	0.0643	0.9450 (0.0056)	0.638 (0.0579)	0.4838
	RUSB	0.9444 (0.0009)	1.957 (0.0141)	0.0104	0.9349 (0.0078)	0.438 (0.0087)	0.0307

to be better when compared with their existing counterparts. The results revealed that the proposed method RSB is performing almost at par with SB when coverage probabilities are compared while lengths of PIs of RUSB are always found to be less than the existing method *i.e.* USB. When the lengths of PIs of two proposed methods RSB and RUSB are compared, by and large, RUSB is always better than RSB both for returns and volatilities. For the proposed method RUSB, the length of PIs across all forecast horizons for both returns and volatilities have been found to be less as compared to those of the existing methods SB and USB and also of the proposed method RSB when the data is contaminated. The proposed methods were able to tackle the inflation of variances and at the same time maintain the

Table 4: Simulated results of ARCH(2) model for sample size 1000 with 5% contaminated normal innovation

h	Method	CVR_{ret} (SE)	LEN_{ret} (SE)	CQ_{ret}	CVR_{vol} (SE)	LEN_{vol} (SE)	CQ_{vol}
1	-	95%	1.768	-	95%	-	-
	USB	0.9644 (0.0016)	2.155 (0.0164)	0.2072	-	-	-
	SB	0.9535 (0.0011)	1.875 (0.0078)	0.0644	-	-	-
	RSB	0.9517 (0.0012)	1.836 (0.0068)	0.0403	-	-	-
	RUSB	0.9444 (0.0022)	1.806 (0.0096)	0.0278	-	-	-
5	-	95%	1.948	-	95%	0.422	-
	USB	0.9675 (0.0007)	2.491 (0.0294)	0.2922	0.9806 (0.0035)	1.141 (0.0592)	1.7386
	SB	0.9557 (0.0005)	2.114 (0.0095)	0.0914	0.9620 (0.0045)	0.594 (0.0083)	0.4220
	RSB	0.9511 (0.0005)	2.013 (0.0071)	0.0344	0.9504 (0.0035)	0.474 (0.0057)	0.1242
	RUSB	0.9454 (0.0007)	1.928 (0.0091)	0.0154	0.9350 (0.0073)	0.397 (0.0055)	0.0732
10	-	95%	1.959	-	95%	0.428	-
	USB	0.9678 (0.0006)	2.519 (0.0333)	0.3103	0.9807 (0.0035)	1.220 (0.0717)	1.8841
	SB	0.9555 (0.0005)	2.134 (0.0107)	0.0950	0.9610 (0.0045)	0.616 (0.0101)	0.4520
	RSB	0.9506 (0.0004)	2.020 (0.0074)	0.0317	0.9488 (0.0035)	0.476 (0.0058)	0.1128
	RUSB	0.9451 (0.0007)	1.921 (0.0088)	0.0249	0.9332 (0.0074)	0.396 (0.0058)	0.0913
15	-	95%	1.959	-	95%	0.431	-
	USB	0.9673 (0.0006)	2.528 (0.0342)	0.3037	0.9807 (0.0035)	1.242 (0.0764)	1.9108
	SB	0.9552 (0.0005)	2.131 (0.0109)	0.0929	0.9606 (0.0045)	0.626 (0.0111)	0.4626
	RSB	0.9503 (0.0005)	2.015 (0.0075)	0.0285	0.9484 (0.0035)	0.478 (0.0059)	0.1090
	RUSB	0.9443 (0.0007)	1.924 (0.0089)	0.0240	0.9340 (0.0073)	0.399 (0.0060)	0.0922
20	-	95%	1.955	-	95%	0.433	-
	USB	0.9676 (0.0007)	2.528 (0.0348)	0.3120	0.9801 (0.0035)	1.245 (0.0780)	1.9076
	SB	0.9555 (0.0005)	2.131 (0.0108)	0.0958	0.9603 (0.0045)	0.626 (0.0116)	0.4562
	RSB	0.9506 (0.0004)	2.014 (0.0071)	0.0309	0.9481 (0.0035)	0.476 (0.0060)	0.1016
	RUSB	0.9452 (0.0007)	1.931 (0.0095)	0.0171	0.9333 (0.0074)	0.399 (0.0064)	0.0961

Table 5: Summary statistics of return series y_t

Mean	Median	SD	Skewness	Kurtosis	Maximum	Minimum
0.0437	0.0158	0.2730	0.8846	5.0058	1.2840	-0.6090

length of PIs. Using the real data set on the monthly onion price of Delhi market, it has been shown that the PIs for returns developed by all methods contained all the future returns and that the proposed methods have smaller lengths as compared to existing methods. Hence the proposed methods can be used as a viable alternative for computing PIs for non-linear

TS models.

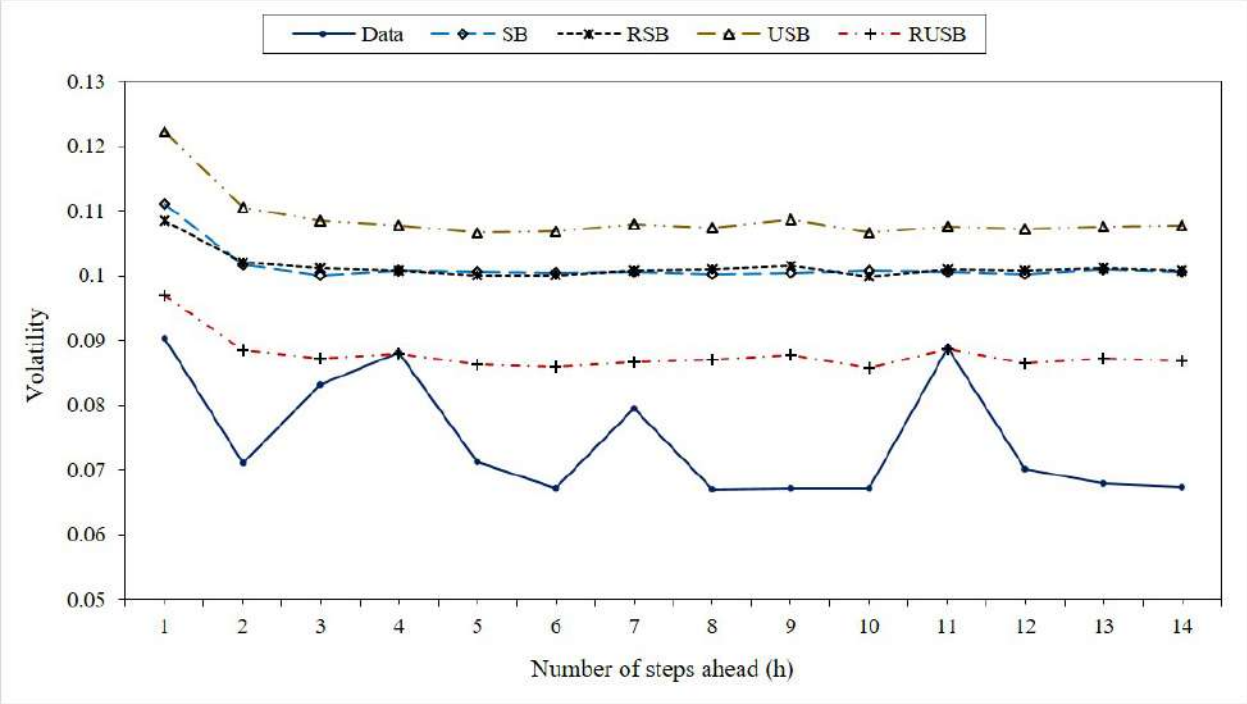


Figure 6: Prediction intervals for volatilities of monthly onion price of Delhi market for forecast horizons $h = 1, 2, \cdots, 14$

Acknowledgements

The facilities provided by ICAR-Indian Agricultural Statistics Research Institute (IASRI), New Delhi and the funding granted to the first author by Indian Council of Agricultural Research in the form of IARI-SRF fellowship is duly acknowledged for carrying out this study, which is part of his doctoral research work being pursued at IASRI. Thanks are also due to the reviewer whose comments have improved the paper to a great extent.

References

Allende, H., Ulloa, G., and Allende-Cid, H. (2015). Prediction intervals in linear and non-linear time series with sieve bootstrap methodology. In: Beran, J., Feng, Y. and Hebbel, H. (Eds.) *Empirical Economic and Financial Research. Advanced Studies in Theoretical and Applied Econometrics*, **48**, Springer, Cham, 255-273. *https* : *//doi.org*/10.1007/978 − 3 − 319 − 03122 − 4_16.

Alonso, A. M., Peña, D., and Romo, J. (2002). Forecasting time series with sieve bootstrap. *Journal of Statistical Planning and Inference*, **100**, 1-11.

Alonso, A. M., Peña, D., and Romo, J. (2003). On sieve bootstrap prediction intervals. *Statistics and Probability Letters*, **65**, 13-20.

Alonso, A. M., Peña, D., and Romo, J. (2004). Introducing model uncertainty in time series bootstrap. *Statistica Sinica*, **14**, 155-174.

Andersen, T. G. and Bollerslev, T. (1998). Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International Economic Review*, **39**, 885-905.

Andersen, T. G., Bollerslev, T., Diebold, F. X., and Labys, P. (2001). The distribution of realized exchange rate volatility. *Journal of the American Statistical Association*, **96**, 42-55.

Baillie, R. T. and Bollerslev, T. (1992). Prediction in dynamic models with time-dependent conditional variances. *Journal of Econometrics*, **52**, 91-113.

Bentes, S. R. (2015). A comparative analysis of the predictive power of implied volatility indices and GARCH forecasted volatility. *Physica A: Statistical Mechanics and its Applications*, **424**, 105-112.

Beyaztas, U. and Shang, H. L. (2020). Robust bootstrap prediction intervals for univariate and multivariate autoregressive time series models. *Journal of Applied Statistics*, 1-24. DOI: 10.1080/02664763.2020.1856351.

Beyaztas, U. and Shang, H. L. (2021). A robust partial least squares approach for function-on-function regression. *arXiv preprint arXiv:2111.01238*.

Bhansali, R. J. (1983). A simulation study of autoregressive and window estimators of the inverse correlation function. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **32**, 141-149.

Bhardwaj, S. P., Paul, R. K., Singh, D. R., and Singh, K. N. (2014). An empirical investigation of ARIMA and GARCH models in agricultural price forecasting. *Economic Affairs*, **59**, 415-428.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, **31**, 307-327.

Bose, A. and Mukherjee, K. (2009). Bootstrapping a weighted linear estimator of the arch parameters. *Journal of Time Series Analysis*, **30**, 315-331.

Bougerol, P. and Picard, N. (1992). Strict stationarity of generalized autoregressive processes. *The Annals of Probability*, **20**, 1714-1730.

Box, G. E., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. (2015). *Time Series Analysis: Forecasting and Control*, John Wiley & Sons.

Bühlmann, P. (1997). Sieve bootstrap for time series. *Bernoulli*, **3**, 123-148.

Bühlmann, P. (2002). Bootstraps for time series. *Statistical Science*, **17**, 52-72.

Chatfield, C. (2000). *Time Series Forecasting*, Chapman and Hall, USA.

Chen, B., Gel, Y. R., Balakrishna, N., and Abraham, B. (2011). Computationally efficient bootstrap prediction intervals for returns and volatilities in ARCH and GARCH processes. *Journal of Forecasting*, **30**, 51-71.

Dyhrberg, A. H. (2016). Bitcoin, gold and the dollar–A GARCH volatility analysis. *Finance Research Letters*, **16**, 85-92.

Engle R. F. (1982). Autoregressive conditional heteroskedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, **50**, 987–1007.

Hannan, E. J. and Rissanen, J. (1982). Recursive estimation of mixed autoregressive-moving average order. *Biometrika*, **69**, 81-94.

Iqbal, F. (2011). A weighted linear estimator of multivariate ARCH parameters. *Communications in Statistics-Simulation and Computation*, **40**, 544-560.

Iqbal, F. and Chand, S. (2013). Efficient bootstrap forecast intervals for return and volatility using the linear estimator of ARCH models. *Middle-East Journal of Scientific Research*, **14**, 1502-1507.

Kreiss, J. P. and Franke, J. (1992). Bootstrapping stationary autoregressive moving-average models. *Journal of Time Series Analysis*, **13**, 297-317.

Lahiri, S. N. (2003). *Resampling Methods for Dependent Data*, Springer, New York.

Lama, A., Jha, G. K., and Paul, R. K. (2015). Modelling and forecasting of price volatility: An application of GARCH and EGARCH models. *Agricultural Economics Research Review*, **28**, 365-382.

Lindsay, B. G. (1994). Efficiency versus robustness: the case for minimum Hellinger distance and related methods. *The Annals of Statistics*, **22**, 1081-1114.

Markatou, M., Basu, A., and Lindsay, B. G. (1998). Weighted likelihood equations with bootstrap root search. *Journal of the American Statistical Association*, **93**, 740-750.

Miguel, J. A. and Olave, P. (1999). Bootstrapping forecast intervals in ARCH models. *Test*, **8**, 345-364.

Mukhopadhyay, P. and Samaranayake, V. A. (2010). Prediction intervals for time series: a modified sieve bootstrap approach. *Communications in Statistics-Simulation and Computation*, **39**, 517-538.

NHRDF (2003-2022). National Horticultural Research and Development Foundation, *https* : $//nhrdf.org/en-us/MonthWiseMarketArrivals$, accessed on March 20, 2022.

Pan, L. and Politis, D. N. (2016). Bootstrap prediction intervals for linear, nonlinear and nonparametric autoregressions. *Journal of Statistical Planning and Inference*, **177**, 1-27.

Pascual, L., Romo, J., and Ruiz, E. (2006). Bootstrap prediction for returns and volatilities in GARCH models. *Computational Statistics and Data Analysis*, **50**, 2293-2312.

Politis, D. N., Romano, J. P., and Wolf, M. (1999). *Subsampling*. Springer, New York.

Politis, D. N. (2003). The impact of bootstrap methods on time series analysis. *Statistical Science*, **28**, 219-230.

Poon S. H. (2005). *A Practical Guide to Forecasting Financial Market Volatility*. Wiley, Chichester.

Reeves, J. J. (2005). Bootstrap prediction intervals for ARCH models. *International Journal of Forecasting*, **21**, 237-248.

Thombs, L. A. and Schucany, W. R. (1990). Bootstrap prediction intervals for autoregression. *Journal of the American Statistical Association*, **85**, 486-492.

Tresch, G. D. (2015). *Sieve Bootstrap-Based Prediction Intervals for GARCH Processes*, Ph.D. Thesis, Ashland University, Ohio, USA.

Trucíos, C., Hotta, L. K., and Ruiz, E. (2017). Robust bootstrap forecast densities for GARCH returns and volatilities. *Journal of Statistical Computation and Simulation*, **87**, 3152-3174.

Tsay, R. S. (1988). Outliers, level shifts, and variance changes in time series. *Journal of Forecasting*, **7**, 1-20.

Tsay R. S. (2010). *Analysis of Financial Time Series*. Wiley-Interscience, New York.

Ulloa, G., Allende-Cid, H., and Allende, H. (2014). Robust sieve bootstrap prediction intervals for contaminated time series. *International Journal of Pattern Recognition and Artificial Intelligence*, **28**, 1460012.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 21, No. 2, 2023 (New Series), pp 59–71
http://www.ssca.org.in/journal

Estimation of AUC of Bi-Generalized Half-Normal ROC Curve

Dashina P. and R. Vishnu Vardhan
Department of Statistics, Pondicherry University, Puducherry, India

Received: 25 July 2022; Revised: 11 October 2022; Accepted: 15 October 2022

Abstract

In ROC literature, there are good number of Bi-distributional ROC curves which are developed to address the practical need and are based on normal and non-normal data. The most widely used ROC form is the Bi-Normal. However, the practical situations in diagnostic medicine and other life testing frameworks, data may not be attributed to make use of the Bi-Normal ROC curve. We have considered such situations using SAPS III dataset, where the data underpins Generalised Half-Normal distribution and not that of any existing bi-distributional ROC forms. The ROC and AUC expressions are derived and these are supported with SAPS III dataset and simulation. The present work is demonstrated by considering minimum (better case), moderate (moderate case) and maximum (worst case) overlapping scenarios at various sample sizes.

Key words: ROC curve; AUC; Non-normal data; Confidence intervals; Generalized Half-Normal distribution.

AMS Subject Classifications: 92B15, 62P10

1. Introduction

In classical statistics and machine learning, the problem of classifying an individual/ object/ image/ voice/ signal has grabbed the attention of researchers from diagnostic medicine, experimental psychology, finance and many more. The statistical tool that supports in explaining the performance of a classifier is the receiver operating characteristic (ROC) curve. Even though the tool originated in early 1950s to analyze the radar signals, researchers from the medical domain started using it in the early 1970s. The theoretical contributions started during mid 1970s wherein the mathematical frame work was proposed by by assuming the data of two populations follow a particular distribution, say 'normal'; hence the name '*binormal ROC model*' Egan (1975). However, basing on the practical need and situations, the theoretical development happened under non-normal data structures. Over the years, many researchers have attempted in proposing the bi-distributional ROC models by considering gamma (Hussain (2012)), logistic (Dorfman and Alf (1969)), half-normal (Vishnu and Kiruthika (2015)), exponential, and Weibull (Vishnu *et al.* (2012)) *etc.*

Corresponding Author: R. Vishnu Vardhan
Email: vrstatsguru@gmail.com

A comprehensive coverage of such bi-distributional ROC models was made by Balaswamy and Vishnu (2016). In understanding the non-normal data, we are well aware that the shape and scale parameters play a crucial role in explaining the tail pattern and asymmetry.

Table 1: One sample KS test for some skewed distributions

Distribution	Status	Parameters	Estimates	KS test value	p-value
Normal	Alive	μ_0	25.53	0.9999	<2.2e-16
		σ_0	17.48		
	Dead	μ_1	33.82	0.9565	<4.44e-16
		σ_1	17.42		
Exponential	Alive	λ_0	0.04	0.1639	0.0575
	Dead	λ_1	0.03	0.2543	0.0059
GHN	Alive	α_0	1.18	0.1141	0.3563
		σ_0	32.66		
	Dead	α_1	1.21	0.1341	0.3936
		σ_1	42.04		

Let us consider a real data namely the Simplified Acute Physiology Score (SAPS) III, which helps in estimating the probability of mortality for ICU patients/subjects. SAPS III score and a status variable (Alive(0); Dead(1)) are the two characteristics recorded for each patient. Figure 1 depicts the density patterns of 'alive' and 'dead' patients indicating the deviation from symmetry. Further, goodness of fit criterion using the one-sample

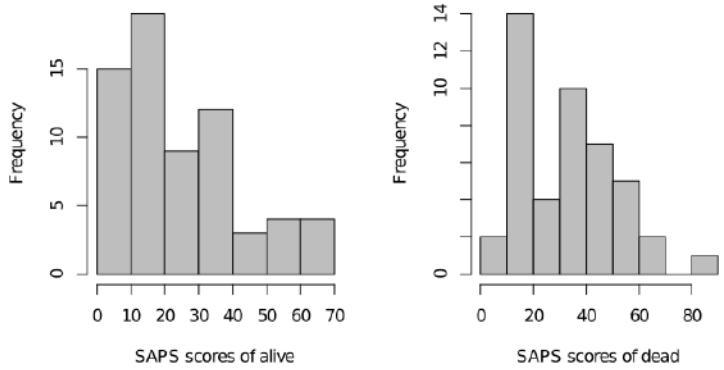


Figure 1: Histogram of SAPS III data

Kolmogorov-Smirnov (KS) test is performed to provide an evidence that the SAPS III data do not follow normality. Along with the normal distribution, exponential and generalised half-normal distribution (GHN) were also considered as competitor distributions. The results of the same are reported in Table 1, clearly indicating that the data is a good fit for GHN distribution. So, the existing bi-normal and bi-exponential ROC models do not support in defining a classifier that helps in classifier or allocating a subject into 'alive' or 'dead' classes of SAPS III data. Hence, the practical situation needs a classifier rule to be defined.

This motivated us to come out with a newer version of ROC model wherein the data of two populations follow GHN distribution.

GHN is a special case of the three-parameter generalized gamma distribution. Even though the GHN distribution is a two-parameter distribution, the hazard rate function can form variety of shapes such as monotonically increasing, monotonically decreasing, and bath-tub shapes. Cooray and Ananda (2008) studied some properties of this family and examples are cited to compare with other commonly used failure time distributions such as Weibull, gamma, lognormal, and Birnbaum-Saunders. Moreover, there is difficulty in developing inference procedures with the generalized gamma distribution, particularly, the maximum likelihood estimation in which the iteration method such as Newton-Raphson fails. Even with samples of size 200 or 300, the algorithms do not converge (Hager and Bain (1970)). Some authors such as Parr and Webster (1965) and Stacy and Mihram (1965) faced problems with the maximum likelihood estimation. In addition, for interval estimation procedures also they faced difficulties. This prompted us to work on GHN with two parameters such as shape and scale and illustrated the features of parameters involved in it with the help of a real data called SAPS III. Simulation studies are also carried out to support the proposed methodology.

The probability density function and cumulative distribution function of GHN distribution are,

$$f(x) = \begin{cases} \sqrt{\frac{2}{\pi}} \left(\frac{\alpha}{x}\right) \left(\frac{x}{\sigma}\right)^{\alpha} exp\left(\frac{-1}{2}\left(\frac{x}{\sigma}\right)^{2\alpha}\right) & if\ x \geq 0 \\ 0 & if\ x < 0 \end{cases} \tag{1}$$

$$F(x) = 2\Phi\left[\left(\frac{x}{\theta}\right)^{\alpha}\right] \qquad x \geq 0,\ \theta, \alpha > 0 \tag{2}$$

where $\Phi(.)$ is the cumulative distribution function of standard normal deviate, α and σ are shape and scale parameters respectively. The expression given in (2) resembles the cumulative distribution function of the half-normal distribution, hence Cooray and Ananda (2008) named this distribution as GHN distribution. The density curves of GHN for different values of shape and scale parameters are shown in Figure 2. For fixed scale parameter, the GHN distribution will be positively skewed if $\alpha \in (0, 2.17)$; symmetric if $\alpha = 0$ and negatively skewed if $\alpha > 2.17$.

2. The Bi-generalised Half-Normal ROC curve

Let us assume that the scores or data points, say S={X,Y} in both populations 1 and 2 follow GHN distribution. Using the probabilistic definitions, the false positive rate (FPR) and true positive rate (TPR) of ROC curve at threshold 't' are given as

$$FPR = P(S > t/0) = 1 - \left[2\left(\Phi\left[\frac{t}{\sigma_0}\right]^{\alpha_0}\right) - 1\right] = 2\left[1 - \Phi\left(\frac{t}{\sigma_0}\right)^{\alpha_0}\right] \tag{3}$$

$$TPR = P(S > t/1) = 1 - \left[2\left(\Phi\left[\frac{t}{\sigma_1}\right]^{\alpha_1}\right) - 1\right] = 2\left[1 - \Phi\left(\frac{t}{\sigma_1}\right)^{\alpha_1}\right] \tag{4}$$

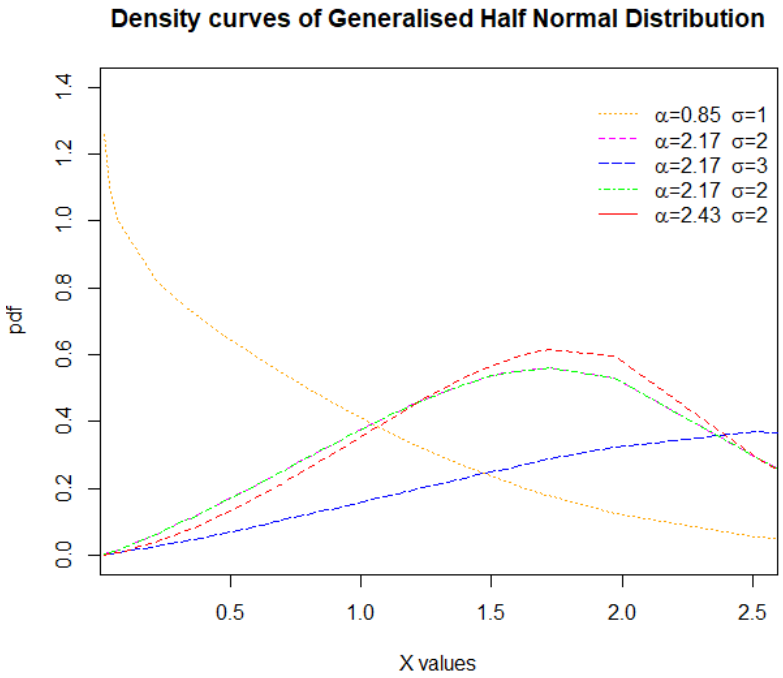


Figure 2: Density curves of GHN distribution

then from equation (3), the threshold can be expressed as,

$$t = \sigma_0 \left(\Phi^{-1} \left[1 - \frac{FPR}{2} \right]^{\frac{1}{\alpha_0}} \right) \tag{5}$$

The ROC expression given in equation (6) is the Bi-Generalised Half-Normal (Bi-GHN) ROC curve, where $\Phi^{-1}(.)$ is the inverse cumulative distribution function of standard normal deviate. Using equation (5) in equation (4), the ROC model is obtained and is given in equation (6).

$$ROC(t) = 2 \left[1 - \Phi \left(\frac{\sigma_H \left(\Phi^{-1} \left[1 - \frac{FPR}{2} \right]^{\frac{1}{\alpha_H}} \right)}{\sigma_D} \right)^{\alpha_D} \right] \tag{6}$$

where σ_0 and σ_1 are the scale parameters and α_0 and α_1 are the shape parameters of the '0' and '1' populations respectively. In next section, the expressions for the area under the curve (AUC) and Youden's index are given.

3. AUC of Bi-GHN ROC curve

The AUC can be interpreted as the average TPRs at all possible TNRs (TNR is the True Negative Rate, which is obtained from 1-FPR). Since ROC curve is only a graphical representation of a classifier it will be always better if we can summarize our findings by a single measure. Such a numerical summary measure of ROC curve is termed as AUC. AUC of an ROC curve explains the accuracy of a diagnostic test. The ability of the test

to discriminate between '1' and '0' groups can be explained by AUC measure. Higher the AUC value, better will be the discriminating power of the test. The value of AUC always lies between 0 and 1. The total area under the ROC curve is always unity because both TPR and TNR values lie between 0 and 1. The line connecting (0,0) and (1,1) in the ROC unit square plot is the diagonal line where the AUC will be equal to 0.5. A test for which AUC < 0.5 need not be considered at all. It means that the test has only 50 percentage or less chance of discriminating the subjects into '1' and '0' categories. Tests with AUC ≥ 0.5 will alone be considered for further classification. AUC of Bi-GHN ROC curve is,

$$AUC = \int_0^1 ROC(t)\, dt$$

$$AUC = \int_0^1 2\left[1 - \Phi\left(\frac{\sigma_0\left(\Phi^{-1}\left[1 - \frac{FPR}{2}\right]^{\frac{1}{\alpha_0}}\right)}{\sigma_1}\right)^{\alpha_1}\right] dt \tag{7}$$

If we consider $\sigma_1 = \sigma_0 = 1$, then it will reduce to one parameter Bi-GHN ROC curve and its AUC will take the following form.

Then the AUC of the one-parameter Bi-GHN can be obtained as

$$AUC = \int_0^1 2\left[1 - \left(1 - \frac{FPR}{2}\right)^{\frac{1}{\alpha_0}}\right]^{\alpha_1} dt$$

$$AUC = \frac{2^{(1-k)}\left[2^k(k-1)+1\right]}{k+1} \quad ; \quad \text{where} \quad k = \frac{\alpha_1}{\alpha_0} \tag{8}$$

In this paper we consider two parameter Bi-GHN distribution. Since, equation (7) does not have a closed form, we need to solve it using numerical integration. Variance of AUC can be obtained using bootstrap method which is described in following section. Another important summary measure of the ROC curve is Youden's index (J). The maximum value of 'J' is the value corresponding to the optimal threshold (cut-off) for the marker in the diagnostic test. The theoretical expression for Youden's index is

$$J = max\{TPR + TNR - 1\}$$

4. Parameter estimation under maximum likelihood method and their confidence intervals

Using the results of maximum likelihood estimates presented in the work of Cooray and Ananda (2008), the expressions for '0' and '1' populations are given in equations (9) and (10) respectively.

$$\frac{n_0}{\widehat{\alpha}_0} + \sum_{i=1}^{n_0} log(x_i) - n_0\left(\sum_{i=1}^{n_0} x_i^{2\widehat{\alpha}_0} log(x_i)\right)\left(\sum_{i=1}^{n_0} x_i^{2\widehat{\alpha}_0}\right)^{-1} \tag{9}$$

$$\frac{n_1}{\widehat{\alpha}_1} + \sum_{j=1}^{n_1} log(y_j) - n_1 \left(\sum_{j=1}^{n_1} y_j^{2\widehat{\alpha}_1} log(y_j) \right) \left(\sum_{j=1}^{n_1} y_j^{2\widehat{\alpha}_1} \right)^{-1} \tag{10}$$

since $\hat{\alpha}_0$ and $\hat{\alpha}_1$ are fixed point solutions of the above non-linear equations, it can be obtained by using a simple iterative scheme as follows: $h(\alpha_{(j)}) = \alpha_{(j+1)}$ where $\lambda_{(j)}$ is the j^{th} iterate of $\hat{\alpha}$. The iteration procedure should be stopped when α_j less than α_{j+1} is sufficiently small. Once we obtain $\widehat{\alpha}_0$ and $\widehat{\alpha}_1$, we can obtain $\widehat{\sigma}_0$ and $\widehat{\sigma}_1$ from below expressions.

$$\widehat{\sigma}_0 = \left(\frac{1}{n_0} \sum_{i=1}^{n_0} x_i^{2\widehat{\alpha}_0} \right)^{\frac{1}{2\widehat{\alpha}_0}} \tag{11}$$

$$\widehat{\sigma}_1 = \left(\frac{1}{n_1} \sum_{j=1}^{n_1} y_j^{2\widehat{\alpha}_1} \right)^{\frac{1}{2\widehat{\alpha}_1}} \tag{12}$$

The $(1-\delta)$ confidence interval for $\widehat{\sigma}_0$ and $\widehat{\sigma}_1$ can be written as

$$\widehat{\sigma}_0 \pm Z_{\frac{\delta}{2}} \sqrt{\frac{(\frac{\pi}{2}) - 2 + (2 - log(2) - \gamma)^2}{n_0(\pi^2 - 4)}}$$

and

$$\widehat{\sigma}_1 \pm Z_{(\frac{\delta}{2})} \sqrt{\frac{(\frac{\pi}{2}) - 2 + (2 - log(2) - \gamma)^2}{n_1(\pi^2 - 4)}}$$

where γ is the Euler's constant (=0.5772156649).

The $(1-\delta)$ confidence interval for $\widehat{\alpha}_0$ and $\widehat{\alpha}_1$ can be written as

$$\widehat{\alpha}_0 \pm Z_{(\frac{\delta}{2})} \frac{2\widehat{\alpha}_0}{\sqrt{n_0(\pi^2 - 4)}}$$

$$\widehat{\alpha}_1 \pm Z_{(\frac{\delta}{2})} \frac{2\widehat{\alpha}_1}{\sqrt{n_1(\pi^2 - 4)}}$$

5. Numerical illustrations

To illustrate the proposed methodology, SAPS III dataset is used. Out of the 111 subjects, 66 (59.46%) belong to alive population and the remaining are of dead population. Table 2 report the parameter estimates along with their confidence limits for both alive and dead populations. Using the expression given in equation (7) the AUC value turns out to be 0.5793. Since the AUC expression do not have the closed form, the $V(\widehat{AUC})$ is obtained using bootstrap method. Upon performing 100 bootstraps, the $\widehat{AUC}_{Boot}$= 0.5629 and its variance is 0.0014. The bootstrap expressions for AUC and its variance are given below.

$$\widehat{AUC}_B = \frac{1}{B} \sum_{b=1}^{B} AUC_b \tag{13}$$

Table 2: The parameters estimates and confidence limits of Bi-GHN ROC curve

n_0	n_1	$\hat{\theta}_0$	$\hat{\theta}_1$	$\hat{\lambda}_0$	$\hat{\lambda}_1$
		(L_0,U_0)	(L_1,U_1)	(L_0,U_0)	(L_1,U_1)
66	45	1.2070	1.2071	32.2297	39.8010
		(0.9666,1.4474)	(0.9158,1.4982)	(32.1977,32.2617)	(39.7623,39.8397)

Table 3: Bootstrap estimates of measures of Bi-GHN ROC curve

$\widehat{AUC}_{Boot}$	$V(\widehat{AUC}_{Boot})$	$\widehat{FPR}_{Boot}$	$\widehat{TPR}_{Boot}$	$\widehat{c}$	$\widehat{J}$
0.5629	0.0014	0.2736	0.4857	36	0.1226

Table 4: Parameter combinations

Scenario	α_0	α_1	σ_0	σ_1
Better	0.75	2.20	0.99	2.11
Moderate	0.53	0.91	0.92	2.61
Worst	1.21	1.52	2.28	2.52

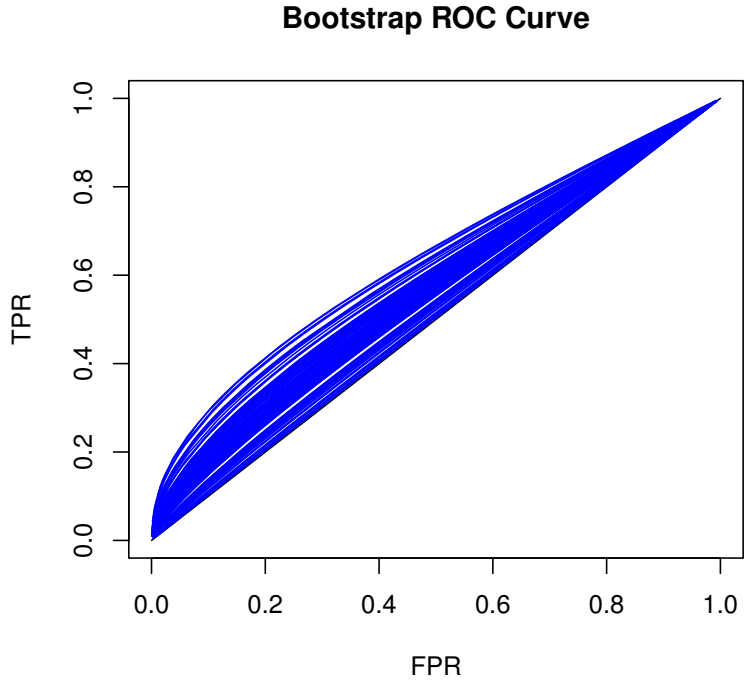


Figure 3: Bootstrap ROC curves for SAPS III dataset

$$V(\widehat{AUC}_B) = \frac{1}{B-1}\sum_{b=1}^{B}(AUC_b - \widehat{AUC}_B)^2 \tag{14}$$

Using the Youden's index, the optimal theshold is determined, that is, t= 36. At this cutoff, the FPR and TPR are observed to be 0.2736 and 0.4857 respectively. The obtained threshold is able to correctly classify 57 subjects out of 100 subjects. It is also noticed that this threshold generates 27% of false positives and truely detects the subject status upto 48% only. Figure 3 depicts the ROC curves generated at each bootstrap.

5.1. Simulation Studies

Further, to give a generalized view on the working methodology of the proposed Bi-GHN ROC curve, sizeable simulations are carried out with various parameter combinations at different sample sizes n= {25, 50, 100, 150, 200, 500}. Three different parameter combinations are considered to illustrate the *better, moderate* and *worst* case scenarios.

The parameter estimates and their confidence intervals of populations '0' and '1' for the combinations (Table 4) at different sample sizes are reported in Tables 5, 7 and 9 respectively. Accordingly, the estimated values of the measures of the proposed ROC curve are reported in Tables 6, 8 and 10 respectively.

Table 5: Parameter estimates at equal sample sizes (*Better case*)

n_0	n_1	$\widehat{\alpha}_0$ (L_0, U_0)	$\widehat{\alpha}_1$ (L_1, U_1)	$\widehat{\sigma}_0$ (L_0, U_0)	$\widehat{\sigma}_1$ (L_1, U_1)
25	25	0.7503	2.1000	0.9823	2.1026
		(0.5086,0.9274)	(1.4215,2.7817)	(0.7363,1.3569)	(2.0499,2.5524)
50	50	0.7499	2.1008	0.9926	2.1078
		(0.6553,0.8586)	(1.5638,2.5805)	(0.8894,1.2258)	(2.0616,2.2741)
100	100	0.7501	2.1016	0.9931	2.1102
		(0.6609,0.8234)	(1.6195,2.4398)	(0.9093,1.1403)	(2.0897,2.2362)
150	150	0.7482	2.1017	0.9936	2.1062
		(0.6642,0.8387)	(1.7602,2.3742)	(0.9247,1.0979)	(2.0924,2.2114)
200	200	0.7499	2.1018	0.9963	2.1115
		(0.6745,0.8049)	(1.8604,2.3484)	(0.9325,1.0089)	(2.0943,2.1940)
500	500	0.7524	2.1021	0.9991	2.1129
		(0.6996,0.7816)	(1.9496,2.3011)	(0.9269,1.0018)	(2.0995,2.1689)

With respect to better case, the following observations can be seen. For n= 100, the optimal cutoff is 1.1563, which is determined at the maximum value of Youden's index $\widehat{J}$= 0.5639. The classification of an individual can be in the following way: An individual is classified into Population '1', if S $>$ 1.1563 and Population '0', if S $\leq$ 1.1563. The optimal cutoff is able to detect around 82.92% of true positive cases with 29.63% of false positives.

Table 6: Accuracy cum intrinsic measures of Bi-GHN ROC (*Better case*)

n_0	n_1	$\widehat{AUC}$	$\widehat{FPR}$	$\widehat{TPR}$	$\hat{c}$	$\hat{J}$	$V(\widehat{AUC})$
25	25	0.9160	0.3282	0.8233	0.9789	0.5111	0.0029
50	50	0.9187	0.3062	0.8265	1.0048	0.5173	0.0132
100	100	0.9218	0.2963	0.8292	1.1563	0.5639	0.0147
150	150	0.9253	0.2923	0.8238	1.0353	0.5315	0.0142
200	200	0.9268	0.2871	0.8226	1.0675	0.5355	0.0068
500	500	0.9283	0.2745	0.8337	0.9992	0.5391	0.0018

The $\widehat{AUC}$ is observed to be 0.9218 which means that, the cutoff will be able to classify the individuals with 92.18% of accuracy. The ROC curves for this situation are shown in Figure 4 with a maximum coverage of area in the unit square plot. Interpretation can be given for the remaining sample sizes in similar manner.

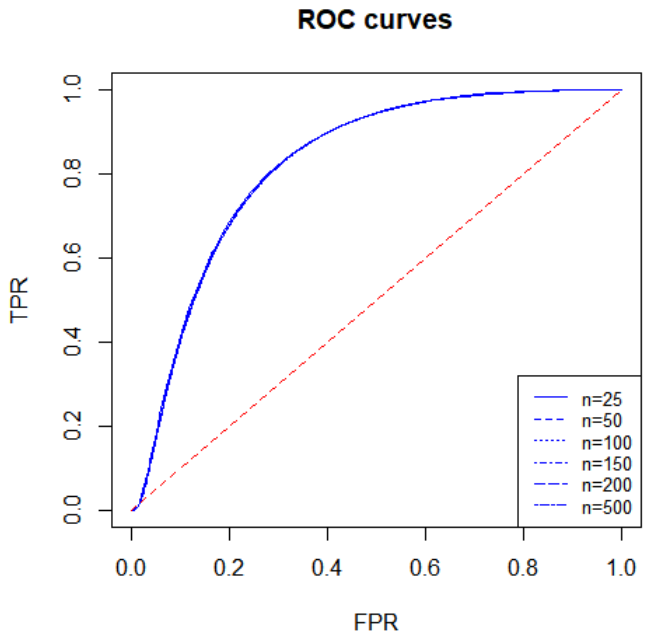


Figure 4: Better case

Now, let us consider the results of moderate case that are reported in Tables 7 and 8. For better understanding, let us consider a sample size from the results reported in Table 8. At $n = 150$, $\widehat{J}$= 0.3526 and the optimal cutoff ($\widehat{c}$) is 0.8886. At this $\widehat{c}$, we can observe 71.44% of true positives and 38.64% of false positives. The $\widehat{AUC}$= 0.7583, which can be interpreted

Table 7: Parameter estimates at equal sample sizes (*Moderate case*)

n_0	n_1	$\widehat{\alpha}_0$ (L_0, U_0)	$\widehat{\alpha}_1$ (L_1, U_1)	$\widehat{\sigma}_0$ (L_0, U_0)	$\widehat{\sigma}_1$ (L_1, U_1)
25	25	0.5314 (0.3618,0.7080)	0.9100 (0.6155,1.2044)	1.2429 (1.1210,1.3487)	2.4724 (2.3095,2.8732)
50	50	0.5348 (0.4086,0.6720)	0.9112 (0.6986,1.1408)	1.2497 (1.1807,1.3414)	2.4775 (2.3338,2.8453)
100	100	0.5365 (0.4099,0.6522)	0.9127 (0.7198,1.1361)	1.2538 (1.2357,1.3301)	2.4798 (2.3546,2.7881)
150	150	0.5397 (0.4286,0.6492)	0.9162 (0.7122,1.1289)	1.2606 (1.2312,1.3283)	2.4805 (2.3645,2.6754)
200	200	0.5329 (0.4319,0.6434)	0.9113 (0.7035,1.1152)	1.2644 (1.2376,1.3242)	2.4844 (2.3938,2.6072)
500	500	0.5222 (0.4691,0.6218)	0.9275 (0.6829,1.1008)	1.2667 (1.2456,1.3091)	2.4881 (2.4123,2.5697)

Table 8: Accuracy cum intrinsic measures of Bi-GHN ROC (*Moderate case*)

n_0	n_1	$\widehat{AUC}$	$\widehat{FPR}$	$\widehat{TPR}$	$\widehat{c}$	$\hat{J}$	$V(\widehat{AUC})$
25	25	0.7508	0.4057	0.7068	0.8338	0.3041	0.0166
50	50	0.7536	0.4044	0.7086	0.8563	0.3519	0.0189
100	100	0.7547	0.3927	0.7104	0.8598	0.3539	0.0251
150	150	0.7583	0.3864	0.7144	0.8886	0.3526	0.0310
200	200	0.7599	0.3514	0.7187	0.8837	0.3571	0.0035
500	500	0.7615	0.3554	0.7198	0.8503	0.3609	0.0012

as, $\widehat{c}$ has the ability to classify the individuals with 75.83% of accuracy. The ROC curves for the moderate case are depicted in Figure 5. Next, we consider the results pertaining to

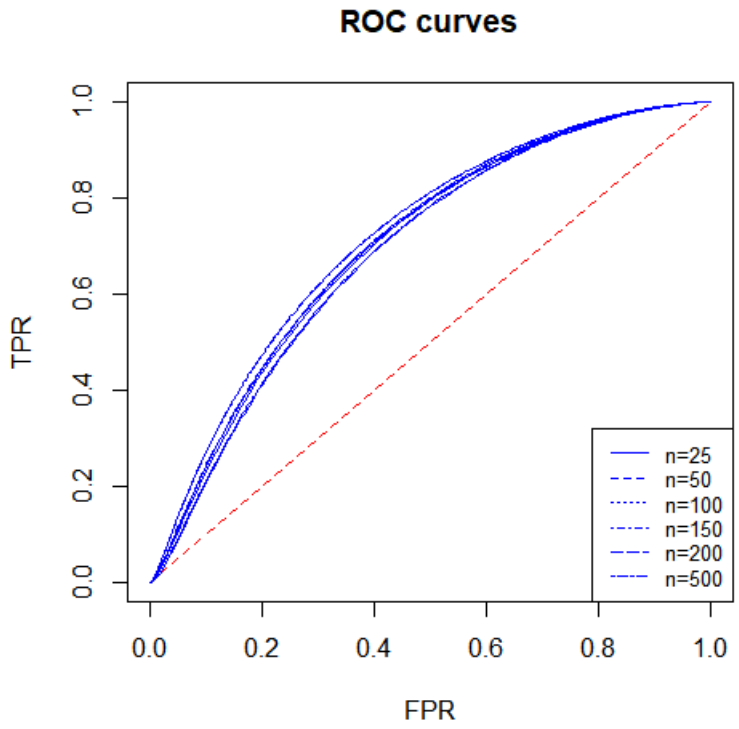


Figure 5: Moderate case

worst classification scenario presented in Tables 9 and 10. The $\widehat{AUC}$ is around 54%. So, this lower $\widehat{AUC}$ will have a maximum overlapping area between the populations '0' and '1'.

For $n = 100$, the $\widehat{AUC}$= 0.5466, where the ROC curve is quite closer to the chance diagonal line indicating random classification. The $\widehat{c} = 1.3878$ is able to detect 66.53% of true positives and 58.57% of false positives. The ROC curves for this case are presented in Figure 6. Since the curves obtained here are closer to the chance diagonal, the classifier fails to classify the subjects into one of the populations with better accuracy.

6. Summary

In this paper, Bi-GHN ROC curve is proposed and accordingly the expressions for AUC, FPR and TPR are derived. Since AUC does not have closed form expression, its variance is obtained using bootstrap. The proposed work is supported with SAPS III dataset and simulations. Better, moderate and worst case scenarios are considered at different sample sizes. For the SAPS III dataset, the optimal threshold is observed to be 36 and $\widehat{AUC}$= 0.5793. The obtained threshold is able to classify the subjects in alive and dead population with 57.93% of accuracy only.

Table 9: Parameter estimates at equal sample sizes (*Worst case*)

n_0	n_1	$\widehat{\alpha}_0$ (L_0, U_0)	$\widehat{\alpha}_1$ (L_1, U_1)	$\widehat{\sigma}_0$ (L_0, U_0)	$\widehat{\sigma}_1$ (L_1, U_1)
25	25	1.2085 (0.8174,1.5996)	1.5200 (1.1281,2.0119)	2.3239 (2.1519,2.5725)	2.6336 (2.4423,2.7656)
50	50	1.2100 (0.9331,1.4868)	1.5183 (1.1609,1.8657)	2.3120 (2.1703,2.5484)	2.6323 (2.4542,2.7614)
100	100	1.2143 (1.0142,1.4014)	1.5169 (1.2072,1.7967)	2.2970 (2.1754,2.5086)	2.5177 (2.4631,2.7547)
150	150	1.2214 (1.0331,1.3857)	1.5210 (1.2740,1.6959)	2.3106 (2.1987,2.4906)	2.6287 (2.4782,2.7512)
200	200	1.2150 (1.1198,1.3241)	1.5200 (1.3461,1.6345)	2.3056 (2.2172,2.4239)	2.6269 (2.5006,2.7154)
500	500	1.2321 (1.1501,1.3098)	1.5288 (1.4378,1.6190)	2.2917 (2.2562,2.3778)	2.6261 (2.5311,2.6921)

Table 10: Accuracy cum intrinsic measures of Bi-GHN ROC (*Worst case*)

n_0	n_1	$\widehat{AUC}$	$\widehat{FPR}$	$\widehat{TPR}$	$\widehat{c}$	$\hat{J}$	$V(\widehat{AUC})$
25	25	0.5319	0.5341	0.6431	1.6358	0.1090	0.0103
50	50	0.5354	0.6101	0.6543	1.3252	0.0908	0.0052
100	100	0.5466	0.5857	0.6653	1.3878	0.0996	0.0050
150	150	0.5490	0.5774	0.6786	1.4026	0.1011	0.0030
200	200	0.5584	0.5540	0.6822	1.5101	0.1082	0.0012
500	500	0.5665	0.5390	0.6909	1.4756	0.1152	0.0003

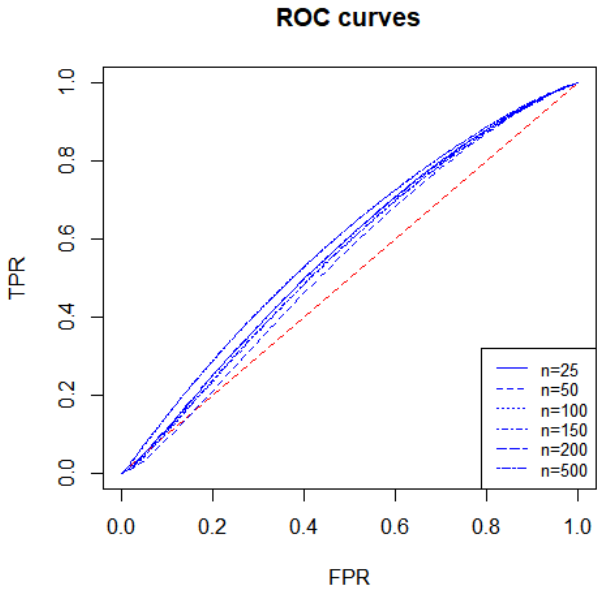


Figure 6: Worst case

References

Balaswamy, S. and Vishnu, R. V. (2016). An anthology of parametric roc models. *Journal of Statistics*, **5**, 32–46.

Cooray, K. and Ananda, M. A. (2008). A generalization of the half-normal distribution with applications to lifetime data. *Communications in Statistics-Theory and Methods*, **37**, 1323–1337.

Dorfman, D. D. and Alf, J. E. (1969). Maximum-likelihood estimation of parameters of signal detection theory and determination of confidence intervals—rating method data. *Journal of Mathematical Psychology*, **6**, 487–496.

Egan, J. P. (1975). *Signal Detection Theory and ROC Analysis*. Academic press.

Hager, H. W. and Bain, L. J. (1970). Inferential procedures for the generalized-gamma distribution. *Journal of the American Statistical Association*, **65**, 1601–1609.

Hussain, E. (2012). The bi-gamma roc curve in a straightforward manner. *Journal of Basic & Applied Sciences*, **8**, 309–314.

Parr, V. B. and Webster, J. T. (1965). A method for discriminating between failure density functions used in reliability predictions. *Technometrics*, **7**, 1–10.

Stacy, E. W. and Mihram, G. A. (1965). Parameter estimation for a generalized-gamma distribution. *Technometrics*, **7**, 349–358.

Vishnu, R. V. and Kiruthika, C. (2015). Properties and the use of half-normal distribution in roc curve analysis. *IAPQR Transactions*, **39**, 169–179.

Vishnu, R. V., Sudesh, P., and Sameera, G. (2012). Estimation of area under the roc curve using exponential and weibull distributions. *Bonfring International Journal of Data Mining*, **2**, 52–56.

Statistics and Application {ISSN 2454-7395 (online)}
Volume 21, No. 2, 2023 (New Series), pp 73-98
http://www.ssca.org.in/journal

Estimation of the Finite Population Mean in Stratified Random Sampling under Non-response

Zakir Hussain Wani and S.E.H. Rizvi
Division of Statistics and Computer Science, Main Campus SKUAST-J, Chatha Jammu-180009, India

Received: 15 November 2021; Revised: 31 May 2022; Accepted: 24 October 2022

Abstract

In this paper we have considered the problem of estimating the population mean using auxiliary information in presence of non-response. This situation is examined under two cases; Case I: when non-response occurs both on study variable and auxiliary variable and population mean of the auxiliary variable is known; Case II: when non-response occurs only on study variable, complete information on auxiliary variable and population mean of auxiliary variable is known. Mathematical properties of proposed estimators such as bias, mean square error and minimum mean square error are separately obtained for both the cases of all the proposed estimators up to the first order of approximation. The proposed estimators have been compared theoretically with the Hansen and Hurwitz (1946) estimator and some other existing estimators. The conditions for which proposed estimators are most efficient are obtained. Moreover, numerical illustrations shows that the proposed estimators perform better than existing estimators in terms of mean square error.

Key words: Non-response; Stratified random sampling; Auxiliary information; Mean square error; Bias; Efficiency.

1. Introduction

In sample surveys, survey statisticians are expected to gather information on each unit of the selected sample in order to provide a precise estimate of the population mean. In many circumstances, information on part of the sample units cannot be collected in the first attempt due to natural interference. The failure of some of the sample units causes the errors to be classified as non-response. Hansen and Hurwitz (1946) introduced a method of sub-sampling to deal with the non-respondents and employed it in a more expensive way using the second trial. They considered two attempts (i) mail questionnaire, and (ii) personal interview to obtain the information from the sample and tried to give the appropriate inference about the population parameter. Later, various authors such as Cochran (1977), Rao (1986), Khare and Srivastava (1997), Singh and Kumar (2008, 2009), Singh and Vishwakarma (2019), Kumar *et al*. (2022) discussed the problem of estimating the population mean of the study variable using information on an auxiliary variable in the presence of non-response following the Hansen and Hurwitz (1946) technique under the simple random sampling without replacement (*SRSWOR*) scheme. When the population units are homogeneous, the *SRSWOR* sampling strategy is usually utilized. However, in practice heterogeneous populations are also commonly encountered. In such cases, stratified random sampling is used. With this in mind,

Corresponding Author: Zakir Hussain Wani
Email: wanizakir927@gmail.com

Chaudhary *et al*. (2009) investigated non-response in stratified random sampling, assuming that non-response happens only on the study variable. Sanaullah *et al*. (2015), Saleem *et al*. (2018), Onyeka *et al*. (2019), Shabbir *et al*. (2019) and Wani *et al*. (2021) have studied the problem of non-response in stratified single and two-phase sampling where non-response occurs on both the study and auxiliary variable, as well as on the study variable only. In this article, we attempted to propose estimators for estimating the population mean of the study variable *Y* using information on the auxiliary variable *X* in the presence of non-response for two cases. Case I occurs when there is non-response on both the study variable *Y* and the auxiliary variable *X*, and the auxiliary variable's population mean ($\bar{X}$) is known, whereas Case II occurs when there is non-response on only the study variable *Y*, and information on the auxiliary variable *X* is obtained from all sample units, and the auxiliary variable's population mean ($\bar{X}$) is known. The mathematical properties of proposed estimators, such as bias, mean square error, and minimum mean square error, were examined using large sample approximation. The proposed estimators have been shown to outperform all other estimators tested in the literature. Numerical illustrations have also been done in support of current investigation.

2. Sampling strategy

Consider a finite heterogeneous population of *N* units organised into *L* homogenous subgroups termed as strata, with the h^{th} stratum containing N_h units, where h = 1, 2, 3, ..., *L* and *N* consists of two mutually exclusive groups, *viz.* response and non-response group. The responding and non-responding units in the h^{th} stratum, respectively, are N_{1h} and N_{2h}. We select a sample of size n_h from N_h units in the stratum by using *SRSWOR* and assume that n_{1h} units respond and n_{2h} units do not respond. We select a sub-sample of size $r_h = (n_{2h} / k_h ;\ k_h > 1)$ from n_{2h} non responding units in the h^{th} stratum. Following is the Hansen and Hurwitz (1946) estimator, $\bar{y}_{st}^{*} = \sum_{h=1}^{L} W_h \bar{y}_h^{*}$ and $\bar{x}_{st}^{*} = \sum_{h=1}^{L} W_h \bar{x}_h^{*}$ be the stratified sample means of *y* and *x* respectively in the h^{th} stratum under non-response, where $\bar{y}_h^{*} = \frac{n_{1h}\bar{y}_{n1h} + n_{2h}\bar{y}_{r2h}}{n_h}$, $\bar{x}_h^{*} = \frac{n_{1h}\bar{x}_{n1h} + n_{2h}\bar{x}_{r2h}}{n_h}$, and $(\bar{y}_{n1h}, \bar{x}_{n1h})$ and $(\bar{y}_{r2h}, \bar{x}_{r2h})$ are the sample means based on n_{1h} units and r_{2h} units, respectively.

The MSE of $\bar{y}_{st}^{*}$ and $\bar{x}_{st}^{*}$ are respectively given by

$$MSE(\bar{y}_{st}^{*}) = \sum_{h=1}^{L} W_h^2 \left\{ \left(\frac{1}{n_h} - \frac{1}{N_h} \right) S_{hy}^2 + \frac{(k_h - 1)}{n_h} W_{2h} S_{hy(2)}^2 \right\} \tag{1}$$

$$MSE(\bar{x}_{st}^{*}) = \sum_{h=1}^{L} W_h^2 \left\{ \left(\frac{1}{n_h} - \frac{1}{N_h} \right) S_{hx}^2 + \frac{(k_h - 1)}{n_h} W_{2h} S_{hx(2)}^2 \right\}$$

where S_{hy}^2 and $S_{hy(2)}^2$ are the population mean squares of entire group and non-response group respectively in the h^{th} stratum for the study variable.

3. Useful notations

Following are some notations used for the theoretical development of present investigation:

$\bar{Y}_h = \frac{1}{N_h}\sum_{i=1}^{N_h} y_{hi}$: Population mean of study variable for h^{th} stratum.

$\bar{X}_h = \frac{1}{N_h}\sum_{i=1}^{N_h} x_{hi}$: Population mean of auxiliary variable for h^{th} stratum.

$\bar{Y} = \sum_{h=1}^{L} W_h \bar{Y}_h$: Population mean of the study variable.

$\bar{X} = \sum_{h=1}^{L} W_h \bar{X}_h$: Population mean of the auxiliary variable.

$S_{hy}^2 = \frac{1}{N_h-1}\sum_{i=1}^{N_h}(y_{hi} - \bar{Y}_h)^2$: Population variance of study variable for h^{th} stratum.

$S_{hx}^2 = \frac{1}{N_h-1}\sum_{i=1}^{N_h}(x_{hi} - \bar{X}_h)^2$: Population variance of auxiliary variable for h^{th} stratum.

$S_{hy(1)}^2 = \frac{1}{N_{1h}-1}\sum_{i=1}^{N_{1h}}(y_{hi} - \bar{Y}_{1h})^2$: Population variance of response group of the study variable for h^{th} stratum

$S_{hx(1)}^2 = \frac{1}{N_{1h}-1}\sum_{i=1}^{N_{1h}}(x_{hi} - \bar{X}_{1h})^2$: Population variance of response group of the auxiliary variable for h^{th} stratum

$S_{hy(2)}^2 = \frac{1}{N_{2h}-1}\sum_{i=1}^{N_{2h}}(y_{hi} - \bar{Y}_{2h})^2$: Population variance of non-response group of the study variable for h^{th} stratum

$S_{hx(2)}^2 = \frac{1}{N_{2h}-1}\sum_{i=1}^{N_{2h}}(y_{hi} - \bar{X}_{2h})^2$: Population variance of non-response group of the auxiliary variable for h^{th} stratum

$\rho_{hxy} = \frac{S_{hxy}}{S_{hy}S_{hx}}$: Correlation coefficient between the auxiliary and study variables in h^{th} stratum.

$\rho_{hxy(2)} = \frac{S_{hxy(2)}}{S_{hy(2)}S_{hx(2)}}$: Correlation coefficient between the auxiliary and study variables of non-response group in h^{th} stratum.

$f_h = \frac{n_h}{N_h}$: The sampling fraction of h^{th} stratum

And also $\sum_{h=1}^{L} N_h = N$; $\theta_h = \frac{1}{n_h} - \frac{1}{N_h}$

To derive the expressions for the bias, mean square error and minimum mean square error of existing and proposed estimators, we consider the following relative error terms along with their expectations.

3.1. For separate estimators

Relative error terms along with their expectations for separate estimators

$\xi_{0h}^* = \frac{\bar{y}_h^* - \bar{Y}_h}{\bar{Y}_h}$, $\xi_{1h}^* = \frac{\bar{x}_h^* - \bar{X}_h}{\bar{X}_h}$, $\xi_{1h} = \frac{\bar{x}_h - \bar{X}_h}{\bar{X}_h}$,

such that $E(\xi_{0h}^*) = E(\xi_{1h}^*) = E(\xi_{1h}) = 0$ and under *SRSWOR*, we have

$$E(\xi_{0h}^{*2}) = \frac{1}{\bar{Y}_h^2}\left[\theta_h S_{hy}^2 + \frac{W_{2h}(k_h-1)}{n_h} S_{hy(2)}^2\right] = A_h$$

$$E(\xi_{1h}^{*2}) = \frac{1}{\bar{X}_h^2}\left[\theta_h S_{hx}^2 + \frac{W_{2h}(k_h-1)}{n_h} S_{hx(2)}^2\right] = B_h$$

$$E(\xi_{0h}^*\xi_{1h}^*) = \frac{1}{\bar{Y}_h\bar{X}_h}\left[\theta_h S_{hxy} + \frac{w_{2h}(k_h-1)}{n_h} S_{hxy(2)}\right] = C_h$$

$$E(\xi_{1h}^2) = \frac{1}{\bar{X}_h^2}\theta_h S_{hx}^2 = D_h \quad , \quad E(\xi_{0h}^*\xi_{1h}) = \frac{1}{\bar{Y}_h\bar{X}_h}\theta_h S_{hxy} = E_h$$

3.2. For combined estimators

Relative error terms along with their expectations for combined estimators

$$\xi_{0st}^* = \frac{\bar{y}_{st}^* - \bar{Y}}{\bar{Y}}, \qquad \xi_{1st}^* = \frac{\bar{x}_{st}^* - \bar{X}}{\bar{X}}, \qquad \xi_{1st} = \frac{\bar{x}_{st} - \bar{X}}{\bar{X}},$$

such that $E(\xi_{0st}^*) = E(\xi_{1st}^*) = E(\xi_{1st}) = 0$, and under *SRSWOR*, we have

$$E(\xi_{0st}^{*2}) = \frac{1}{\bar{Y}^2}\sum_{h=1}^{L} W_h^2\left[\theta_h S_{hy}^2 + \frac{W_{2h}(k_h-1)}{n_h} S_{hy(2)}^2\right] = A$$

$$E(\xi_{1st}^{*2}) = \frac{1}{\bar{X}^2}\sum_{h=1}^{L} W_h^2\left[\theta_h S_{hx}^2 + \frac{W_{2h}(k_h-1)}{n_h} S_{hx(2)}^2\right] = B$$

$$E(\xi_{0st}^*\xi_{1st}^*) = \frac{1}{\bar{Y}\bar{X}}\sum_{h=1}^{L} W_h^2\left[\theta_h S_{hxy} + \frac{w_{2h}(k_h-1)}{n_h} S_{hxy(2)}\right] = C$$

$$E(\xi_{1st}^2) = \frac{1}{\bar{X}^2}\sum_{h=1}^{L} W_h^2\,\theta_h S_{hx}^2 = D \quad , \quad E(\xi_{0st}^*\xi_{1st}) = \frac{1}{\bar{Y}\bar{X}}\sum_{h=1}^{L} W_h^2\,\theta_h S_{hxy} = E$$

4. Existing estimators in the literature

This section gives a brief introduction of some well-known estimators/ classes of estimators from the literature.

For the simple random sampling method, we can mention some important studies in literature when there is a complete information on the study and auxiliary variable for homogenous populations. For estimating the population mean, Cochran (1977) proposed the classical ratio type estimator as

$$t_R = \frac{\bar{y}}{\bar{x}}\bar{X} \tag{2}$$

where $\bar{X}$ refers population mean of the auxiliary variable, $\bar{y}$ and $\bar{x}$ represents the sample means of the study and auxiliary variable respectively

Bhul and Tuteja (1991) are the first to suggest an estimator using the exponential function to estimate the population mean and is given by

$$t_{EX} = \bar{y}exp\left(\frac{\bar{X}-\bar{x}}{\bar{X}+\bar{x}}\right) \tag{3}$$

Motivated by Bahl and Tuteja (1991), Ozel Kadilar (2016) proposed an exponential type estimator as

$$t_O = \bar{y}\left(\frac{\bar{x}}{\bar{X}}\right)^{\delta_1} exp\left(\frac{\bar{X}-\bar{x}}{\bar{X}+\bar{x}}\right) \tag{4}$$

Motivated by Bahl and Tuteja (1991), Upadhyaya *et al*. (2011) proposed a ratio type exponential estimator and product type exponential estimator and is given as

$$t_{UP1} = \bar{y}exp\left[\frac{\bar{X}-\bar{x}}{\bar{X}+(\delta_2-1)\bar{x}}\right] \tag{5}$$

$$t_{UP2} = \bar{y}exp\left[\frac{\bar{x}-\bar{X}}{\bar{X}+(\delta_3-1)\bar{x}}\right] \tag{6}$$

Motivated by Bahl and Tuteja (1991), Vishwakarma *et al*. (2016) proposed exponential type estimator and is given as

$$t_V = \delta_4\bar{y} + (1-\delta_4)\bar{y}exp\left[\frac{\bar{X}-\bar{x}}{\bar{X}+\bar{x}}\right] \tag{7}$$

The mean square error of t_R , t_{EX} , t_{OK} t_{UP1} , t_{UP2} and t_v are given as

$$MSE(t_R) = \theta\bar{Y}^2\left(C_y^2 + C_x^2 - 2C_{xy}\right) \tag{8}$$

$$MSE(t_{EX}) = \theta\bar{Y}^2\left(C_y^2 + \frac{C_x^2}{4} - C_{xy}\right) \tag{9}$$

$$MSE_{min}(t_O) = MSE_{min}(t_{UP1}) = MSE_{min}(t_{UP2}) = MSE_{min}(t_v) = \theta\bar{Y}^2C_y^2\left(1-\rho_{xy}^2\right) \tag{10}$$

where $\theta = \frac{1}{n} - \frac{1}{N}$, $C_y = \frac{S_y}{\bar{Y}}$ and $C_x = \frac{S_x}{\bar{X}}$

Because of the various reasons, the required correct information cannot be obtained completely at all times which is named as a case of non-response. In order to solve this problem, a method is considered and a new technique of sub-sampling the non-respondents is introduced by Hansen and Hurwitz (1946). In various real-life situations, the population under study is heterogeneous, and in that case, we adopt stratified random sampling to obtain precise estimators for the population parameter(s) of the study variable. Considering this fact an attempt was made in this paper to develop some improved estimators in presence of non-

response using stratified random sampling. So, the Hansen and Hurwitz (1946) estimator in stratified random sampling is given as

$$\bar{y}_{st}^{*} = \sum_{h=1}^{L} W_h \bar{y}_h^{*} \tag{11}$$

The mean square error of $\bar{y}_{st}^{*}$ is given as

$$MSE(\bar{y}^{*}) = \sum_{h=1}^{L} W_h^2 \, A_h = \bar{Y}^2 A \tag{12}$$

The usual separate ratio estimator when non-response occurs both on study variable and auxiliary variable and the population mean of the auxiliary variable is known is given by

$$\bar{y}_{SR}^{*} = \sum_{h=1}^{L} W_h \frac{\bar{y}_h^{*}}{\bar{x}_h^{*}} \bar{X}_h \tag{13}$$

The mean square error of $\bar{y}_{SR}^{*}$ is given as

$$MSE(\bar{y}_{SR}^{*}) = \sum_{h=1}^{L} W_h^2 \, \bar{Y}_h^2 [A_h + B_h - 2C_h \,] \tag{14}$$

The separate ratio estimator when non-response occurs only on study variable, complete information on auxiliary variable and the population mean of the auxiliary variable is known is given by

$$\bar{y}_{SR}' = \sum_{h=1}^{L} W_h \frac{\bar{y}_h^{*}}{\bar{x}_h} \bar{X}_h \tag{15}$$

The mean square error of $\bar{y}_{SR}'$ is given as

$$MSE(\bar{y}_{SR}') = \sum_{h=1}^{L} W_h^2 \, \bar{Y}_h^2 [A_h + D_h - 2E_h \,] \tag{16}$$

The usual combined ratio estimator when non-response occurs both on study variable and auxiliary variable and the population mean of the auxiliary variable is known is given by

$$\bar{y}_{CR}^{*} = \frac{\bar{y}_{st}^{*}}{\bar{x}_{st}^{*}} \bar{X} \tag{17}$$

The mean square error $\bar{y}_{CR}^{*}$ is given as

$$MSE(\bar{y}_{CR}^{*}) = \bar{Y}^2 [A + B - 2C \,] \tag{18}$$

The combined ratio estimator when non-response occurs only on study variable, complete information on auxiliary variable and the population mean of the auxiliary variable is known is given by

$$\bar{y}'_{CR} = \frac{\bar{y}^*_{st}}{\bar{x}_{st}}\bar{X} \tag{19}$$

The mean square error of $\bar{y}'_{CR}$ is given as

$$MSE(\bar{y}'_{CR}) = \bar{Y}^2\,[A + D - 2E\,] \tag{20}$$

The following are the stratified modified estimators in presence of non-response developed by Onyeka *et al*. (2019) using known values of coefficient of correlation, kurtosis, and coefficient of variation when non- response occurs both on the study variable and auxiliary variable and the population mean of the auxiliary variable is known.

$$\bar{y}^{*(i)}_{ok} = \sum_{h=1}^{L} W_h \bar{y}^*_h exp\left[\frac{\alpha_h(\bar{X}_h - \bar{x}^*_h)}{\alpha_h(\bar{X}_h - \bar{x}^*_h) + 2\beta_h}\right] \tag{21}$$

The mean square error of $\bar{y}^{*(i)}_{ok}$ is given as

$$MSE\left(\bar{y}^{*(i)}_{ok}\right) = \sum_{h=1}^{L} W_h^2\,\bar{Y}_h^2\left[A_h + \frac{1}{4}\varphi_{hi}^2 B_h - \varphi_{hi}C_h\,\right] \tag{22}$$

When non-response occurs only on the study variable, complete information on the auxiliary variable and the population mean of the auxiliary variable is known the Onyeka *et al*. (2019) estimators are

$$\bar{y}'^{(i)}_{ok} = \sum_{h=1}^{L} W_h \bar{y}^*_h exp\left[\frac{\alpha_h(\bar{X}_h - \bar{x}_h)}{\alpha_h(\bar{X}_h - \bar{x}_h) + 2\beta_h}\right] \tag{23}$$

The mean square error of $\bar{y}'^{(i)}_{ok}$ is given as

$$MSE\left(\bar{y}'^{(i)}_{ok}\right) = \sum_{h=1}^{L} W_h^2\,\bar{Y}_h^2\left[A_h + \frac{1}{4}\varphi_{hi}^2 D_h - \varphi_{hi}E_h\,\right] \tag{24}$$

where

$$\varphi_{hi} = \frac{\alpha_h\bar{X}_h}{\alpha_h\bar{X}_h + \beta_h}$$

$$\varphi_{h1} = 1, \quad \varphi_{h2} = \frac{\bar{X}_h}{\bar{X}_h + C_{(x)h}}, \quad \varphi_{h3} = \frac{\bar{X}_h}{\bar{X}_h + \beta_{2(x)h}}, \quad \varphi_{h4} = \frac{C_{(x)h}\bar{X}_h}{C_{(x)h}\bar{X}_h + \rho_{yxh}}$$

$$\varphi_{h5} = \frac{\beta_{2(x)h}\bar{X}_h}{\beta_{2(x)h}\bar{X}_h + C_{(x)h}}, \quad \varphi_{h6} = \frac{\bar{X}_h}{\bar{X}_h + \rho_{yxh}}, \quad \varphi_{h7} = \frac{\rho_{yxh}\bar{X}_h}{\rho_{yxh}\bar{X}_h + \beta_{2(x)h}}, \quad \varphi_{h8} = 0$$

5. Proposed estimators

In this section we propose exponential estimators for estimating the population mean in presence of non-response under stratified random sampling motivated from Upadhyaya *et al*. (2011), Vishwakarma *et al*. (2016) and Ozel Kadilar (2016). The situation of non-response is examined under two cases; Case I: when non-response occurs both on study variable and auxiliary variable and population mean of the auxiliary variable is known; Case II: when non-response occurs only on study variable, complete information on auxiliary variable and population mean of auxiliary variable is known.

Proposed Estimator 1: Based on Upadhyaya *et al*. (2011), we propose a ratio type exponential estimator for estimating the population mean in stratified random sampling in presence of non-response for case I and case II.

Case I: When non-response occurs both on the study variable and auxiliary variable and population mean of the auxiliary variable is known. The proposed estimator t_1^* is given as

$$t_1^* = \bar{y}_{st}^* exp\left[\frac{\bar{X} - \bar{x}_{st}^*}{\bar{X} + (a-1)\bar{x}_{st}^*}\right] \tag{25}$$

Now, we express the equation (25) in terms of ξ_{0st}^* and ξ_{1st}^* up to first order of approximation and is given as

$$t_1^* = \bar{Y}(1 + \xi_{0st}^*)\left[\frac{\bar{X} - \bar{X}(1 + \xi_{1st}^*)}{\bar{X} + (a-1)\bar{X}(1 + \xi_{1st}^*)}\right]$$

$$t_1^* = \bar{Y}(1 + \xi_{0st}^*)\left[\frac{-\xi_{1st}^*}{a}\left\{1 + \frac{(a-1)^{-1}}{a}\xi_{1st}^*\right\}\right]$$

$$t_1^* - \bar{Y} = \bar{Y}\left[\xi_{0st}^* - \frac{\xi_{1st}^*}{a} + \frac{\xi_{1st}^{*2}}{a^2}\left(a - \frac{1}{2}\right) - \frac{\xi_{0st}^*\xi_{1st}^*}{a}\right] \tag{26}$$

Taking expectation in equation (26), the bias of t_1^* to first order of approximation is given as

$$E(t_1^* - \bar{Y}) = \bar{Y}\left[E(\xi_{0st}^*) - \frac{E(\xi_{1st}^*)}{a} + \frac{E(\xi_{1st}^{*2})}{a^2}\left(a - \frac{1}{2}\right) - \frac{E(\xi_{0st}^*\xi_{1st}^*)}{a}\right]$$

$$E(t_1^* - \bar{Y}) = \bar{Y}\left[\frac{E(\xi_{1st}^{*2})}{a^2}\left(a - \frac{1}{2}\right) - \frac{E(\xi_{0st}^*\xi_{1st}^*)}{a}\right]$$

$$Bias(t_1^*) = \frac{\bar{Y}}{a}\left[\frac{(2a-1)}{2a}B - C\right] \tag{27}$$

Squaring up to first order of approximation and then taking expectation in equation (26), The *MSE* of t_1^* is given as

$$E(t_1^* - \bar{Y})^2 = \bar{Y}^2 E\left[\xi_{0st}^* - \frac{\xi_{1st}^*}{a} + \frac{\xi_{1st}^{*2}}{a^2}\left(a - \frac{1}{2}\right) - \frac{\xi_{0st}^*\xi_{1st}^*}{a}\right]^2$$

$$E(t_1^* - \bar{Y})^2 = \bar{Y}\left[E(\xi_{0st}^{*2}) + \frac{E(\xi_{1st}^{*2})}{a^2} - 2\frac{1}{a}E(\xi_{0st}^*\xi_{1st}^*)\right]$$

$$MSE(t_1^*) = \bar{Y}^2\left[A + \frac{1}{a^2}B - 2\frac{1}{a}C\right] \tag{28}$$

For obtaining the optimal values of a , differentiating equation (28) w.r.t a and equating to zero we have

$$\frac{\partial MSE(t_1^*)}{\partial\, a} = 0$$

$$a_{opt} = \frac{B}{C}$$

Substituting the optimal value of a in equation (28), we have MSE as

$$MSE(t_1^*) = \bar{Y}^2\left[A + \frac{1}{\left(\frac{B}{C}\right)^2}B - 2\frac{1}{\left(\frac{B}{C}\right)}C\right] \tag{29}$$

Simplifying equation (29), we have the minimum mean square error of the proposed estimator t_1^*

$$MSE(t_1^*)_{min} = \bar{Y}^2\left[A - \frac{C^2}{B}\right] \tag{30}$$

Case II: When non-response occurs only on study variable, complete information on auxiliary variable and population mean of the auxiliary variable is known. The proposed estimator t_1' is given as

$$t_1' = \bar{y}_{st}^* exp\left[\frac{\bar{X} - \bar{x}_{st}}{\bar{X} + (a' - 1)\bar{x}_{st}}\right] \tag{31}$$

Now, we express the equation (31) in terms of ξ_{0st}^* and ξ_{1st} up to first order of approximation and is given as

$$t_1' = \bar{Y}(1 + \xi_{0st}^*)\left[\frac{\bar{X} - \bar{X}(1 + \xi_{1st})}{\bar{X} + (a' - 1)\bar{X}(1 + \xi_{1st})}\right]$$

$$t_1' = \bar{Y}(1 + \xi_{0st}^*)\left[\frac{\xi_{1st}}{a'}\left\{1 + \frac{(a' - 1)^{-1}}{a'}\,\xi_{1st}\right\}\right]$$

$$t_1' - \bar{Y} = \bar{Y}\left[\xi_{0st}^* - \frac{\xi_{1st}}{a'} + \frac{\xi_{1st}^2}{a'^2}\left(a' - \frac{1}{2}\right) - \frac{\xi_{0st}^*\xi_{1st}}{a'}\right] \tag{32}$$

Taking expectation in equation (32), the bias of t_1' to first order of approximation to get bias and is given as

$$E(t_1' - \bar{Y}) = \bar{Y}\left[E(\xi_{0st}^*) - \frac{E(\xi_{1st})}{a'} + \frac{E(\xi_{1st}^2)}{a'^2}\left(a' - \frac{1}{2}\right) - \frac{E(\xi_{0st}^*\xi_{1st})}{a'}\right]$$

$$E(t_1' - \bar{Y}) = \bar{Y}\left[\frac{E(\xi_{1st}^2)}{a'^2}\left(a' - \frac{1}{2}\right) - \frac{E(\xi_{0st}^*\xi_{1st})}{a'}\right]$$

$$Bias(t_1') = \frac{\bar{Y}}{a'}\left[\frac{(2a'-1)}{2a'}D - E\right] \tag{33}$$

Squaring up to first order of approximation and then taking expectation in equation (32), The *MSE* of t_1' is given as

$$E(t_1' - \bar{Y})^2 = \bar{Y}^2 E\left[\xi_{0st}^* - \frac{\xi_{1st}}{a'} + \frac{\xi_{1st}^2}{a'^2}\left(a' - \frac{1}{2}\right) - \frac{\xi_{0st}^*\xi_{1st}}{a'}\right]^2$$

$$E(t_1' - \bar{Y})^2 = \bar{Y}\left[E(\xi_{0st}^{*2}) + \frac{E(\xi_{1st}^2)}{a'^2} - 2\frac{1}{a'}E(\xi_{0st}^*\xi_{1st})\right]$$

$$MSE(t_1') = \bar{Y}^2\left[A + \frac{1}{a'^2}D - 2\frac{1}{a'}E\right] \tag{34}$$

For obtaining the optimal values of a' , differentiating equation (34) w.r.t a' and equating to zero we have

$$\frac{\partial MSE(t_1')}{\partial a'} = 0$$

$$a'_{opt} = \frac{D}{E}$$

Substituting the optimal value of a' in equation (34), we have *MSE* as

$$MSE(t_1') = \bar{Y}^2\left[A + \frac{1}{\left(\frac{D}{E}\right)^2}D - 2\frac{1}{\left(\frac{D}{E}\right)}E\right] \tag{35}$$

Simplifying equation (35), we have the minimum mean square error of the proposed estimator t_1'

$$MSE(t_1')_{min} = \bar{Y}^2\left[A - \frac{E^2}{D}\right] \tag{36}$$

Proposed Estimator 2: Based on Upadhyaya *et al*. (2011), we propose a product type exponential estimator for estimating the population mean in stratified random sampling in presence of non–response for case I and case II.

Case 1: When non-response occurs both on the study variable and auxiliary variable and population mean of the auxiliary variable is known. The proposed estimator t_2^* is given as

$$t_2^* = \bar{y}_{st}^* exp\left[\frac{\bar{x}_{st}^* - \bar{X}}{\bar{X} + (b-1)\bar{x}_{st}^*}\right] \tag{37}$$

Now, we express the equation (37) in terms of ξ_{0st}^* and ξ_{1st}^* up to first order of approximation and is given as

$$t_2^* = \bar{Y}(1+\xi_{0st}^*)\left[\frac{\bar{X}(1+\xi_{1st}^*)-\bar{X}}{\bar{X}+(b-1)\bar{X}(1+\xi_{1st}^*)}\right]$$

$$t_2^* = \bar{Y}(1+\xi_{0st}^*)\left[\frac{\xi_{1st}^*}{b}\left\{1+\frac{(b-1)^{-1}}{b}\xi_{1st}^*\right\}\right]$$

$$t_2^* - \bar{Y} = \bar{Y}\left[\xi_{0st}^* + \frac{\xi_{1st}^*}{b} - \frac{\xi_{1st}^{*2}}{b^2}\left(b-\frac{3}{2}\right) + \frac{\xi_{0st}^*\xi_{1st}^*}{b}\right] \quad (38)$$

Taking expectation in equation (38), the bias of t_2^* to first order of approximation is given as

$$E(t_2^* - \bar{Y}) = \bar{Y}\left[E(\xi_{0st}^*) + \frac{E(\xi_{1st}^*)}{b} - \frac{E(\xi_{1st}^{*2})}{b^2}\left(b-\frac{3}{2}\right) + \frac{E(\xi_{0st}^*\xi_{1st}^*)}{b}\right]$$

$$E(t_2^* - \bar{Y}) = \bar{Y}\left[\frac{E(\xi_{0st}^*\xi_{1st}^*)}{b} - \frac{E(\xi_{1st}^{*2})}{b^2}\left(b-\frac{3}{2}\right)\right]$$

$$Bias(t_2^*) = \frac{\bar{Y}}{b}\left[C - \frac{(2b-3)}{2b}B\right] \quad (39)$$

Squaring up to first order of approximation and then taking expectation in equation (38), The *MSE* of t_2^* is given as

$$E(t_2^* - \bar{Y})^2 = \bar{Y}^2 E\left[\xi_{0st}^* + \frac{\xi_{1st}^*}{b} - \frac{\xi_{1st}^{*2}}{b^2}\left(b-\frac{3}{2}\right) + \frac{\xi_{0st}^*\xi_{1st}^*}{b}\right]^2$$

$$E(t_2^* - \bar{Y})^2 = \bar{Y}^2\left[E(\xi_{0st}^{*2}) + \frac{E(\xi_{1st}^{*2})}{b^2} + 2\frac{1}{b}E(\xi_{0st}^*\xi_{1st}^*)\right]$$

$$MSE(t_2^*) = \bar{Y}^2\left[A + \frac{1}{b^2}B + 2\frac{1}{b}C\right] \quad (40)$$

For obtaining the optimal values of b , differentiating equation (40) w.r.t b and equating to zero we have

$$\frac{\partial MSE(t_2^*)}{\partial\, b} = 0$$

$$b_{opt} = -\frac{B}{C}$$

Substituting the optimal value of b in equation (40), we have *MSE* as

$$MSE(t_2^*) = \bar{Y}^2\left[A + \frac{1}{\left(-\frac{B}{C}\right)^2}B + 2\frac{1}{\left(-\frac{B}{C}\right)}C\right] \quad (41)$$

Simplifying equation (41), we have the minimum mean square error of the proposed estimator t_2^*

$$MSE(t_2^*)_{min} = \bar{Y}^2\left[A - \frac{C^2}{B}\right] \tag{42}$$

Case II: When non-response occurs only on study variable, complete information on auxiliary variable and population mean of the auxiliary variable is known. The proposed estimator t_2' is given as

$$t_2' = \bar{y}_{st}^* exp\left[\frac{\bar{x}_{st} - \bar{X}}{\bar{X} + (b' - 1)\bar{x}_{st}}\right] \tag{43}$$

Now, we express the equation (43) in terms of ξ_{0st}^* and ξ_{1st} up to first order of approximation and is given as

$$t_2' = \bar{Y}(1 + \xi_{0st}^*)\left[\frac{\bar{X}(1 + \xi_{1st}) - \bar{X}}{\bar{X} + (b' - 1)\bar{X}(1 + \xi_{1st})}\right]$$

$$t_2' = \bar{Y}(1 + \xi_{0st}^*)\left[\frac{\xi_{1st}}{b'}\left\{1 + \frac{(b-1)^{-1}}{b'}\ \xi_{1st}\right\}\right]$$

$$t_2' - \bar{Y} = \bar{Y}\left[\xi_{0st}^* + \frac{\xi_{1st}}{b'} - \frac{\xi_{1st}^2}{b'^2}\left(b' - \frac{3}{2}\right) + \frac{\xi_{0st}^*\xi_{1st}}{b'}\right] \tag{44}$$

Taking expectation in equation (44), the bias of t_2' to first order of approximation is given as

$$E(t_2' - \bar{Y}) = \bar{Y}\left[E(\xi_{0st}^*) + \frac{E(\xi_{1st})}{b'} - \frac{E(\xi_{1st}^2)}{b'^2}\left(b' - \frac{3}{2}\right) + \frac{E(\xi_{0st}^*\xi_{1st})}{b'}\right]$$

$$E(t_2' - \bar{Y}) = \bar{Y}\left[\frac{E(\xi_{0st}^*\xi_{1st})}{b'} - \frac{E(\xi_{1st}^2)}{b'^2}\left(b' - \frac{3}{2}\right)\right]$$

$$Bias(t_2') = \frac{\bar{Y}}{b'}\left[E - \frac{(2b' - 3)}{2b'}D\right] \tag{45}$$

Squaring and then taking expectation in equation (44), The *MSE* of t_2' is given as

$$E(t_2' - \bar{Y})^2 = \bar{Y}^2 E\left[\xi_{0st}^* + \frac{\xi_{1st}}{b'} - \frac{\xi_{1st}^2}{b'^2}\left(b' - \frac{3}{2}\right) + \frac{\xi_{0st}^*\xi_{1st}}{b'}\right]^2$$

$$E(t_2' - \bar{Y})^2 = \bar{Y}^2\left[E(\xi_{0st}^{*2}) + \frac{E(\xi_{1st}^2)}{b'^2} + 2\frac{1}{b'}E(\xi_{0st}^*\xi_{1st})\right]$$

$$MSE(t_2') = \bar{Y}^2\left[A + \frac{1}{b'^2}D + 2\frac{1}{b'}E\right] \tag{46}$$

For obtaining the optimal values of b' , differentiating equation (46) w.r.t b' and equating to zero we have

$$\frac{\partial MSE(t_2')}{\partial\, b'} = 0$$

$$b'_{opt} = -\frac{D}{E}$$

Substituting the optimal value of b' in equation (46), we have *MSE* as

$$MSE(t_2') = \bar{Y}^2\left[A + \frac{1}{\left(-\frac{D}{E}\right)^2}D + 2\frac{1}{\left(-\frac{D}{E}\right)}E\right] \tag{47}$$

Simplifying equation (47), we have the minimum mean square error of the proposed estimator t_2'

$$MSE(t_2')_{min} = \bar{Y}^2\left[A - \frac{E^2}{D}\right] \tag{48}$$

Proposed Estimator 3: Based on Vishwakarma *et al*. (2016), we propose a stratified exponential estimator in presence of non-response for case I and case II.

Case 1: When non-response occurs both on the study variable and auxiliary variable and population mean of the auxiliary variable is known. The proposed estimator t_3^* is given as

$$t_3^* = c\bar{y}_{st}^* + (1-c)\bar{y}_{st}^* exp\left[\frac{\bar{X} - \bar{x}_{st}^*}{\bar{X} + \bar{x}_{st}^*}\right] \tag{49}$$

Now, we express the equation (49) in terms of ξ_{0st}^* and ξ_{1st}^* up to first order of approximation and is given as

$$t_3^* = \bar{Y}(c + c\xi_{0st}^*) + \bar{Y}(1 + \xi_{0st}^* - c - c\xi_{0st}^*)exp\left[\frac{-\xi_{1st}^*}{2}\left(1 + \frac{\xi_{1st}^*}{2}\right)^{-1}\right]$$

$$t_3^* = \bar{Y}\left(1 + \xi_{0st}^* - \frac{\xi_{1st}^*}{2} + \frac{c\xi_{1st}^*}{2} + \frac{3\xi_{1st}^{*2}}{8} - \frac{3c\xi_{1st}^{*2}}{8} - \frac{\xi_{0st}^*\xi_{1st}^*}{2} + \frac{c\xi_{0st}^*\xi_{1st}^*}{2}\right)$$

$$t_3^* - \bar{Y} = \bar{Y}\left(\xi_{0st}^* + \left(\frac{c}{2} - \frac{1}{2}\right)\xi_{1st}^* + \left(\frac{3}{8} - \frac{3c}{8}\right)\xi_{1st}^{*2} + \left(\frac{c}{2} - \frac{1}{2}\right)\xi_{0st}^*\xi_{1st}^*\right) \tag{50}$$

Taking expectation in equation (50), the bias of t_3^* to first order of approximation is given as

$$E(t_3^* - \bar{Y}) = \bar{Y}\left(E(\xi_{0st}^*) + \left(\frac{c}{2} - \frac{1}{2}\right)E(\xi_{1st}^*) + \left(\frac{3}{8} - \frac{3c}{8}\right)E(\xi_{1st}^{*2}) + \left(\frac{c}{2} - \frac{1}{2}\right)E(\xi_{0st}^*\xi_{1st}^*)\right)$$

$$E(t_3^* - \bar{Y}) = \bar{Y}\left(\left(\frac{3}{8} - \frac{3c}{8}\right)E(\xi_{1st}^{*2}) + \left(\frac{c}{2} - \frac{1}{2}\right)E(\xi_{0st}^*\xi_{1st}^*)\right)$$

$$Bias(t_3^*) = \bar{Y}\left[\left(\frac{3}{8} - \frac{3c}{8}\right)B + \left(\frac{c}{2} - \frac{1}{2}\right)C\right] \tag{51}$$

Squaring up to first order of approximation and then taking expectation in equation (50), The *MSE* of t_3^* is given as

$$E(t_3^* - \bar{Y})^2 = \bar{Y}^2 E\left(\xi_{0st}^* + \left(\frac{c}{2} - \frac{1}{2}\right)\xi_{1st}^* + \left(\frac{3}{8} - \frac{3c}{8}\right)\xi_{1st}^{*2} + \left(\frac{c}{2} - \frac{1}{2}\right)\xi_{0st}^*\xi_{1st}^*\right)^2$$

$$E(t_3^* - \bar{Y})^2 = \bar{Y}^2\left(E(\xi_{0st}^*) + \left(\frac{c}{2} - \frac{1}{2}\right)^2 E(\xi_{1st}^{*2}) + (c-1)E(\xi_{0st}^*\xi_{1st}^*)\right)^2$$

$$MSE(t_3^*) = \bar{Y}^2\left[A + \left(\frac{c^2}{4} - \frac{c}{2} + \frac{1}{4}\right)B + (c-1)C\right] \tag{52}$$

For obtaining the optimal values of c, differentiating equation (52) w.r.t c and equating to zero we have

$$\frac{\partial MSE(t_3^*)}{\partial\, c} = 0$$

$$c_{opt} = \frac{B - 2C}{B}$$

Substituting the optimal value of c in equation (52), we have *MSE* as

$$MSE(t_3^*) = \bar{Y}^2\left[A + \left(\frac{\left(\frac{B-2C}{B}\right)^2}{4} - \frac{\left(\frac{B-2C}{B}\right)}{2} + \frac{1}{4}\right)B + \left(\left(\frac{B-2C}{B}\right) - 1\right)C\right] \tag{53}$$

Simplifying equation (53), we have the minimum mean square error of the proposed estimator t_3^*

$$MSE_{min}(t_3^*) = \bar{Y}^2\left[A - \frac{C^2}{B}\right] \tag{54}$$

Case II: When non-response occurs only on study variable, complete information on auxiliary variable and population mean of the auxiliary variable is known. The proposed estimator t_3' is given as

$$t_3' = c'\bar{y}_{st}^* + (1-c')\bar{y}_{st}^*\left[\frac{\bar{X} - \bar{x}_{st}}{\bar{X} + \bar{x}_{st}}\right] \tag{55}$$

Now, we express the equation (55) in terms of ξ_{0st}^* and ξ_{1st} up to first order of approximation and is given as

$$t_3' = \bar{Y}(c' + c'\xi_{0st}^*) + \bar{Y}(1 + \xi_{0st}^* - c' - c'\xi_{0st}^*)exp\left[\frac{-\xi_{1st}}{2}\left(1 + \frac{\xi_{1st}}{2}\right)^{-1}\right]$$

$$t_3' = \bar{Y}\left(1 + \xi_{0st}^* - \frac{\xi_{1st}}{2} + \frac{c'\xi_{1st}}{2} + \frac{3\xi_{1st}^2}{8} - \frac{3c'\xi_{1st}^2}{8} - \frac{\xi_{0st}^*\xi_{1st}}{2} + \frac{c'\xi_{0st}^*\xi_{1st}}{2}\right)$$

$$t_3' - \bar{Y} = \bar{Y}\left(\xi_{0st}^* + \left(\frac{c'}{2} - \frac{1}{2}\right)\xi_{1st} + \left(\frac{3}{8} - \frac{3c'}{8}\right)\xi_{1st}^2 + \left(\frac{c'}{2} - \frac{1}{2}\right)\xi_{0st}^*\xi_{1st}\right) \tag{56}$$

Taking expectation in equation (56), the bias of t_3' to first order of approximation is given as

$$E(t_3' - \bar{Y}) = \bar{Y}\left(E(\xi_{0st}^*) + \left(\frac{c'}{2} - \frac{1}{2}\right)E(\xi_{1st}) + \left(\frac{3}{8} - \frac{3c'}{8}\right)E(\xi_{1st}^2) + \left(\frac{c'}{2} - \frac{1}{2}\right)E(\xi_{0st}^*\xi_{1st})\right)$$

$$E(t_3' - \bar{Y}) = \bar{Y}\left(\left(\frac{3}{8} - \frac{3c'}{8}\right)E(\xi_{1st}^2) + \left(\frac{c'}{2} - \frac{1}{2}\right)E(\xi_{0st}^*\xi_{1st})\right)$$

$$Bias(t_3') = \bar{Y}\left[\left(\frac{3}{8} - \frac{3c'}{8}\right)D + \left(\frac{c'}{2} - \frac{1}{2}\right)E\right] \tag{57}$$

Squaring up to first order of approximation and then taking expectation in equation (56), The *MSE* of t_3' is given as

$$E(t_3' - \bar{Y})^2 = \bar{Y}^2 E\left(\xi_{0st}^* + \left(\frac{c}{2} - \frac{1}{2}\right)\xi_{1st} + \left(\frac{3}{8} - \frac{3c}{8}\right)\xi_{1st}^2 + \left(\frac{c}{2} - \frac{1}{2}\right)\xi_{0st}^*\xi_{1st}\right)^2$$

$$E(t_3' - \bar{Y})^2 = \bar{Y}^2\left(E(\xi_{0st}^*) + \left(\frac{c}{2} - \frac{1}{2}\right)^2 E(\xi_{1st}^2) + (c-1)E(\xi_{0st}^*\xi_{1st})\right)^2$$

$$MSE(t_3') = \bar{Y}^2\left[A + \left(\frac{c'^2}{4} - \frac{c'}{2} + \frac{1}{4}\right)D + (c'-1)E\right] \tag{58}$$

For obtaining the optimal values of c', differentiating equation (58) w.r.t c' and equating to zero we have

$$\frac{\partial MSE(t_3')}{\partial\, c'} = 0$$

$$c_{opt}' = \frac{D - 2E}{D}$$

Substituting the optimal value of c' in equation (58), we have *MSE* as

$$MSE(t_3') = \bar{Y}^2\left[A + \left(\frac{\left(\frac{D-2E}{D}\right)^2}{4} - \frac{\left(\frac{D-2E}{D}\right)}{2} + \frac{1}{4}\right)D + \left(\left(\frac{D-2E}{D}\right) - 1\right)E\right] \tag{59}$$

Simplifying equation (59), we have the minimum mean square error of the proposed estimator t_3'

$$MSE_{min}(t_3') = \bar{Y}^2\left[A - \frac{E^2}{D}\right] \tag{60}$$

Proposed Estimator 4: Based on Ozel Kadilar (2016), we propose a stratified exponential estimator in presence of non-response for case I and case II.

Case 1: When non-response occurs both on the study variable and auxiliary variable and population mean of the auxiliary variable is known. The proposed estimator t_4^* is given as

$$t_4^* = \bar{y}_{st}^* \left(\frac{\bar{x}_{st}^*}{\bar{X}}\right)^d exp\left[\frac{\bar{X} - \bar{x}_{st}^*}{\bar{X} + \bar{x}_{st}^*}\right] \tag{61}$$

Now, we express the equation (61) in terms of ξ_{0st}^* and ξ_{1st}^* up to first order of approximation and is given as

$$t_4^* = \bar{Y}(1 + \xi_{0st}^*)(1 + \xi_{1st}^*)^d exp\left[\frac{\bar{X} - \bar{X}(1 + \xi_{1st}^*)}{\bar{X} + \bar{X}(1 + \xi_{1st}^*)}\right]$$

$$t_4^* = \bar{Y}(1 + \xi_{0st}^*)(1 + \xi_{1st}^*)^d exp\left[\frac{-\xi_{1st}^*}{2}\left(1 + \frac{\xi_{1st}^*}{2}\right)^{-1}\right]$$

$$t_4^* - \bar{Y} = \bar{Y}\left[\left(\frac{d^2}{2} - d + \frac{3}{8}\right)\xi_{1st}^{2*} + \left(d - \frac{1}{2}\right)\xi_{0st}^*\xi_{1st}^* + \xi_{0st}^* - \frac{\xi_{1st}^*}{2} + d\xi_{1st}^*\right] \tag{62}$$

Taking expectation in equation (62), the bias of t_4^* to first order of approximation to get bias and is given as

$$E(t_4^* - \bar{Y}) = \bar{Y}\left[\begin{array}{c}\left(\frac{d^2}{2} - d + \frac{3}{8}\right)E(\xi_{1st}^{2*}) + \left(d - \frac{1}{2}\right)E(\xi_{0st}^*\xi_{1st}^*) \\ +E(\xi_{0st}^*) - \frac{E(\xi_{1st}^*)}{2} + dE(\xi_{1st}^*)\end{array}\right]$$

$$E(t_4^* - \bar{Y}) = \bar{Y}\left[\left(\frac{d^2}{2} - d + \frac{3}{8}\right)E(\xi_{1st}^{2*}) + \left(d - \frac{1}{2}\right)E(\xi_{0st}^*\xi_{1st}^*)\right]$$

$$Bias(t_4^*) = \bar{Y}\left[\left(\frac{d^2}{2} - d + \frac{3}{8}\right)B + \left(d - \frac{1}{2}\right)C\right] \tag{63}$$

Squaring up to first order of approximation and then taking expectation in equation (62), The *MSE* of t_4^* is given as

$$E(t_4^* - \bar{Y})^2 = \bar{Y}^2 E\left[\left(\frac{d^2}{2} - d + \frac{3}{8}\right)\xi_{1st}^{2*} + \left(d - \frac{1}{2}\right)\xi_{0st}^*\xi_{1st}^* + \xi_{0st}^* - \frac{\xi_{1st}^*}{2} + d\xi_{1st}^*\right]^2$$

$$E(t_4^* - \bar{Y})^2 = \bar{Y}^2\left[(\xi_{0st}^{*2}) + \left(d^2 - d + \frac{1}{4}\right)E(\xi_{1st}^{2*}) + (2d - 1)E(\xi_{0st}^*\xi_{1st}^*)\right]$$

$$MSE(t_4^*) = \bar{Y}^2\left[A + \left(d^2 - d + \frac{1}{4}\right)B + (2d - 1)C\right] \tag{64}$$

For obtaining the optimal values of d, differentiating equation (64) w.r.t d and equating to zero we have

$$\frac{\partial MSE(t_4^*)}{\partial\, d} = 0$$

$$d_{opt} = \frac{B - 2C}{2B}$$

Substituting the optimal value of d in equation (64), we have *MSE* as

$$MSE(t_4^*) = \bar{Y}^2 \left[A + \left(\left(\frac{B - 2C}{2B} \right)^2 - \left(\frac{B - 2C}{2B} \right) + \frac{1}{4} \right) B + \left(2 \left(\frac{B - 2C}{2B} \right) - 1 \right) C \right] \quad (65)$$

Simplifying equation (65), we get the minimum mean square error of the proposed estimator t_4^*

$$MSE_{min}(t_4^*) = \bar{Y}^2 \left[A - \frac{C^2}{B} \right] \quad (66)$$

Case II: When non-response occurs only on study variable, complete information on auxiliary variable and population mean of the auxiliary variable is known. The proposed estimator t_4' is given as

$$t_4' = \bar{y}_{st}^* \left(\frac{\bar{x}_{st}}{\bar{X}} \right)^{d'} \left[\frac{\bar{X} - \bar{x}_{st}}{\bar{X} + \bar{x}_{st}} \right] \quad (67)$$

Now, we express the equation (67) in terms of ξ_{0st}^* and ξ_{1st} up to first order of approximation and is given as

$$t_4' = \bar{Y}(1 + \xi_{0st}^*)(1 + \xi_{1st})^{d'} exp \left[\frac{\bar{X} - \bar{X}(1 + \xi_{1st})}{\bar{X} + \bar{X}(1 + \xi_{1st})} \right]$$

$$t_4' = \bar{Y}(1 + \xi_{0st}^*)(1 + \xi_{1st})^{d'} exp \left[\frac{-\xi_{1st}}{2} \left(1 + \frac{\xi_{1st}}{2} \right)^{-1} \right]$$

$$t_4' - \bar{Y} = \bar{Y} \left[\left(\frac{d'^2}{2} - d' + \frac{3}{8} \right) \xi_{1st}^2 + \left(d' - \frac{1}{2} \right) \xi_{0st}^* \xi_{1st} + \xi_{0st}^* - \frac{\xi_{1st}^*}{2} + d' \xi_{1st} \right] \quad (68)$$

Taking expectation in equation (68), the bias of t_4' to first order of approximation to get bias and is given as

$$E(t_4' - \bar{Y}) = \bar{Y} \left[\begin{array}{c} \left(\frac{d'^2}{2} - d' + \frac{3}{8} \right) E(\xi_{1st}^2) + \left(d' - \frac{1}{2} \right) E(\xi_{0st}^* \xi_{1st}) \\ + E(\xi_{0st}^*) - \frac{E(\xi_{1st})}{2} + d' E(\xi_{1st}) \end{array} \right]$$

$$E(t_4' - \bar{Y}) = \bar{Y} \left[\left(\frac{d'^2}{2} - d' + \frac{3}{8} \right) E(\xi_{1st}^2) + \left(d' - \frac{1}{2} \right) E(\xi_{0st}^* \xi_{1st}) \right]$$

$$Bias(t_4') = \bar{Y} \left[\left(\frac{d'^2}{2} - d' + \frac{3}{8} \right) D + \left(d' - \frac{1}{2} \right) E \right] \quad (69)$$

Squaring up to first order of approximation and then taking expectation in equation (68), The *MSE* of t_4' is given as

$$E(t_4' - \bar{Y})^2 = \bar{Y}^2 E\left[\left(\frac{d'^2}{2} - d' + \frac{3}{8}\right)\xi_{1st}^2 + \left(d' - \frac{1}{2}\right)\xi_{0st}^*\xi_{1st} + \xi_{0st}^* - \frac{\xi_{1st}}{2} + d'\xi_{1st}\right]^2$$

$$E(t_4' - \bar{Y})^2 = \bar{Y}^2\left[(\xi_{0st}^{*2}) + \left(d'^2 - d' + \frac{1}{4}\right)E(\xi_{1st}^2) + (2d' - 1)E(\xi_{0st}^*\xi_{1st})\right]$$

$$MSE(t_4') = \bar{Y}^2\left[A + \left(d'^2 - d' + \frac{1}{4}\right)D - (2d' - 1)E\right] \tag{70}$$

For obtaining the optimal values of d' , differentiating equation (70) w.r.t d' and equating to zero we have

$$\frac{\partial MSE(t_{14}')}{\partial\, d'} = 0$$

$$d'_{opt} = \frac{D - 2E}{2D}$$

Substituting the optimal value of d' in equation (70), we have *MSE* as

$$MSE(t_4') = \bar{Y}^2\left[A + \left(\left(\frac{D - 2E}{2D}\right)^2 - \left(\frac{D - 2E}{2D}\right) + \frac{1}{4}\right)D - \left(2\left(\frac{D - 2E}{2D}\right) - 1\right)E\right] \tag{71}$$

Simplifying equation (71), we have the minimum mean square error of the proposed estimator t_4'

$$MSE_{min}(t_4') = \bar{Y}^2\left[A - \frac{E^2}{D}\right] \tag{72}$$

Interesting Note: We have proposed four estimators for case I and four estimators for case II having same MSE respectively and are given as

$$MSE_{min}(t_i^*) = \bar{Y}^2\left[A - \frac{C^2}{B}\right]; i = 1 - 4 \tag{73}$$

$$MSE_{min}(t_i') = \bar{Y}^2\left[A - \frac{E^2}{D}\right]; i = 1 - 4 \tag{74}$$

6. Efficiency comparison

Now we will investigate the efficiencies of $t_i^*, i = 1 - 4$ and $t_i', i = 1 - 4$ given in equation (73) and (74) with various estimators from the literature.

6.1. Efficiency comparison for case I

Using equations (12), (14), (18), (22) and (73) we find the efficiency conditions of $t_i^*, i = 1 - 4$ as follows

1. t_i^*, i $= 1 - 4$ Perform better than $\bar{y}^*$ if:

$$MSE(t_i^*) < MSE(\bar{y}^*)$$

$$\bar{Y}^2\left[A - \frac{C^2}{B}\right] < \bar{Y}^2 A$$

$$-\frac{C^2}{B} < 0$$

which is obviously true because $C > 0$ and $B > 0$

2. $t_i^*,\ \mathrm{i} = 1 - 4$ Perform better than $\bar{y}_{SR}^*$ if:

$$MSE(t_i^*) < MSE(\bar{y}_{SR}^*)$$

$$\bar{Y}^2\left[A - \frac{C^2}{B}\right] < \sum_{h=1}^{L} W_h^2\,[A_h + R_h^2 B_h - 2R_h C_h\,]$$

$$\bar{Y}^2\left[A - \frac{C^2}{B}\right] - \sum_{h=1}^{L} W_h^2\,\bar{Y}_h^2[A_h + B_h - 2C_h\,] < 0$$

3. $t_i^*,\ \mathrm{i} = 1 - 4$ Perform better than $\bar{y}_{CR}^*$ if:

$$MSE(t_i^*) < MSE(\bar{y}_{CR}^*)$$

$$\bar{Y}^2\left[A - \frac{C^2}{B}\right] < \bar{Y}^2[A + B - 2C\,]$$

$$\left[2C - \frac{C^2}{B} - B\right] < 0$$

4. $t_i^*,\ \mathrm{i} = 1 - 4$ Perform better than $\bar{y}_{ok}^{*(i)}$ if:

$$MSE(t_i^*) < MSE\left(\bar{y}_{ok}^{*(i)}\right)$$

$$\bar{Y}^2\left[A - \frac{C^2}{B}\right] < \sum_{h=1}^{L} W_h^2\,\bar{Y}_h^2\left[A_h + \frac{1}{4}\varphi_{hi}^2 B_h - \varphi_{hi} C_h\,\right]$$

$$\bar{Y}^2\left[A - \frac{C^2}{B}\right] - \sum_{h=1}^{L} W_h^2\,\bar{Y}_h^2\left[A_h + \frac{1}{4}\varphi_{hi}^2 B_h - \varphi_{hi} C_h\,\right] < 0$$

6.2. Efficiency comparison for case II

Using equations (12), (16), (20), (24) and (74) we find the efficiency conditions of $t_i', i = 1 - 4$ as follows.

1. $t_i',\ i = 1-4$ Perform better than $\bar{y}^*$ if:

$$MSE(t_i') < MSE(\bar{y}^*)$$

$$\bar{Y}^2\left[A - \frac{E^2}{D}\right] < \bar{Y}^2 A$$

$$-\frac{E^2}{D} < 0$$

which is obviously true because $E > 0$ and $D > 0$

2. $t_i',\ i = 1-4$ Perform better than $\bar{y}_{SR}'$ if:

$$MSE(t_i') < MSE(\bar{y}_{SR}')$$

$$\bar{Y}^2\left[A - \frac{E^2}{D}\right] < \sum_{h=1}^{L} W_h^2\,\bar{Y}_h^2[A_h + D_h - 2E_h\,]$$

$$\bar{Y}^2\left[A - \frac{E^2}{D}\right] - \sum_{h=1}^{L} W_h^2\,\bar{Y}_h^2[A_h + D_h - 2E_h\,] < 0$$

3. $t_i',\ i = 1-4$ Perform better than $\bar{y}_{CR}'$ if:

$$MSE(t_i') < MSE(\bar{y}_{CR}')$$

$$\bar{Y}^2\left[A - \frac{E^2}{D}\right] < \bar{Y}^2 \sum_{h=1}^{L} W_h^2\,[A + D - 2E\,]$$

$$\left[2E - \frac{E^2}{D} - D\right] < 0$$

4. $t_i',\ i = 1-4$ Perform better than $\bar{y}_{ok}^{\prime(i)}$ if:

$$MSE(t_i') < MSE\left(\bar{y}_{ok}^{\prime(i)}\right)$$

$$\bar{Y}^2\left[A - \frac{E^2}{D}\right] < \sum_{h=1}^{L} W_h^2\,\bar{Y}_h^2\left[A_h + \frac{1}{4}\varphi_{hi}^2 D_h - \varphi_{hi}E_h\,\right]$$

$$\bar{Y}^2\left[A - \frac{E^2}{D}\right] - \sum_{h=1}^{L} W_h^2\,\bar{Y}_h^2\left[A_h + \frac{1}{4}\varphi_{hi}^2 D_h - \varphi_{hi}E_h\,\right] < 0$$

7. Empirical study

To illustrate the performance of the proposed estimators $t_i^*, i = 1-4$ and $t_i', i = 1-4$ over other existing estimators using the real data set given as

Data Set Source: Koyuncu and Kadilar (2009)

We consider number of teachers as study variable *(Y)* and number of classes in primary and secondary schools as auxiliary variable *(X),* for 923 districts at six 6 regions (1) Marmara; (2) Agean; (3) Mediterranean; (4) Central Anatolia; (5) Black Sea; and (6) East and Southeast Anatolia in Turkey in 2007.

Table 1: The descriptive statistics

h	1	2	3	4	5	6
N_h	127	117	103	170	205	201
n_h	31	21	29	38	22	39
S_{hy}	883.84	644.92	1033.40	810.58	403.65	771.72
S_{hx}	555.58	365.46	612.95	458.03	260.85	397.05
$\bar{Y}_h$	703.74	413	573.17	424.66	267.03	393.84
$\bar{X}_h$	498.28	318.33	431.36	311.32	227.2	313.71
ρ_{hxy}	0.979	0.976	0.984	0.983	0.964	0.983
$\mathbf{W_{2h} = 10\%\ Non - response}$						
$S_{hy(2)}$	510.57	386.77	1872.88	1603.3	264.19	497.84
$S_{hx(2)}$	303.92	278.51	960.71	821.29	190.85	287.99
$\rho_{hxy(2)}$	0.9931	0.9871	0.9972	0.9942	0.985	0.9647
$\mathbf{W_{2h} = 20\%\ Non - response}$						
$S_{hy(2)}$	396.77	406.15	1654.40	1333.35	335.83	903.91
$S_{hx(2)}$	244.56	274.42	965.42	680.28	214.49	469.86
$\rho_{hxy(2)}$	0.9898	0.9798	0.9846	0.9940	0.9818	0.9874
$\mathbf{W_{2h} = 30\%\ Non - response}$						
$S_{hy(2)}$	500.26	356.95	1383.70	1193.47	289.41	825.24
$S_{hx(2)}$	284.44	247.63	811.21	631.28	188.30	437.90
$\rho_{hxy(2)}$	0.9739	0.9793	0.9839	0.9904	0.9799	0.9829

Table 2: The percent relative efficiency of the existing and proposed estimators with respect to Hansen and Hurwitz estimator for Case 1

$\mathbf{W_{2h} = 10\%\ Non - response}$					
Estimators	*K*=2	*K*=2.5	*K*=3	K=3.5	*K*=4
$\bar{y}^*$	100	100	100	100	100
$\bar{y}^*_{SR}$	1021.72	1022.33	1019.68	1023.33	1023.73
$\bar{y}^*_{CR}$	1031.44	1023.39	1013.42	1010.73	1005.66
$\bar{y}^{*(1)}_{ok}$	245.18	244.57	243.28	243.58	243.19
$\bar{y}^{*(2)}_{ok}$	244.13	243.52	242.24	242.54	242.14
$\bar{y}^{*(3)}_{ok}$	237.43	236.89	235.70	236.04	235.69
$\bar{y}^{*(4)}_{ok}$	244.56	243.95	242.68	242.98	242.59
$\bar{y}^{*(5)}_{ok}$	245.05	244.44	243.15	243.46	243.06
$\bar{y}^{*(6)}_{ok}$	244.39	243.78	242.50	242.80	242.41

$\bar{y}_{ok}^{*(7)}$	237.27	236.74	235.55	235.89	235.55
$\bar{y}_{ok}^{*(8)}$	1021.71	1022.35	1019.68	1023.33	1023.72
$t_i^*;$ $i = 1 - 4$	2564.04	2644.96	2717.59	2803.98	2884.16
$\mathbf{W_{2h} = 20\%\ Non - response}$					
$\bar{y}^*$	100.00	100.00	100.00	100.00	100.00
$\bar{y}_{SR}^*$	1021.71	1037.16	1040.68	1043.60	1046.05
$\bar{y}_{CR}^*$	1034.36	1043.18	1041.25	1039.66	1038.42
$\bar{y}_{ok}^{*(1)}$	243.05	245.32	222.94	244.78	244.58
$\bar{y}_{ok}^{*(2)}$	242.00	244.26	244.43	243.72	243.52
$\bar{y}_{ok}^{*(3)}$	235.28	237.51	237.24	237.02	236.84
$\bar{y}_{ok}^{*(4)}$	242.43	244.71	244.42	244.18	243.98
$\bar{y}_{ok}^{*(5)}$	242.92	245.29	244.90	244.66	244.46
$\bar{y}_{ok}^{*(6)}$	242.26	244.53	244.23	243.99	243.79
$\bar{y}_{ok}^{*(7)}$	235.13	237.35	237.09	236.87	236.69
$\bar{y}_{ok}^{*(8)}$	1021.71	1037.16	1040.68	1043.60	1046.05
$t_i^*;$ $i = 1 - 4$	2618.32	2748.12	2839.69	2922.43	2997.38
$\mathbf{W_{2h} = 30\%\ Non - response}$					
$\bar{y}^*$	100.00	100.00	100.00	100.00	100.00
$\bar{y}_{SR}^*$	1027.85	1030.21	1032.04	1033.48	1034.66
$\bar{y}_{CR}^*$	1043.22	1040.35	1038.15	1036.42	1035.02
$\bar{y}_{ok}^{*(1)}$	245.42	245.01	244.70	244.45	244.25
$\bar{y}_{ok}^{*(2)}$	244.36	243.95	243.63	243.35	243.18
$\bar{y}_{ok}^{*(3)}$	237.49	237.09	236.78	236.54	236.34
$\bar{y}_{ok}^{*(4)}$	244.80	244.39	244.08	243.84	243.64
$\bar{y}_{ok}^{*(5)}$	245.29	244.88	244.57	244.32	244.12
$\bar{y}_{ok}^{*(6)}$	244.62	244.21	243.90	243.66	243.46
$\bar{y}_{ok}^{*(7)}$	237.33	236.93	236.63	236.39	236.19
$\bar{y}_{ok}^{*(8)}$	1027.85	1030.21	1032.04	1033.48	1034.66
$t_i^*;$ $i = 1 - 4$	2595.87	2664.30	2722.05	2771.39	2814.02

Table 3: The percent relative efficiency of the existing and proposed estimators with respect to Hansen and Hurwitz estimator for Case II

$\mathbf{W_{2h} = 10\%\ Non - response}$					
Estimators	*K*=2	*K*=2.5	*K*=3	*K*=3.5	*K*=4
$\bar{y}^*$	100.00	100.00	100.00	100.00	100.00
$\bar{y}_{SR}'$	407.23	330.47	283.52	253.68	231.73
$\bar{y}_{CR}'$	411.50	333.06	285.29	255.00	232.77
$\bar{y}_{ok}^{\prime(1)}$	199.01	185.15	174.15	166.53	159.97

$\bar{y}_{ok}^{\prime(2)}$	198.44	184.69	173.77	166.20	159.69
$\bar{y}_{ok}^{\prime(3)}$	194.63	181.63	171.25	164.05	157.82
$\bar{y}_{ok}^{\prime(4)}$	198.66	184.87	173.92	166.33	159.80
$\bar{y}_{ok}^{\prime(5)}$	198.94	185.09	174.11	166.48	159.93
$\bar{y}_{ok}^{\prime(6)}$	198.57	184.80	173.86	166.28	159.75
$\bar{y}_{ok}^{\prime(7)}$	194.54	181.56	171.19	163.99	157.77
$\bar{y}_{ok}^{\prime(8)}$	407.23	330.47	283.52	253.68	231.73
$t_i^{\prime}$; $i = 1-4$	504.55	386.38	320.64	280.77	252.63
$\mathbf{W_{2h} = 20\%\ Non-response}$					
$\bar{y}^{*}$	100.00	100.00	100.00	100.00	100.00
$\bar{y}_{SR}^{\prime}$	314.60	255.12	221.46	199.80	184.70
$\bar{y}_{CR}^{\prime}$	316.89	256.47	222.37	200.48	185.23
$\bar{y}_{ok}^{\prime(1)}$	181.76	166.94	156.66	149.12	143.35
$\bar{y}_{ok}^{\prime(2)}$	181.33	166.61	156.40	148.91	143.17
$\bar{y}_{ok}^{\prime(3)}$	178.45	164.43	154.67	147.47	141.95
$\bar{y}_{ok}^{\prime(4)}$	181.50	166.74	156.50	148.99	143.24
$\bar{y}_{ok}^{\prime(5)}$	181.71	166.90	156.63	149.09	143.33
$\bar{y}_{ok}^{\prime(6)}$	181.43	166.68	156.46	148.96	143.21
$\bar{y}_{ok}^{\prime(7)}$	178.38	164.38	154.63	147.44	141.92
$\bar{y}_{ok}^{\prime(8)}$	314.60	255.12	221.46	199.80	184.70
$t_i^{\prime}$; $i = 1-4$	363.59	282.6541	239.75	213.16	195.08
$\mathbf{W_{2h} = 30\%\ Non-response}$					
$\bar{y}^{*}$	100.00	100.00	100.00	100.00	100.00
$\bar{y}_{SR}^{\prime}$	269.24	220.20	193.19	176.09	164.30
$\bar{y}_{CR}^{\prime}$	270.79	221.09	193.80	176.55	164.66
$\bar{y}_{ok}^{\prime(1)}$	170.80	156.24	146.65	139.85	134.78
$\bar{y}_{ok}^{\prime(2)}$	170.45	155.99	146.45	139.69	134.65
$\bar{y}_{ok}^{\prime(3)}$	168.09	154.27	145.11	138.60	133.73
$\bar{y}_{ok}^{\prime(4)}$	170.58	156.09	146.53	139.75	134.70
$\bar{y}_{ok}^{\prime(5)}$	170.75	156.21	146.62	139.83	134.77
$\bar{y}_{ok}^{\prime(6)}$	170.53	156.04	146.50	139.73	134.68
$\bar{y}_{ok}^{\prime(7)}$	168.04	154.23	145.08	138.57	133.70
$\bar{y}_{ok}^{\prime(8)}$	269.24	220.20	193.19	176.09	164.30
$t_i^{\prime}$; $i = 1-4$	301.25	238.17	205.20	184.93	171.21

Table 2 presents the empirical comparison based on percent relative efficiencies (*PREs*) of the proposed class of combined exponential type of estimators t_1^* , t_2^* , t_3^* and t_4^* and it clearly shows that the proposed estimators are more efficient than the Hansen and

Hurwitz estimator as well as from the other existing estimators taken in literature when non-response occurs both on the study variable and on the auxiliary variable and the population mean of the auxiliary variable is known. The proposed estimators t_1^* , t_2^* , t_3^* and t_4^* are equally efficient. The *PREs* of the proposed estimators $t_i^*, i = 1 - 4$ at 10% non-response rate and at $K_h = 2$ is (2564.04). Similarly, the *PREs* of the proposed estimators $t_i^*, i = 1 - 4$ at 20% non-response rate and at $K_h = 2$ is (2618.32) and also, the *PREs* of the proposed estimators $t_i^*, i = 1 - 4$ at 30% non-response rate and at $K_h = 2$ is (2595.87). Further an increasing trend has been observed in *PREs* with increase in the value of K_h at 10%, 20% and 30% non-response rates.

Table 3 presents the empirical comparison based on percent relative efficiencies *PRE*s of the proposed class of combined exponential type of estimators t_1' , t_2' , t_3' and t_4' clearly shows that the proposed estimators are more efficient than the Hansen and Hurwitz estimator as well as from the other existing estimators taken in literature when non-response occurs only on the study variable, complete information on the auxiliary variable and the population mean of the auxiliary variable is known. The proposed estimators t_1' , t_2' , t_3' and t_4' are equally efficient. The *PREs* of the proposed estimators $t_i', i = 1 - 4$ at 10% non-response rate and at $K_h = 2$ is (504.55). Similarly, the *PREs* of the proposed estimators $t_i', i = 1 - 4$ at 20% non-response rate and at $K_h = 2$ is (363.59) and also, the *PREs* of the proposed estimators $t_i', i = 1 - 4$ at 30% non-response rate and at $K_h = 2$ is (301.25). Further a decreasing trend has been observed in *PREs* with increase in the value of K_h at 10%, 20% and 30% non-response rates.

8. Conclusion

In this paper, we have discussed the problem of estimating the population mean using auxiliary information in stratified random sampling under non-response. The situation of non-response is examined under two cases; Case I: when non-response occurs both on study variable and auxiliary variable and population mean of the auxiliary variable is known; Case II: when non-response occurs only on study variable, complete information on auxiliary variable and population mean of auxiliary variable is known. Four exponential estimators t_1^* , t_2^* , t_3^* and t_4^* have been proposed in the Case I of non-response when non-response occurs both on study variable and auxiliary variable and population mean of the auxiliary variable is known. Similarly, four exponential estimators t_1' , t_2' , t_3' and t_4' have been proposed in Case II, when non-response occurs only on study variable, complete information on auxiliary variable and population mean of auxiliary variable is known. Expression of bias and mean square error of the proposed estimators $t_i^*, i = 1 - 4$ and $t_i', i = 1 - 4$ are obtained separately for all the proposed estimators. Optimum conditions of the proposed estimators are obtained at which the mean squared error of the proposed estimators $t_i^*, i = 1 - 4$ and $t_i', i = 1 - 4$ are minimized. The proposed estimators compared with Hansen and Hurwitz (1946) estimator and some other existing estimators theoretically. We have also carried out empirical study to validate the performance of the proposed estimators $t_i^*, i = 1 - 4$ and $t_i', i = 1 - 4$ over the existing estimators. Thus, the proposed study is recommended for its use in practice.

Acknowledgement

The authors are very thankful to the Editor-in-Chief and the anonymous learned referees for their valuable suggestions regarding the improvement of the paper.

References

Chaudhary, M. K. and Kumar, A. (2015). Estimating the population mean in stratified sampling using two phase sampling in the presence of non-response. *World Applied Sciences Journal*, **33**, 874–82.

Chaudhury, M. K., Singh, R., Shukla, R. K., Kumar, M., and Smarandache, F. (2009). A family of estimators for estimating population mean in stratified sampling under non-response. *Pakistan Journal of Statistics and Operation Research*, **5**, 47-54.

Cochran, W. G. (1977). *Sampling Techniques*. 3rd Edition, John Wiley, New York.

Hansen, M. H. and Hurwitz, W. N. (1946). The problem of non-response in sample surveys. *Journal of the American Statistical Association*, **41**, 517-529.

Ismail, M. and Shahbaz, M. Q. (2017). Combination of ratio and regression estimator of population mean in presence of non–response. *Gazi University Journal of Sci*ence, **30**, 634-642.

Khare, B. B. and Srivastava, S. (1997). Transformed ratio-type estimators for the population mean in the presence of non-response. *Communications in Statistics-Theory and Methods*, **26**, 1779- 1791.

Kumar, S., Kour, S. P., and Sharma, V. (2022). Modified exponential estimators using auxiliary information under response and non-response. *Revista Investigación Operacional*, **43**, 491-504.

Onyeka, A. C., Ogbumuo, D. T., and Izunobi, C. (2019). Estimation of population mean in stratified random sampling when using auxiliary information in the presence of non-response. *Far East Journal of Theoretical Statistics*, **55**, 151-167.

Rao, P. S. R. S. (1986). Ratio estimation with sub-sampling the non-respondents. *Survey Methodology*, **12**, 217-230.

Saleem, I., Sanaullah, A., and Hanif, M. (2018). A generalized class of estimators for estimating population mean in the presence of non-response. *Journal of Statistical Theory and Applications*, **17**, 616-626.

Sanaullah, A., Noor-Ul-Amin, M., and Hanif, M. (2015). Generalized exponential-type ratio-cum-ratio and product-cum-product estimators for population mean in the presence of non-response under stratified two-phase random sampling. *Pakistan Journal of Statistics and Operation Research*, **31**, 71-94.

Shabbir, J., Gupta, S., and Ahmed, S. (2019). A generalized class of estimators under two-phase stratified sampling for non-response. *Communications in Statistics-Theory and Methods*, **48**, 3761- 3777.

Singh, H. P. and Kumar, S. (2008). A general family of estimators of finite population ratio, product and mean using two-phase sampling scheme in presence of non-response. *Journal of Statistical Theory and Practice*, **2**, 677-692.

Singh, H. P. and Kumar, S. (2009). A general procedure of estimating the population mean in the presence of non-response under double sampling using auxiliary information. *SORT*, **33**, 71-84.

Singh, N. and Vishwakarma, G. K. (2019). A generalized class of estimator of population mean with the combined effect of measurement errors and non-response in sample survey. *Investigación Operacional*, **40**, 275-285.

Singh, R. and Sharma, P. (2015). Method of estimation in the presence of non-response and measurement errors simultaneously. *Journal of Modern Applied Statistical Methods*, **14**, 107-114.

Singh, R., Singh, R., and Bouza, C. N. (2018). Effect to measurement error and non-response on estimation of population mean. *Revista Investigación Operacional*, **39**, 108-120.

Ozel, G. K. (2016). A new exponential type estimator for the population mean in simple random sampling. *Journal of Modern Applied Statistical Methods*, **15**, 207-214.

Upadhyaya, L. N., Singh, H. P., Chatterjee, S., and Yadav, R. (2011). Improved ratio and product exponential type estimators. *Journal of Statistical Theory and Practice*, **5**, 285–302.

Vishwakarma, G. K., Singh, R., Gupta, P. C., and Pareek, S. (2016). Improved ratio and product type estimators of finite population mean in simple random sampling. *Revista Investigacion Operacional*, **37**, 70–76.

Wani, Z. H., Rizvi, S. E. H., Sharma, M., and Bhat, M. I. J. (2021). Efficient class of combined ratio type estimators for estimating the population mean under non-response. *International Journal of Scientific Research in Mathematical and Statistical Sciences*, **8**, 01-06.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 21, No. 2, 2023 (New Series), pp 99-106
http://www.ssca.org.in/journal

COVID-19 Cumulative Death Prediction in Two Most Populated Countries by Fitting ARIMA Model and Linear Regression

Shagun Sachdeva and Ravinder Singh
Department of Statistics, Central University of Haryana, Haryana, India

Received: 26 August 2022; Revised: 01 October 2022; Accepted: 08 November 2022

Abstract

COVID-19 (Coronavirus) has caused widespread disruption and hindered economic growth worldwide. Since the entire world has been hit by two dangerous waves of this epidemic, it has become increasingly vital for us to analyze COVID-19 casualties in order to forecast our future days. As a result, in this work, an attempt has been made to do a time series analysis and fit linear regression on the cumulative death of the two most populated countries, China and India. The research utilizes a simple yet powerful and objective approach, called autoregressive integrated moving average (ARIMA) to forecast the number of cumulative deaths. We have also fitted linear regression on the data to predict future values. The forecasted values have also been compared with the original cumulative death values. In conclusion, ARIMA model forecasted better results in comparison with regression model. As a result, ARIMA(0,2,1) and ARIMA(1,2,0) turns out to be the best model for China and India respectively. So, in the future, the government and health personnel can use these models to take desirable action to control the death count.

Key words: ARIMA model; Linear regression; Epidemic forecast; Cumulative deaths.

AMS Subject Classifications: 62J05, 62M10

1. Introduction

Coronavirus disease 2019 (COVID-19) is a serious, long-lasting contagious disease caused by the severe acute respiratory syndrome coronavirus-2 (SARS-CoV-2) that first arose in December 2019 in China. The pandemic has infected 511,749,262 people in 195 nations around the World, with 6,228,622 cases of death as of 28th April 2020. COVID-19's first case was found on 12th December 2019 which is now spreading exponentially and causing a large number of casualties. COVID-19 has been identified as the third largest cause of death in the World. More than half of the world's population lives in one of the 10 most populous countries in the world. China is the most populous country on earth with a population of around 1,397,715,000. The United States has the highest number of COVID-19

Corresponding Author: Ravinder Singh
Email: ravinderstats@cuh.ac.in

deaths (1,019,774), followed by India (523,693). Even though China is among the world's most populous countries, the number of fatalities by COVID-19 in China is lower. How a country faces such tough situations as COVID-19 gives an idea about the development of the country in terms of the medical sector, so it is important to study different countries and compare them.

To compare the most populated country in terms of COVID-19 deaths, we have used ARIMA model and simple linear regression. The regression and SEIR model were fitted on COVID-19 cases previously in (Panday *et al.*, 2020) and the SEIR model predicted closer results. Katoch and Sindhu (2021) also proposed a time series model based on genetic programming for the analysis of confirmed death and cases across the three most pretentious states of India - Maharashtra, Andhra Pradesh, Tamil Nadu, and Karnataka as well as for the whole India. Ding *et al.* (2020) studied the epidemic data from February 24 to March 30, 2020, and concluded that an inflection point was expected in early April in Italy. Bayyurt and Bayyurt (2020) compared the lag between COVID-19 cases and deaths with the help of the ARIMA model. Gambhir *et al.*(2020) applied regression on COVID-19 to study future patterns. An attempt was made by Hengjian and Tao (2020) to fit Non-linear regression on COVID-19 data. Batista (2020) studied the second phase of the coronavirus COVID-19 epidemic by the logistic model. It is vital to model and predict the deaths to deal with their consequences. Forecasting future COVID-19 deaths using statistical models is critical for breaking the transmission cycle in highly populated nations like China and India.

2. Mathematical background

We have used secondary data from World Health Organization (WHO) COVID-19 dashboard from 1st April 2021 to 30th June 2021 considering the number of cumulative deaths per day. The data includes confirmed cases and deaths along with their cumulative counts of all the countries. Excel was used in building the database of time series.

2.1. ARIMA model

ARIMA model in equations:

$$Y_t = \alpha + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_n Y_{t-n} \epsilon_t + \phi_1 \epsilon_{t-1} + \phi_2 \epsilon_{t-2} + \cdots + \phi_q \epsilon_{t-q} \tag{1}$$

ARIMA, short for 'Auto-Regressive Integrated Moving Average' explains the time series based on its previous values (Box *et al.*, 2015). In ARIMA *(p,d,q)* the *'p'* denotes the order of the 'Auto Regressive' (AR) term which refers to the number of lags of Y, which is used as predictors. The 'Moving Average' (MA) term's order is *'q'* showing the number of lagged forecast errors that can be included in the ARIMA model. The value of *'d'* is the smallest number of differencing required to get the series stationary. There are different ARIMA models that we can fit, to select the best one we can choose the criteria like AIC (Akaike Information Criteria), and log-likelihood. To check the accuracy of the model, we have calculated the change percentage of both ARIMA and linear regression and compared them. We have also plotted the Auto Correlation Function (ACF) and Partial Auto Correlation Function (PACF) of residuals in order to check the normality.

2.1.1. Analysis of data - I

An attempt has been made in this study to analyze the data collected with the COVID-19 cumulative deaths in China and India. In this case, the methodology described below was applied, and conclusions were drawn from the study. The study was of about 91 days covering second wave data starting from 1st April 2021 to 30th June 2021. During the second wave, India faced more death losses when compared to China. The data is plotted in time series shown in Figure 1 and 2, demonstrating stochastic trends.

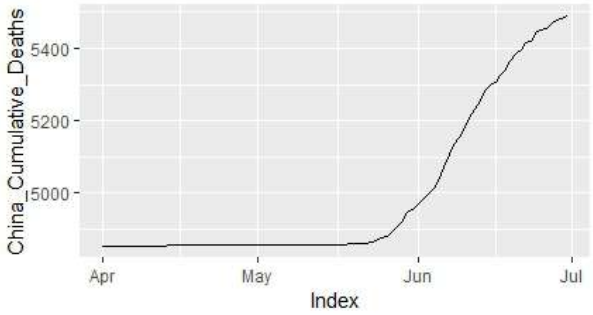


Figure 1: China cumulative death plot

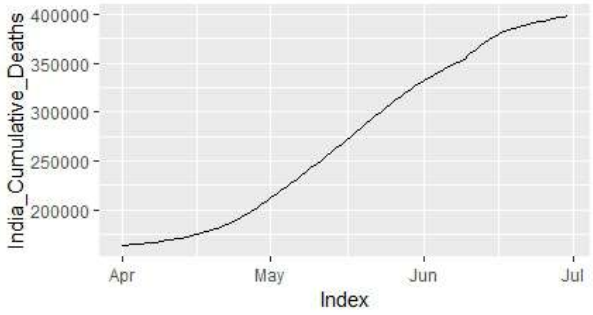


Figure 2: India cumulative death plot

Figure 1 and 2 depict that the COVID-19 peak started to reach India in April whereas for China it was after two months in June. With the use of software R, we plotted the Auto Correlation Function (ACF) and Partial Auto Correlation Function (PACF) to test the stationarity of China and India's time series graphically, as shown in Figure 3 to 6.

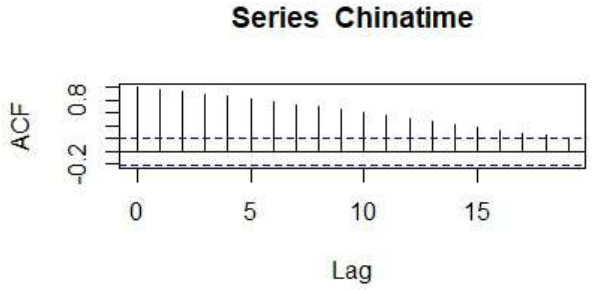


Figure 3: ACF China

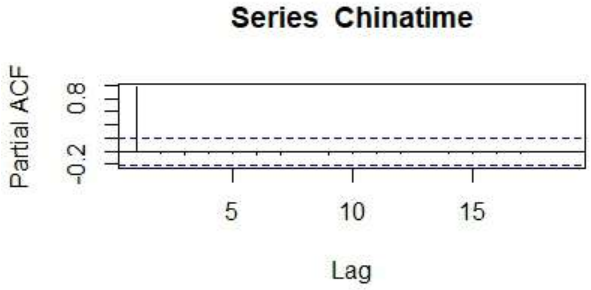


Figure 4: PACF China

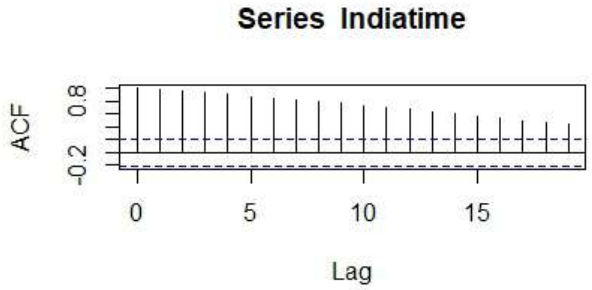


Figure 5: ACF India

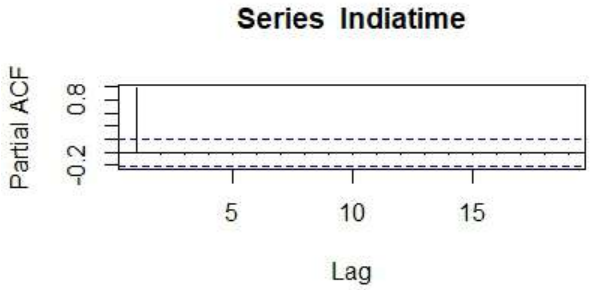


Figure 6: PACF India

Figure 3 and 5 depict the ACF of China and India respectively, since most of the bars in the ACF plot are above the upper limit so it has been concluded that data depends on the past value and thus it is not stationary. In the PACF plot, (Figure 4 and 6) one bar is not within the limit which also suggests the non-stationarity of data.

To check statistically whether data is already stationary or not, we have also applied Augmented Dickey-Fuller Test (ADF) and the results are demonstrated below in Table 1:

Table 1: Augmented Dickey- Fuller test results

Country	**Hypothesis**	**Test statistic**	***p*-value**
China	H_0 : The data is not stationary H_1 : The data is stationary	-1.9889	0.5808
India	H_0 : The data is not stationary H_1 : The data is stationary	-1.9825	0.5834

Table 2: ARIMA models

Country	**Model**	**ARIMA order**	**AIC**
China	1.1	ARIMA (2,2,2)	530.772
	1.2	ARIMA (0,2,0)	538.1234
	1.3	ARIMA (1,2,0)	528.0987
	1.4	**ARIMA (0,2,1)**	**525.1874**
	1.5	ARIMA (1,2,1)	526.9417
	1.6	ARIMA (0,2,2)	526.9914
	1.7	ARIMA (1,2,2)	528.8983
India	**2.1**	**ARIMA (1,2,0)**	**1374.284**
	2.2	ARIMA (2,2,1)	1377.411
	2.3	ARIMA (1,2,1)	1375.423
	2.4	ARIMA (0,2,1)	1375.489
	2.5	ARIMA (2,2,0)	1375.602
	2.6	ARIMA (0,2,0)	1394.637
	2.7	ARIMA (2,2,2)	1379.341

By checking the p-values we can conclude whether the data is stationary or not. The null hypothesis can be rejected if the p-value is less than 0.05; else, the null hypothesis will stand. In both China and India p-value is greater than 0.05 so we can't reject the null hypothesis and conclude that the time series of both countries is not stationary. Now we have to make them stationary for that we can take differences or log them. Here we are using R Studio which suggests various models and we can select the best ARIMA Model according to AIC (Akaike Information Criteria).

The model that has the minimum AIC (mentioned in bold) will be the best fit for our data. For India ARIMA model (1,2,0) and the China ARIMA model (0,2,1) satisfies this criterion as shown in Table 2. So we have selected these models and use them to forecast future cumulative death counts. In both China and India, the value of $d = 2$ means that the data is differenced two times to make it stationary. To ensure that the ARIMA model's residuals are normal, we have plotted the ACF and PACF of the Model's residuals, shown from Figure 7 to 10. The bars were coming within the limit which concludes that residuals follow Normal distribution.

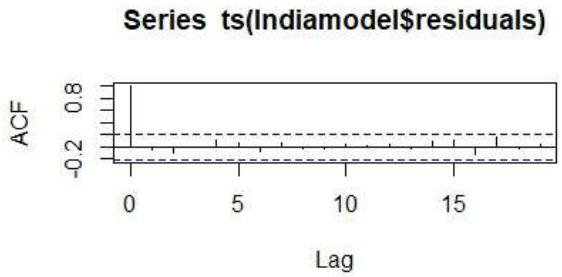


Figure 7: ACF of India residuals

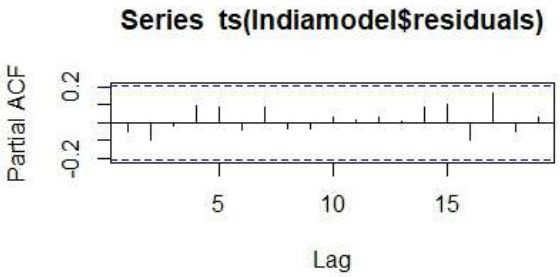


Figure 8: PACF of India residuals

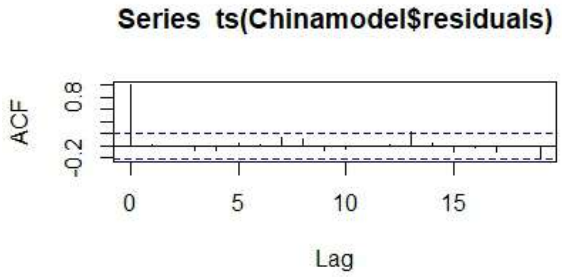


Figure 9: ACF of China residuals

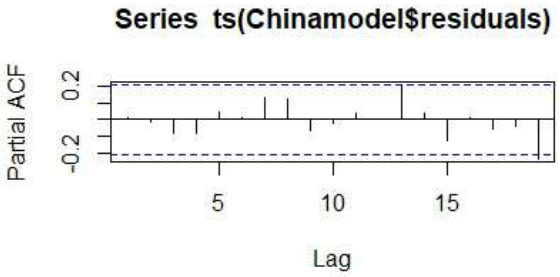


Figure 10: PACF of China residuals

2.2. Regression model

Simple linear regression is a method for predicting a response Y by using only one predictor variable X. The regression model estimates independent variables by using continuous dependent variables provided in the list. The model can be written as :

$$Y = \alpha + \beta\ X + \epsilon \tag{2}$$

where X is the independent variable that is used to estimate the value of Y by following the above relation. In our study cumulative deaths is the dependent variable whose values are being estimated by taking days as X variable or independent variable. In the above equation, α and β are the intercept term and slope parameter respectively, which are also known as regression coefficients. The unobservable error component ϵ indicates the gap between the true and observed values of Y and accounts for the failure of data to lie on a straight line. For statistical inferences we assume ϵ to be an independent and identically distributed random variable with mean zero and constant variance σ^2. We have used R Studio to fit the simple regression model for both China and India.

2.2.1. Analysis of data - II

COVID-19 is spreading exponentially in the world and many people are losing their close ones. The number of deaths due to COVID-19 was increasing in counts as the days were passing especially during the second wave of COVID-19. This study includes the COVID-19 second wave data set from WHO. The collected data was analyzed in R Studio. In India the number of deaths has been increasing day by day while in China the loss of deaths due to COVID-19 has been almost constant, this fact matches our calculated result.

3. Results and conclusion

In ARIMA time series analysis ARIMA (0,2,1) and ARIMA(1,2,0) turn out to be the best model for China and India respectively. The predicted values for the models are shown in Figure 11 and 12. Referring to Table 3 and 4, it has been concluded that ARIMA is the better model for forecasting future cumulative death as the change percentage for the ARIMA model is only 0.01 to 0.2% and 0.01 to 0.3 % for China and India respectively but on the other hand we can see a higher percentage difference for regression model of about 3.4 to 3.7 % for China and 4.0 to 9.9% for India. It is also observed that the second wave of COVID-19 has turned out to be less severe for China as compared to India. The government and associated departments can use ARIMA model to forecast COVID-19 deaths rather than regression model. The comprehensive study designed on the cumulative deaths of COVID-19 can help to see the severity of the situation by predicting the mortality rate. This will help the policymakers to take preventive measures and actions such as fulfilling the oxygen and vaccination demand, arranging beds and medical experts for controlling the COVID-19 situation. This research will also help policymakers to keep a track of how different decisions such as quarantine, lockdown, vaccination *etc.* are helpful in reducing the death count. This can be done by monitoring the difference between predicted deaths and actual deaths taking place after the implementation of policies.

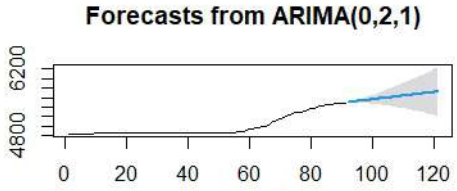


Figure 11: China forecast plot

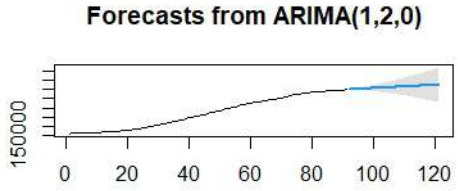


Figure 12: India forecast plot

Table 3: Forecasted and actual value comparison of China

Date	ARIMA actual data	ARIMA forecasted data	ARIMA percentage change	Regression forecasted data	Regression percentage change
01-07-2021	5495	5497.672	0.048626024	5305.5117	-3.448376706
02-07-2021	5508	5505.344	-0.04822077	5312.3607	-3.551911765
03-07-2021	5523	5513.016	-0.18077132	5319.2097	-3.689847909
04-07-2021	5533	5520.688	-0.222519429	5326.0587	-3.740128321
05-07-2021	5535	5528.36	-0.119963866	5332.9077	-3.651170732
06-07-2021	5537	5536.032	-0.017482391	5339.7567	-3.562277407
07-07-2021	5554	5543.704	-0.185379906	5346.6057	-3.73414296
08-07-2021	5563	5551.376	-0.208952004	5353.4547	-3.766767931
09-07-2021	5566	5559.048	-0.124901186	5360.3037	-3.695585699
10-07-2021	5578	5566.72	-0.202223019	5367.1527	-3.77998028
11-07-2021	5584	5574.392	-0.172063037	5374.0017	-3.760714542
12-07-2021	5588	5582.064	-0.106227631	5380.8507	-3.707038296
13-07-2021	5589	5589.736	0.013168724	5387.6997	-3.601723027
14-07-2021	5595	5597.407	0.043020554	5394.5487	-3.582686327
15-07-2021	5601	5605.079	0.072826281	5401.3977	-3.563690412

Table 4: Forecasted and actual value comparison of India

Date	ARIMA actual data	ARIMA forecasted data	ARIMA percentage change	Regression forecasted data	Regression percentage change
01-07-2021	399459	399313.2	-0.036499365	413929.61	4.01359566
03-07-2021	401050	401001.3	-0.012143124	420171.85	4.550959328
04-07-2021	402005	401845.6	-0.039651248	423292.97	5.029133841
05-07-2021	402728	402692	-0.008939036	426414.09	5.554715605
06-07-2021	403281	403537.4	0.063578497	429535.21	6.112236992
07-07-2021	404211	404383.3	0.042626252	432656.33	6.574578488
08-07-2021	405028	405229	0.049626199	435777.45	7.056227898
09-07-2021	405939	406074.7	0.033428668	438898.57	7.509609794
10-07-2021	407145	406920.4	-0.055164622	442019.69	7.889849884
11-07-2021	408040	407766.2	-0.067101265	445140.81	8.334623375
12-07-2021	408764	408611.9	-0.037209735	448261.93	8.811350542
13-07-2021	410784	409457.6	-0.322894757	451383.05	8.994367423
14-07-2021	411408	410303.4	-0.268492591	454504.17	9.482018614
15-07-2021	411989	411149.1	-0.203864666	457625.29	9.972414331

References

Batista, M. (2020). Estimation of the final size of the second phase of the coronavirus COVID-19 epidemic by the logistic model. *medrxiv.* DOI:10.1101/2020.02.16.20023606

Bayyurt, L. and Bayyurt, B. (2020). Forecasting of COVID-19 cases and deaths using ARIMA models. *medrxiv.* DOI:10.1101/2020.04.17.20069237

Box, G. E., Jenkins, G. M., Reinsel, G. C., and Ljung, G. M. (2015). *Time Series Analysis: Forecasting and Control*. John Wiley Sons.

Ding, G., Li, X., Jiao, F., and Shen, Y. (2020). Brief analysis of the ARIMA model on the COVID-19 in Italy. *medRxiv.* DOI:10.1101/2020.04.08.20058636

Gambhir, E., Jain, R., Gupta, A., and Tomer, U. (2020, September). Regression analysis of COVID-19 using machine learning algorithms. *International Conference on Smart Electronics and Communication (ICOSEC),* 65-71. DOI:10.1109/ICOSEC49089.2020.9215356

Hengjian, C. and Tao, H. (2021). Nonlinear regression in COVID-19 forecasting. *Scientia Sinica Mathematica*, **51**. DOI:10.1360/SSM-2020-0055

https://covid19.who.int/data

Katoch, R. and Sidhu, A. (2021). An application of ARIMA model to forecast the dynamics of COVID-19 epidemic in India. *Global Business Review.* DOI:10.1177/0972150920988653

Pandey, G., Chaudhary, P., Gupta, R., and Pal, S. (2020). SEIR and regression model based COVID-19 outbreak predictions in India. *arXiv preprint arXiv:2004.00958.* DOI:10.48550/arXiv.2004.00958

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 21, No. 2, 2023 (New Series), pp 107–120
http://www.ssca.org.in/journal

Dealing with the Imperfect Frame Arising Due to Rare Outdated Units from Finite Population

Neelam Kumar Singh
Brahmanand Post Graduate College, Rath (Hamirpur), UP, India

Received: 01 October 2021; Revised: 26 April 2022; Accepted: 15 November 2022

Abstract

Frame is quite often incomplete and imperfect by the time actual survey starts. Imperfection of frame also arises due to some of the rare units being out-of–scope, out-dated and missing from the sampling frame which may be of considerable measure in size and weight and may lead to deviation between sampled and target population. The estimators from such imperfect frame will not give unbiased results. Unbiased estimator is being devised for target population total and variance in the present study with appropriate sampling design considering the finite population. It is considered that, in the case of finite population, the phenomenon of some of the rare units missing, out of scope or out-dated units follow some probability distribution function (p.d.f) and unbiased estimate of population parameter is devised and developed.

Key words: Imperfect frame; Target population; Inverse sampling.

1. Introduction

In some situations, there are some rare units in the sampled population which are out-dated from the target population at the time of actual survey. The rare units of the old frame for which sampled population correspond, may be out-dated, may be out-of-scope or may be missing from the frame at the time of actual survey for which the information and observations are desired. Such rare units missing from the target population will lead to the imperfection of the frame. Thus there is deviation between sampled population and target population due to imperfection in the frame arising due to some rare sampling units being out-dated from the frame required for the study of statistical results of the desired target population. Therefore, in such cases some rare units of the sampled population do not belong to the target population because frame prepared at some time may contain some rare units which may not exist in the target population at the time of actual survey and enumeration. This happens because during passage of time, the frame prepared at some time will be out-dated by the time the actual enumeration starts.

For example in a frame of list of Agriculture labourers in a district, some rare number

Corresponding Author: Neelam Kumar Singh
Email: neelamkumar099@gmail.com

of labourers may migrate to other districts in search of work at the time, the survey actually starts. Therefore, frame of Agricultural labourers become imperfect at the time of enquiry.

The list of irrigated land holdings in a Tehesil prepared at some time corresponding to the sampled population, may contain some rare number of fields which may not in-fact be irrigated discovered at the time of actual enquiry of crop survey due to failure of some canals or tube wells. In a list of cultivated area under some crop designed for the purpose of survey from source at hand, may contain some cultivated area under some other crop so that some of the rare number of cultivated area may not belong to the crop-area desired for the study. The frame becomes imperfect due to some rare sampling units not belonging to the target population which are unknown at the time of sample selection but is discovered because investigator visits particular cultivated land selected in the sample.

There are also some rare phenomena in the nature when some rare units change rapidly in their geographic ordering and location. For example, in case of shifting cultivation adopted in North-East of India, some of the rare area listed under forest land may be found to be under shifting cultivation due to unprecedental customs and traditions of tribes in that region because of dependence of tribes on nature and nurture. They are discovered only at the time of investigation. Similarly, frame of fields under Jhooming cultivation obtained from some source may contain some rare field which may be discovered as no more belonging under same pattern of agriculture at the time of actual enumeration as is evident in number of Anthropological studies.

The frame, available for some nomadic tribal families will become rapidly imperfect because of some rare number of nomadic tribal families migrating from one place to other, thus making their demographic studies difficult and complex at the time of actual survey. In a frame of mango trees for estimation of total amount of mango fruits, some rare number of trees may not be found bearing fruits discovered at the time of enquiry, In a list of fields for tomato, potato, wheat, gram, Arahar etc, some rare field may contain damaged crops due to adverse weather conditions and estimation of total production of these crops would be difficult on the basis of available old frame as frame become imperfect for desired target population under study. For estimating total amount of Tendu leaves in a forest, list of Tendu tree may contain rare trees which may not bear leaves during passage of time making the frame imperfect. For estimating total amount of water in a district, the list of water tanks available from some official sources, may not correspond to the target population as some rare tanks may be dry and barren by the time of survey.

Thus more often, frame may very soon become out–dated as some rare units of the frame may go quantitative and qualitative change in the sampled population.

Estimates on the basis of sample selected from such imperfect frame would not give unbiased results. These rare out-dated units from the frame will also contribute to the bias of target population results. The correct, complete, and up-to-date frame is rather impossible in practice because some rare units of the sampled population rapidly cease to exist or are observed as non- existent in the target population at the time of actual survey.

Seal (1962) discussed the use of out-dated frames in large scale surveys and considered the changes in the population as a continuous stochastic process. Hartley (1962) proposed the use of two or more frames to overcome the problem of incomplete frames. Hansen *et*

al. (1963) discussed various procedures for the use of incomplete frame and proposed the predecessor-successor method to obtain information on missing units in the frame. Szamsitat and Schaffer (1963) discussed about consequences of imperfect frame in sampling. Singh (1983) gave a mathematical formulation of predecessor-successor method for estimating total number of units missing from the frame. Singh (1989) proposed suitable method of estimation when sampling is done from imperfect frame and a geographic ordering of units can be established. Singh *et al.* (1997) discussed imperfection in the frame of finite population and proposed estimators for domain of study considering probability distribution of the out-dated units in the incomplete frame. Singh *et al.* (2001) discussed the imperfection of frame arising due to omission of some of the units from the frame and also frame containing some units which no more belongs to target population and proposed appropriate estimation procedure for the population, its variance when sampling is done from two frames.

Agarwal and Gupta (2008) developed a method of estimation of population total, mean and variance from incomplete frame in case of SRSWOR and SRSWR.

Singh (2020) discussed the frame error as error due to imperfection of the frame in detail because of deviation of the target population from sample population. Singh (2020) discussed that frames are often imperfect in any sample survey which arises due to some of the rare sampling units being out-dated at the time of actual survey. He further devised unbiased estimator for the imperfection of frame arising due to rare out-of-scope units considering population size to be large. Suitable estimators for the proportion of out-dated units from the population and target population total for a character with their variance was developed considering probability distribution function (p.d.f.) of rare out- dated units in large population.

Agarwal and Singh (2021) considered every finite of population as a constituent part of some super population in which complete frame is quite often not available under consideration. They providers estimators under more realistic situation as compared to finite population concepts.

Singh (2021) discussed that the existence of the frame is pre-requisite for any sample survey or census of a large population. Frames are quite often imperfect due to dynamic nature of sampling units. Frames become incomplete by the time actual survey and enumeration starts which affect the statistical results desired for the target population. He reported and considered imperfection in the frame of large population arising due to the qualitative change of units from one class to other. He considered incomplete frame assuming the nature of units following dynamic change from class one to other which follows a probability distribution function. Suitable estimator for proportion of units belonging to a particular domain and unbiased estimate of target population for a class was proposed along with its estimate of variance. The estimates are evolved so as to eliminate error caused due to the deviation of sampled population from target population.

Singh (2022) gave appraisal of problem of incomplete frame in different situations. The cause and type of imperfection of frame was discussed covering various aspects and review of work done by different scholars was explained along with measures and suggestions to deal with imperfection of frame.

More troublesome are those cases when missing or out-dated units although rare in

number are exceedingly large in their measure of size and such rare units were discovered because units were selected. For example, in list of business establishments, some large establishments although rare in number, may no longer be active at the time of enumeration.

These large sized rare units missing from the frame would lead to the imperfection of the frame and would contribute much to the bias of sampling results for target population parameters.

Therefore, this study deals with imperfect frame arising due to missing or out- dated rare units from the target population of finite size. The objective in the present study is to design a sampling procedure to devise an estimator which is unbiased for target population parameters in such cases of imperfection in the frame.

2. Method of estimation

Consider a finite sampled population of size N as listed in the available frame for selecting the sample at some time. However, during passage of time, population structure has undergone some change in the sense that some rare units of the available original frame *i.e.* of sampled population, have ceased to exist in the target population. Let N_1 denotes the number of units in the frame of size N, actually belonging to the target population. N_2 denotes number of rare sampling units which have gone out of the target population. So that, $N = (N_1 + N_2)$. The N_2 units, though rare in the number may attribute for high measure of their size. Assume that the rare units which have ceased to exist in the target population are not identifiable and hence these rare units although being out-of scope and out-dated units, cannot be deleted from the frame. Therefore, actual frame of target population will be of size N_1 which is unknown at the time of actual survey. N_2 units are rare in number but these are also not identifiable which are attributing for their high measure of sizes with considerable importance and significant for population parameters in their observations of characteristics under study.

Situation of incomplete frame arises when units under go qualitative change and becomes out-dated units for the desired targeted population, during passage of time. The units respond but their measures/ values are out-dated for target population. In case of non-response, frame is complete but some of the units do not respond.

The information and observations attributed to the N_2 rare units of the sampled population are, although, of significant importance but will not correspond the characteristics of the target population. This phenomenon occurs when there are changes in the target population when population is subject to continuous change and if there is rather long interval between the dates to which sampled population relates and date or time for which the information is to be collected.

One procedure to select the sample may be to select a random number from 1 to N keeping old numbering as such. The unit corresponding to this number is selected, provided it is not of N_2 rare units which have become non-existent in the target population. If non-existent rare unit is selected, draw is rejected and the procedure is repeated. This gives equal chance of selection to $N - N_2 = N_1$ units of the target population.

However, this procedure assumes that the information is available about the rare out-

dated units from the old frame at the time of sample selection. But, most often, we do not know about the out-dated units from the old frame at the time of sample selection unless the actual enumeration starts. It is only when enumerator visits a particular rare out-dated unit that he finds that the units no more exist in the target population.

Therefore, we propose an alternative method of sample selection and sampling plan for estimation procedure.

3. Notations

Let p denotes the proportion of rare sampling units from the available original frame of sampled population which are out-dated, out-of-scope or missing units from the target population leading to the imperfection of the frame. Hence, $p = N_2/N$. Evidently Np units will be rare units which are missing from the target population. We have $N = N_1 + Np$ and $N_1 = Nq, q = 1 - p$. Hence, Nq units actually belong to the target population.

Let Y denotes the character under study and Y_i denote the value of Y for ith unit of the sampled population

$$U = (U_1, U_2, U_3, ..., U_N).$$

Let $\bar{Y}_1 = \frac{1}{N_1}\sum_{i=1}^{N_1} Y_i$: population mean of the target population

$$S_1^2 = \frac{1}{N_1}\sum_{i=1}^{N_1}(Y_i - \bar{Y}_{N_1})^2$$

$p = N_2/N$: proportion of rare sampling units in the sampled population which are non-existent in the target population.
$q = N_1/N$: proportion of sampling units in the available frame of sampled population actually belonging to the target population.

The population total for the target population can be easily written as

$$Y_{N_1} = N_1\bar{Y}_1.$$

In order to estimate target population total for Y, we have to estimate N_1 and $\bar{Y}_1$.

4. Proposed sampling procedure: Inverse simple random sampling without replacement (I.S.R.S.W.O.R.)

Since, the number of out-dated units is rare in number; the proportion p is very small. In such situation, method of Inverse-sampling can be used with advantage. In this method, the sampled size n is not fixed in advance. Instead sampling is continued until a predetermined number of rare units out-dated from the frame have been drawn. Let p denotes the proportion of rare units missing from the frame. Evidently Np units will be missing or out-dated units in the frame of sampled population and $N - Np = N_1$ units will exist in the target population so that $N_1 = Nq, q = 1 - p$. For estimating the population p, the sampling units are drawn one by one with equal probability of selection and without replacement. This procedure is called I.S.R.S.W.O.R. sampling is discontinued as soon as

the number of units in the sample possessing the rare units missing from the frame is some predetermined number.

In some situations, the statistical investigation may demand a sample to include a required representation from the category of rare missing units. It may be required of a sample to include a specified number of rare units which are out-dated. In such situation direct sampling procedure based on fixed sample size may not be appropriate. Rather, inverse sampling procedure discussed by Haldane (1945), Finney (1949), Chapman (1952) and Chikkagudar (1969) among others are expected to be more appropriate. Suppose, the statistical enquiry requires that the sample should include n_2 units from the rare missing units from the frame.

We continue selecting units, one by one, with equal probability and without replacement from the sampled population $U = \{U_1, U_2, ..., U_N\}$ until there are exactly n_2 units (given) discovered at the enumeration stage represented from the missing units. The total number of units, n in the sample is obviously a random variable. Method of continuing sample, called I.S.R.S.W.O.R has one important advantage. This case of I.S.R.S.W.O.R can be used with advantage when proportion $p \leq 10\%$ (Haldane, 1945). In such situation p is small but not well known in advance. The value of n is large if p is small. Thus sample of size n will contain n_2 units which do not exist in the target population and $n - n_2 = n_1$ (say) units belong in the target population after enumerator visits each unit of the sample of size n. As such, there can be no observations obtained for such n_2 (given) non-existent units in the frame. The observations for $n - n_2$ units can be obtained for which observations are available. The n_2 units are non-existent or even if they exist, the enumerator can identify them as not belonging to the target population and hence cannot be observed.

Therefore, corresponding probability distribution $P(n)$ for random variable n is given by

$$p(n) = p\left\{\begin{array}{c}\text{In a sample of } (n-1) \text{ units drawn} \\ \text{from } N, n_2 - 1 \text{ units will be} \\ \text{discovered to be missing or out-dated} \\ \text{units in the target population}\end{array}\right\} . p\left\{\begin{array}{c}\text{the unit drawn at the} \\ n\text{th draw will be rare} \\ \text{missing or out-dated unit} \\ \text{from the target population}\end{array}\right\}$$

$$= \frac{\binom{N_2}{n_2-1}\binom{N_1}{n-n_2}}{\binom{N}{n-1}} . \frac{N_2 - (n_2 - 1)}{N - (n-1)}$$

where $n = n_2, n_2 + 1, ..., (n_2 + N–N_2)$
or

$$p(n) = \frac{\binom{Np}{n_2-1} . \binom{Nq}{n-n_2}}{\binom{N}{n-1}} . \frac{Np - (n_2 - 1)}{N - (n-1)}$$

where $n = n_2, n_2 + 1, ..., n_2 + Nq$. Here, we also have $\sum p(n) = 1$, where $n \geq n_2$.

5. Estimation of proportion and its variance

It can be shown that an unbiased estimate of p can be given by

$$Est\ p = \frac{n_2 - 1}{n - 1} = \hat{p},\ \text{say.} \tag{1}$$

Thus,

$$\begin{aligned} E\left(\frac{n_2-1}{n-1}\right) &= \sum\left(\frac{n_2-1}{n-1}\right) \times \frac{\binom{Np}{n_2-1}\cdot\binom{Nq}{n-n_2}}{N\binom{N-1}{n-1}} \cdot \frac{Np-n_2+1}{N-n+1} \\ &= \sum \frac{(n_2-1)}{(n-1)}\frac{(n-1)}{(n_2-1)} Np \frac{\binom{Np-1}{n_2-2}\cdot\binom{Nq}{n-n_2}}{N\binom{N-1}{n-2}} \frac{Np-n_2+1}{N-n+1} \\ &= p \sum_{n\geq n_2} \frac{\binom{Np-1}{n_2-2}\cdot\binom{Nq}{n-n_2}}{\binom{N-1}{n-2}} \frac{Np-n_2+1}{N-n+1} \\ &= p \end{aligned}$$

We shall now determine the estimate of $V(\hat{p})$. We know that

$$V(\hat{p}) = E(\hat{p}^2) - (E(\hat{p}))^2 = E(\hat{p}^2) - p^2. \tag{2}$$

Therefore, an biased estimate of $V(\hat{p})$ is given by

$$Est\ V(\hat{p}) = \hat{p}^2 - Est\ p^2$$

Now to determine $Est\ p^2$ we have

$$E\frac{(n_2-1)(n_2-2)}{(n-1)(n-2)} = \frac{Np(Np-1)}{N(N-1)} \sum_{n\geq n} \frac{\binom{Np-2}{n_2-3}\binom{Nq}{n-n_2}}{\binom{N-2}{n-3}} \frac{Np-n_2+1}{N-n+1} \quad = \frac{N}{N-1}p^2 - \frac{1}{N-1}p$$

Therefore, $Est\ \frac{N}{N-1}p^2 - \frac{1}{N-1}Est\ p = \frac{(n_2-1)(n_2-2)}{(n-1)(n-2)}$ or,

$$Est\ p^2 = \frac{N-1}{N}\frac{(n_2-1)(n_2-2)}{(n-1)(n-2)} + \frac{1}{N}\hat{p}. \tag{3}$$

Therefore, from (2) and (3), we have

$$\begin{aligned} Est\ V(\hat{p}) &= \hat{p}^2 - \frac{N-1}{N}\frac{(n_2-1)(n_2-2)}{(n-1)(n-2)} - \frac{1}{N}\hat{p} \\ &= \hat{p}\left\{\hat{p} - \left(\frac{N-1}{N}\right)\left(\frac{n_2-2}{n-2}\right) - \frac{1}{N}\right\} \end{aligned} \tag{4}$$

or,

$$\hat{V}(\hat{p}) = \frac{(n_2-1)^2}{(n-1)^2} - \frac{(N-1)(n_2-1)(n_2-1)}{N(n-1)(n-2)} - \frac{(n_2-1)}{N(n-2)} \tag{5}$$

Similarly $V(\hat{p})$ can be obtained as

$$V(\hat{p}) = \frac{(n_2-1)p + np(1-p)}{N-1} \tag{6}$$

Therefore, we have an unbiased estimate of $\hat{Y}_{N_1}$ (target population total), as given by $Est\ \hat{Y}_{N_1} = Est\ (Nq\bar{Y}_1) = Est\ (Nq)Est(\bar{Y}_1) = \hat{Y}_{N_1}$ say.
Thus,

$$\begin{aligned}\hat{Y}_{N_1} &= (N\hat{q})(\hat{\bar{Y}}_1) = N\hat{q}\bar{y}_1 \\ &= N(1-\hat{p})\bar{y}_1 \\ &= N\left(\frac{n-n_2}{n-1}\right)\bar{y}_1\end{aligned} \tag{7}$$

$\hat{Y}_{N_1}$ is an unbiased estimate of Y_{N_1}, because
$E(\hat{Y}_{N_1}) = E(N\hat{q}\bar{y}_1) = Nq\bar{Y}_1 =$ Target population total.

6. Variance of $\hat{Y}_{N_1}$

The variance of estimate of the target population total is given by

$$\begin{aligned}V(\hat{Y}_{N_1}) &= V(N\hat{q}\bar{y}_1) \\ &= N^2V\{(1-\hat{p})\bar{y}_1\} \\ &= N^2[V(\bar{y}_1) + E\{V(\hat{p}\bar{y}_1)|\hat{p}\} + V(E(\hat{p}\bar{y}_1|\hat{p}))] \\ &= N^2[V(\bar{y}_1) + E\{\hat{p}^2V(\bar{y}_1)\} + V(\bar{Y}_1\hat{p})] \\ &= N^2[V(\bar{y}_1) + V(\bar{y}_1)E(\hat{p}^2) + \bar{Y}_1^2V(\hat{p})]\end{aligned} \tag{8}$$

as $E(\bar{y}_1) = \bar{Y}_1$ and also $E(\hat{p}^2) = V(\hat{p}) + p^2$. Here $\bar{y}_1 = \frac{1}{n_1}\sum_{i=1}^{n_1} y_i$ which is sample mean for y based on n_1 observations and $n_1 = n - n_2$ for which observation are available in the sample belonging to the target population. Again number of units belonging to the target population are $Nq = N_1$ whose variance is given by S_1^2. Sample mean square error for units in the target population can be given by

$$s_1^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(y_i - \bar{y}_1)^2.$$

Also, $\bar{Y}_1$ = Mean of the target population so that

$$\bar{Y}_1 = \frac{1}{N_1}\sum_{i=1}^{N_1} Y_i.$$

We can also have

$$V(\bar{y}_1) = \frac{N_1 - n_1}{N_1 n_1}S_1^2.$$

Thus, again we have from (8)

$$\begin{aligned}V(\hat{Y}_{N_1}) &= N^2[V(\bar{y}_1) + V(\bar{y}_1)\{V(\hat{p}) + \hat{p}^2\} + \bar{Y}_1^2V(\hat{p})] \\ &= N^2[V(\bar{y}_1)\{1 + \hat{p}^2 + V(\hat{p})\} + \bar{Y}_1^2V(\hat{p})].\end{aligned} \tag{9}$$

Putting values of $V(\bar{y}_1)$ and $V(\hat{p})$ we can also have

$$V(\hat{Y}_{N_1}) = N^2\left[\frac{N_1-n_1}{N_1 n_1}S_1^2\left\{1+p^2+\frac{(n_2-1)p+Np(1-p)}{N-1}\right\}\right] + N^2\bar{Y}_1^2\left\{\frac{(n_2-1)p+Np(1-p)}{N-1}\right\}. \tag{10}$$

If $\frac{N}{N-1} \cong 1$ then we have after simplification

$$V(\hat{Y}_{N_1}) = \frac{N(N_1-n_1)}{N_1 n_1}S_1^2\{N-1+p(N+n_2p-1)\} + Np\bar{Y}_1^2\{N(1-p)+n_2-1\} \tag{11}$$

and as $N_1 = Nq = N(1-p)$, we can also have

$$V(\hat{Y}_{N_1}) = \frac{N(1-p)-n_1}{(1-p)n_1}S_1^2\{N-1+p(N+n_2-p-1)\} + Np\bar{Y}_1^2\{N(1-p)+n_2-1\}. \tag{12}$$

Thus $V(\hat{Y}_{N_1})$ is function of N, p, n_1, S_1^2, n_2, and $\bar{Y}_1^2$. Second term in the $V(\hat{Y}_{N_1})$ is independent of n but first term is not independent of n because $n_1 = n - n_2$. For fixed n_2, n_1 increases as n (random variable) increases. Therefore, for a given n_2, $V(\hat{Y}_{N_1})$ *i.e.* variance of the estimate of target population total decreases as sample size n increases. However, $V(\hat{Y}_{N_1})$ increases as n decreases. But $n \geq n_2$ so that

$$n = n_2, n_2+1, n_2+2, ..., n_2+N(1-p).$$

For $n_2 = n$, *i.e.*, when $n_1 = 0$, it is not possible to determine $V(\hat{Y}_{N_1})$.

However, if $n = n_2 + 1$ then $n_1 = 1$ and $V(\hat{Y}_{N_1})$ is maximum and is given by

$$V(\hat{Y}_{N_1}) = \frac{N(1-p)-1}{((1-p)}S_1^2\{N-1+p(N+n-p-2)\} + Np\bar{Y}_1^2\{N(1-p)+n_2-2\}.$$

If n is the maximum value which is given as $n = n_2 + N(1-p)$ then $n_2 = n - N(1-P)$ and $n_1 = N(1-p)$. In this case variance of $V(\hat{Y}_{N_1})$ is given by

$$V(\hat{Y}_{N_1}) = N(n-1)p\bar{Y}_1^2.$$

Again $V(\hat{Y}_{N_1})$ also depends on the nature of p. The behavior of the frame depends on the proportion p. It can be seen that, in case of perfect frame $p = 0$. Therefore, when there is no imperfection in the frame $V(\hat{Y}_{N_1}) = \frac{(N-n_1)}{n_1}S_1^2(N-1)$, which is approximately equal to the $V(\hat{Y}_{N_1})$. In case of perfect frame, when there are no rare missing units in the sampled population. We have

$$V(\hat{Y}_{N_1}) = V(\hat{Y}_N) = \frac{(N-n_1)}{n_1}(N-1)S^2$$

as $S_1^2 = S^2$ and $n_1 = n$. The approximation occurs as we have assumed $\frac{N}{(N-1)} \cong 1$ in (11) earlier. However, if $p = 1$ *i.e.* when there is total imperfection in the frame then $V(\hat{Y}_{N_1})$ cannot be determined.

We know that $0 \leq p \leq 1$. It can be seen that $V(\hat{Y}_{N_1})$ is maximum as $p \to 0$.

7. Estimation of $V(\hat{Y}_{N_1})$

We have from (9)

$$V(\hat{Y}_{N_1}) = N^2[V(\bar{y}_1)(1+p^2) + V(\bar{p})\{V(\bar{y}_1) + \bar{Y}_1^2\}].$$

Estimate of $V(\hat{Y}_{N_1})$ can be obtained by estimating each of the right hand side terms. Therefore,

$$Est\ V(\hat{Y}_{N_1}) = N^2[Est\ V(\bar{y}_1)(1+p^2) + Est\ V(\bar{p})Est\ \{V(\bar{y}_1) + \bar{Y}_1^2\}].$$

As we know that

$$V(\bar{y}_1^2) = E(\bar{y}_1^2) - \bar{Y}_1^2$$

therefore,

$$E(\bar{y}_1^2) = V(\bar{y}_1^2) + \bar{Y}_1^2$$

and

$$Est\ V(\bar{y}_1^2) + Est\ (\bar{Y}_1^2) = \bar{y}_1^2.$$

Similarly,

$$Est\ (1+p^2) = 1 + Est\ p^2.$$

But we know that

$$V(\hat{p}^2) = E(\hat{p}^2) - p^2$$

and

$$p^2 = E(\hat{p}^2) - V(\hat{p}).$$

Hence,

$$Est\ p^2 = \hat{p}^2 - \hat{V}(\hat{p}). \tag{13}$$

Putting these values in (9) we can obtain

$$\begin{aligned} \hat{V}(\hat{Y}_{N_1}) &= N^2\left[Est\ V(\bar{y}_1)\{1+\hat{p}^2 - \hat{V}(\hat{p})\} + Est\ V(\hat{p})(\bar{Y}_1^2)\right] \\ &= N^2\left[\frac{N_1-n_1}{N_1 n_1}s_1^2\{1+\hat{p}^2-\hat{V}(\hat{p})\} + \hat{V}(\hat{p})\bar{Y}_1^2\right]. \end{aligned} \tag{14}$$

Since n_1 is also selected with S.R.S.W.O.R. in I.S.R.S.W.O.R. with n_2 fixed and n random so that $n_1 = n - n_2$ and we have $s_1^2 = \frac{1}{(n_1-1)}\sum_{i=1}^{n_1}(y_i - \bar{y}_1^2)$ Putting value of $\hat{V}(\hat{p})$ from equation (5) in (14), we have

$$\hat{V}(\hat{Y}_{N_1}) = N^2\left[\frac{N_1-n_1}{N_1 n_1}s_1^2\left\{1 + \frac{N-1}{N}\hat{p}\left(\frac{n_2-2}{n-2}\right) + \frac{1}{N}\hat{p}\right\}\right] + \bar{y}_1^2\left\{\hat{p}^2 - \frac{N-1}{N}\hat{p}\left(\frac{n_2-2}{n-2}\right) - \frac{1}{N}\hat{p}\right\} \tag{15}$$

or

$$\hat{V}(\hat{Y}_{N_1}) = N\left\{\frac{N_1-n_1}{N_1 n_1}s_1^2\right\}\left\{N + (N-1)\hat{p}\frac{n_2-2}{n-2} + \hat{p}\right\} + N\bar{y}_1^2\left\{N\hat{p}^2 - (N-1)\hat{p}\frac{n_2-2}{n-2} - \hat{p}\right\}. \tag{16}$$

8. Estimation of $\frac{1}{N_1}$

As N_1 is unknown, it has to be substituted with its estimate, therefore, we can have

$$\frac{N_1 - n_1}{N_1 n_1} = \left(\frac{1}{n_1} - \frac{1}{N_1}\right).$$

Thus,

$$Est\ \left(\frac{N_1 - n_1}{N_1 n_1}\right) = Est\ \left(\frac{1}{n_1} - \frac{1}{N_1}\right) = \frac{1}{n_1} - Est\ \frac{1}{N_1}.$$

Now, $Est\ \frac{1}{N_1} = Est(N_1^{-1})$, let $Est\ N_1 = \hat{N}_1$ (say). But we know that

$$\hat{N}_1 = Est\ (Nq)$$

or

$$\hat{N}_1 = Nq = N(1 - \hat{p}).$$

Assume $\hat{N}_1 = N_1 + \epsilon$ where $E(\epsilon) = 0$. We may write $(\hat{N}_1)^{-1} = \{1 + \frac{\epsilon}{N_1}\}^{-1} \cong (N_1)^{-1}\{1 - \frac{\epsilon}{N_1}\}$, neglecting the power of higher than one. Thus,

$$E\left(\frac{1}{N_1}\right) \cong E\left(\frac{1}{N}\right) - \frac{E(\epsilon)}{N_1^2}.$$

Therefore,

$$Est\ \frac{1}{N_1} \cong \frac{1}{\hat{N}_1}\frac{1}{N\hat{q}} \text{ as } \hat{N}_1 = N\hat{q}.$$

Hence,

$$Est\ \frac{N_1 - n_1}{N_1 n_1} \cong \frac{N\hat{q} - n_1}{N\hat{q}n_1} = \frac{N(1-\hat{p}) - n_1}{N(1-\hat{p})n_1}.$$

Putting these values in (16) we obtain

$$\begin{aligned}\hat{V}(\hat{Y}_{N_1}) &= N\left\{\frac{N(1-\hat{p}) - n_1}{N(1-\hat{p})n_1}s_1^2\right\}\left\{N + (N-1)\hat{p}\left(\frac{n_2 - 2}{n-2}\right) + \hat{p}\right\} \\ &\quad + N\bar{y}_1^2\{N\hat{p}^2 - (N-1)\hat{p}\left(\frac{n_2-2}{n-2}\right) - \hat{p}\} \\ &= N\left\{\frac{N(1-\hat{p}) - n_1}{N(1-\hat{p})n_1}s_1^2\right\}\left\{N + (N-1)\hat{p}\left(\frac{n_2 - 2}{n-2}\right) + \hat{p}\right\} \\ &\quad + N\hat{p}\bar{y}_1^2\{N\hat{p} - (N-1)\left(\frac{n_2-2}{n-2}\right) - 1\}\end{aligned} \tag{17}$$

or,

$$\begin{aligned}\hat{V}(\hat{Y}_{N_1}) &= \frac{N-n+1}{n_1}s_1^2\left\{N + \frac{(N-1)(n_2-1)(n_2-2)}{(n-1)(n-2)} - \frac{n_2-1}{n-1}\right\} \\ &\quad + N\bar{y}_1^2\left(\frac{n_2-1}{n-1}\right)\left\{N\left(\frac{n_2-1}{n-1}\right) - \frac{(N-1)(n_2-2)}{n-2} - 1\right\}\end{aligned}. \tag{18}$$

As

$$(1-\hat{p}) = \frac{n - n_2}{n-1} = \frac{n_1}{n-1}$$

for $n = n_1 + n_2$ Therefore, estimate of the $V(\hat{Y}_{N_1})$ is function of n, n_1, n_2, N, s_1^2, and $\bar{y}_1^2$. These values can be obtained with the help of the samples. Estimate of variance of the estimate of the target population total can be obtained in case of imperfect frame arising due to rare missing units. Estimate of $V(\hat{Y}_{N_1})$ increases as N and n_1 increases but decreases as n increases. If we have $\hat{p} = 1$ so that $n = n_2$ and $n_1 = 0$ then s_1^2, in such case, can not be estimated because $1/n_1$ tends to infinity in the first term of the above equation. This case may arise when there is total imperfection in the frame.

In case of $n_1 = 1$, we have $n_2 = n+1$. In such case $s_1^2 = 0$. Because, s_1^2 can not be determined for one observation. Therefore

$$V(\hat{Y}_{N_1}) = N\bar{y}_1^2 \left\{ N\hat{p}^2 - (N-1)\hat{p}\left(\frac{n_2 - 2}{n-2}\right) - \hat{p} \right\}.$$

As first term vanishes and $\bar{y}_1^2 = y_1^2$. However, if $\hat{p} = 0$ which may be the case of no imperfection in the frame, then $\hat{V}(\hat{Y}_{N_1})$ is obtained as

$$\hat{V}(\hat{Y}_{N_1}) = \frac{N - n_1}{n_1} s_1^2 N = \frac{N(N - n_1)}{n_1} s_1^2.$$

Since, when $\hat{p} = 0, n_1 = n$ and $s_1^2 = s^2$ therefore

$$V(\hat{Y}_{N_1}) = \frac{N(N-n)}{n} s^2$$

which is obtained in case of perfect frame with S.R.S.W.O.R., this case arises when there is no rare unit in the sample which is non-existent in the target population.

Example: In Chhattisgarh State, the Chhattisgarh Renewable Energy Development Agency (CREDA) had installed 23953 biogas plants by year 2010. During passage of time, it was indicated that there are some plants which become in non-working conditions, annually. The sampler desires to estimate total working plants and total biogas production in the state. Since, number of non-working plants was not known in advance, therefore, the sampling frame is incomplete. The inverse sampling methodology was used to estimate working and non-working plants and total biogas production along with total biogas loss was estimated. with the help of imperfect frame. The non-working plants to be selected were fixed as 18, which were not identified in advance. To get 18 non-working plants, 117 plants were observed. The average production of working biogas plants was found to be 2.8 m^3. Here,

$$n_2 = 18, n_1 = 99, n = 117, N = 23953$$

thus,

$$\hat{p} = 0.146, \hat{q} = 0.854, \bar{y}_1 = 2.8m^3.$$

Therefore, estimated total number of working biogas plants, $N\hat{q} = 20442$ and non-working plants as $N\hat{p} = 3497$. From equation (7), we get total production of biogas in the state as 57276.4 m^3 and the total loss of biogas due to non-performance of plants is found to be 9791.6 m^3.

9. Conclusion

While planning sample survey or census, the existence of sample frame comprising a list of the all the sampling units is pre-requisite. But unfortunately, this situation is hardly achieved in practice and frames are quite often imperfect and incomplete. This may also arise when some of the rare units are missing out- dated or may be out of scope at the time, sampler desire to use. The appropriate and suitable method of estimation is proposed when sampling is done from imperfect frame as elucidated in the formulae explained and illustrated above in section 5 - 8.

Acknowledgements

Thanks are due to a referee for helpful comments that led to a much improved version of the paper.

References

Agarwal, B. and Gupta, P. C. (2008). Estimation from incomplete sampling frame in case of simple random sampling. *Model Assisted Statistics and Applications*, **3**, 113-117.

Agarwal, B. and Singh, S. (2017). Estimation from super-population in case of finite population with incomplete sampling frame. *International Journal of Research and Scientific Innovation*, **4**, 1-4.

Champman, D. G. (1952). Inverse, multiple and sequential sample censuses. *Biometrika*, **8**, 286-288.

Chikkagudar, M. S. (1969). Inverse sampling without replacement. *Australian Journal of Statistics*, **11**, 155-165.

Finney, D. J. (1949). On the method of estimating frequencies. *Biometrika*, **36**, 233-234.

Haldane, J. B. S. (1945). On a method of estimating frequencies. *Biometrika*, **33**, 222-225.

Hansen, M. H., Hurwitz, W. N., and Jabine, T. B. (1963). Use of imperfect tests for probability sampling at the U.S. Bureau of Census. *Bulletin of the International Statistical Institute*, **40**, 497-517.

Hartley, H. O. (1962). *Multiple Frame Surveys.* In Proc. Social Statistics Section Amer. *Statist. Assoc., Annual Meeting, Minneapolis, Minnesota*, 203-206.

Seal, K. C. (1962). Use of outdated frames in large scale sample surveys. *Calcutta Statistical Association Bulletin*, **11**, 68-84.

Singh, N. K. (2020). Frame error in sample survey. *International Research Journal of Agricultural Economics and Statistics*, **11**, 240-244.

Singh, N. K. (2020). Sampling with imperfect frame in large population. *International Research Journal of Agriculture Science and Research*, **10**, 105-112.

Singh, N. K. (2021). Domain studies with imperfect frame in large population. *International Journal of Agricultural Sciences*, **17**, 522-527.

Singh, N. K. (2022). Review On dealing with the problem of imperfect frame in sample survey. *Journal of Emerging Technologies and Innovative Research*, **9(1)**, a690-a699.

Singh, N. K., Kumar, R., and Sehgal, V. K. (2001). Use of incomplete frame on large scale sample survey. *Gujarat Statistical Review*, **28**, 3-10.

Singh, N. K., Sehgal, V. K., and Kumar, R. (1997). Use of incomplete frame for domain studies. *Journal of Indian Statistical Association*, **35**, 71-81.

Singh, R. (1983). On the use of incomplete frame in sample survey. *Biometrical Journal*, **25**, 545-549.

Singh, R. (1989). Method of estimation for sampling from incomplete frames. *Australian Journal of Statistics*, **31**, 269-276.

Szameitat, K. and Schaffer, K. A. (1963). Imperfect frames in statistics and consequences of their use of sampling. *Bulletin of International Statistical Institute*, **40**, 517-538.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 21, No. 2, 2023 (New Series), pp 121-140
http://www.ssca.org.in/journal

Heterogeneous Auto-Regressive Modeling based Realised Volatility Forecasting

G. Avinash[1], Ramasubramanian V.[2] and Badri Narayanan Gopalakrishnan[3]
[1]*The Graduate School, ICAR-Indian Agricultural Research Institute, New Delhi*
[2]*ICAR-Indian Agricultural Statistics Research Institute, New Delhi*
[3]*Lead Adviser & Head, Trade and Commerce, NITI Aayog, GoI, New Delhi*

Received: 08 June 2022; Revised: 13 November 2022; Accepted: 17 November 2022

Abstract

Volatility forecasting is a critical task in the financial markets. It exhibits persistence, which is implicit in option prices. In this study, estimation of Realised Volatility (RV) through high frequency data on the basis of realised variance measures by Heterogeneous Auto-Regressive (HAR) modeling termed as HAR-RV is discussed. This volatility cascade leads to a simple AR-type model in the realised volatility with the feature of considering different volatility components realised over different time horizons successfully capturing the main characteristics of finance data. The HAR model can be extended by adding different decompositions of volatility components into the benchmark model. Thus HAR-RV and its extensions namely, HAR with the simple jump measure (HAR-J), HAR augmented with Quarticity component (HAR-Q), Bi-power variation (BPV) to separate the continuous and jump component named as HAR with continuous and jump component (HAR-CJ), HAR with Quarticity and Jump component (HAR-QJ), Without Jump component (CHAR) and along with Quarticity component (CHAR-Q) models were studied. HAR models have been widely used to forecast crude oil futures volatility, agricultural commodities, stock returns *etc.* An attempt has been done on real dataset relating to Standard and Poor's 500 (S&P 500) stock market high frequency data and its volatility was estimated by using HAR models and its extensions and were compared on different horizons with their volatility studied. The results revealed that CHAR-Q models perform well in the estimation period compared to all other models.

Key words: Bi-power variation; Continuous component; Jump component; High frequency data; Quarticity; Standard and Poor's 500.

1. Introduction

Volatility modeling and forecasting are integral to finance, and are used in a variety of financial applications such as risk management and hedging, because volatility plays an

Corresponding Author: Ramasubramanian V.
Email: ram.vaidhyanathan@gmail.com

important role in asset pricing, portfolio construction, risk management, and trading strategy creation. Researchers and practitioners continuously strive for improving the forecast accuracy of asset return volatility. Engle (1982), Bollerslev (1986), Nelson (1991) and others have done extensive and in-depth research on the measurement and modeling of the volatility of asset price, and they believe that the volatility of financial markets has a particular time-varying nature. Later, they introduced the Auto-Regressive Conditional Heteroscedasticity (ARCH) or Generalized ARCH (GARCH) model to capture the aggregation effect on market volatility and achieved better results. Taylor (1994) worked on Stochastic Volatility (SV) model, which is more elastic than the ARCH type for representing the time-varying character of market volatility. The classic GARCH model, SV model, and other research outcomes based on low frequency financial data on asset price fluctuations have been immensely recognised by domestic and international research institutions. Almost all GARCH models are associated with daily, close-to-close returns, or with even lower-frequency data requirements. Though these models perform well in predicting volatility, they fail to capture the intraday activity patterns. Once high-frequency data available, researchers recognized that these data are even more informative regarding volatility, and the concept of realised volatility emerged (Barndorff-Nielsen and Shephard, 2002). However, daily squared returns are a noisy proxy for true volatility (Molnar, 2012). Realised volatility quickly found its way into the volatility modeling and forecasting literature (Andersen *et al.*, 2003) and became popular, not only in volatility models but also in price forecasting (Degiannakis and Filis, 2017).

With the availability and broad application of high-frequency financial data, the Realised Volatility (RV) and the realised double power variation based on high-frequency data measurement contain more market information than the low-frequency model volatility. An attempt has been made by combining their research to model high-frequency volatility from different perspectives. Based on the theory of heterogeneous markets, Corsi (2009) presented an article that discussed the HAR-RV model. The first order autoregressive volatility process is implemented, which represents the market's heterogeneous trading behaviour. Also, constructed a new HAR model (HAR-RV-CJ) based on the original one, decomposing realised volatility into continuous sample path variance and jump variance to study the impact of volatility. Specifically, the current applications of the HAR models follow the (1, 5, 22) time horizon structure originally proposed for developed markets, using daily (1 day), weekly (5 days), and monthly (22 days) periods to represent the short-term, medium-term, and long-term investors trading frequencies, respectively. However, investors cultural backgrounds and investment habits, as well as the alternative investment choices, differ largely across markets, which will probably result in different heterogeneous structures across markets. Furthermore, investors trading frequencies may be affected by financial and economic policies as well as market conditions, which will probably lead to a market's heterogeneous structure varying over time.

It is well known that stock market prices fluctuate the most during and in the early moments of bubbles and crashes due to uncertainty in the markets. Volatility forecasting, therefore, plays a crucial role in determining the distress of an asset or a market and the research in this area has grown over time. Even till date, forecasting volatility still "remains very much an art rather than a science" quoted two decades earlier by Figelwski (2004).

Traditionally, the multivariate volatility models include the multivariate GARCH and

the multivariate SV models. Despite the numerous modifications to multivariate volatility models, such models consider covariance as a latent variable and suffer from intraday information loss due to the use of low frequency data. However, this resulted in a considerable loss of information on inter-day trading data and also caused bias in estimating and forecasting the conditional volatility. Hence, with the availability of reliable high-frequency intraday asset prices, researchers were motivated to conduct further research aiming to primarily produce short-run volatility forecasts better. In this study, the existing HAR-type models and its extensions have been studied empirically to infer about their predictive power for forecasting realised volatilities by taking the case of S&P 500 futures. These findings add to the concepts of financial risk management and volatility forecasting. When faced with high equity market uncertainty, the findings will assist market participants to choose appropriate strategies to limit risk and maximize returns.

The structure of paper is as follows. Section 2 deals with genesis of HAR modelling. Some preliminaries and methodology are given in Section 3. A case study on real data of S&P 500 market is given in Section 4 followed by concluding remarks in Section 5.

2. Genesis of HAR modeling

Volatility is arguably referred to as a quantitative measure of risk where the higher the volatility, higher the risk of a specific asset and therefore it's forecast becomes crucial in areas such as portfolio management and asset allocation. Most available studies apply models based on low-frequency transaction data, such as GARCH, SV, and ARMA to forecast the volatility of crude oil futures (Chang *et al.*, 2010). Although these models perform well in predicting volatility in crude oil futures markets, they fail to capture the intraday activity patterns, the macroeconomic announcements and the volatility persistence that are separately quantified and have been shown to account for a substantial fraction of return variability, both at the intraday and daily level.

Giot and Laurent (2003) have employed GARCH-type models to create estimates for cocoa, coffee, and sugar futures price volatility. Tian *et al.* (2017) and Yang *et al.* (2017) on the other hand, used high-frequency data and enhanced Corsi's (2009) HAR model to create short-run volatility projections (up to 20 days ahead) motivated by the 'Heterogeneous Market Hypothesis' and the measure of RV.

The HAR-RV model uses high-frequency transaction data to successfully capture the main characteristics of financial data. Hence, many scholars extend the HAR-RV model by adding different decompositions of volatility components into the benchmark model (Andersen *et al.*, 2007; Patton and Sheppard, 2015; Gong and Lin, 2018). HAR-type models have been widely used to forecast crude oil futures volatility, and have been proven to be better than the traditional models which are based on low-frequency transaction data (Haugom *et al.*, 2014 and Andersen *et al.*, 2007) further proposed the use of BPV to separate the realised volatilities into continuous and jump components termed as HAR-CJ model. Corsi and Reno (2012) extended HAR model by adding the jump component and termed it as HAR-J model and also introduced without jump component (CHAR) and Quarticity component (CHAR-Q). Bollerslev *et al.* (2016) introduced the HAR-Q models using realised Quarticity (RQ) as an estimator of Integrated Quarticity (IQ) to capture temporal variation in the measurement error.

Degiannakis *et al.* (2022) used variants of the HAR model, and forecasted the realised volatility of agricultural commodities. They obtained data from Chicago Mercantile Exchange (CME)/ Intercontinental Exchange (ICE) with tick-by-tick data on five widely traded agricultural commodities (corn, rough rice, soybean, sugar, and wheat) during the period January 01, 2010 to June 30, 2017. The data was divided as In-sample estimation period and Out-sample forecasting period. Their results revealed that HAR model performed well when the variations in volatility measurements were decomposed into their continuous path and jump components.

Although the HAR-type models discussed above offer good predictive capacity for volatility forecasting, higher the prediction accuracy, better for risk management, financial asset pricing and portfolio optimization. Hence, it will be of interest to fit various HAR models on a given dataset (here, S&P 500 prices) and ascertain about which model better represent the underlying pattern and also their forecasting performance.

3. Preliminaries and methodology

3.1. Terminologies

The terminologies relating to volatility are described briefly. Implied volatility represents the current market price for volatility, or the fair value of volatility based on the market's expectation for movement over a defined period of time. Realised volatility is nothing but the assessment of variation in returns for an investment product when its historical returns within a defined time period are analysed. Analysts make use of high-frequency intraday data to determine measures of volatility at hourly/ daily/ weekly/ monthly frequency. Hence, volatility traders obviously care not only about what is expected but also what actually transpired. Note that in econometrics, sum of squared returns is called as realised volatility (Barndorff-Nielsen and Sheppard, 2004). Stochastic volatility models are similar to GARCH models but introduce a stochastic innovation term to the equation that describes the evolution of the conditional variance σ_t^2. To ensure positiveness of the conditional variances, stochastic volatility models are defined in terms of $ln\sigma_t^2$ instead of σ_t^2. If the autocorrelation function ρ_k of stationary ARMA(p, q) process decreases rapidly as $k \to \infty$, processes then it is often referred to as short memory processes. Stationary processes with much more slowly decreasing autocorrelation function are known as long memory processes. High-frequency data are mostly used in financial analysis and in high frequency trading which basically contain intraday observations that can be used to understand market behavior, dynamics, and micro-structures. Tick-by-tick market data, in which each single 'event' (transaction, quote, price movement, *etc.*) is characterised by a "tick" was first used to create high frequency data collections. The quantity of daily data acquired in 30 years can be equalled by high frequency observations over one day of a liquid market.

3.2. Tests used for realised variance measures

The Ljung-Box statistic is computed under the null hypothesis that there is no autocorrelation in the residuals in order to see whether the best-fitted model residuals are white noise or not. Normality of residuals can be tested by employing Shapiro-Wilk's (W) test. Jarque-Bera test is a goodness-of-fit test to test whether sample data have the skewness and kurtosis matching a normal distribution. Augmented Dickey-Fuller (ADF) test is used for

testing the presence of a unit root in a time series by under the assumption that the time series is non-stationary.

3.3. Realised volatility (RV)

RV is a model free measurement of financial market volatility and was proposed by Andersen *et al.* (2001, 2003) and Barndorff-Nielsen and Shephard (2002) by defining a continuous time diffusion process. Andersen *et al.* (2003) showed that, under suitable conditions, including the absence of serial correlation in the intraday returns, RV is a consistent estimator of Integrated Volatility (IV_t). Hence

$$RV_t = \sum_{i=1}^{m} r_{t,i}^2 \xrightarrow{P} \int_{t-1}^{t} \sigma_s^2 ds$$

at day $t = 1, 2, ..., M$ for $i = 2, 3, ..., m$ with the number of intraday observations as m and the total number of observation days as M. Following Andersen and Bollerslev (1998), discretizing the data by equidistant sampling, might introduce intraday price jumps which translate into higher realised variances. In order to obtain a more robust measure of the realised volatility, Barndorff-Nielsen and Sheppard (2004) introduced the concept of the BPV for separating the realised variance into a continuous part and a discontinuous (jump) part. Using the approach of Huang (2004), the jump component is identified. Hence RV provides an ex-post measure of the true total variation including the discontinuous jump part.

3.4. HAR models

With the widespread availability of high-frequency intraday data, the recent literature has focused on employing RV to build forecasting models for time-varying return volatility. Among these forecasting models, the HAR model proposed by Corsi (2009) has gained popularity due to its simplicity and consistent forecasting performance in applications. The formulation of the HAR model is based on a straightforward extension of the heterogeneous ARCH (HARCH) class of models dealt by Muller *et al.* (1997). Under this approach, the conditional variance of the discretely sampled returns is parameterized as a linear function of lagged squared returns over the same horizon together with the squared returns over longer and/or shorter horizons.

The original HAR model specifies RV as a linear function of daily, weekly and monthly realised variance components, and can be expressed as

$$RV_t = \beta_o + \beta_1 RV_{t-1}^d + \beta_2 RV_{t-1}^w + \beta_3 RV_{t-1}^m + \varepsilon_t$$

where β_j ($j = 0, 1, 2, 3$) are unknown parameters that need to be estimated, RV_t is the realised variance of day t, and $RV_{t-1}^d = RV_{t-1}, RV_{t-1}^w = \frac{1}{5}\sum_{i=1}^{5} RV_{t-i}, RV_{t-1}^m = \frac{1}{22}\sum_{i=1}^{22} RV_{t-i}$ denote the daily, weekly and monthly lagged realised variance, respectively. This specification of RV parsimoniously captures the high persistence observed in most realised variance series. The various types of HAR models are discussed subsequently.

3.4.1. Standard HAR model

$$RV_{t+h}^{(h)} = \beta_0^{(t)} + \beta_1^{(t)} RV_t + \beta_2^{(t)} RV_t^{(5)} + \beta_3^{(t)} RV_t^{(22)} + \varepsilon_{t+h}^{(h)}$$

where RV_t denotes the previous day's volatility $RV_t^{(5)}$ denotes the averaged volatility during the previous week, and $RV_t^{(22)}$ denotes the averaged volatility over the previous month, h denotes the forecasting horizon.

3.4.2. HAR-J model

Augmenting the above standard HAR with the simple jump measure forms HAR-J model.

$$RV_{t+h}^{(h)} = \beta_0^{(t)} + \beta_1^{(t)} RV_t + \beta_2^{(t)} RV_t^{(5)} + \beta_3^{(t)} RV_t^{(22)} + \varepsilon_{t+h}^{(h)} + \beta_4^{(t)} RJ_t + \varepsilon_{t+h}^{(h)}$$

where RJ_t is the daily discontinuous jump variation.

3.4.3. HAR-Q model

It is obtained by using Realised Quarticity (RQ) as an estimator of Integrated Quarticity (IQ) to capture temporal variation in the measurement error by Bollerslev *et al.* (2016).

$$RV_{t+h}^{(h)} = \beta_0 + \beta_1^{(t)} RV_t + Q^{(1)} \underline{RQ_t^{1/2}} + \beta_2^{(t)} RV_t^{(5)} + \beta_3^{(t)} RV_t^{(22)} + \varepsilon_{t+h}^{(h)}$$

where $\underline{RQ_t^{1/2}}$ is the daily lagged realised quarticity and it is useful as most of the attenuation bias in the forecasts (due to RV_t being less persistent than unobserved IV_t) is due to the estimation error in RV_{t-1} . In other words, RQ_t as an estimator of IQ_t to capture temporal variation in the measurement error with $\underline{RQ_t^{1/2}}$ as the de-meaned values of $RQ_t^{1/2}$ for easy interpretation.

3.4.4. HAR-CJ model

Andersen *et al.* (2007) further proposed the use of BPV to separate the realised volatilities into continuous and jump components, which model is resulted as HAR-CJ and defined as

$$RV_{t+h}^{(h)} = \beta_0^{(t)} + \beta_1^{(t)} C_t + \beta_2^{(t)} C_t^{(5)} + \beta_3^{(t)} C_t^{(22)} + J^{(1)} RJ_t + J^{(5)} RJ_t^{(5)} + J^{(22)} RJ_t^{(22)} + \varepsilon_{t+h}^{(h)}$$

where C_t and RJ_t are continuous and discontinuous jump components respectively.

3.4.5. HAR-QJ model

It is obtained by using standard HAR along with previous day's Quarticity and jump component respectively.

$$RV_{t+h}^{(h)} = \beta_0 + \beta_1^{(t)} RV_t + Q^{(1)} \underline{RQ_t^{1/2}} + J^{(1)} RJ_t + \beta_2^{(t)} RV_t^{(5)} + \beta_3^{(t)} RV_t^{(22)} + \varepsilon_{t+h}^{(h)}$$

where RQ as an estimator of IQ to capture temporal variation in the measurement error. Using $\underline{RQ_t^{1/2}}$ as the de-meaned values of $RQ_t^{1/2}$ and RJ_t is the daily discontinuous jump variation.

3.4.6. CHAR model

$$RV_{t+h}^{(h)} = \beta_0^{(t)} + \beta_1^{(t)} C_t + \beta_2^{(t)} C_t^{(5)} + \beta_3^{(t)} C_t^{(22)} + \varepsilon_{t+h}^{(h)}$$

where, C_t, $C_t^{(5)}$ and $C_t^{(22)}$ are respectively, the daily continuous path variation, the daily average over the past five days and daily average over the past 22 days at time t. Without jump component, it is better at capturing volatility persistence and long memory than RV in HAR model.

3.4.7. CHAR-Q model

$$RV_{t+h}^{(h)} = \beta_0^{(t)} + \beta_1^{(t)} C_t + \beta_2^{(t)} C_t^{(5)} + \beta_3^{(t)} C_t^{(22)} + \beta_4^{(t)} (TPQ)^{1/2} + \varepsilon_{t+h}^{(h)}$$

where, $TPQ^{1/2}$ is Tri-power quarticity, which is consistent for the integrated quarticity in the presence of jumps.

3.5. Forecasting and evaluation

To quantitatively evaluate the forecasting of each model, three popular accuracy measures, namely the Mean Squared Prediction Error (MSPE), the Mean Absolute Prediction Error (MAPE), and Quasi Likelihood (QLIKE) by Patton (2011) have been used (and multiplied by 100 to express in percentages):

$$\text{MSPE} = \sqrt{N^{-1} \sum_{t=1}^{N} \left(RV_t - \widehat{RV}_t \right)^2}$$

$$\text{MAPE} = N^{-1} \sum_{t=1}^{N} \frac{\left| RV_t - \widehat{RV}_t \right|}{RV_t}$$

$$\text{QLIKE} = N^{-1} \sum_{t=1}^{N} \left(\log \widehat{RV}_t + \frac{\left| \widehat{RV}_t \right|}{RV_t} \right)$$

where RV_t and $\widehat{RV}_t$ are the actual and the forecasted RV respectively at the different forecasting horizons, and N is the number of real out-of-sample forecasts.

4. Case study

The Standard and Poor's 500 (S&P 500) is usually referred as leading indicator of the stock market in the United States. The S&P 500 index is made up of 500 large-cap stocks that represent the most important industries in the US economy. Furthermore, because of their high liquidity, they can easily be bought or sold in the market without influencing the asset price. Forecasting of asset return volatility S&P 500 index futures prices from Tick Data Inc (http://public.econ.duke.edu/ ap172/code.html), during April 8, 1997 to August 30, 2013 (4096 trading days) has been considered in this case study. These data points

Table 1: Descriptive statistics for S&P 500 market realised volatility measures, relative jump component and bi-power variation

Descriptive Statistics	RV	RJ	BPV
Mean	1.17	0.09	1.11
Minimum	0.04	0.00	0.03
Maximum	60.56	10.25	50.31
Standard Deviation	2.31	0.36	2.21
Skewness	10.02	17.6	9.36
Kurtosis	166.92	398.99	134.74
Tests used for checking assumptions for required analysis			
Ljung-Box test Q (1)	1735.60**	24.68**	1852.90**
Ljung-Box test Q (5)	7117.20**	135.04**	7747.80**
Ljung-Box test Q (10)	11989.00**	531.26**	12821.00**
Ljung-Box test Q (22)	20042.00**	884.56**	21112.00**
Shapiro-Wilk's (W) test	0.35**	0.19**	0.36**
J-B test	4654549.00**	26974481.00**	3021571.00**
ADF test	-6.86**	-8.87**	-7.09**
PP test	-2042.90**	-4966.50**	-1864.00**

Note: Asterisks (** and *) indicate statistical significance at 1% and 5% respectively.

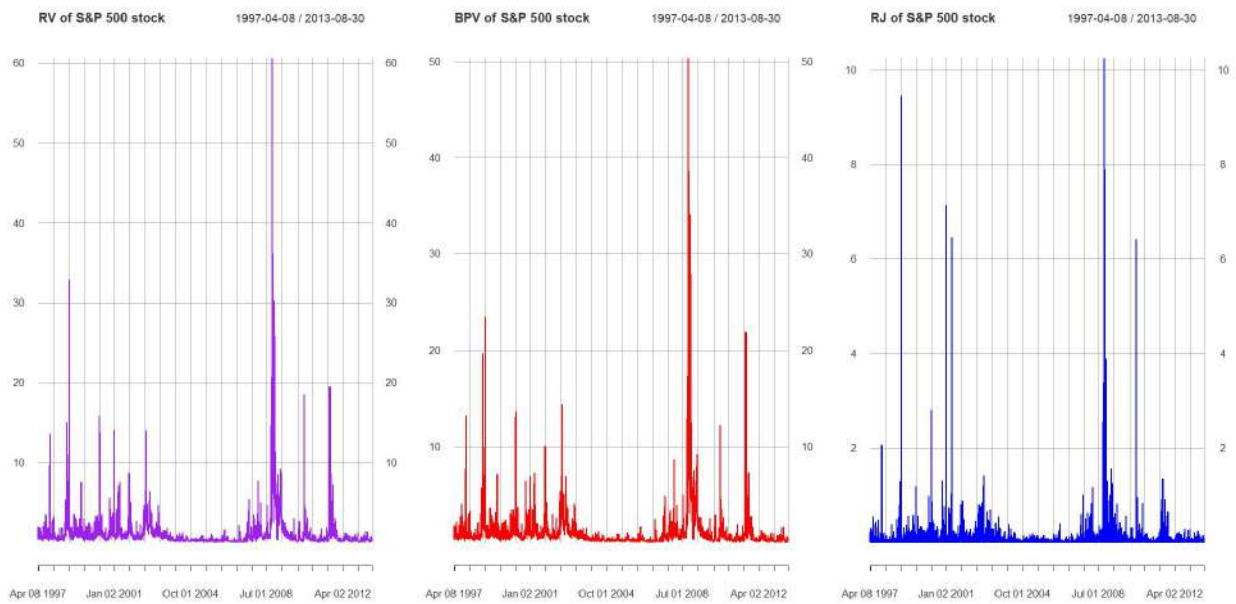


Figure 1: Realised volatility (RV), bi-power variation (BPV) and realised jump (RJ) components in S&P 500 stock

available as tick-by-tick data prices, which were 23,400 data points for each trading day resulted in 9,58,46,400 data points. In order to avoid microstructure noise, aggregating the data to 5 min prices led to 3,19,488 data points. For in-sample analysis, April 8, 1997 to January 06, 2012 (3686 trading days) period was considered which is 90 per cent of 4096

trading days for estimation period, whereas, the remaining period till August 30, 2013 were used as real out-of-sample forecasts based on rolling window approach. The use of rolling window approach works best to capture changes in the market conditions as suggested by Degiannakis and Filis (2017), Degiannakis *et al.* (2018), and Engle *et al.* (1990). The same data which has been used in this study has been used as a default data in the R package for fitting HAR models, but, in this study, the data has been aggregated to 5 min prices and also more variants of HAR models have been tried for comparison purposes.

4.1. Empirical results

Table 1 provides an overview of descriptive statistics and test statistics of the Ljung–test for one, five, ten and 22 lags (trading days). The descriptive statistics like skewness and kurtosis indicate the data considered were very much erratic. All the data series had positive skewness and were highly leptokurtic in nature. W and J-B tests for normality showed that all the series deviate from normality. Phillip-Perron (PP) test employed to test a unit root in a time series indicated presence of stationarity in the RV, RJ and BPV time series.

Table 2: In-sample HAR results for S&P 500 with RV

HAR	$h = 1$	$h = 5$	$h = 10$	$h = 22$	$h = 44$	$h = 66$
β_0	0.12** (3.23)	0.18** (7.64)	0.24** (10.46)	0.37** (15.53)	0.59** (22.89)	0.72** (27.90)
$\beta_1^{(1)}$	0.22** (3.24)	0.18** (13.11)	0.13** (10.05)	0.10** (7.49)	0.08* (5.59)	0.06** (4.24)
$\beta_2^{(5)}$	0.49** (3.33)	0.39** (16.55)	0.37* (16.51)	0.33* (14.22)	0.30** (12.31)	0.21** (8.68)
$\beta_3^{(22)}$	0.18** (3.04)	0.26** (12.50)	0.28** (14.03)	0.26** (12.55)	0.14** (6.44)	0.15** (7.15)
$AdjR^2$	0.51	0.63	0.62	0.54	0.38	0.30
AIC	3852.83	1444.90	1134.06	1301.31	1809.72	1676.70
BIC	3883.86	1475.92	1165.07	1332.30	1809.74	1676.70
RMSE	1.69	1.22	1.17	1.19	1.28	1.26
Q-LIKE	0.15	0.12	0.13	0.17	0.22	0.25

Note: Parenthesis in the above table indicates test statistic value. Asterisks (** and *) indicate statistical significance at 1% and 5% respectively.

The measures for realised volatilities for S&P 500 stock index have shown significant autocorrelations at 1, 5, 10, 22 lags, tested with Ljung-Box chi-square test. This motivated further for the application of autoregressive models such as HAR and its extensions. Astonishingly, even the jump components (J_t) showed autoregressive behavior of jumps indicating that because of the impact of major economic events, there were structural breaks in the volatility of returns of financial assets, which feature may help in improving the predictive ability of the HAR-type models. As the continuous component refers to the realised volatility that remained after discarding jumps, the Ljung Box test statistics, ADF, W, J-B test and PP test were naturally much higher and had similar patterns like realised volatility components. Figure 1 depicts the Realised Volatilities RV, BPV and RJ components in S&P 500 stock price index considered. It can be seen from Figure 1 that RV plot subsumes BPV along with other components whereas BPV consists of both continuous and jump components. In

Figure 1, the third plot relating to jump component arises due to the intra-day variations in the data which occur on a daily basis whose magnitudes and ranges are much smaller than the other two components as can be seen in the plot.

Huge spike of realised volatility in 2009 can be observed in Figure 1 through Figure 8. This is so because S&P 500 market price bottomed out during 2008-2009 owing to financial crisis that resulted in great U.S market recession. S&P 500 lost approximately 50% of its value due to market crash and took two years to recover from it. As a result, squared returns increased irrespective of direction of their original values leading to sudden increase in realised volatility.

4.2. In-sample parameter estimation results

In-sample analysis results are presented in Tables 2 through 8 for RV of S&P 500 market with Figures 2 through 8 depicting these results. In Tables 2 through 8, the estimation results and model performance accuracy measures have been reported for the seven models considered *viz.*, HAR, HAR-J, HAR-Q, HAR-CJ, HAR-QJ, CHAR and CHAR-Q at six prediction horizons (h = 1, 5, 10, 22, 44 and 66 days). The analysis was done using R software. For fitting the models, Ordinary Least Squares (OLS) estimation was employed. It can be seen that most of the parameters were significant at the 1% level, suggesting strong persistence in the realised volatility. The Adjusted R^2 (higher values), AIC, BIC, RMSE and Q-LIKE (comparatively smaller) measures at 5 days and 10 days ahead prediction horizons revealed that the fitted models performed well for these days as compared to 1, 22, 44, 66 days prediction horizons. In HAR and HARJ models, all the parameters were significant at all horizons. Most of the parameters of the HAR-Q, HAR-CJ, HAR-QJ models were significant at short and medium horizons, but for long memory horizons some of the components like Quarticity (especially for HAR-QJ) and jump components showed non-significance. CHAR model showed significant contribution by all continuous components.

When all the seven HAR and its extension models fitted were compared based on their prediction performances, CHAR-Q type of HAR model came out to be the best model at horizons $h = 5$ and $h = 10$ and hence can be considered superior with regard to model fit. This shows that the continuous component along with the Quarticity component work better as compared to all other models for S&P 500 stock market price index data.

In-sample analysis results showed that as the h-day-ahead horizon increases, the HAR and its extension models fail to estimate well compared to the short and medium memory realised volatility.

4.3. Out-of-sample forecasting results

The out-of-sample predictive performance of seven models were compared by using a rolling window prediction method for forecasting the volatility of S&P 500 stock price returns over the multi-period horizons (1, 5, 10, 22, 44 and 66 days). For this, firstly, the whole sample was divided into two sub-samples called "estimation sample" and "prediction sample". Estimation sample is the estimation window containing the 3686 days at any given time starting from the first day (with rolling window method, the period shifts by one day every time, but the sample size will remain 3686), the prediction sample contained days from

Table 3: In-sample HAR-J results for S&P 500 with RV

HAR-J	$h=1$	$h=5$	$h=10$	$h=22$	$h=44$	$h=66$
β_0	0.13** (3.86)	0.19* (7.92)	0.25* (10.66)	0.38** (15.63)	0.59** (22.95)	0.72** (27.95)
$\beta_1^{(1)}$	0.35** (15.71)	0.27** (16.44)	0.19** (12.32)	0.13** (8.40)	0.10** (6.10)	0.08** (4.74)
$\beta_2^{(5)}$	0.43** (13.14)	0.35** (15.04)	0.35** (15.31)	0.31** (13.47)	0.29** (11.76)	0.20** (8.24)
$\beta_3^{(22)}$	-0.18** (6.27)	0.26** (12.65)	0.28** (14.11)	0.26** (12.57)	0.14** (6.44)	0.15** (7.15)
$J^{(1)}$	-1.00** (3.39)	-0.64** (4.98)	-0.45** (4.86)	-0.25** (1.90)	-0.18** (1.55)	-0.15** (1.55)
$AdjR^2$	0.53	0.64	0.63	0.55	0.39	0.30
AIC	3733.44	1351.50	1086.37	1288.54	1731.86	1674.01
BIC	3770.66	1388.71	1123.58	1325.72	1769.03	1711.12
RMSE	1.66	1.20	1.16	1.19	1.27	1.26
Q-LIKE	0.16	0.12	0.13	0.17	0.22	0.25

Note: Parenthesis in the above table indicates test statistic value. Asterisks (** and *) indicate statistical significance at 1% and 5% respectively.

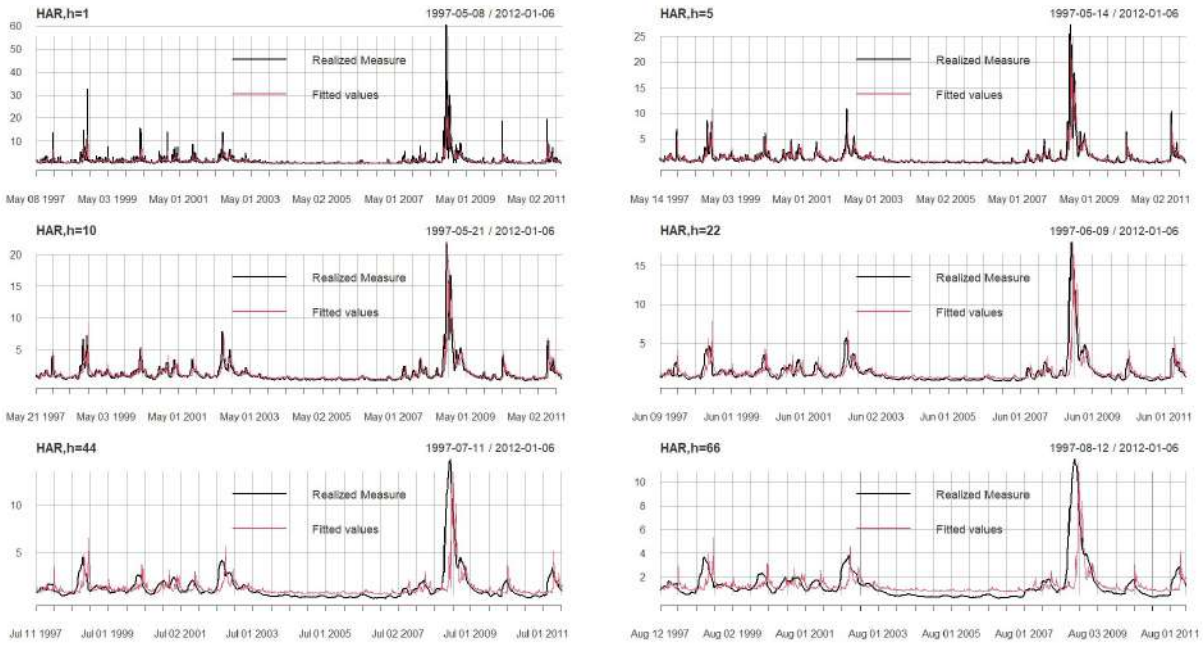


Figure 2: Plots for the fitted HAR model at different horizons

the 3687^{th} day till the end of data period. In this way, on estimation samples, models were fitted to compute the predicted values for each of the subsequent 1-day, 5-days, 10-days, 22-days, 44-days and 66-days periods for the given samples.

It is noted here that the estimation sample was moved forward every time by one

Table 4: In-sample HAR-Q results for S&P 500 with RV

HAR-Q	$h = 1$	$h = 5$	$h = 10$	$h = 22$	$h = 44$	$h = 66$
β_0	-0.01	0.02	0.18*	0.32**	0.55**	0.68**
	(0.25)	(0.82)	(1.77)	(2.86)	(3.71)	(3.00)
$\beta_1^{(1)}$	0.59**	0.18**	0.30**	0.24**	0.19**	0.16**
	(21.55)	(13.73)	(5.35)	(2.55)	(2.03)	(1.95)
$\beta_2^{(5)}$	0.35**	0.68**	0.31*	0.27*	0.26**	0.17**
	(11.00)	(22.30)	(1.84)	(1.88)	(2.08)	(2.02)
$\beta_3^{(22)}$	0.09**	0.15**	0.24	0.22	0.11	0.13
	(3.35)	(7.21)	(1.12)	(1.16)	(0.78)	(1.13)
$Q^{(1)}$	-0.36**	-0.56**	-0.16**	-0.14**	-0.11**	-0.10*
	(18.28)	(14.41)	(5.04)	(2.12)	(1.70)	(1.68)
$AdjR^2$	0.56	0.65	0.64	0.56	0.40	0.30
AIC	3519.61	1224.52	991.53	1201.99	1638.03	1635.26
BIC	3556.85	1261.76	1028.75	1239.19	1720.20	1672.38
RMSE	1.61	1.18	1.14	1.18	1.26	1.25
Q-LIKE	0.14	0.11	0.11	0.14	0.20	0.24

Note: Parenthesis in the above table indicates test statistic value. Asterisks (** and *) indicate statistical significance at 1% and 5% respectively.

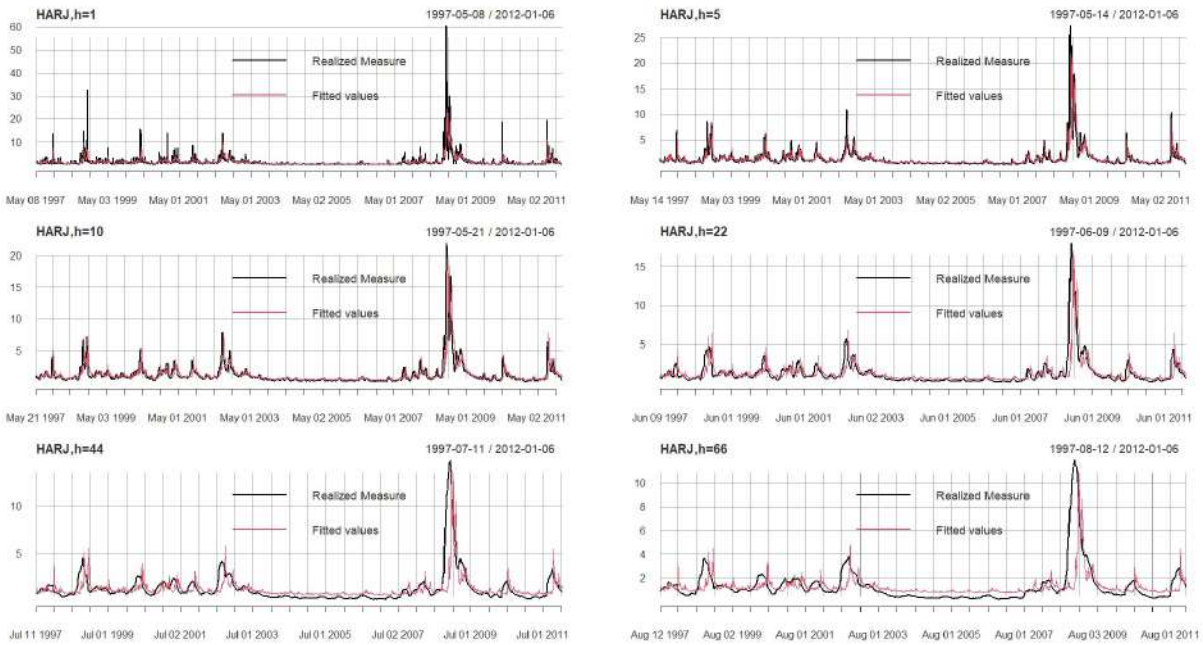


Figure 3: Plots for the fitted HAR-J model at different horizons

day. The estimation sample still contained 3686 observations, the last estimation sample with same number of observations but with the last observation in it belonging to the 4095^{th} day. The predicted values of the 1, 5, 10, 22, 44 and 66 days were obtained from the fitted models on each of these 410 estimation samples. The forecasting accuracies of each model were measured using MSPE, MAPE and Q-LIKE functions to evaluate the deviation between

Table 5: In-sample HAR-CJ results for S&P 500 with RV

HAR-CJ	$h=1$	$h=5$	$h=10$	$h=22$	$h=44$	$h=66$
β_0	0.07	0.17**	0.27**	0.39**	0.57**	0.66**
	(1.30)	(2.15)	(2.44)	(2.41)	(2.49)	(2.10)
$C_1^{(1)}$	0.33**	0.22**	0.18**	0.13**	0.10	0.08
	(3.58)	(4.20)	(3.02)	(1.95)	(1.49)	(1.29)
$C_2^{(5)}$	0.58**	0.56**	0.43**	0.35**	0.31**	0.22*
	(2.19)	(3.26)	(2.49)	(2.26)	(2.15)	(1.92)
$C_3^{(22)}$	0.05	0.06	0.14	0.19	0.13	0.21*
	(0.32)	(0.35)	(0.61)	(0.77)	(0.66)	(1.83)
$J_1^{(1)}$	-0.56*	-0.14	-0.20	-0.12	-0.09	-0.05
	(1.67)	(0.45)	(0.94)	(0.77)	(0.67)	(0.61)
$J_2^{(5)}$	-1.11	-1.83	-0.59	-0.04	0.06	0.09
	(0.90)	(1.57)	(0.85)	(0.06)	(0.10)	(0.16)
$J_3^{(22)}$	1.66	2.64	2.18	1.21	0.24	-0.67
	(1.07)	(1.19)	(0.96)	(0.62)	(0.20)	(0.53)
$AdjR^2$	0.54	0.66	0.63	0.55	0.39	0.30
AIC	3692.10	1288.01	1055.27	1283.86	1731.00	1667.71
BIC	3714.75	1277.65	1104.90	1333.47	1780.56	1717.22
RMSE	1.65	1.18	1.15	1.19	1.27	1.26
Q-LIKE	0.15	0.10	0.11	0.16	0.21	0.24

Note: Parenthesis in the above table indicates test statistic value. Asterisks (** and *) indicate statistical significance at 1% and 5% respectively.

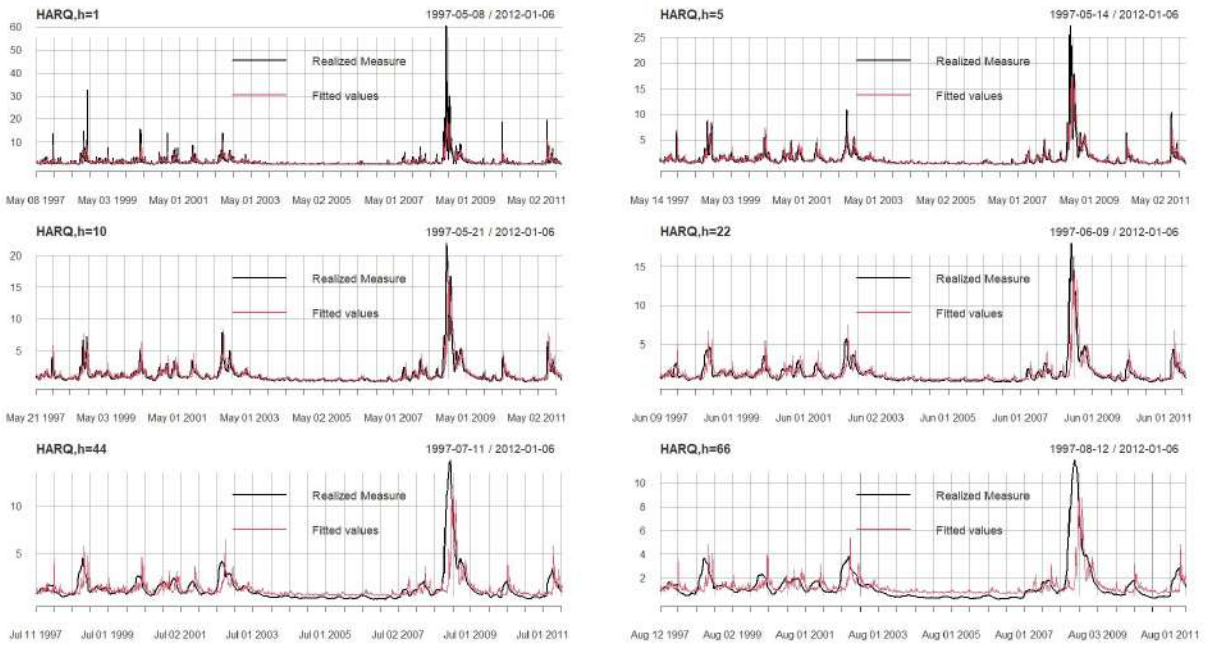


Figure 4: Plots for the fitted HAR-Q model at different horizons

Table 6: In-sample HAR-QJ results for S&P 500 with RV

HAR-QJ	$h = 1$	$h = 5$	$h = 10$	$h = 22$	$h = 44$	$h = 66$
β_0	0.00	0.12*	0.19*	0.32**	0.54**	0.68**
	(0.17)	(1.72)	(1.95)	(2.88)	(3.88)	(3.15)
$\beta_1^{(1)}$	0.60**	0.40**	0.30**	0.24**	0.19**	0.15**
	(21.83)	(5.26)	(5.17)	(2.63)	(2.16)	(1.92)
$\beta_2^{(5)}$	0.35**	0.31**	0.31*	0.28*	0.26**	0.17**
	(10.80)	(2.94)	(1.87)	(1.88)	(2.09)	(1.13)
$\beta_3^{(22)}$	0.10**	0.22*	0.25	0.22	0.11	0.13
	(3.62)	(1.79)	(1.15)	(1.11)	(0.77)	(1.13)
$Q^{(1)}$	-0.33 **	-0.25	-0.13	0.05	0.06	0.06
	(14.83)	(0.89)	(0.82)	(0.25)	(0.33)	(0.41)
$J^{(1)}$	-0.33**	-0.19**	-0.15**	-0.15*	-0.12	-0.10
	(3.39)	(4.98)	(4.86)	(1.90)	(1.55)	(1.55)
$AdjR^2$	0.56	0.66	0.64	0.56	0.40	0.30
AIC	3508.58	1214.08	989.88	1203.43	1684.29	1636.49
BIC	3552.03	1257.52	1033.31	1246.84	1727.66	1679.81
RMSE	1.61	1.18	1.14	1.18	1.26	1.25
Q-LIKE	0.13	1.1	0.12	0.15	0.21	0.23

Note: Parenthesis in the above table indicates test statistic value. Asterisks (** and *) indicate statistical significance at 1% and 5% respectively.

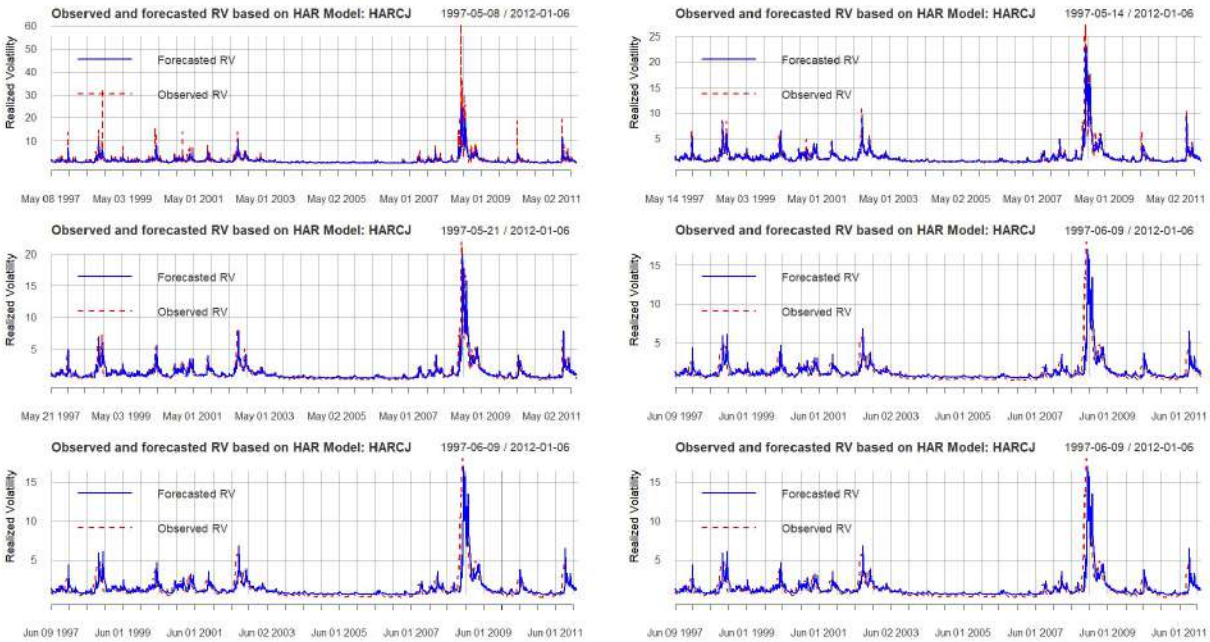


Figure 5: Plots for the fitted HAR-CJ model at different horizons

the predicted values and the true values of realised volatilities.

Table 9 and Figure 9 report the values of forecasting performance measures of all the

Table 7: In-sample CHAR results for S&P 500 with RV

CHAR	$h = 1$	$h = 5$	$h = 10$	$h = 22$	$h = 44$	$h = 66$
β_0	0.14** (4.36)	0.21** (8.85)	0.27** (11.71)	0.40** (16.61)	0.61** (23.76)	0.73** (28.69)
$\beta_1^{(1)}$	0.26** (12.62)	0.20** (13.91)	0.16** (11.05)	0.12** (8.03)	0.09** (5.94)	0.07** (4.49)
$\beta_2^{(5)}$	0.49** (14.54)	0.42** (17.15)	0.37** (15.92)	0.32** (13.43)	0.30** (11.66)	0.21** (8.18)
$\beta_3^{(22)}$	0.17** (5.75)	0.24** (11.58)	0.28** (13.78)	0.27** (12.68)	0.15** (6.56)	0.16** (7.32)
$AdjR^2$	0.53	0.65	0.63	0.54	0.39	0.29
AIC	3757.68	1315.29	1099.14	1309.77	1746.68	1682.51
BIC	3788.71	1346.29	1130.16	1340.77	1777.63	1713.43
RMSE	1.67	1.2	1.16	1.2	1.27	1.26
Q-LIKE	0.15	0.12	0.14	0.17	0.22	0.24

Note: Parenthesis in the above table indicates test statistic value. Asterisks (** and *) indicate statistical significance at 1% and 5% respectively.

Table 8: In-sample CHAR-Q results for S&P 500 with RV

CHAR-Q	$h = 1$	$h = 5$	$h = 10$	$h = 22$	$h = 44$	$h = 66$
β_0	0.03 (0.90)	0.14** (0.03)	0.22** (2.40)	0.35** (3.36)	0.57** (3.63)	0.70** (3.82)
$\beta_1^{(1)}$	0.55** (20.24)	0.37 ** (4.89)	0.28** (4.69)	0.22** (2.23)	0.17* (1.77)	0.14** (1.23)
$\beta_2^{(5)}$	0.40** (12.16)	0.36** (3.28)	0.33* (1.91)	0.29* (1.86)	0.27** (1.99)	0.18** (1.72)
$\beta_3^{(22)}$	0.10 (3.60)	0.21 (1.53)	0.25 (1.16)	0.24 (1.21)	0.13 (0.82)	0.14 (1.16)
$\beta_4^{(1)}$	-0.35** (15.72)	-0.20** (4.43)	-0.16** (4.06)	-0.13* (1.72)	-0.10 (1.37)	-0.09 (1.40)
$AdjR^2$	0.56	0.66	0.64	0.55	0.39	0.30
AIC	3510.40	1159.47	997.35	1244.92	1712.05	1653.44
BIC	3547.64	1196.68	1034.55	1282.12	1749.22	1690.57
RMSE	1.61	1.17	1.14	1.18	1.26	1.26
Q-LIKE	0.13	0.10	0.12	0.17	0.22	0.23

Note: Parenthesis in the above table indicates test statistic value. Asterisks (** and *) indicate statistical significance at 1% and 5% respectively.

models for forecasting realised volatilities at 1, 5, 10, 22, 44 and 66 days. These results showed that extensions of HAR-type models using BPV, jump and quarticity components tend to have the good prediction accuracies. Moreover, it can be seen that the forecasting accuracy decreases with increase in prediction horizon, which indicates that HAR-type models are more accurate in predicting realised volatilities in the short and medium runs. For forecasting horizon $h = 1$, 5, 10 and 66-days, as per the Q-LIKE function, CHAR-Q performed better

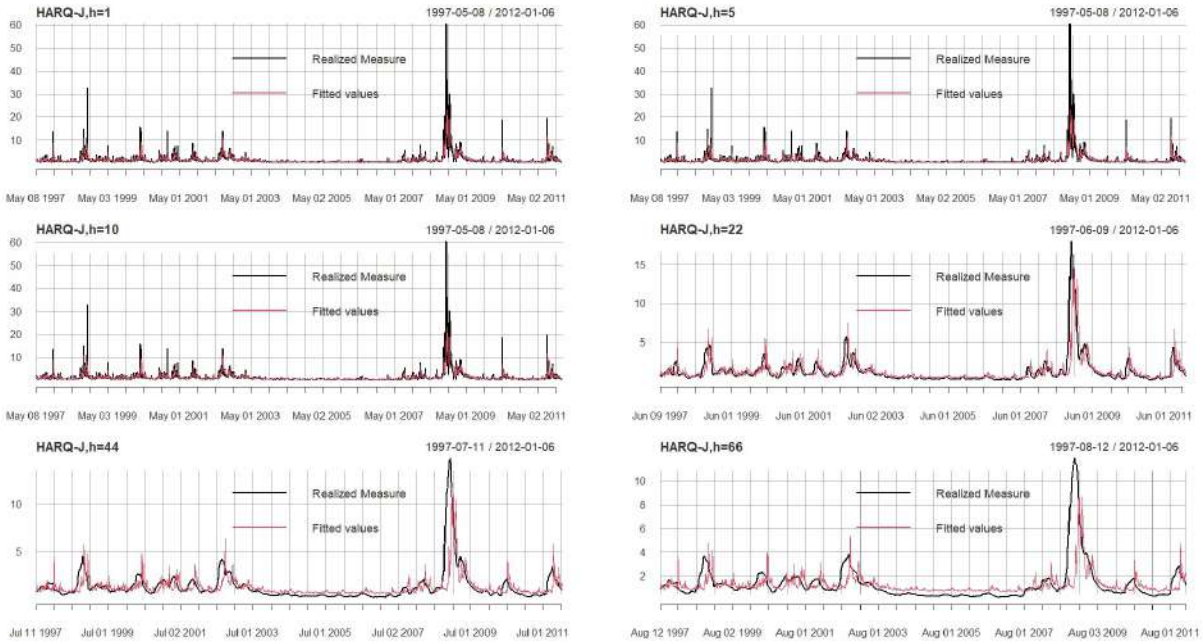


Figure 6: Plots for the fitted HAR-QJ model at different horizons

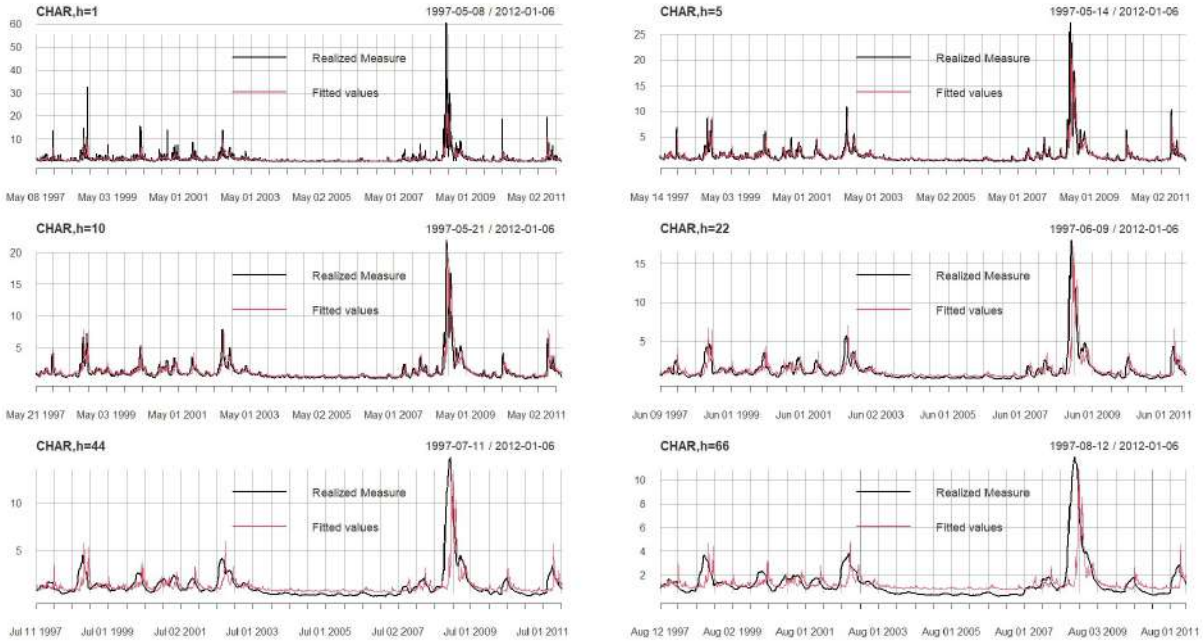


Figure 7: Plots for the fitted CHAR model at different horizons

than all other HAR model types while only for $h = 22$ and 44 horizons HARQJ-model performed well. When MSPE is considered for $h = 1$, 5, CHAR-Q performed well whereas at $h = 22$ and 66 days, CHAR performed better and for $h = 10$, HAR-QJ model performed better. At $h = 44$, HARQ model performed well for forecasting realised volatility. Overall, while considering all these measures, CHARQ, HAR-QJ, CHAR and HAR-Q performed well as compared to all other HAR and its extensions.

Table 9: Forecasting evaluation for S&P market with RV

Accuracy measures	h	HAR	HAR-J	HAR-Q	HAR-QJ	HARCJ	CHAR	CHAR-Q
MSPE	1	5.17	5.10	4.99	4.89	4.42	4.4	4.29
	5	3.92	3.84	3.15	2.89	5.64	5.62	2.21
	10	4.96	4.82	3.88	3.79	5.60	6.31	5.47
	22	9.44	9.36	8.12	7.26	6.88	6.85	7.27
	44	18.96	18.82	15.4	16.24	17.04	16.83	18.63
	66	27.09	27.07	24.06	24.62	25.23	17.23	20.63
MAPE	1	17.96	17.52	15.38	15.37	14.98	14.99	14.77
	5	22.55	22.1	19.58	18.84	17.27	17.25	17.11
	10	26.39	25.94	23.86	23.59	17.38	18.43	18.37
	22	34.85	34.72	32.96	31.63	19.75	19.67	20.28
	44	47.83	47.67	43.85	44.79	30.33	19.92	42.38
	66	57.49	57.42	53.66	54.53	42.13	20.61	44.29
Q-LIKE	1	14.87	14.55	13.98	13.19	14.73	14.88	12.00
	5	11.49	11.15	8.50	8.47	21.17	21.07	8.21
	10	12.65	12.37	9.90	9.74	21.52	25.21	9.79
	22	16.95	16.81	14.51	13.209	26.23	26.19	13.80
	44	23.32	23.21	19.35	20.29	28.79	28.15	27.69
	66	26.19	26.12	22.4	23.35	35.38	29.35	23.15

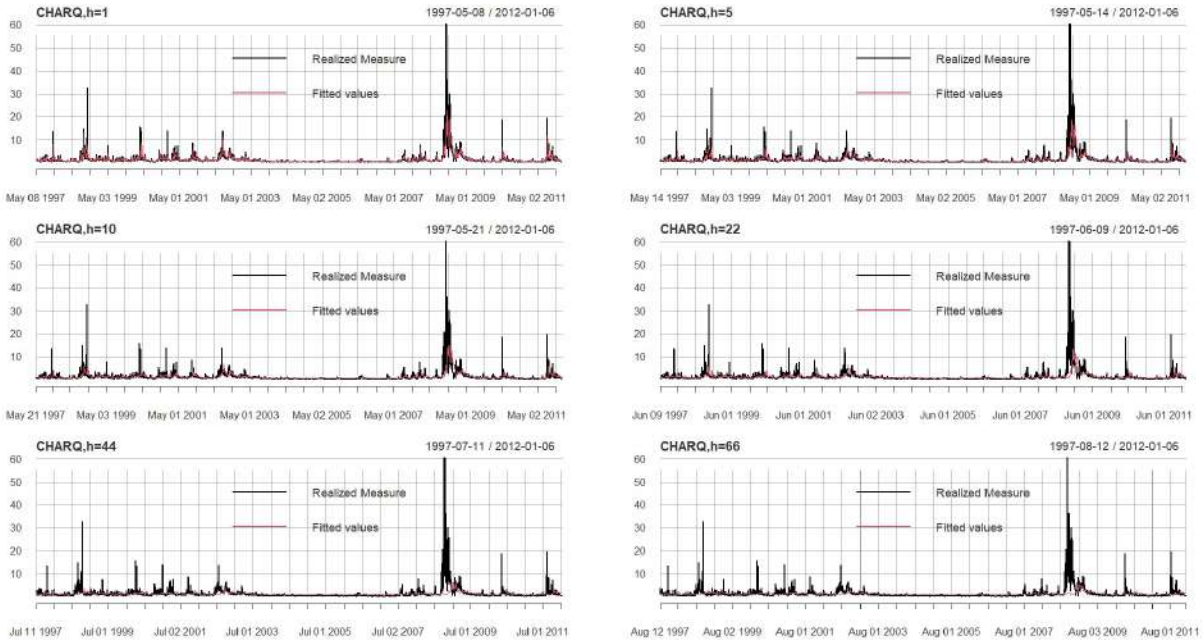


Figure 8: Plots for the fitted CHARQ model at different horizons

5. Concluding remarks

HAR models were studied along with their extensions for dynamic modeling of realised variance behaviour and its advantages over intraday were brought out which is widely used in high frequency data structure in order to capture the noise present in the intraday.

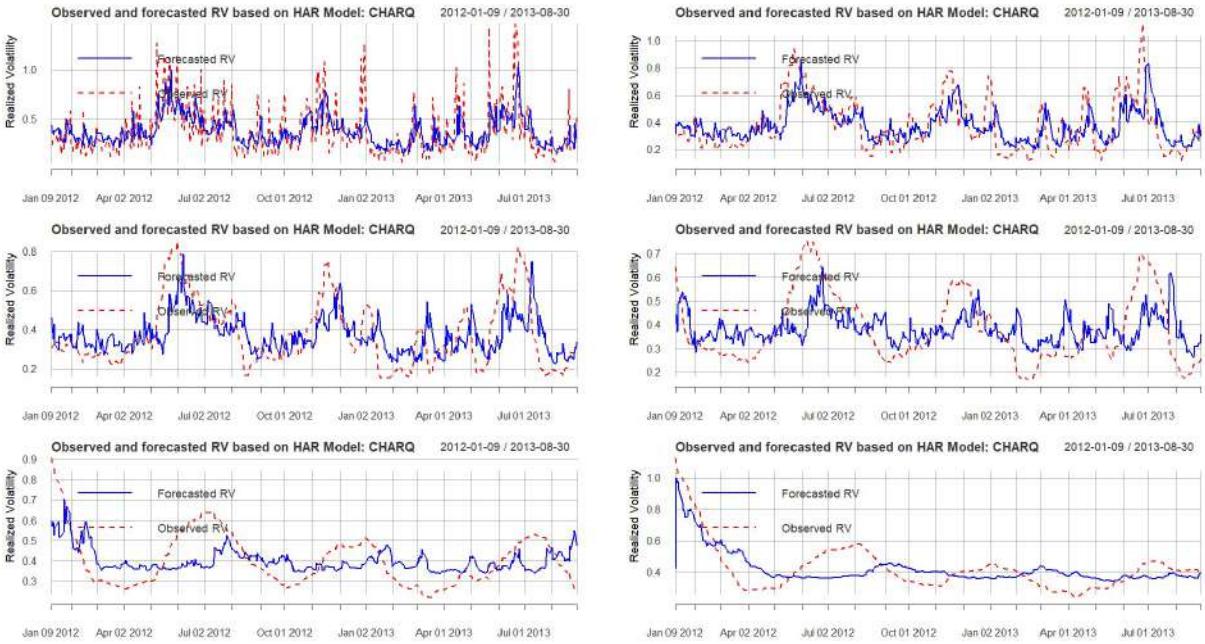


Figure 9: Plots for the forecasted CHARQ model at different horizons

An attempt has been done on a real dataset relating to Standard and Poor's 500 (S&P 500) stock market high frequency data and its volatility was estimated by using HAR models and its extensions and were compared on different horizons with their volatility were studied. The In-sample estimation results revealed that CHAR-Q models performed well in the estimation period compared to all other models. The out-sample forecasting results revealed that extensions of HAR-type models using BPV, jump and quarticity components tend to have the good prediction accuracies, more so for short run periods. In short, by way of an example, volatility on monetary policy announcement today will be more sensitive to the market mood on the pre-announcement day than on other days.

Acknowledgements

The facilities provided by ICAR-Indian Agricultural Statistics Research Institute (IASRI), New Delhi and the funding granted to the first author by Indian Council of Agricultural Research in the form of IARI-SRF fellowship is duly acknowledged for carrying out this study, which is his credit seminar delivered as part of the PhD course curriculum being pursued at ICAR-IASRI. The authors also thank the Chair Editor and reviewer for helpful comments which led to considerable improvement in the paper.

References

Andersen, T. G. and Bollerslev, T. (1998). Answering the skeptics: Yes, standard volatility models do provide accurate forecasts. *International Economic Review*, **39**, 885–905.

Andersen, T. G., Bollerslev, T., and Diebold, F. X. (2007). Roughing it up: Including jump components in the measurement, modeling, and forecasting of return volatility. *The Review of Economics and Statistics*, **89**, 701–720.

Andersen, T. G., Bollerslev, T., Diebold, F. X., and Labys, P. (2001). The distribution of realized exchange rate volatility. *Journal of the American Statistical Association*, **96**, 42–55.

Andersen, T. G., Bollerslev, T., Diebold, F.X., and Labys, P. (2003). Modeling and forecasting realized volatility. *Econometrica*, **71**, 579–625.

Asai, M., McAleer, M., and Yu, J. (2006). Multivariate stochastic volatility: a review. *Econometric Reviews*, **25**, 145-175.

Baillie, R. T., Calonaci, F., Cho, D., and Rho, S. (2019). Long memory, realized volatility and heterogeneous autoregressive models. *Journal of Time Series Analysis*, **40**, 609-628.

Barndorff-Nielsen, O. E. and Shephard, N. (2002). Estimating quadratic variation using realized variance. *Journal of Applied Econometrics*, **17**, 457–477.

Barndorff-Nielsen, O. E. and Sheppard, N. (2004). Power and bipower variation with stochastic volatility and jumps. *Journal of Financial Economics*, **2**, 1–37.

Bauwens, L., Laurent, S., and Rombouts, J. V. (2006). Multivariate GARCH models: a survey. *Journal of Applied Econometrics*, **21**, 79-109.

Bollerslev, T. (1986). Generalized autoregressive conditional heteroskedasticity. *Journal of Econometrics*, **31**, 307-327.

Bollerslev, T., Patton, A. J., and Quaedvlieg, R. (2016). Exploiting the errors: A simple approach for improved volatility forecasting. *Journal of Econometrics*, **192**, 1-18.

Chang, C. L., McAleer, M., and Tansuchat, R. (2010). Analyzing and forecasting volatility spill overs, asymmetries, and hedging in major oil markets. *Energy Economics*, **32**, 1445-1455.

Corsi, F. (2009). A simple approximate long-memory model of realized volatility. *Journal of Financial Economics*, **7**, 174–196.

Corsi, F. and Reno, R. (2012). Discrete-time volatility forecasting with persistent leverage effect and the link with continuous-time volatility modelling. *Journal of Business and Economic Statistics*, **30**, 368–380.

Degiannakis, S. and Filis, G. (2017). Forecasting oil price realized volatility using information channels from other asset classes. *Journal of International Money and Finance*, **76**, 28–49.

Degiannakis, S., Filis, G., and Hassani, H. (2018). Forecasting global stock market implied volatility indices. *Journal of Empirical Finance*, **46**, 111–129.

Degiannakis, S., Filis, G., Klein, T., and Walther, T. (2022). Forecasting realized volatility of agricultural commodities. *International Journal of Forecasting*, **38**, 74-96.

Engle, R. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation. *Econometrica*, **50**, 391-407.

Engle, R., Hong, C. H., and Kane, A. (1990). Valuation of variance forecast with simulated option markets. Cambridge, National Bureau of Economic Research.

Figlewski, S. (2004). *Forecasting Volatility*. Manuscript, New York University Stern School of Business.

Giot, P. and Laurent, S. (2007). The information content of implied volatility in light of the jump/continuous decomposition of realized volatility. *Journal of Futures Markets*, **27**, 337–359.

Gong, X. and Lin, B. (2018). Structural breaks and volatility forecasting in the copper futures market. *Journal of Futures Markets*, **38**, 290–339.

Greene, M. and Fielitz, B. (1977). Long-term dependence in common stock returns. *Journal of Financial Economics*, **4**, 339–349.

Hansen, P. R., Lunde, A., and Nason, J. M. (2011). The model confidence set. *Econometrica*, **79**, 453–497.

Haugom, E. H. Langeland, Molnar, P., and Westgaard, S. (2014). Forecasting Volatility of the US Oil Market. *Journal of Banking and Finance*, **47**, 1–14.

Hurst, H. E. (1951). Long-term storage capacity of reservoirs. *Transactions of the American Society of Civil Engineers*, **116**, 770–799.

Hurst, H. E. (1957). A suggested statistical model of some time series that occur in nature. *Nature*, **180**, 494-494.

Liu, L. Y., Patton, A. J., and Sheppard, K. (2015). Does anything beat 5- minute RV? A comparison of realized measures across multiple asset classes. *Journal of Econometrics*, **187**, 293–311.

Mandelbrot, B. B. and Wallis, J. (1968). Noah, Joseph, and operational hydrology. *Water Resources Research*, **4**, 909–918.

Molnar, P. (2016). High-low range in GARCH models of stock return volatility. *Applied Economics*, **48**, 4977–4991.

Muller, U. A., Dacorogna, M. M., Dave, R. D., Olsen, R. B., Pictet, O. V., and Von Weizsacker, J. E. (1997). Volatilities of different time resolutions analyzing the dynamics of market components. *Journal of Empirical Finance*, **4**, 213-239.

Patton, A. J. (2011). Volatility forecast comparison using imperfect volatility proxies. *Journal of Econometrics*, **160**, 246–256.

Patton, A. J. and Sheppard, K. (2015). Good volatility, bad volatility, Signed jumps, and the persistence of volatility. *The Review of Economics and Statistics*, **97**, 683–697.

Taylor, S. J. (1986). *Modeling Financial Time Series*, John Wiley & sons. Great Britain.

Taylor, S. J. (1994). Modeling stochastic volatility: A review and comparative study. *Mathematical Finance*, **4**, 183-204.

Tian, F., Yang, K., and Chen, L. (2017). Realized volatility forecasting of agricultural commodity futures using the HAR model with time-varying sparsity. *International Journal of Forecasting*, **33**, 132–152.

Xiong, Q. (2021). Forecast on S&P 500 Index Based on HAR-RV Model. *International Conference on Economic Management and Cultural Industry, Atlantis Press*, 1333-1338.

Yang, K., Tian, F., Chen, L., and Li, S. (2017). Realized volatility forecast of agricultural futures using the HAR models with bagging and combination approaches. *International Review of Economics and Finance*, **49**, 276–291.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 21, No. 2, 2023 (New Series), pp 141-154
http://www.ssca.org.in/journal

Resolvable and 2-Replicate PBIB Designs based on Higher Association Schemes Using Polyhedra

Vinayaka[1] **and Rajender Parsad**[2]
[1] *The Graduate School, ICAR-Indian Agricultural Research Institute, New Delhi-110012*
[2] *ICAR-Indian Agricultural Statistics Research Institute, New Delhi-110012*

Received: 15 May 2022; Revised: 29 November 2022; Accepted: 01 December 2022

Abstract

In this article, three new association schemes and construction of partially balanced incomplete block (PBIB) designs based on these association schemes in three and four associate classes using polyhedra have been proposed. Construction methods use polyhedra such as icosahedron, octahedron and pentagonal prism. PBIB designs based on icosahedral and octahedral association schemes are resolvable block designs whereas designs based on pentagonal prism association scheme are 2-replicate PBIB designs. A simple analysis of these designs is outlined including generalized forms of canonical efficiency factors (CEFs) and average variances ($\bar{V}$). A catalogue of PBIB designs for k (size of each block) ≤ 20 is given along with computed efficiencies.

Key words: Icosahedral association scheme; Octahedral association scheme; Pentagonal prism association scheme; Resolvable partially balanced incomplete block design.

1. Introduction

PBIB designs based on 2-associate classes have been extensively studied in the literature and for a comprehensive catalogue of these designs; one may refer to Clatworthy (1973); Dey (1977); Sinha (1991); Ghosh and Divecha (1995); and Saurabh and Sinha (2022). A lot of literature is available on PBIB designs based on 3- or higher class association schemes. PBIB designs based on rectangular (3-class) association scheme (known as rectangular designs) are an important class of block designs with factorial structure for experiments with two factors [see *e.g.*, Vartak (1955), Sharma and Das (1985), Suen (1989), Srivastava *et al.* (2000), Parsad *et al.* (2007a, 2007b) and references cited therein]. The nested group divisible designs, a class of PBIB(3) designs, useful for 3-factor experiments was introduced by Roy (1953) were subsequently studied by Raghavarao (1960); Bhagwandas *et al.* (1992); Duan and Kageyama (1993); Miao *et al.* (1996); and Mitra *et al.* (2002). More generalized association scheme called extended group divisible association scheme and designs based on this scheme are known as extended group divisible (EGD) designs was introduced by Hinkelmann (1964). Many useful applications of these designs and their catalogue are given in Parsad *et*

Corresponding Author: Vinayaka
Email: vinayaka.b3vs@gmail.com

al. (2007a, 2007b). Rao (1956) developed circular lattices which were essentially PBIB(3) designs for $v = 2n^2$ treatments, where $n \geq 2$ and these were further generalized by Varghese and Sharma (2004) to accommodate $2sn^2$ treatments; $n, s \geq 2$. Also, Varghese *et al.* (2004) gave some PBIB(3) designs and their applications to partial diallel crosses. Sharma *et al.* (2010) introduced 3-associate-class tetrahedral and cubical association schemes and methods of constructions of PBIB(3) designs based on these schemes using polyhedra such as tetrahedron and cube (hexahedron). On the similar lines, Vinayaka and Vinaykumar (2021) extended the work on graph based 2- and 3-associate class schemes of Garg and Farooq (2014) to 3- and 4-class graphical association schemes and constructions of related PBIB designs.

Some work on investigations of 4-associate class PBIB designs was carried out by several authors such as Nair (1951), Tharthare (1963, 1965), Garg *et al.* (2011), and others. Further, investigations on 2-replicate PBIB designs are limited to only Varghese and Sharma (2004); Sharma *et al.* (2010); and Kipkemoi *et al.* (2013, 2015).

For some parameters neither a BIB design nor a PBIB design with 2-associate classes is available. The best alternative design for such situations is higher associate PBIB design, if such design exists. Hence, in this investigation, we extend the work on 3- and 4-associate class PBIB designs further by proposing three new association schemes called icosahedral association scheme with 4-associate classes; octahedral association scheme with 3-associate classes and pentagonal prism association scheme with 3-associate classes and methods of constructing related PBIB designs based on these schemes. First two schemes produces 3- and 4-class PBIB designs belongs to the resolvable block designs which are also used in information theory *i.e.*, constructing A^2-codes and low density parity-check (LDPC) codes [see *e.g.*, Pei (2006); and Xu *et al.* (2015, 2020)] and in sequential experimentation over space and time [see *e.g.*, Patterson and Silvery (1980); John and Williams (1995); and Morgan and Reck (2007)]. The third scheme give rise to the two-replicate PBIB design which is beneficial in the situation of limited resources and also for developing mating plans in the area of plant breeding experiments like Narain (1993), Kaushik (1999), and others. We can also find applications of PBIB design in cryptology; see for example, Adhikari *et al.* (2007).

However, several authors such as Harshbarger (1949); Bose and Nair (1962); David (1967); Patterson and Williams (1976); Williams *et al.* (1976, 1977); Jarrett and Hall (1978); Varghese and Sharma (2004); Sharma *et al.* (2010); *etc.*, fostered detailed information on problems of construction and analysis of resolvable incomplete block designs.

Flowchart of the article as follows: In Section 2, three new association schemes *viz.*, icosahedral association scheme, octahedral association scheme and pentagonal prism association scheme are defined along with numerical illustrations. Section 3 deals with the constructions of PBIB designs using icosahedron, octahedron, and pentagonal prism along with examples. An outline of analysis of these designs is established in Section 4. Section 5 reveals a brief discussion. A catalogue of efficient PBIB designs has been obtained for $k \leq 20$ and is presented in the Appendix.

2. Definition of association schemes and numerical illustrations

It is well known that any polyhedron is a three-dimensional shape with V number of vertices, E number of edges and F number of faces. Polyhedra satisfy the Euler characteristic

χ which relates the V, F and E as $\chi = \mathrm{V} + \mathrm{F} - \mathrm{E}$, for details, one may refer to Richeson (2019). Further, convex polyhedra where every face is the same kind of regular polygon with n number of edges may be found among three families *viz.*, formerly triangles: these polyhedra are called deltahedra. There are only eight strictly-convex deltahedra out of which three are regular polyhedra (such as tetrahedron, octahedron and icosahedron are indeed platonic solids), and five are Johnson solids. Secondly, squares: the hexahedron is the only convex example and thirdly, pentagons: the regular dodecahedron is the only convex example. These are useful for constructions of PBIB designs; a reference can be made to Sharma *et al.* (2010).

Now we define three association schemes using icosahedron, octahedron and pentagonal prism in the sequel.

2.1. Icosahedral association scheme

Let the number of symbols (treatments) be $v = 12m$ ($m \geq 2$). Arrange these symbols on the twelve vertices of an icosahedron such that each vertex contains exactly m distinct symbols and intersected by five distinct edges. We define the four associates of a particular treatment ϕ as follows:

(i) Treatments except ϕ appearing in the same vertex with ϕ are the first associates;

(ii) Treatments appearing in different vertices that directly meet a vertex of ϕ through single edge are the second associates;

(iii) Treatments appearing in the end vertex that is exactly opposite to the vertex of ϕ are the third associates;

(iv) The remaining treatments are the fourth associates.

The parameters of first kind and second kind (association matrices) of the association scheme are delineated in continuation. *i.e.*, $v(= 12m)$, $n_1 = m-1$, $n_2 = 5m$, $n_3 = m$, $n_4 = 5m$, and

$$\boldsymbol{P_1} = \begin{bmatrix} m-2 & 0 & 0 & 0 \\ 0 & 5m & 0 & 0 \\ 0 & 0 & m & 0 \\ 0 & 0 & 0 & 5m \end{bmatrix}, \boldsymbol{P_2} = \begin{bmatrix} 0 & m-1 & 0 & 0 \\ m-1 & 2m & 0 & 2m \\ 0 & 0 & 0 & m \\ 0 & 2m & m & 2m \end{bmatrix}$$

$$\boldsymbol{P_3} = \begin{bmatrix} 0 & 0 & m-1 & 0 \\ 0 & 0 & 0 & 5m \\ m-1 & 0 & 0 & 0 \\ 0 & 5m & 0 & 0 \end{bmatrix}, \boldsymbol{P_4} = \begin{bmatrix} 0 & 0 & 0 & m-1 \\ 0 & 2m & m & 2m \\ 0 & m & 0 & 0 \\ m-1 & 2m & 0 & 2m \end{bmatrix}.$$

Here, n_i is the number of i^{th} (i =1, 2, 3, 4) associates of a given treatment. Given any two treatments that are mutually i^{th} associates, the number of treatments common to the j^{th} associates of the first and k^{th} associates of the second is p^i_{jk} (i, j, k =1, 2, 3, 4) reflected in P_i matrices.

Moreover, this association scheme may also be defined alternatively as follows: Arrange $v = 12m$ ($m \geq 2$) treatments in 12 columns and m rows then the treatment δ, say, is the

first associate of specific treatment ϕ, say, if δ belongs to the same column of ϕ; the second associate, if δ occur either in second or third or fourth or fifth or sixth column; the third associate, if δ occur in seventh column; and the fourth associate, otherwise.

Illustration 1: Let $v = 24(= 12 \times 2)$ treatments arranged on the twelve vertices of an icosahedron such that each vertex comprises exactly two distinct treatments are shown in Figure 1 or arrange these $v = 24(= 12 \times 2)$ treatments in 12 columns and 2 rows as given below.

1	3	5	7	9	11	13	15	17	19	21	23
2	4	6	8	10	12	14	16	18	20	22	24

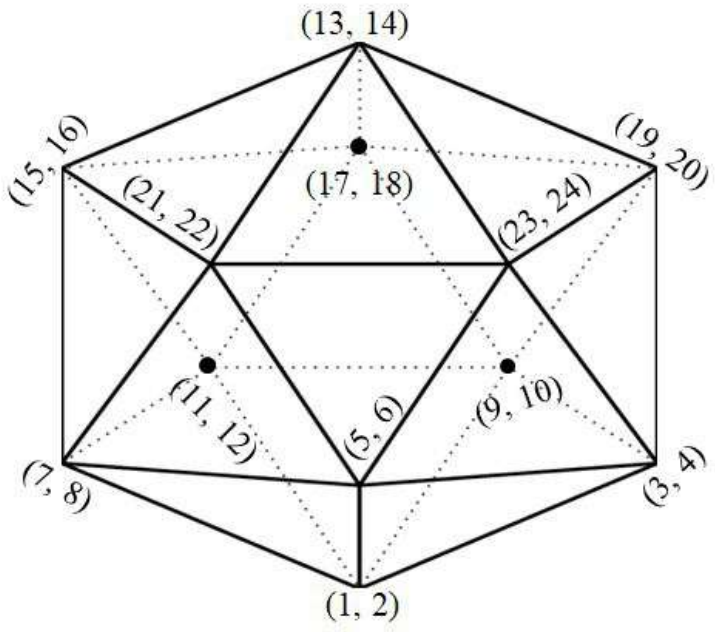


Figure 1: Arrangement of 24 treatments on the vertices of an icosahedron

The parameters of this association scheme are $v = 24$, $n_1 = 1$, $n_2 = 10$, $n_3 = 2$, $n_4 = 10$, and association matrices as:

$$\boldsymbol{P_1} = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 10 & 0 & 0 \\ 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 10 \end{bmatrix}, \boldsymbol{P_2} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 1 & 4 & 0 & 4 \\ 0 & 0 & 0 & 2 \\ 0 & 4 & 2 & 4 \end{bmatrix}, \boldsymbol{P_3} = \begin{bmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 10 \\ 1 & 0 & 0 & 0 \\ 0 & 10 & 0 & 0 \end{bmatrix}, \boldsymbol{P_4} = \begin{bmatrix} 0 & 0 & 0 & 1 \\ 0 & 4 & 2 & 4 \\ 0 & 2 & 0 & 0 \\ 1 & 4 & 0 & 4 \end{bmatrix}.$$

2.2. Octahedral association scheme

Let the number of treatments be $v = 6m$ $(m \geq 2)$. Arrange these $v = 6m$ treatments on the six vertices of an octahedron such that each vertex filled with m number of distinct treatments. Now we define the three associates of a specific treatment θ as follows:

(i) Treatments other than θ present in the same vertex of θ are the first associates;

(ii) Treatments present in different vertices that intersect the vertex of θ through their respective edges are the second associates;

(iii) The remaining treatments are the third associates.

The parameters of first kind of this association scheme are $v(=6m)$, $n_1 = m-1$, $n_2 = 4m$, $n_3 = m$. Further, association matrices (parameters of second kind) are as follows:

$$\boldsymbol{P_1} = \begin{bmatrix} m-2 & 0 & 0 \\ 0 & 4m & 0 \\ 0 & 0 & m \end{bmatrix}, \boldsymbol{P_2} = \begin{bmatrix} 0 & m-1 & 0 \\ m-1 & 2m & m \\ 0 & m & 0 \end{bmatrix}, \boldsymbol{P_3} = \begin{bmatrix} 0 & 0 & m-1 \\ 0 & 4m & 0 \\ m-1 & 0 & 0 \end{bmatrix}.$$

The alternative definition for the above association scheme is as follows: Arrange $v = 6m$ $(m \geq 2)$ treatments in six columns and m rows then the treatment δ, say, is the first associate of specific treatment θ, say, if δ belongs to the same column of θ; the second associate, if δ appears in any column except fourth column; and the third associate, otherwise.

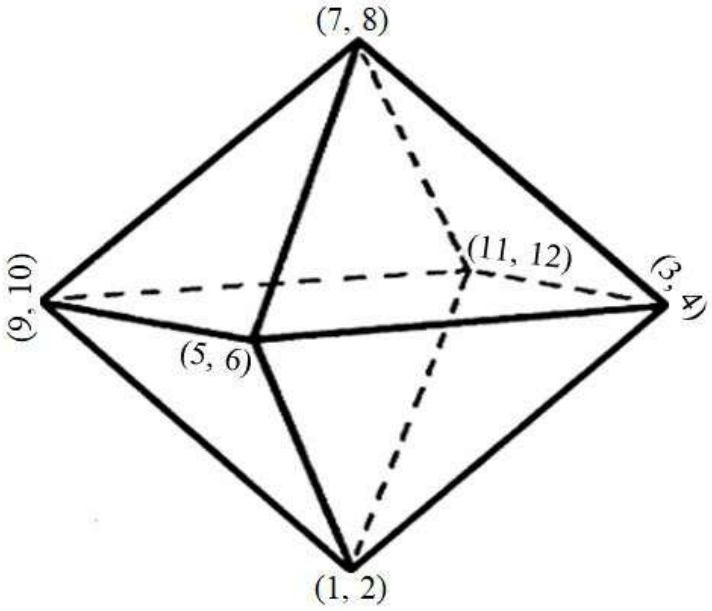


Figure 2: Arrangement of 12 treatments on the vertices of an octahedron

Illustration 2: Let $v = 12(= 6 \times 2)$ treatments arranged on the six vertices of an octahedron such that each vertex contains 2 distinct treatments are shown in Figure 2 or arrange these $v = 12(= 6 \times 2)$ treatments in six columns and two rows as given below.

1	3	5	7	9	11
2	4	6	8	10	12

The parameters of this association scheme are $v = 12$, $n_1 = 1$, $n_2 = 8$, $n_3 = 2$, and

$$\boldsymbol{P_1} = \begin{bmatrix} 0 & 0 & 0 \\ 0 & 8 & 0 \\ 0 & 0 & 2 \end{bmatrix}, \boldsymbol{P_2} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 4 & 2 \\ 0 & 2 & 0 \end{bmatrix}, \boldsymbol{P_3} = \begin{bmatrix} 0 & 0 & 1 \\ 0 & 8 & 0 \\ 1 & 0 & 0 \end{bmatrix}.$$

2.3. Pentagonal prism association scheme

A pentagonal prism is also polyhedron and a type of three-dimensional solid objects which comprises the two identical five sided pentagonal bases (ends) and remaining five faces are rectangles or parallelograms. Interestingly, two identical five sided pentagons contact each other with five edges respectively. Let $v = 10m$ $(m \geq 1)$ be the number of treatments. Arrange these treatments on the ten vertices of a pentagonal prism such that each vertex contains exactly m distinct treatments. Now we define the three associates of a specific treatment ψ as follows:

(i) Treatments other than ψ present in the two vertices of the same edge $E_y \; \forall \; y = 1, 2, 3, 4, 5$ are the first associates;

(ii) Treatments present in the different vertices of any two rectangles that contain common edge E_y except treatments lie on both terminals of E_y are the second associates;

(iii) The remaining treatments are the third associates.

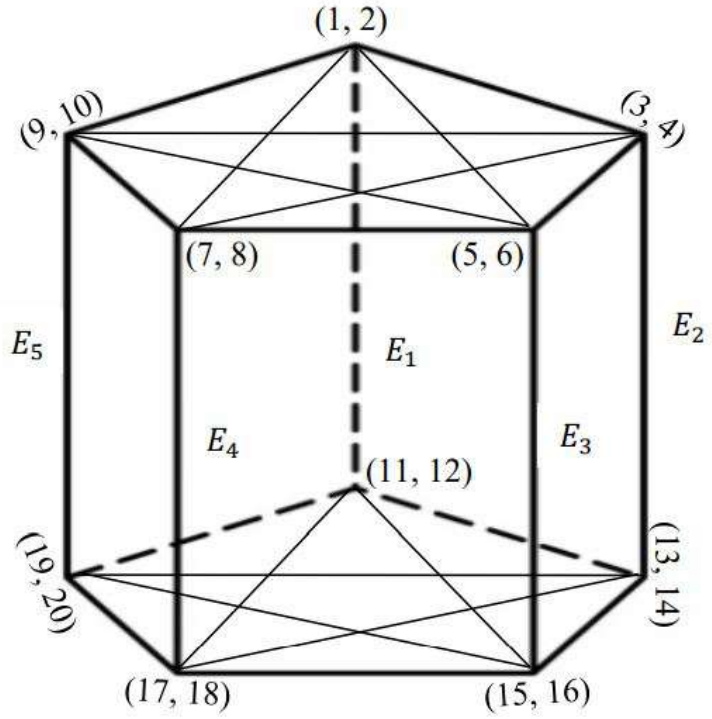


Figure 3: Arrangement of 20 treatments on the vertices of pentagonal prism

Here, the edges within identical five sided pentagons (both upper and lower) are of not interested, hence these are not named in Figure 3. The parameters of the association scheme are: $v(= 10m)$, $n_1 = 2m - 1$, $n_2 = 4m$, $n_3 = 4m$, and

$$\boldsymbol{P_1} = \begin{bmatrix} 2(m-1) & 0 & 0 \\ 0 & 4m & 0 \\ 0 & 0 & 4m \end{bmatrix}, \boldsymbol{P_2} = \begin{bmatrix} 0 & 2m-1 & 0 \\ 2m-1 & 0 & 2m \\ 0 & 2m & 2m \end{bmatrix}, \boldsymbol{P_3} = \begin{bmatrix} 0 & 0 & 2m-1 \\ 0 & 2m & 2m \\ 2m-1 & 2m & 0 \end{bmatrix}$$

Alternatively, the above association scheme may be defined as follows: arrange $v = 10m$ ($m \geq 1$) treatments in 10 columns and m rows then the treatment δ, say, is the first associate of particular treatment ψ, say, if δ belongs to either same column of ψ or sixth column; the second associate, if δ appears either in second or fifth or seventh column; and the third associate, otherwise.

Illustration 3: Let $v = 20(= 10 \times 2)$ treatments arranged on the ten vertices of a pentagonal prism such that each vertex contains $m = 2$ distinct treatments as shown in Figure 3 or arrange these treatments in 10 columns and 2 rows as given below.

1	3	5	7	9	11	13	15	17	19
2	4	6	8	10	12	14	16	18	20

The parameters of this association scheme are: $v = 20$, $n_1 = 3$, $n_2 = 8$, $n_3 = 8$, and

$$\boldsymbol{P_1} = \begin{bmatrix} 2 & 0 & 0 \\ 0 & 8 & 0 \\ 0 & 0 & 8 \end{bmatrix}, \boldsymbol{P_2} = \begin{bmatrix} 0 & 3 & 0 \\ 3 & 0 & 4 \\ 0 & 4 & 4 \end{bmatrix}, \boldsymbol{P_3} = \begin{bmatrix} 0 & 0 & 3 \\ 0 & 4 & 4 \\ 3 & 4 & 0 \end{bmatrix}.$$

3. Construction methods of PBIB designs

In this section, we give two construction methods of resolvable PBIB designs and a construction method of 2-replicate PBIB design based on aforesaid association schemes *i.e.*, icosahedral association scheme, octahedral association scheme, and pentagonal prism association scheme, respectively.

3.1. Method of constructing icosahedral PBIB(4) design

An arrangement $v = 12m$ ($m \geq 2$) treatments on the vertices of an icosahedron such that each vertex contains m number of distinct treatments is given in Figure 1. Evidently, each vertex is intersected by five edges. Let $v = 12m$ treatments are defined on the icosahedral association scheme. In order to form a block, combine the treatments of a chosen vertex and five distinct vertices which intersect this chosen vertex. Applying this process to all twelve vertices of an icosahedron yields a PBIB(4) design based on icosahedral association scheme with parameters $v = 12m$, $b = 12$, $r = 6$, $k = 6m$, $\lambda_1 = 6$, $\lambda_2 = 4$, $\lambda_3 = 0$, $\lambda_4 = 2$.

Example 1: Let $v = 24(= 12 \times 2)$ treatments are defined on the icosahedral association scheme. One can get an idea about arrangement of treatments on vertices of icosahedron with the help of Figure 1. Now, by following the procedure of Method 3.1, one gets a PBIB(4) design based on the icosahedral association scheme with parameters are as $v = 24$, $b = 12$, $r = 6$, $k = 12$, $\lambda_1 = 6$, $\lambda_2 = 4$, $\lambda_3 = 0$, $\lambda_4 = 2$. This design is a resolvable class of incomplete block designs wherein twelve blocks can be grouped into six sets of two blocks each, that is, {(B1, B2); (B3, B4); (B5, B6); (B7, B8); (B9, B10); (B11, B12)} such that every treatment appears in each set exactly once. The block structure of the design is given below.

Remark 1: For $m = 1$, this scheme also reduced to 3-associate class rectangular association scheme. The PBIB(3) design so obtained is symmetric rectangular design with parameters $v = 12 = b$, $r = 6 = k$, $\lambda_1 = 4$, $\lambda_2 = 0$, $\lambda_3 = 2$. This design seems to be new and not reported in the Varghese *et al.* (2004) and Parsad *et al.* (2007b).

Replication No.	Block No.	Block Contents
I	B1	(1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12)
	B2	(13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24)
II	B3	(1, 2, 3, 4, 5, 6, 9, 10, 19, 20, 23, 24)
	B4	(7, 8, 11, 12, 13, 14, 15, 16, 17, 18, 21, 22)
III	B5	(1, 2, 3, 4, 5, 6, 7, 8, 21, 22, 23, 24)
	B6	(9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20)
IV	B7	(1, 2, 5, 6, 7, 8, 11, 12, 15, 16, 21, 22)
	B8	(3, 4, 9, 10, 13, 14, 17, 18, 19, 20, 23, 24)
V	B9	(1, 2, 7, 8, 9, 10, 11, 12, 15, 16, 17, 18)
	B10	(3, 4, 5, 6, 13, 14, 19, 20, 21, 22, 23, 24)
VI	B11	(1, 2, 3, 4, 9, 10, 11, 12, 17, 18, 19, 20)
	B12	(5, 6, 7, 8, 13, 14, 15, 16, 21, 22, 23, 24)

3.2. Method of constructing octahedral PBIB(3) design

An octahedron has eight triangular faces and twelve edges, each face enclosed by the three vertices. Let $v = 6m$ ($m \geq 2$). Arrangement of these v treatments on the six vertices of an octahedron such that each vertex contains m number of distinct treatments as indicated in the association scheme. Now form the contents of a block by taking treatments that lie on three vertices of specific triangular face. Likewise, obtain the other seven blocks using remaining triangular faces of octahedron. The eight blocks thus obtained, each corresponding to one triangular face. This process results in a PBIB(3) design based on the octahedral association scheme with parameters $v = 6m$, $b = 8$, $r = 4$, $k = 3m$, $\lambda_1 = 4$, $\lambda_2 = 2$, $\lambda_3 = 0$.

Example 2: Let $v = 12(= 6 \times 2)$ treatments are defined on the octahedral association scheme. Figure 2 gives an idea about arrangement of treatments on vertices of octahedron. Now applying the procedure of Method 3.2, we can get a PBIB(3) design based on the octahedral association scheme with parameters as $v = 12$, $b = 8$, $r = 4$, $k = 6$, $\lambda_1 = 4$, $\lambda_2 = 2$, $\lambda_3 = 0$. This design is resolvable as its eight blocks can be grouped into four sets of two blocks each, that is, {(B1, B2); (B3, B4); (B5, B6); (B7, B8)} such that every treatment appears in each set exactly once. The block layout of the design is displayed below.

Replication No.	Block No.	Block Contents
I	B1	(1, 2, 3, 4, 5, 6)
	B2	(7, 8, 9, 10, 11, 12)
II	B3	(1, 2, 3, 4, 11, 12)
	B4	(5, 6, 7, 8, 9, 10)
III	B5	(1, 2, 5, 6, 9, 10)
	B6	(3, 4, 7, 8, 11, 12)
IV	B7	(1, 2, 9, 10, 11, 12)
	B8	(3, 4, 5, 6, 7, 8)

Remark 2: For $m = 1$, this scheme also reduced to two-class group divisible (GD) association scheme. The design so obtained is a semi-regular group divisible (SRGD) design with parameters as $v = 6$, $b = 8$, $r = 4$, $k = 3$, $\lambda_1 = 0$, $\lambda_2 = 2$, $n_1 = 1$, $n_2 = 4$ which is SR19 in the Clatworthy (1973).

3.3. Method of constructing pentagonal prism PBIB(3) design

Arrange $v = 10m$ ($m \geq 1$) treatments on the vertices of pentagonal prism such that each vertex contains m number of distinct treatments. Let $v = 10m$ treatments are defined on the pentagonal prism association scheme. Evidently, one can form five distinct rectangular shapes through diagonals using upper and lower pentagons given in Figure 3, so these are named as diagonal rectangles. Form five blocks of the design each one corresponding to a diagonal rectangular shape by combining the treatments situated on four vertices of that diagonal rectangle as the block contents. This process yields a PBIB(3) design based on pentagonal prism association scheme with parameters as $v = 10m$, $b = 5$, $r = 2$, $k = 4m$, $\lambda_1 = 2$, $\lambda_2 = 0$, $\lambda_3 = 1$.

Example 3: Let $v = 20(= 10 \times 2)$ treatments are defined on the pentagonal prism association scheme. For the arrangement of the treatments given in Figure 3, Now, by following the procedure of Method 3.3, one can get a PBIB(3) design based on pentagonal prism association scheme with block contents are given below:

Block No.	Block Contents
B1	$(1, 2, 5, 6, 11, 12, 15, 16)$
B2	$(1, 2, 7, 8, 11, 12, 17, 18)$
B3	$(3, 4, 7, 8, 13, 14, 17, 18)$
B4	$(3, 4, 9, 10, 13, 14, 19, 20)$
B5	$(5, 6, 9, 10, 15, 16, 19, 20)$

The design so obtained is a pentagonal prism design with parameters as $v = 20$, $b = 5$, $r = 2$, $k = 8$, $\lambda_1 = 2$, $\lambda_2 = 0$, $\lambda_3 = 1$.

4. Analysis

The above designs *viz.*, icosahedral, octahedral, and pentagonal prism designs can be analyzed as general PBIB designs. For completeness, simple steps for method of analysis are as follows: we know that the liner additive fixed effect model *i.e.*,

$$\mathbf{y} = \mu\mathbf{1} + \mathrm{Z}_1'\boldsymbol{\alpha} + \mathrm{Z}_2'\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

where, $\mathbf{y}$ = vector of n observations, μ = general mean, $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \ldots, \alpha_v)$ = vector of treatment effects, $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_b)$ = vector of block effects, $\mathbf{1}$ = vector of unities with order $(n \times 1)$, $\mathbf{Z_1'}$ = treatments vs observations incidence matrix with order $(v \times n)$, $\mathbf{Z}_2'$ = blocks vs observations incidence matrix with order $(b \times n)$ and $\boldsymbol{\varepsilon} \sim \mathrm{N}(\mathbf{0}, \sigma^2\mathbf{I}_n)$ = vector of errors with order $(n \times 1)$.

The general expressions of the information (C) matrices, Eigen values (η_l, $\forall\ l = 1$, 2, 3, 4) and corresponding multiplicities (ω_l, $\forall\ l = 1$, 2, 3, 4) of these information matrices for aforementioned designs (*i.e.*, icosahedral, octahedral, and pentagonal prism designs) are displayed in the Table 1. Here, $\boldsymbol{C}_{id}$, $\boldsymbol{C}_{od}$, and $\boldsymbol{C}_{pd}$ are the information matrices of icosahedral, octahedral, and pentagonal prism designs, respectively and also their corresponding incidence matrices denoted as $\mathbf{N_1}$, $\mathbf{N_2}$, and $\mathbf{N_3}$. Further, concurrence matrices and associates using these incidence matrices are also mentioned in the Table 2.

Table 1: Eigen values and corresponding multiplicities of C-matrices of designs

Particulars	C-matrix	Eigen values		Multiplicities	
Icosahedral design	$\boldsymbol{C}_{id} = 6\mathbf{I}_{12m} - (6m)^{-1}\mathbf{N_1}\mathbf{N_1'}$	η_1	6	ω_1	$12m-7$
		η_2	5.745	ω_2	3
		η_3	4.255	ω_3	3
		η_4	0	ω_4	1
Octahedral design	$\boldsymbol{C}_{od} = 4\mathbf{I}_{6m} - (3m)^{-1}\mathbf{N_2}\mathbf{N_2'}$	η_1	4	ω_1	$2(3m-2)$
		η_2	2.667	ω_2	3
		η_3	0	ω_3	1
Pentagonal prism design	$\boldsymbol{C}_{pd} = 2\mathbf{I}_{10m} - (4m)^{-1}\mathbf{N_3}\mathbf{N_3'}$	η_1	2	ω_1	$5(2m-1)$
		η_2	1.809	ω_2	2
		η_3	0.691	ω_3	2
		η_4	0	ω_4	1

It is well known that the canonical efficiency factors (CEFs) is $1/r$ times of harmonic mean of non-zero and positive Eigen values of the information matrix for a given block design. *i.e.*,

$$CEFs = \frac{1}{r}\left[\frac{(\omega_1+\omega_2+\ldots+\omega_l)}{\left(\frac{\omega_1}{\eta_1}+\frac{\omega_2}{\eta_2}+\ldots+\frac{\omega_l}{\eta_l}\right)}\right]$$

Table 2: Concurrence matrices and associates using incidence matrices of designs

Particulars	$\mathbf{N_1}\mathbf{N_1'} = ((n_{ii'}))$	$\mathbf{N_2}\mathbf{N_2'} = ((n_{ii'}))$	$\mathbf{N_3}\mathbf{N_3'} = ((n_{ii'}))$
if $i = i'(=1,2,\ldots,v)$	$= r(=6)$	$= r(=4)$	$= r(=2)$
if i and i' are the 1^{st} associates	$= \lambda_1(=6)$	$= \lambda_1(=4)$	$= \lambda_1(=2)$
if i and i' are the 2^{nd} associates	$= \lambda_2(=4)$	$= \lambda_2(=2)$	$= \lambda_2(=0)$
if i and i' are the 3^{rd} associates	$= \lambda_3(=0)$	$= \lambda_3(=0)$	$= \lambda_3(=1)$
if i and i' are the 4^{th} associates	$= \lambda_4(=2)$	–	–

Suppose for icosahedral design, there are four Eigen values (η_l) and their corresponding multiplicities (ω_l) as in Table 1, then its canonical efficiency factors are derived as follows:

$$CEFs = \frac{1}{6}\left[\frac{(12m-7+3+3)}{\left(\frac{12m-7}{6}+\frac{3}{5.745}+\frac{3}{4.255}\right)}\right] = \left[\frac{(12m-1)}{(12m+0.364)}\right] = \frac{11(12m-1)}{4(33m+1)}$$

Similarly, expressions of canonical efficiency factors (CEFs) and average variances ($\bar{V}$) of these designs are generalized in Table 3.

Table 3: Canonical efficiency factors (CEFs) and average variances ($\bar{V}$)

Particulars	CEFs	$\bar{V}$
Icosahedral design	$11(12m-1)/4(33m+1)$	$4(33m+1)/33(12m-1)$
Octahedral design	$2(6m-1)/(12m+1)$	$(12m+1)/4(6m-1)$
Pentagonal prism design	$(10m-1)/(10m+3)$	$(10m+3)/(10m-1)$

For more details and a comprehensive bibliography on canonical efficiency factors (CEFs), one may refer to Dey (2008). At last, a list of these designs using aforementioned three methods of construction is given along with computed efficiencies as Table 4, Table 5, and Table 6 respectively in the Appendix.

5. Discussion

The designs obtained from the icosahedral and octahedral association schemes fall into the resolvable class of incomplete block designs with minimal replications (*i.e.*, $r \leq 6$). The benefit of resolvable design is that its replications can be applied over different locations or over distinct time periods. Further, pentagonal prism association scheme provide 2-replicate PBIB designs which are beneficial when the experimenters facing the situation of constraint of resources. Additionally, efficiencies of these designs are quite high. Hence, these designs can be used to test a large number of cultivars in agricultural trials. The association schemes of these designs also find application in obtaining efficient partial diallel cross plans in plant/animal breeding experiments.

Acknowledgements

The authors thank the Editor-in-Chief and anonymous reviewers for their valuable suggestions and helpful comments which led to considerable improvement in the article.

References

Adhikari, A., Bose, M., Kumar, D., and Roy, B. (2007). Applications of partially balanced incomplete block designs in developing (2, n) visual cryptographic schemes. *IEICE Transactions on Fundamentals of Electronics, Communications and Computer Sciences*, **90**, 949-951.

Bhagwandas, Sinha, K., and Kageyama, S. (1992). Constructions of PBIB designs based on nested group divisible association scheme. *Utilitas Mathematica*, **41**, 169-174.

Bose, R. C. and Nair, K. R. (1962). Resolvable incomplete block designs with two replications. *Sankhya: The Indian Journal of Statistics, Series A*, 9-24.

Clatworthy, W. H. (1973). *Tables of Two-associate-class Partially Balanced Designs*. National Bureau of Standards, Applied Mathematics, Series No. 63, Washington D.C.

David, H. A. (1967). Resolvable cyclic designs. *Sankhya: The Indian Journal of Statistics, Series A*, 191-198.

Dey, A. (1977). Construction of regular group divisible designs. *Biometrika*, **64**, 647-649.

Dey, A. (2008). Canonical efficiency factors and related issues revisited. *Journal of Indian Society of Agricultural Statistics*, **62**, 169-173.

Duan, X. and Kageyama, S. (1993). Constructions of nested group divisible designs. *Statistics and Probability Letters*, **18**, 41-48.

Garg, D. K. and Farooq, S. A. (2014). Construction of PBIB designs through chosen lines and triangles of graphs. *International Journal of Mathematics Trends and Technology*, **8**, 25-32.

Garg, D. K., Jhaji, H. S., and Mishra, G. (2011). Construction of some new triangular and four associate class PBIB designs with two replicates. *International Journal of Mathematical Sciences and Applications*, **1**, 808-821.

Ghosh, D. K. and Divecha, J. (1995). Some new semi-regular group divisible designs. *Sankhya: The Indian Journal of Statistics, Series B*, **57**, 453-455.

Harshbarger, B. (1949). Triple rectangular lattices. *Biometrics*, **5**, 1-13.

Hinkelmann, K. (1964). Extended group divisible partially balanced incomplete block designs. *The Annals of Mathematical Statistics*, **35**, 681-695.

Jarrett, R. G. and Hall, W. B. (1978). Generalized cyclic incomplete block designs. *Biometrika*, **65**, 397-401.

John, J. A. and Williams, E. R. (1995). *Cyclic and Computer Generated Designs*. 2nd Edition. Chapman and Hall London.

Kaushik, L. S. (1999). Partial diallel crosses based on three associate class association schemes. *Journal of Applied Statistics*, **26**, 195-201.

Kipkemoi, E. C., Koske, J. K., and Mutiso, J. M. (2013). Construction of three-associate class partially balanced incomplete block designs in two replicates. *American Journal of Mathematical Science and Applications*, **1**, 61-65.

Kipkemoi, E. C., Mutiso, J. M., and Bekolle, D. (2015). Construction of some new three associate class partially balanced incomplete block designs in two replicates. *International Journal of Academic Research in Progressive Education and Development*, **1**, 188-192.

Miao, Y., Kageyama, S., and Duan, X. (1996). Further constructions of nested group divisible designs. *Journal of the Japan Statistical Society*, **26**, 231-239.

Mitra, R. K., Sinha, K., Kageyama, S., and Singh, M. K. (2002). Constructions of group divisible and nested group divisible designs. *Utilitas Mathematica*, **61**, 167-174.

Morgan, J. P. and Reck, B. H. (2007). Resolvable designs with large blocks. *The Annals of Statistics*, **35**, 747-771.

Nair, K. R. (1951). Rectangular lattices and partially balanced incomplete block designs. *Biometrics*, **7**, 145-154.

Narain, P. (1993). *Statistical Genetics*. Wiley Eastern Ltd., New Delhi.

Parsad, R., Gupta, V. K., and Srivastava, R. (2007a). Designs for cropping systems research. *Journal of Statistical Planning and Inference*, **137**, 1687-1703.

Parsad, R., Kageyama, S., and Gupta, V. K. (2007b). Use of complementary property of block designs in PBIB designs. *Ars Combinatoria*, **85**, 173-182.

Patterson, H. D. and Silvey, V. (1980). Statutory and recommended list trials of crop varities in the united kingdom. *Journal of the Royal Statistical Society: Series A*, **143**, 219-252.

Patterson, H. D. and Williams, E. (1976). A new class of resolvable incomplete block designs. *Biometrika*, **63**, 83-92.

Pei, D. (2006). A survey of construction for A2-codes. *In Authentication Codes and Combinatorial Designs. Discrete Mathematics and its Applications*, ed. K. H. Rosen, 215-30. Chapman and Hall/CRC, New York.

Raghavarao, D. (1960). A generalization of group divisible designs. *The Annals of Mathematical Statistics*, 756-771.

Rao, C. R. (1956). A general class of quasifactorial and related designs. *Sankhya: The Indian Journal of Statistics*, **17**, 165-174.

Richeson, D. S. (2019). *Euler's Gem: the Polyhedron Formula and the Birth of Topology*. Princeton University Press, Princeton, New Jersey.

Roy, P. M. (1953). Hierarchical group divisible incomplete block designs with m-associate classes. *Science and Culture*, **19**, 210-211.

Saurabh, S. and Sinha, K. (2022). A list of new partially balanced designs. *Communications in Statistics-Theory and Methods*, 1-4. doi: 10.1080/03610926.2022.2059685.

Sharma, V. K. and Das, M. N. (1985). On resolvable incomplete block designs 1. *Australian Journal of Statistics*, **27**, 298-302.

Sharma, V. K., Varghese, C., and Jaggi, S. (2010). Tetrahedral and cubical association schemes with related PBIB(3) designs. *Model Assisted Statistics and Applications*, **5**, 93-99.

Sinha, K. (1991). A list of new group divisible designs. *Journal of Research of the National Institute of Standards and Technology*, **96**, 613-615.

Srivastava, R., Parsad, R., and Gupta, V. K. (2000). Structure resistant factorial designs. *Sankhya: The Indian Journal of Statistics, Series B*, **62**, 257-265.

Suen, C. Y. (1989). Some rectangular designs constructed by the method of differences. *Journal of Statistical Planning and Inference*, **21**, 273-276.

Tharthare, S. K. (1963). Right angular designs. *The Annals of Mathematical Statistics*, 1057-1067.

Tharthare, S. K. (1965). Generalized right angular designs. *The Annals of Mathematical Statistics*, 1535-1553.

Varghese, C. and Sharma, V. K. (2004). A series of resolvable PBIB(3) designs with two replicates. *Metrika*, **60**, 251-254.

Varghese, C., Sharma, V. K., Jaggi, S., and Sharma, A. (2004). *Three-Associate Class Partially Balanced Incomplete Block Designs and their Application to Partial Diallel Crosses*. Project report, IASRI publication, New Delhi.

Vartak, M. N. (1955). On an application of Kronecker product of matrices to statistical designs. *The Annals of Mathematical Statistics*, **26**, 420-438.

Vinayaka and Vinaykumar, L. N. (2021). Higher associate class partially balanced incomplete block designs using graphs for agricultural experiments. *The Pharma Innovation Journal*, **10**, 967-972. doi: 10.22271/tpi.2021.v10.i12Sn.11252.

Williams, E. R., Patterson, H. D., and John, J. A. (1976). Resolvable designs with two replications. *Journal of the Royal Statistical Society. Series B (Methodological)*, 296-301.

Williams, E. R., Patterson, H. D., and John, J. A. (1977). Efficient two-replicate resolvable designs. *Biometrics*, 713-717.

Xu, H., Feng, D., Sun, C., and Bai, B. (2015). Construction of LDPC codes based on resolvable group divisible designs. *In 2015 International Workshop on High Mobility Wireless Communications (HMWC)*, 111-115.

Xu, H., Yu, Z., Feng, D., and Zhu, H. (2020). New construction of partial geometries based on group divisible designs and their associated LDPC codes. *Physical Communication*, **39**, 1-38. doi: https://doi.org/10.1016/j.phycom.2019.100970.

Appendix

Table 4: PBIB(4) designs based on Icosahedral association scheme with $k \leq 20$ using Method 3.1

SI. No.	m	v	b	r	k	λ_1	λ_2	λ_3	λ_4	E_1	E_2	E_3	E_4	E
1	2	24	12	6	12	6	4	2	0	1	0.9649	0.8979	0.9282	0.9440
2	3	36	12	6	18	6	4	2	0	1	0.9763	0.9295	0.9510	0.9625

Table 5: PBIB(3) designs based on Octahedral association scheme with $k \leq 20$ using Method 3.2

SI. No.	m	v	b	r	k	λ_1	λ_2	λ_3	E_1	E_2	E_3	E
1	2	12	8	4	6	4	2	0	1	0.8889	0.8000	0.8800
2	3	18	8	4	9	4	2	0	1	0.9230	0.8571	0.9189
3	4	24	8	4	12	4	2	0	1	0.9411	0.8889	0.9388
4	5	30	8	4	15	4	2	0	1	0.9524	0.9090	0.9508
5	6	36	8	4	18	4	2	0	1	0.9600	0.9231	0.9589

Table 6: PBIB(3) designs based on Pentagonal prism association scheme with $k \leq 20$ using Method 3.3

SI. No.	m	v	b	r	k	λ_1	λ_2	λ_3	E_1	E_2	E_3	E
1	1	10	5	2	4	2	0	1	1	0.5882	0.7692	0.6923
2	2	20	5	2	8	2	0	1	1	0.7407	0.8695	0.8261
3	3	30	5	2	12	2	0	1	1	0.8108	0.9090	0.8788
4	4	40	5	2	16	2	0	1	1	0.8511	0.9302	0.9070
5	5	50	5	2	20	2	0	1	1	0.8772	0.9433	0.9245

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 21, No. 2, 2023 (New Series), pp 155-173
http://www.ssca.org.in/journal

A Bimodal Extension of Suja Distribution with Applications

Samuel U. Enogwe[1], Emmanuel W. Okereke[1] and Gabriel C. Ibeh[2]
[1]*Department of Statistics*
Michael Okpara University of Agriculture, Umudike, Nigeria
[2]*Department of Mathematics/Statistics*
Federal Polytechnic Nekede, Owerri, Nigeria

Received: 02 May 2022; Revised: 30 September 2022; Accepted: 15 December 2022

Abstract

This paper introduces a new lifetime distribution called the Bimodal Extension of Suja (BES) distribution using the Quadratic Rank Transmutation Map. The proposed distribution has Suja distribution as a special case. Some statistical and reliability properties of the new distribution were derived and the method of maximum likelihood was employed for estimating the model parameters. The usefulness and flexibility of the BES distribution were illustrated with two real lifetime data sets. Results based on the log-likelihood and goodness of fit statistics values showed that the BES provides a better fir to the data than the other competing (lifetime) distributions considered in this study. Also, the consistency of the parameters of the new distribution was demonstrated through a simulation study. The BES distribution is therefore recommended foe effective modelling of the unimodal or bimodal continuous lifetime data with a non decreasing or bathtub shaped hazard rate function ...

Key words: Bimodal data; Hazard rate function; Maximum likelihood method; Quadratic rank transformation map; Suja distribution; BES distribution.

AMS Subject Classifications: 62B15, 60E05

1. Introduction

One of the activities of statisticians is to make informed decisions about a population on the basis of a sample drawn from that population. Obviously, several phenomena upon which decisions are taken often occur by chance and the best way to account for uncertainties surrounding them is to adopt probabilistic models. Probability models serve as mathematical structures for describing physical phenomena. A necessary step in the use of probabilistic models for modelling real-life problems is to ensure that the observed sample data follow certain probability distribution(s). Standard probability distributions commonly used for modelling several real-life problems include exponential, Weibull, gamma, two-parameter Odoma (Enogwe *et al.*, 2020) and so on. Unfortunately, so many datasets do not come from the existing probability distributions and this has engendered a demand for alternative

Corresponding Author: Samuel U. Enogwe
Email: enogwe.samuel@mouau.edu.ng

distributions, especially for the extension of the existing distributions which can be moore appropriate for fitting real-life data.

Recently, Shanker (2017) introduced and studied a new distribution, called the Suja distribution with probability density function (PDF) and cumulative distribution function (CDF) given, respectively, by

$$g(x;\eta) = \frac{\eta^5}{\eta^4 + 24}(1 + x^4)e^{-\eta x}; \quad x > 0, \eta > 0 \tag{1}$$

and

$$G(x;\eta) = 1 - \left[1 + \frac{\eta x(\eta^3 x^3 + 4n^2 x^2 + 12\eta x + 24)}{\eta^4 + 24}\right]e^{-\eta x}; \quad x > 0, \eta > 0 \tag{2}$$

An application of the Suja distribution to lifetime analysis of engineering data was presented by Shanker (2017) which showed that the Suja distribution outperforms the Akash (Shanker, 2015a), Shanker (Shanker, 2015b), Amarendra (Shanker, 2016a), Aradhana (Shanker, 2016b), Devya (Shanker, 2016c), Sujatha (Shanker, 2016d), Lindley (Ghitany, *et al.*, 2008) and exponential distributions in modelling lifetime data.

In spite of the utility of the Suja distribution, it cannot be used for statistical modelling of datasets with varieties of tails due its dependency on only one parameter. This limitation of Suja distribution can be overcome by obtaining some of its generalization so as to provide greater flexibility in modelling observed data. The work of Al-Omari and Alsmairan (2019) introduced a length-biased Suja distributionn. Also, a power lenght-biased Suja distribtion was developed by Al-Omari *et al.* (2019). Further, Alsmairan and Al-Omari (2020) used the weighted method to extend the Suja distribution, which was applied to ball bearing data to show that the weighted Suja distribution is better than the Suja distribution. It is evident that these extensions of Suja distribution cannot be used to model data with bimodal shape. To obtain an extension of Suja distribution that can model bimodal data, the quadratic rank transformatio map (QRTM) proposed by Shaw and Buckey (2007) is utilized.

According to Shaw and Buckley (2007), the QRTM provides distributions that are more flexible than baseline distributions in modelling real-life datasets with complex structure. The cumulative distribution function (CDF) and probability density function (PDF) of the quadratic transmuted family of distributions may be written as

$$F(x) = (1 + \lambda)G(x) - \lambda G^2(x) \tag{3}$$

and

$$f(x) = g(x)((1 + \lambda) - 2\lambda G(x)) \tag{4}$$

respectively, where $|\lambda| \leq 1, G(x)$ is the baseline CDF of X and $g(x) = dG(x)/dx$, the baseline PDF of X. Observe from (3) and (4) that if $\lambda = 0$, the quadratic transmuted family of distributions reduces to the baseline distribution.

Apart from the work of Shaw and Buckley (2007), other researchers have explored some members of the quadratic transmuted family of distributions. The members of the family of distributions include transmuted extreme value distribution (Aryal and Tsokos, 2009), transmuted

Weibull distribution (Aryal and Tsokos, 2011) transmuted log-logistic distribution (Aryal, 2013), transmuted Lindley distribution (Merovci, 2013a) and transmuted Rayleigh distribution by Merovci (2013b), transmuted Lomax distribution (Ashour and Eltehiwy, 2013), transmuted Pareto distribution (Merovci and Puka, 2014), transmuted two-parameter Lindley distribution due to Al-khazaleh *et al.* (2016), transmuted Dagum distribution (Shahzad and Asghar, 2016), transmuted Janardan distribution by Al-Omari *et al.* (2016), transmuted Burr XII distribution (Maurya *et al.*, 2017), transmuted Mukherjee-Islam (Rather and Subramanian, 2018), transmuted ArcSine distribution (Bleed and Abdelali, 2018), transmuted Ishita distribution (Gharaibeh and Al-Omari, 2019), transmuted Pranav distribution (Odom *et al.*, 2019), transmuted Garima distribution (Mohiuddin *et al.*, 2020), transmuted Aradhana (Gharaibeh, 2020), among others.

The aim of this article is to propose a new distribution, called a BES distribution, which is more flexible than the Suja distribution and some other competing lifetime distributions for modelling complex lifetime datasets. Specifically, this study reveals that the QRTM can be used to generalize a one-parameter continuous distribution to obtain a bimodal two-parameter distribution that has a monotone or non-monotone hazard rate function, especially the bathtub shape. As expected in the proposed distribution, the QRTM has been adopted in previous researches to generate new distributions that are more flexible than the baseline distributions. In Section 2, we define the expressions for the PDF and CDF of the BES distribution. The statistical and reliability properties of the BES distribution are discussed in Section 3. The quantile function and entropies of the BES distribution are given in Section 4. Section 5 provides the distribution of order statistics. In Section 6, the parameters of the BES distribution are estimated through the method of maximum likelihood estimation. Section 7 discusses the asymptotic confidence intervals of the parameters of the BES distribution. A simulation study is conducted in Section 8. In Section 9, two real datasets, methods of model selection, applications of the BES distribution to the data sets and the results are presented. In Section 10, we give the concluding remarks.

2. Definition of BES distribution

Inserting (2) into (3), we get the CDF of the new distribution. Also, inserting (1) and (2) into (4), we obtain the PDF of the new distribution. Consequently, a random variable X is said to have the BES distribution if its CDF and PDF are defined as

$$F_{BES}(x;\eta.\lambda) = (1+\lambda)\left[1-\left(1+\frac{\eta x(\eta^3x^3+4\eta^2x^2+12\eta x+24)}{\eta^4+24}\right)e^{\eta x}\right] - \lambda\left[1-\left(1+\frac{\eta x(\eta^3x^3+4\eta^2x^2+12\eta x+24)}{\eta^4+24}\right)e^{\eta x}\right]^2 \tag{5}$$

and

$$f_{BES}(x;\eta,\lambda) = \frac{\eta^5}{\eta^4+24}(1+x^4)e^{-\eta x}\left[1-\lambda+2\lambda\left(1+\frac{\eta x(\eta^3x^3+4\eta^2x^2+12\eta x+24)}{\eta^4+24}\right)e^{\eta x}\right] \tag{6}$$

respectively, for $x > 0$, $\eta > 0$ and $\mid \lambda \mid \leq 1$. The BES distribution reduces to the Suja distribution when $\lambda = 0$. Figure 1 shows the plots of the PDF of the BES variable based

on several sets of values of the parameters of the distribution. Figure 1 indicates that the PDF of the BES distribution has unimodal shape if $\lambda = 0.1, \eta = 0.6, \lambda = 0.3, \eta = 0.7$. The biomodal shape of the BES distribution is observed when $\lambda = -0.9, \eta = 2.0, \lambda = 0.4, \eta = 1.6$, among others. Again, the shape of the BES is nondecreasing if $\lambda = 0.9, \eta = 0.1$.

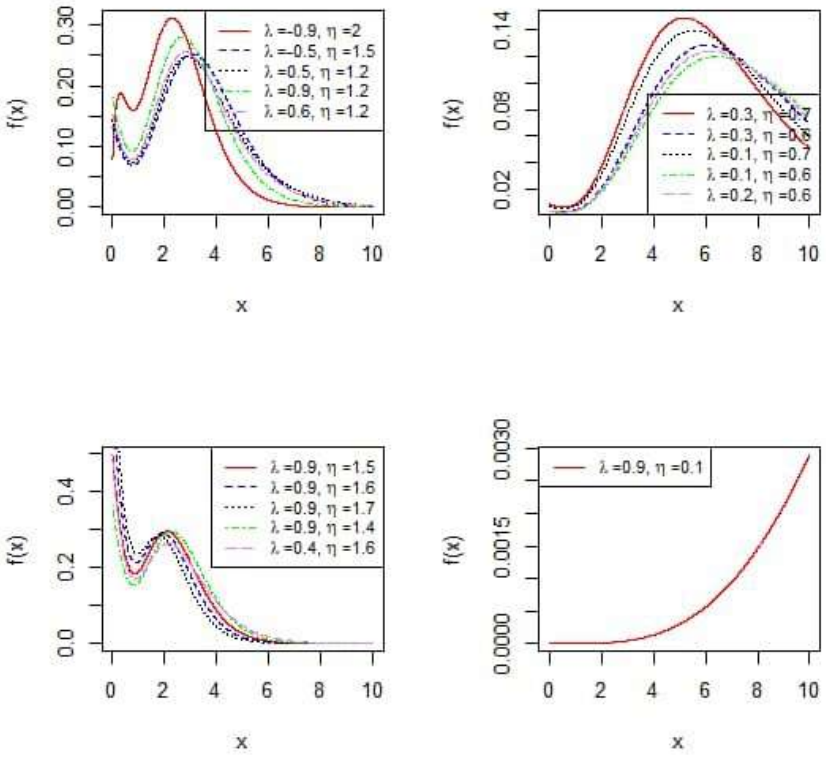


Figure 1: Various shapes of the PDF of BES

The graphs depicted as Figure 2 show that the Cumulative Distribution of BES is nondecreasing.

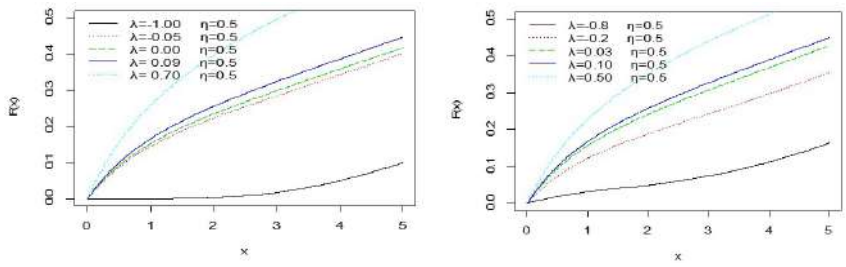


Figure 2: Various shapes of CDF of BES

3. Statistical and reliability properties of BES distribution

3.1. Statistical properties

The moment generating function of $X \sim BES(\eta, \lambda)$ is given by

$$
\begin{aligned}
M_X(t) &= \int_0^\infty e^{tx} f_{BES}(x;\eta,\lambda)dx \\
&= \frac{\eta^5}{\eta^4+24}\int_0^\infty e^{tx}(1+x^4)e^{-\eta x}\left[1-\lambda+2\lambda\left(1+\frac{\eta x(\eta^3x^3+4\eta^2x^2+12\eta x+24)}{\eta^4+24}\right)e^{-\eta x}\right]dx \\
&= \frac{\eta^5}{\eta^4+24}\int_0^\infty \left[2\lambda(1+x^4)\left(1+\frac{24}{\eta^4+24}\sum_{r=1}^{4}\frac{(\eta x)^r}{r!}\right)e^{-(2\eta-t)x}+(1-\lambda)(1+x^4)e^{-(\eta-t)x}\right]dx \\
&= \frac{2\lambda\eta^5}{\eta^4+24}\left[\int_0^\infty e^{-(2\eta-t)x}+\frac{24}{\eta^4+24}\sum_{r=1}^{4}\frac{\eta^r}{r!}\left(\int_0^\infty x^r e^{-(2\eta-t)x}+\int_0^\infty x^{r+4}e^{-(2\eta-t)x}\right)\right. \\
&\quad \left.+\int_0^\infty x^4 e^{-(2\eta-t)x}\right]dx+\frac{(1-\lambda)\eta^5}{\eta^4+24}\left[\int_0^\infty e^{-(2\eta-t)x}+\int_0^\infty x^4e^{-(2\eta-t)x}\right]dx \\
&= \frac{2\lambda\eta^5}{\eta^4+24}\left[\frac{1}{(2\eta-t)}+\frac{24}{\eta^4+24}\sum_{r=1}^{4}\frac{\eta^r}{r!}\left(\frac{\Gamma(r+1)}{(2\eta-t)^{r+1}}+\frac{\Gamma(r+5)}{(2\eta-t)^{r+5}}\right)+\frac{24}{(2\eta-t)^5}\right] \\
&\quad +\frac{(1-\lambda)\eta^5}{n^4+24}\left[\frac{1}{(\eta-t)}+\frac{24}{(\eta-t)^5}\right] \qquad (7)
\end{aligned}
$$

The r^{th} non-central moment of $X \sim BES(\eta, \lambda)$ is given by

$$\mu_r' = E(X^r) = \int_0^\infty x^r f_{BES}(x;\eta,\lambda)dx$$

$$= \frac{\eta^5}{(\eta^4+24)^2}\int_0^\infty x^r(1+x^4)e^{-\eta x}\left[1-\lambda+2\lambda\left(1+\frac{\eta x(\eta^3x^3+4\eta^2x^2+12\eta x+24)}{\eta^4+24}\right)e^{-\eta x}\right]dx$$

$$= \frac{\eta^5}{(\eta^4+24)^2}\int_0^\infty \begin{bmatrix}(1-\lambda)(\eta^4+24)x^re^{-\eta x}+(1-\lambda)(\eta^4+24)x^{r+4}e^{-\eta x} \\ +2\lambda(\eta^4+24)x^re^{-2\eta x}+2\lambda(\eta^4+24)x^{r+4}e^{-2\eta x}+8\lambda\eta^3x^{r+3}e^{-2\eta x} \\ +24\lambda\eta^2x^{r+2}e^{-2\eta x}+48\lambda\eta x^{r+1}e^{-2\eta x}+2\lambda\eta^4x^{r+8}e^{-2\eta x}+8\lambda\eta^3x^{r+7}e^{-2\eta x} \\ +24\lambda\eta^2x^{r+6}e^{-2\eta x}+48\lambda\eta x^{r+5}e^{-2\eta x}\end{bmatrix}dx$$

$$= \frac{\eta^5}{(\eta^4+24)^2}\begin{bmatrix}(1-\lambda)(\eta^4+24)\left(\frac{\Gamma(r+1)}{\eta^{r+1}}\right)+(1-\lambda)(\eta^4+24)\left(\frac{\Gamma(r+5)}{\eta^{r+5}}\right)+2\lambda(\eta^4+24)\left(\frac{\Gamma(r+1)}{(2\eta)^{r+1}}\right) \\ +2\lambda(\eta^4+24)\left(\frac{\Gamma(r+5)}{(2\eta)^{r+5}}\right)+8\lambda\eta^3\left(\frac{\Gamma(r+4)}{(2\eta)^{r+4}}\right)+24\lambda\eta^2\left(\frac{\Gamma(r+3)}{(2\eta)^{r+3}}\right)+48\lambda\eta\left(\frac{\Gamma(r+2)}{(2\eta)^{r+2}}\right) \\ 2\lambda\eta^4+\left(\frac{\Gamma(r+9)}{(2\eta)^{r+9}}\right)+8\lambda\eta^3\left(\frac{\Gamma(r+8)}{(2\eta)^{r+8}}\right)+24\lambda\eta^2\left(\frac{\Gamma(r+7)}{(2\eta)^{r+7}}\right)+48\lambda\eta\left(\frac{\Gamma(r+6)}{(2\eta)^{r+6}}\right)\end{bmatrix}$$

$$\therefore \mu'_r = \frac{\eta^5}{(\eta^4+24)^2}\begin{bmatrix}(1-\lambda)(\eta^4+24)\left(\frac{\Gamma(r+1)}{\eta^{r+1}}+\frac{\Gamma(r+5)}{\eta^{r+5}}\right)+\frac{\lambda(\eta^4+24)\Gamma(r+1)}{2^r\eta^{r+1}}\\ +\frac{\lambda(\eta^4+24)\Gamma(r+5)}{2^{r+4}\eta^{r+5}}+\frac{\lambda\Gamma(r+4)}{2^{r+1}\eta^{r+1}}+\frac{3\lambda\Gamma(r+3)}{2^r\eta^{r+1}}+\frac{12\lambda\Gamma(r+2)}{2^r\eta^{r+1}}\\ +\frac{\lambda\Gamma(r+9)}{2^{r+8}\eta^{r+4}}+\frac{\lambda\Gamma(r+8)}{2^{r+5}\eta^{r+5}}+\frac{3\lambda\Gamma(r+7)}{2^{r+4}\eta^{r+5}}+\frac{\lambda\Gamma(r+6)}{2^{r+2}\eta^{r+5}}\end{bmatrix} \tag{8}$$

Substituting $r = 1, 2, 3, 4$ in (8), yields the first four crude moments of the BES distribution as

$$\mu'_1 = \frac{(\theta^4+24)[(\theta^4+120)-\lambda(\theta^4+103)]+\lambda(4725\theta+3600)+108\theta^2(\theta^4+24)^2}{4\theta(\theta^4+24)^2} \tag{9}$$

$$\mu'_2 = \frac{(\theta^4+24)(8\theta^4-6\lambda\theta^4-2835\lambda+2880)-408\theta^4\lambda+263655\lambda}{4\theta^2(\theta^4+24)^2} \tag{10}$$

$$\mu'_3 = \frac{2(\theta^4+24)(54\theta^4+48\lambda\theta^4-40005\lambda+40320)-2106\theta^4+2835\theta\lambda+423360\lambda}{160^3(\theta^4+24)} \tag{11}$$

$$\mu'_4 = \frac{16(\theta^4+24)(48\theta^4-45\lambda\theta^4-80325\lambda+80640)+12240\theta^4\lambda+3742200\Theta\lambda+4399920\lambda}{32\Theta^4(\theta^4+24)^2} \tag{12}$$

The rth central moment of $X \sim BES(\eta, \lambda)$ can be obtained from the relation

$$\mu_r = \sum_{j=0}^{r}(-1)^j\binom{r}{j}\mu'_j(\mu)^{r-1} \tag{13}$$

where μ'_j is deduced from (8) by replacing r with j and μ is defined in (9). The following central moments are obtained by letting $r = 2, 3, 4$ in (13):

$$\mu_2 = \sum_{j=0}^{2}(-1)^j\binom{2}{j}\mu'_j(\mu)^{2-1} \tag{14}$$

$$\mu_3 = \sum_{j=0}^{3}(-1)^j\binom{3}{j}\mu'_j(\mu)^{3-1} \tag{15}$$

$$\mu_4 = \sum_{j=0}^{4}(-1)^j\binom{4}{j}\mu_j(\mu)^{4-1} \tag{16}$$

The coefficient of variation (γ_0), skewness (γ_1) and kurtosis (γ_2) of the BES distribution could be obtained by evaluating

$$\gamma_0 = \frac{(\mu_2)^{\frac{1}{2}}}{\mu} \tag{17}$$

$$\gamma_1 = \frac{\mu_3}{(\mu_2)^{\frac{3}{2}}} \tag{18}$$

$$\gamma_2 = \frac{\mu_4}{(\mu_2)^2} \tag{19}$$

3.2. Reliability properties

Suppose $X \sim BES(x;\eta,\lambda)$, then the reliability function may be written as

$$\begin{aligned} R_{BES}(x;\eta,\lambda) &= 1 - F_{BES}(x;\eta,\lambda) \\ &= \frac{(\lambda-1)}{\eta^4+24}(\eta x(\eta^3x^3+4\eta^2x^2+12\eta x+24)+\eta^4+24)e^{-\eta x} \\ &+ \frac{\lambda}{(\eta^4+24)^2}(\eta x(\eta^3x^3+4\eta^2x^2+12\eta x+24)+\eta^4+24)^2e^{-2\eta x} \end{aligned} \tag{20}$$

Taking the ratio of (6) to (20), we obtain the hazard rate function for $X \sim BES(\eta,\lambda)$ as

$$h_{BES}(x;\eta,\lambda) = \frac{f_{BES}(x;\eta,\lambda)}{R_{BES}(x;\eta,\lambda)}$$

$$h_{BES}(x;\eta,\lambda) = \frac{\eta^5\begin{bmatrix}(1-\lambda)(\eta^4+24)e^{-\eta x}+(1-\lambda)(\eta^4+24)x^4e^{-\eta x} \\ +2\lambda(\eta^4+24)e^{-2\eta x}+2\lambda(\eta^4+24)x^4e^{-2\eta x}+8\lambda\eta^3x^3e^{-2\eta x} \\ +24\lambda\eta^2x^2e^{-2\eta x}+48\lambda\eta xe^{-2\eta x}+2\lambda\eta^4x^8e^{-2\eta x}+8\lambda\eta^3x^7e^{-2\eta x} \\ +24\lambda\eta^2x^6e^{-2\eta x}+48\lambda\eta x^5e^{-2\eta x}\end{bmatrix}}{\begin{bmatrix}(\lambda-1)(\eta^4+24)(\eta^4x^4+4\eta^3x^3+12\eta^2x^2+24\eta x+\eta^4+24)e^{-\eta x} \\ +\lambda(\eta^4x^4+4\eta^3x^3+12\eta^2x^2+24\eta x+\eta^4+24)^2e^{-2\eta x}\end{bmatrix}} \tag{21}$$

The graphical representation of the hazard rate function of the BES distribution is presented as Figure 3. In accordance with Figure 3, the distribution is quite flexible as its hazard rate function is capable of possessing different shapes depending on the values of the associated parameters. Specifically, the figure reveals that the hazard rate function can be nondecreasing or bathtub shaped. It can also have an s-shaped curve or be a bimodal function.

The cumulative hazard function of $X \sim BES(\eta,\lambda)$ can be written as

$$\begin{aligned} Ch_{BES}(x;\eta,\lambda) &= -\ln(1-F_{BES}(x;\eta,\lambda)) = -\ln(R_{BES}(x;\eta,\lambda)) \\ &= -\ln[(\lambda-1)(\eta^4+24)+\lambda(\eta^4x^4+4\eta^3x^3+12\eta^2x^2+24\eta x+\eta^4+24)e^{-\eta x}] \\ &+ 2\ln(\eta^4+24)-\ln(\eta^4x^4+4\eta^3x^3+12\eta^2x^2+24\eta x+\eta^4+24)+\eta x \end{aligned} \tag{22}$$

The reverse hazard function of $X \sim BES(\eta,\lambda)$ is given by

$$Rh_{BES}(x;\eta,\lambda) = \frac{f_{BES}(x;\eta,\lambda)}{F_{BES}(x;\eta,\lambda)}$$

$$h_{BES}(x;\eta,\lambda) = \frac{\eta^5\begin{bmatrix}(1-\lambda)(\eta^4+24)e^{-\eta x}+(1-\lambda)(\eta^4+24)x^4e^{-\eta x} \\ +2\lambda(\eta^4+24)x^4e^{-2\eta x}+2\lambda(\eta^4+24)x^4e^{-2\eta x}+8\lambda\eta^3x^3e^{-2\eta x} \\ +24\lambda\eta^2x^2e^{-2\eta x}+48\lambda\eta xe^{-2\eta x}+2\lambda\eta^4x^8e^{-2\eta x}+8\lambda\eta^3x^7e^{-2\eta x} \\ +24\lambda\eta^2x^6e^{-2\eta x}+48\lambda\eta x^5e^{-2\eta x}\end{bmatrix}}{\begin{bmatrix}(\eta^4+24)-\lambda(\eta^4x^4+4\eta^3x^3+12\eta^2x^2+24\eta x+\eta^4+24)^2e^{-2\eta x} \\ -(1-\lambda)(\eta^4+24)(\eta^4x^4+4\eta^3x^3+12\eta^2x^2+24\eta x+\eta^4+24)e^{-\eta x}\end{bmatrix}} \tag{23}$$

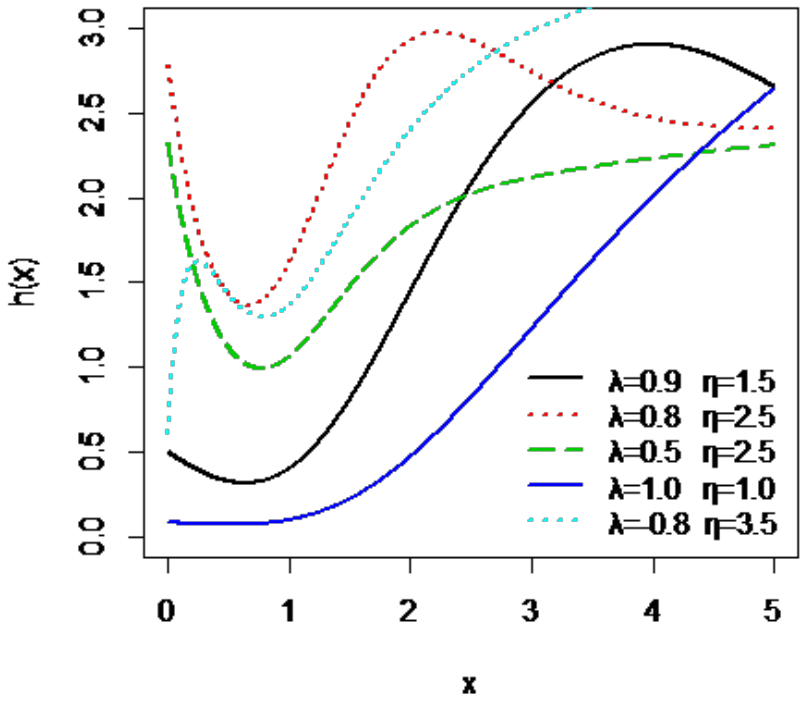


Figure 3: Various shapes of the hazard function of the BES distribution

The odds function of $X \sim BES(\eta, \lambda)$ is given by

$$O_{BES}(x;\eta,\lambda) = \frac{F_{BES}(x;\eta,\lambda)}{1 - F_{BES}(x;\eta,\lambda)}$$

$$O_{BES}(x;\eta,\lambda) = \begin{bmatrix} (1-\lambda)(\eta^4+24)^{-1}(\eta^4x^4+4\eta^3x^3+12\eta^2x^2+24\eta x+\eta^4+24)e^{-\eta x} \\ +\lambda(\eta^4+24)^{-2}(\eta^4x^4+4\eta^3x^3+12\eta^2x^2+24\eta x+\eta^4+24)^2e^{-2\eta x} \end{bmatrix}^{-1} - 1 \tag{24}$$

4. Quantile function and entropy measures of BES distribution

4.1. Quantile function of BES distribution

The x_ω^{th} quantile function of BES distribution satisfies the equation

$$F_{BES}(x;\eta,\lambda) = \omega, \quad 0 < \omega < 1 \tag{25}$$

Plugging (5) into (25), we have

$$(1+\lambda)\left[1-\left(1+\frac{\eta x(\eta^3x^3+4\eta^2x^2+12\eta x+24)}{\eta^4+24}\right)e^{-\eta x}\right]^2 - \lambda\left[1-\left(1+\frac{\eta x(\eta^3x^3+4\eta^2x^2+12\eta x+24)}{\eta^4+24}\right)e^{-\eta x}\right]^2=\omega \tag{26}$$

Let

$$z=1-\left(1+\frac{\eta x(\eta^3x^3+4\eta^2x^2+12\eta x+24)}{\eta^4+24}\right)e^{-\eta x} \tag{27}$$

Then

$$\begin{aligned}(1+\lambda)z-\lambda z^2&=\omega\\ \lambda z^2-(1+\lambda)z+\omega&=0\end{aligned} \tag{28}$$

Applying the quadratic formula on (28), we obtain

$$z=\frac{1+\lambda\pm\sqrt{(1+\lambda)^2-4\lambda\omega}}{2\lambda} \tag{29}$$

Substituting (29) into (27), one obtains

$$\frac{1+\lambda\pm\sqrt{(1+\lambda)^2-4\lambda\omega}}{2\lambda}=1-\left(1+\frac{\eta x_\omega(\eta^3x_\omega^3+4\eta^2x_\omega^2+12\eta x_\omega+24)}{\eta^4+24}\right)e^{-\eta x_\omega}$$

Thus, the quantile is obtained by solving the equations:

$$\frac{1+\lambda\pm\sqrt{(1-\lambda)^2}-4\lambda\omega}{2\lambda}=\left(1+\frac{24}{n^4+24}\sum_{r=1}^{4}\frac{(\eta x_\omega)^r}{r!}\right)e^{-\eta x_\omega} \tag{30}$$

Therefore, the ωth quantile, denoted by x_ω, for BES distribution, is a positive solution of (30), which can be found by numerical method.

4.2. Entropy measures of the BES distribution

The Renyi entropy may be defined for the BES as

$$E_R=\frac{1}{1-\beta}\log\left(\int_0^\infty f_{BES}^{\beta}(x;\eta,\lambda)dx\right),\quad \beta\neq 1,\quad \beta>0$$

$$=\frac{1}{1-\beta}log\left[\begin{array}{l}\left(\frac{\eta^5}{\eta^4+24}\right)^\beta\int_0^\infty(1+x^4)^\beta e^{-\beta\eta x}\\ \times\left((1-\lambda)+2\lambda\left(1+\frac{\eta x(\eta^3x^3+4\eta^2x^2+12\eta x+24)}{\eta^4+24}\right)e^{\eta x}\right)^\beta dx\end{array}\right] \tag{31}$$

Applying binomial expansion to the terms in (31) and simplifying, one gets

$$E_R = \frac{1}{1-\beta} log \left[\begin{array}{l} \left(\frac{\eta^5}{\eta^4+24}\right)^{\beta} \sum_{i=0}^{\infty} \sum_{j=0}^{\beta} \sum_{k=0}^{j} \sum_{l=0}^{k} \sum_{m=0}^{l} \sum_{n=0}^{m} \binom{\beta}{i}\binom{\beta}{j}\binom{j}{k}\binom{k}{l}\binom{l}{m}\binom{m}{n} \eta^{k+l+m+n} \\ \frac{(24)^{k-l}(12)^{l-m}(4)^{m-n}(1-\lambda)^j(2\lambda)^{\beta-j}}{(\eta^4+24)^k} \int_0^{\infty} x^{k+l+m+n} e^{-\eta(\beta+j)x} dx \end{array} \right]$$

$$= \frac{1}{1-\beta} log \left[\begin{array}{l} \left(\frac{\eta^5}{\eta^4+24}\right)^{\beta} \sum_{i=0}^{\infty} \sum_{j=0}^{\beta} \sum_{k=0}^{j} \sum_{l=0}^{k} \sum_{m=0}^{l} \sum_{n=0}^{m} \binom{\beta}{i}\binom{\beta}{j}\binom{j}{k}\binom{k}{l}\binom{l}{m}\binom{m}{n} \\ \frac{(24)^{k-l}(12)^{l-m}(4)^{m-n}(1-\lambda)^j(2\lambda)^{\beta-j}\eta^{k+l+m+n-1}\Gamma(k+l+m++1)}{(\eta^4+24)^k(\beta+j)^{k+l+m+n+1}} \end{array} \right] \tag{32}$$

The Tsallis entropy for the BES distribution may be defined as

$$E_S = \frac{1}{\beta-1}\left(1 - \int_0^{\infty} f_{BES}^{\beta}(x;\eta,\lambda)dx\right), \quad \beta \neq 1, \quad \beta > 0$$

$$= \frac{1}{\beta-1}\left(1 - \left[\begin{array}{l} \left(\frac{\eta^5}{\eta^4+24}\right)^{\beta} \sum_{i=0}^{\infty} \sum_{j=0}^{\beta} \sum_{k=0}^{j} \sum_{l=0}^{k} \sum_{m=0}^{l} \sum_{n=0}^{m} \binom{\beta}{i}\binom{\beta}{j}\binom{j}{k}\binom{k}{l}\binom{l}{m}\binom{m}{n} \\ \frac{(24)^{k-l}(12)^{l-m}(4)^{m-n}(1-\lambda)^j(2\lambda)^{\beta-j}\eta^{k+l+m+n-1}\Gamma(k+l+m++1)}{(\eta^4+24)^k(\beta+j)^{k+l+m+n+1}} \end{array} \right]\right) \tag{33}$$

5. Distributions of order statistics of BES distribution

The PDF of the rth order statistic for $X \sim BES(\eta,\lambda)$ is given by

$$f_{X_{(r)}}(x) = \frac{n!}{(r-1)!(n-r)!}[F(x)]^{r-1}[1-F(x)]^{n-r}f(x)$$

$$= \frac{r\binom{n}{r}\eta^5(1+x^4)e^{-(n-r+1)x}}{\eta^4+24}\left(1+\frac{\eta x(\eta^3x^3+4\eta^2x^2+12\eta x+24)}{\eta^4+24}\right)^{\eta-r}$$
$$\times\left[(1-\lambda)+2\lambda\left(1+\frac{\eta x(\eta^3x^3+4\eta^2x^2+12\eta x+24)}{\eta^4+24}\right)e^{-\eta x}\right]$$

$$\times\left[\begin{array}{l} 1-(1-\lambda)\left(1+\frac{\eta x(\eta^3x^3+4\eta^2x^2+12\eta x+24)}{\eta^4+24}\right)e^{-\eta x} \\ -\lambda\left(1+\frac{\eta x(\eta^3x^3+4\eta^2x^2+12\eta x+24)}{\eta^4+24}\right)^2 e^{-2\eta x} \end{array}\right]^{r-1}$$

$$\times\left[(1-\lambda)+\lambda\left(1+\frac{\eta x(\eta^3x^3+4\eta^2x^2+12\eta x+24)}{\eta^4+24}\right)e^{-\eta x}\right]^{n-r} \tag{34}$$

Putting $r = 1$ in (34), we get the PDF of the first order statistic $X_{(1)}$ as

$$f_{X_{(1)}}(x) = \frac{\eta^5(1+x^4)ne^{-\eta nx}}{\eta^4+24}\left(1+\frac{\eta x(\eta^3x^3+4\eta^2x^2+12\eta x+24)}{\eta^4+24}\right)^{n-1} \times \left[(1-\lambda)+\lambda\left(1+\frac{\eta x(\eta^3x^3+4\eta^2x^2+12\eta x+24)}{\eta^4+24}\right)e^{-\eta x}\right]^{n-1} \times \left[(1-\lambda)+2\lambda\left(1+\frac{\eta x(\eta^3x^3+4\eta^2x^2+12\eta x+24)}{\eta^4+24}\right)e^{-\eta x}\right] \tag{35}$$

Putting $n = r$ in (34), we get the PDF of the largest order statistic $X_{(n)}$ as

$$f_{X_{(n)}}(x) = \frac{\eta^5(1+x^4)ne^{(-\eta-r+1)x}}{\eta^4+24}\left((1-\lambda)+2\lambda\left(1+\frac{\eta x(\eta^3x^3+4\eta^2x^2+12\eta x+24)}{\eta^4+24}\right)\right)^{n-1} \times \left[\begin{array}{l} 1-(1-\lambda)\left(1+\frac{\eta x(\eta^3x^3+4\eta^2x^2+12\eta x+24)}{\eta^4+24}\right)e^{-\eta x} \\ -\lambda\left(1+\frac{\eta x(\eta^3x^3+4\eta^2x^2+12\eta x+24)}{\eta^4+24}\right)^2 e^{-2\eta x} \end{array}\right]^{n-1} \tag{36}$$

6. Maximum likelihood estimates of parameters of BES distribution

Consider a random sample of a sample size, n, $X_1, X_2, ..., X_n$ drawn from the BES distribution. Obviously, the likelihood function of the random sample is

$$L(\eta,\lambda) = \prod_{i=1}^{n} f_{BES}(x_i;\eta,\lambda)$$

$$= \left(\frac{\eta^5}{\eta^4+24}\right)^n e^{-n\sum_{i=1}^n x_i}\prod_{i=1}^{n}(1+x_i^4)\left[(1-\lambda)+2\lambda\left(1+\frac{\eta x(\eta^3x^3+4\eta^2x^2+12\eta x+24)}{\eta^4+24}\right)e^{-\eta x_i}\right] \tag{37}$$

The log-likelihood function is

$$\ln L(\eta,\lambda) = \sum_{i=1}^{n}\ln\left[(1-\lambda)+2\lambda\left(1+\frac{\eta x(\eta^3x^3+4\eta^2x^2+12\eta x+24)}{\eta^4+24}\right)e^{-\eta x}\right] + \sum_{i=1}^{n}\ln(1+x_i^4) + n[5\ln(\eta)-\ln(\eta^4+24)] - \eta\sum_{i=1}^{n}x_i \tag{38}$$

Taking the partial derivatives of (38) with respect to η and λ, and equating the results to zero, yields

$$\frac{\partial \ln L(\eta,\lambda)}{\partial\eta} = \sum_{i=1}^{n}\frac{2\lambda((\eta^3x_i^3+4\eta^2x_i^2+12\eta x_i+24)+\eta(3\eta^2x_i^2+8\eta x_i^2+12x_i))x_ie^{-\eta x_i}}{2\lambda(\eta x(\eta^3x_i^3+4\eta^2x_i^2+12\eta x_i+24)+\eta^4+24)e^{-\eta x_i}+(1-\lambda)(\eta^4+24)}$$

$$
\begin{aligned}
&-\sum_{i=1}^{n}\frac{8\lambda\eta^4x_i((\eta^3x_i^3+4\eta^2x_i^2+12\eta x_i+24)e^{-\eta x_i}}{(\eta^4+24)[2\lambda(\eta x(\eta^3x_i^3+4\eta^2x_i^2+12\eta x_i+24)\eta^4+24)e^{-\eta x_i}+(1-\lambda)(\eta^4+24)]} \\
&-\sum_{i=1}^{n}\frac{2\lambda x_i(4\eta^4x_i(\eta^3x_i^3+4\eta^2x_i^2+12\eta x_i+24)+(\eta^4+24))e^{-\eta x_i}}{2\lambda(\eta x(\eta^3x_i^3+4\eta^2x_i^2+12\eta x_i+24)+\eta^4+24)e^{-\eta x_i}+(1-\lambda)(\eta^4+24)]}+\frac{2(\eta^4+120)}{\eta(\eta^4+24)} \\
&-\sum_{i=1}^{n}x_i=0
\end{aligned}
\tag{39}
$$

$$
\frac{\partial \ln L(\eta,\lambda)}{\partial\lambda}=\sum_{i=1}^{n}\frac{2(\eta x(\eta^3x_i^3+4\eta^2x_i^2+12\eta x_i+24)+\eta^4+24)e^{-\eta x_i}-(\eta^4+24)}{(1-\lambda)(\eta^4+24)+2\lambda(\eta x(\eta^3x_i^3+4\eta^2x_i^2+12\eta x_i+24)+\eta^4+24)e^{-\eta x_i}}=0 \tag{40}
$$

Due to the complex nature of (39) and (40), an iterative method such as Newton-Raphson method is adopted for finding its solution.

7. Asymptotic confidence intervals of the parameters of BES distribution

Let $\hat{\Theta}=(\hat{\eta},\hat{\lambda})^T$ be the MLE of $\Theta=(\eta,\lambda)^T$ for the BES distribution. To construct the confidence intervals, the Fisher information, denoted by $I(\Theta)$ is required. Consequently

$$
I(\Theta)=\begin{pmatrix} I_{\hat{\eta}\hat{\eta}} & I_{\hat{\eta}\hat{\lambda}} \\ I_{\hat{\lambda}\hat{\eta}} & I_{\hat{\lambda}\hat{\lambda}} \end{pmatrix} \tag{41}
$$

The elements of (41) are the second derivatives of (38) with respect to the parameters of the BES distribution. Notice that the asymptotic distribution of $\sqrt{n}(N_2(1,I^{-1}(\Theta))$, under certain regularity conditions. Consequently, the approximate $100(1-\omega)\%$ two sided confidence intervals for η and λ are given, respectively, by

$$
\hat{\eta}\pm Z_{\omega/2}\sqrt{I_{\eta\eta}^{-1}(\hat{\Theta})} \quad \text{and} \quad \hat{\lambda}\pm Z_{\omega/2}\sqrt{I_{\lambda\lambda}^{-1}(\hat{\Theta})} \tag{42}
$$

where $I_{\eta\eta}^{-1}(\hat{\Theta})$ and $I_{\lambda\lambda}^{-1}(\hat{\Theta})$ are diagonal elements of the matrix $I_n^{-1}(\hat{\Theta})$ and $Z_{r/2}$ is the upper $(\omega/2)th$ percentile of a standard normal distribution.

8. Monte-Carlo simulation study of the BES distribution

To investigate the effect of sample size on the maximum likelihood estimates of parameters of the BES distribution and assess the stability of the parameter estimates, it is essential to conduct a Monte-Carlo simulation on the BES distribution.
The simulation procedure as outlined below was performed using R package:

Step 1: Simulate a random sample of size n from the BES distribution with parameters $\lambda=0.8$ and $\eta=1.4$ using the inversion of the CDF method with Equation (30)

Step 2: Set intial values for the parameters of the BES distribution.

Step 3: Compute the MLE of the parameters of the BES distribution.

Step 4: Repeat steps 1-3 $N = 10,000$ times.

Step 5: Compute the mean, standard error, average bias and average mean square error (MSE) of the 10,000 maximum likelihood estimates of each parameter λ and η. The mean estimate of the maximum likelihood estimator $\hat{\tau}$ of the parameter $\tau = (\lambda, \eta)$ is given by

$$\bar{\hat{\tau}} = \frac{1}{N}\sum_{i=1}^{N}\hat{\tau}_i \tag{43}$$

The standard error of $\bar{\hat{\tau}}$ is given by

$$SE_{\bar{\hat{\tau}}} = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\hat{\tau}_i - \bar{\hat{\tau}})^2} \tag{44}$$

The Bias of $\bar{\hat{\tau}}$ is given by

$$Bias(\bar{\hat{\tau}}) = \bar{\hat{\tau}}_i - \tau, \quad i = 1, 2, ..., n \tag{45}$$

The average bias of the MLE $\hat{\tau}$ of the parameter $\tau = (\lambda, \eta)$ is given by

$$Ave.Bias(\hat{\tau}) = \frac{1}{N}\sum_{i=1}^{N}(\hat{\tau}_i - \tau) \tag{46}$$

The average mean square error (MSE) of the MLE $\hat{\tau}$ of the parameter $\tau = (\lambda, \eta)$ is given by

$$Ave.MSE(\hat{\tau}) = \frac{1}{N}\sum_{i=1}^{N}(\hat{\tau}_i - \tau)^2 \tag{47}$$

Step 6: Repeat Steps 1-5 with different sample sizes ($n = 20, 30, 50, 100, 500, 1000$).

Table 1: Simulation results of the estimates, bias and mean square error of the BES distribution parameters for different sample sizes

n	$\hat{\eta}$	$\hat{\lambda}$	$Bias(\hat{\eta})$	$MSE(\hat{\eta})$	$Bias(\hat{\lambda})$	$MSE(\hat{\lambda})$
20	1.5334	0.4270	0.1338	0.0458	-0.3730	0.1951
30	1.5977	0.2602	0.1977	0.0645	-0.1718	0.1879
50	1.4559	0.6282	0.0559	0.0237	-01718	0.1879
100	1.4559	0.7406	0.0220	0.0192	-0.0594	0.1514
500	1.4091	0.7761	0.0091	0.0042	-0.0239	0.0306
1000	1.4030	0.7917	0.0003	0.0014	-0.0083	0.0091

As shown in Table 1, the parameter estimates tend toward the actual parameter values as the sample size increases. Also, average bias and mean squared error tend to zero with increasing sample size.

9. Applications

In this section, we illustrate the flexibility of the BES distribution with two real datasets. The first dataset comprises the failure times of mechanical components as reported in Javed *et al.* (2018).

0.040, 1.866, 2.385, 3.443, 0.301, 1.876, 2.481, 3.467, 0.309, 1.899, 2.610, 3.478, 0.557, 1.911, 2.625, 3.578, 0.943, 1.912, 2.632, 3.595, 1.070, 1.914, 2.646, 3.699, 1.124, 1.981, 2.661, 3.779, 1.248, 2.010, 2.688, 3.924, 1.281, 2.038, 2.82,3, 4.035, 1.281, 2.085, 2.890, 4.121, 1.303, 2.089, 2.902, 4.167, 1.432, 2.097, 2.934, 4.240, 1.480, 2.135, 2.962, 4.255, 1.505, 2.154, 2.964, 4.278, 1.506, 2.190, 3.000, 4.305, 1.568, 2.194, 3.103, 4.376, 1.615, 2.223, 3.114, 4.449, 1.619, 2.224, 3.117, 4.485, 1.652, 2.229, 3.166, 4.570, 1.652, 2.300, 3.344, 4.602, 1.757, 2.324, 3.376, 4.663

The second dataset depicts the fatigue life of some aluminium coupons cut in specific manner reported in Birnbaum and Saunders (1969). The dataset (after subtracting 65) is:

5, 25, 31, 32 ,34 ,35 ,38, 39, 39, 40, 42, 43, 43, 43, 44, 44, 47, 47, 48, 49, 49, 49, 51, 54, 55, 55, 55, 56, 56, 56, 58, 59, 59, 59, 59, 59, 63, 63, 64, 64, 65, 65, 65, 66, 66, 66, 66, 66, 67, 67, 67, 68, 69, 69, 69, 69, 71, 71, 72, 73, 73, 73, 74, 74, 76, 76, 77, 77, 77, 77, 77, 77, 79, 79, 80, 81, 83, 83, 84, 86, 86, 87, 90, 91, 92, 92, 92, 92, 93, 94, 97, 98, 98, 99, 101, 103, 105, 109, 136, 147

Consequently, we fit the BES distribution (BESD) as well as the competing distributions, such as gamma distribution (GD) and each of the following distributions (in each case $g(x)$ is the PDF while $G(x)$ is the CDF of the concerned distribution) to each of the two data sets listed above. The reason for choosing these distributions is because they all belong to the same family of the proposed distribution; so we chose them for comparison to illustrate the flexibility achieved as a result of the generalization. (1) Transmuted Lindley distribution (TLD) (Merovci, 2013a)

$$g(x) = \frac{\eta^2}{\eta+1}(1+x)e^{-\eta x}\left(1-\lambda+2\lambda\left(\frac{\eta+1+\eta x}{\eta+1}\right)e^{-\eta x}\right) \tag{48}$$

and

$$G(x) = \left(1-\frac{\eta+1+\eta x}{\eta+1}e^{-\eta x}\right)\left(1+\lambda\left(\frac{\eta+1+\eta x}{\eta+1}\right)e^{-\eta x}\right) \tag{49}$$

(2) Transmuted Exponential distribution (TED) (Owoloko, *et al.*, 2015)

$$g(x) = \frac{1}{\eta}e^{-\eta x}(1-\lambda+2\lambda e^{-nx}) \tag{50}$$

and

$$G(x) = (1-e^{-\eta x})(1+\lambda e^{-\eta x}), \quad x>0, \eta>0, |\lambda| \leq 1 \tag{51}$$

(3) Transmuted Aradhana distribution (TAD) (Gharaibeh, 2020)

$$g(x) = \frac{\eta^3}{\eta^2+2\eta+2}(1+x)^2e^{-\eta x}\left(1-\lambda+2\lambda\left(\frac{\eta x(\eta x+2\eta+2)}{\eta^2+2\eta+2}\right)e^{-\eta x}\right) \tag{52}$$

and

$$G(x) = (1+\lambda)\left(1-\left(1+\frac{\eta x(\eta x+2\eta+2)}{\eta^2+2\eta+2}\right)e^{-\eta x}\right)-\lambda\left(1-\left(1+\frac{\eta x(\eta x+2\eta+2)}{\eta^2+2\eta+2}\right)e^{-\eta x}\right)^2 \quad (53)$$

(4) Transmuted Ishita distribution (TID) (Sharaibeh and Al-Omari, 2019)

$$g(x) = \frac{\eta^3}{\eta^3+2}(\eta+x^2)e^{-\eta x}\left(1-\lambda+2\lambda\left(1+\frac{\eta x(\eta x+2)}{\eta^3+2}\right)e^{-\eta x}\right) \quad (54)$$

and

$$G(x) = (1+\lambda)\left(1-\left(1+\frac{\eta x(\eta x+2)}{\eta^2+2\eta+2}\right)e^{-\eta x}\right)-\lambda\left(1-\left(1+\frac{\eta x(\eta x+\lambda)}{\eta^3+2}\right)e^{-\eta x}\right)^2 \quad (55)$$

(5) Transmuted Pranav distribution (TPD) (Odom *et al.*, 2019)

$$g(x) = \frac{\eta^4}{\eta^4+6}(\eta+x^3)e^{-\eta x}\left(1-\lambda+2\lambda\left(1+\frac{\eta x(\eta^2x^2+3\eta x+6)}{\eta^4+6}\right)e^{-\eta x}\right) \quad (56)$$

and

$$G(x) = (1+\lambda)\left(1-\left(1+\frac{\eta x(\eta^2x^2+3\eta x+6)}{\eta^4+6}\right)e^{-\eta x}\right)-\lambda\left(1-\left(1+\frac{\eta x(\eta^2x^2+3\eta x+6)}{\eta^4+6}\right)e^{-\eta x}\right)^2 \quad (57)$$

Comparison of the fitted models was basd on the following goodness-of-fit measures: the Akaike Information Criterion (AIC) due to Akaike (1992), given by

$$AIC = -2l + 2k, \quad (58)$$

the Bayesian Information Criterion (BIC) due to Schwarz (1978), given by

$$BIC = k\ln(n) - 2l, \quad (59)$$

and the generalized Carmer-von Mises W* statistics; due to Chen and BAlakrishnan (1995), given by

$$CVM = \frac{1}{12n} + \sum\left[\frac{2i-1}{2n} - \hat{F}(x_i)\right] \quad (60)$$

where k is the number of parameters in the BES distribution, l is the maximized value of the log-likelihood function of the BES distribution,$\hat{F}(x_i)$ is the value of the CDF of the BES distribution and n is the sample size. The smaller the criterion statistics the better the model.

Maximum likelihood estimates of the parameters of the BES distribution and the other seven distributions fitted to both data and the associated results are given in Table 2 and Table 3 for the first and second data respectively.

A comparison of AIC and BIC values of the eight lifetime distributions in Tables 2 and 3 shows that the BES distribution gives a better fit for the lifetime datasets as it has smaller AIC and BIC values than the others. The estimated parameters also satisfy the theoretical range of the parameters as expected.

Table 2: Maximum likelihood fit of the failure times of mechanical components data

Models	Estimates	SE	ℓ	AIC	BIC	KS	CVM	AD
BES $\hat{\eta}$ $\hat{\lambda}$	1.9390 -0.9844	0.0657 0.0762	-131.4167	266.9234	271.8087	0.0811	0.0910	0.8088
SD $\hat{\eta}$	1.6098	0.0667	-138.8172	279.6344	282.0770	0.1488	0.3886	2.8274
GD a b	3.5284 1.3769	0.5177 0.2171	-138.3853	280.7907	285.6760	0.1037	0.1779	1.4110
TLD $\hat{\eta}$ $\hat{\lambda}$	0.8495 -0.9686	0.0538 0.0622	-140.1180	284.2359	289.1212	0.1440	0.4764	3.0160
TED $\hat{\eta}$ $\hat{\lambda}$	0.5718 -0.9970	0.0470 0.5029	-146.8122	297.6244	302.5097	0.1855	0.8876	5.0645
TID $\hat{\eta}$ $\hat{\lambda}$	1.1561 -0.9720	0.0503 0.0689	-136.8579	277.7159	282.6012	0.9735	26.4719	241.0491
TAD $\hat{\eta}$ $\hat{\lambda}$	1.1838 -0.9528	0.0646 0.0678	-155.8115	315.6229	320.5082	0.9548	23.1526	195.3768
TPD $\hat{\eta}$ $\hat{\lambda}$	1.4740 -0.9871	0.0518 0.0672	-135.5607	275.1213	280.0066	0.9877	27.4735	306.8742

10. Conclusion

This paper introduces a new lifetime distribution, named the BES distribution. The new distribution generalizes the Suja distribution. We have provided explicit mathematical expressions for some of its basic statistical properties such as the probability density function, cumulative distribution function, rth crude and central moments, variance, coefficient of variation, skewness, kurtosis, and quantile function and some reliability characteristics like the survival, hazard rate, cumulative hazard and reverse hazard functions. Rényi and Tsallis entropies were discussed. Also, the distributions of rth, first and largest order statistics were introduced. Estimation of the model parameters was approached through the method of maximum likelihood estimates. A Monte-Carlo simulation was performed to verify the stability of the maximum likelihood estimates of the model parameters. The flexibility and applicability of the new lifetime distribution were illustrated with two real data sets and the results obtained revealed that the BES distribution provides the best fit among all the compared related distributions. We recommend the transmuted distribution for modelling unimodal or bimodal continuous lifetime data with a nondecreasing or bathtub shaped hazard rate function and hope that it would receive significant applications in the future.

Table 3: Maximum likelihood fit of the fatigue life of some aluminium coupons data

Models	Estimates	SE	ℓ	AIC	BIC	KS	CVM	AD
BES $\hat{\eta}$ $\hat{\lambda}$	0.0888 -0.8960	0.0035 0.0967	- 454.9401	913.8802	919.0906	0.0950 7	0.1431	0.8679
SD $\hat{\eta}$	0.0732	0.0327	- 462.1056	926.2112	928.8163	0.1360	0.5334	3.2141
GD a b	1.0000 0.9677	0.9677 1.0000	- 457,8804	919.7608	924.9712	0.0998	0.1629	0.9871
TLD $\hat{\eta}$ $\hat{\lambda}$	0.0390 -0.1010	0.0021 0.0389	- 471.1273	946.2546	951.465	0.1695	1.0042	5.9380
TED $\hat{\eta}$ $\hat{\lambda}$	0.0231 -1.2477	0.0016 0.0444	-488.351	980.702	985.9123	0.2304	2.0347	
TID	0.5627 -0.9565	0.0026 0.0026	- 558.2248	112045	1125.66	0.9036	26.5551	
TAD $\hat{\eta}$ $\hat{\lambda}$	1.1838 -0.9528	0.0646 0.0678	- 155.8115	315.6229	320.5082	0.9548	25.1526	
TPD $\hat{\eta}$ $\hat{\lambda}$	0.7284 -0.9348	0.0029 0.0662	-456.6	9.7200	922.4104	0.9342	28.2863	200.1088

References

Abdul Moniem, I. B. and Seham, M. (2015). Transmuted Gompertz distribution. *Computational and Applied Mathematics Journal*, **1**, 36-38.

Akaike, H. (1992). Information theory and an extension of the maximum likelihood principle. in: Breakthroughs in Statistics, *Springer, New York*. 610-624

Al-khazaleh, M., Al-Omari, A. I., and Al-khazaleh, A. H. (2016). Transmuted two-parameter Lindley distribution. *Journal of Statistics Applications and Probability*, **5**, 1-11.

Al-Omari, A. I. and Alsmairan, I. K. (2019). Length-biased Suja distribution and its application. *Journal of Applied Probability and Statistics*, **14**, 95-116.

Al-Omari, A. I., Alhyasat, K. M., Ibrahim, K., and Bakar, A. A. M. (2019). Power length-biased Suja distribution: properties and application. *Electronic Journal of Applied Statistical Analysis,* **12**, 429-452.

Alsmairan, I. K. and Al-Omari, A. I. (2020). Weighted Suja distribution with applications

to ball bearings data. *Life Cycle Reliability and safety Engineering,* **9**, 195-211.

Aryal, G. R. (2013). Transmuted log-Logistic distribution. *Journal of Statistics Applications and Probability,* **2**, 11–20.

Aryal, G. R. and Tsokos C. P. (2011). Transmuted Weibull distribution: A generalization of Weibull probability distributions. *European Journal of Pure and Applied Mathematics*,**4**, 89–102.

Aryal, G. R. and Tsokos, C. P. (2009). On the transmuted extreme value distribution with applications. *Nonlinear Analysis: Theory, Methods and Applications*, **71**, 1401-1407.

Ashour, S. K. and Eltehiwy, M. A. (2013). Transmuted Lomax distribution. *American Journal of Applied Mathematics and Statistics,* **1**, 121-127.

Birnabaum, Z. W. and Saunders, S. C. (1969). Estimation for a family of life distributions with applications to fatigue. Journal of Applied Probability, **6**, 328-347.

Bleed, S. O. and Abdelali, A. E. A. (2018). Transmuted Arcsine distribution properties and application. *International Journal of Research-Granthaalayah*, **6**, 38-47.

Bose, R. C. and Nair, K. R. (1939). Partially balanced incomplete block designs. *SankhyA: The Indian Journal of Statistics*, **4**, 337-372.

Chen, G. and Balakrishnan, N. (1995). A general purpose approximate goodness-of-fit test. *Journal of Quality Technology*, **27**, 154-161.

Enogwe, S. U., Nwosu, D. F., Ngome, E. C., Onyekwere, C. K., and Omeje, I. L. (2020). Two-Parameter Odoma distribution with applications. *Journal of Xidian University,* **14**, 740-764.

Gharaibeh, M. M. (2020). Transmuted Aradhana distribution: Properties and Application. *Jordan Journal of Mathematics and Statistics*, **13**, 287-304.

Gharaibeh, M. M. and Al-Omari, A. (2019). Transmuted Ishita distribution and its applications. *Journal of Statistics Applications and Probability*, **8**, 67-81.

Ghitany, M. E., Atieh, B., and Nadarajah, S. (2008). Lindley distribution and its applications. *Mathematics and Computers in Simulation,* **78**, 493-506.

Javed, M., Nawaz, T., and Irfan, M. (2018). The Marshall-Olkin Kappa distribution: properties and applications. *Journal of King Saudi University-Science*, **31**, 684-691.

Mahmoud, M. R. and Mandouh, R. M. (2013). On the transmuted Frechet distribution. *Journal of Applied Sciences Research*, **9**, 5553-5561.

Maurya, R. K., Tripathi, Y. M., and Rastogi, M. K. (2017). Transmuted Burr XII distribution. *Journal of the Indian Society for Probability and Statistics*, **18**, 177-193.

Merovci, F. (2013a). Transmuted Lindley distribution. *International Journal of Open Problems in Computer Science and Mathematics,* **6**, 63-72.

Merovci, F. (2013b). Transmuted Rayleigh distribution. *Austrian Journal of Statistics,* **22**, 21-30.

Merovci, F. and Puka, I. (2014). Transmuted Pareto distribution. *ProbStat Forum*, **7**, 1-11.

Mohiuddin, M., Rather, A. A., Subramanian, C., and Dar, S. A. (2020). Transmuted Garima distribution: Properties and Applications. *Journal of Xidian University*, **14**, 112-123.

Odom, C. C., Nduka, E. C., and Ijomah, M. A. (2019). A modification of Pranav distribution using quadratic rank transmutation map approach. *International Journal of Scientific Research in Mathematical and Statistical Sciences*, **6**, 193-202.

Owoloko, E. A., Oguntunde, P. E., and Adejumo, A. O. (2015). Performance rating of the transmuted exponential distribution: an analytical approach. *Springer Plus*, **4**, 1-15.

Rather, A. A. and Subramanian, C. (2018). Transmuted Mukherjee-Islam failure model, *Journal of Statistics Applications and Probability*, **7**, 343-347.

Schwarz, G. (1978). Estimating the dimension of a model. *Annals of Statistics,* **6**, 461–464.

Shahzad, M. N. and Asghar, Z. (2016). Transmuted Dagum distribution: A more flexible and broad shaped hazard function model. *Hacettepe Journal of Mathematics and Statistics*, **45**, 227-244.

Shanker, R. (2015a). Akash distribution and its applications. *International Journal of Probability and Statistics,* **4**, 65-75.

Shanker, R. (2015b). Shanker distribution and its applications. *International Journal of Statistics and Applications,* **5**, 338-348.

Shanker, R. (2016a). Amarenha distribution and its applications. *American Journal of Mathematics and Statistics*, **6**, 44-56.

Shanker, R. (2016b). Aradhana distribution and its applications. *International Journal of Statistics and Applications,* **6**, 23-34.

Shanker, R. (2016c). Devya distribution and its applications. *International Journal of Statistics,* **6**, 189-202.

Shanker, R. (2016d) Sujatha distribution and its applications. *Statistics in Transition New Series,* **17**, 391-410.

Shanker, R. (2017). Suja distribution and its application. *International Journal of Probability and Statistics*, **6**, 11-19.

Shaw, W. and Buckley, I. (2007). The alchemy of probability distributions: beyond Gram-Charlier expansions, and a skew-kurtotic normal distribution from a rank transmutation map. *Research Report.*

Statistics and Application {ISSN 2454-7395 (online)}
Volume 21, No. 2, 2023 (New Series), pp 175-192
http://www.ssca.org.in/journal

Evaluation of the Status of Frigate Tuna *Auxis thazard* (Lacepède, 1800) Fishery in the Tamil Nadu Coast of India

R. Abinaya and M. K. Sajeevan

Department of Fisheries Resource Management, Kerala University of Fisheries and Ocean Studies, Kochi, 682506, Kerala, India.

Received: 16 October 2022; Revised: 22 January 2023; Accepted: 22 February 2023

Abstract

The frigate tuna *Auxis thazard* (Lacepède, 1800) is one of the commercially important tuna species that contribute a major share to tuna fisheries of Tamil Nadu. Information on biological reference points and stock status is necessary for effective fishery management. Hence, a stock assessment study was carried out to understand the status of the stock. The present study was used the Catch and Effort Data Analysis (CEDA) software to investigate stock dynamics by running surplus production models with catch and effort data. Reconstructed time series catch and effort data from 1998 to 2018 were used for the study. Annual landings fluctuated between 555 and 2,523 metric tonnes (MT) with an average catch of 1,732 MT year^{-1}. Based on the diagnostic graph, high R^2 and low root mean square error (*RMSE*) value, the Fox log-normal model was selected as the best-fit model for further analysis of biological reference points (BRPs). The best-fitted Fox log-normal model estimated maximum sustainable yield (*MSY*), biomass yield *MSY* (B_{MSY}) and fishing mortality yield *MSY* (F_{MSY}) as 2,543 MT, 3,723 MT and 0.69 MT, respectively. F_{MSY} and B_{MSY} values were compared with current fishing mortality (F) and biomass (B). A lower F/F_{MSY} value (0.41) and higher B/B_{MSY} value (1.66) indicated that the frigate tuna stock of Tamil Nadu has not reached to overfishing or overfished status. However, an overall reduction trend of catch per unit effort (*CPUE)* since 2012 indicates that stock is exploited very close to *MSY*. Results from the BRPs showed that the frigate tuna resource off Tamil Nadu were optimally exploited and an increase in effort will lead to the collapse of the fishery in future. Hence, it is recommended to maintain the fishing effort to the present level for ensuring sustainable exploitation.

Key words: Biological reference points; Catch and effort data analysis; Maximum sustainable yield; Stock exploitation; Tuna stock assessment.

1. Introduction

The sustainable development of marine fisheries is an important activity from a social, environmental and economic view. Total catch is an important metric for monitoring and assessing the status of a fishery (George and Gopalakrishnan, 2013). India's fisheries have long been accessible to the public, with limited control, leading to unsustainable

Corresponding Author: R. Abinaya
Email: abinayabfsc@gmail.com

expansion and development (Devaraj and Vivekanandan, 1999; Satyanarayana *et al*., 2008; Bhathal, 2014; Ansell, 2020). Tamil Nadu is the southeastern maritime state of India, where marine and inland fish production has steadily increased (Tabitha and Gunalan, 2012). Tamil Nadu's marine fisheries have proliferated since 1950, due to the introduction of innovative fishing vessels, new fishing gear, fishing methods and infrastructural facilities. This rapid growth in exploitation resulted in an increased fish catch.

In fisheries management, the concept of sustainable development is always the baseline. However, sustainable management of these renewable but exhausting natural fish stocks is challenging. Most of the world's fishing is biologically and economically unsustainable, much against the belief that fish stocks are inexhaustible (FAO, 1994). Because of the intensive fisheries and the dramatic collapse of fish stocks in India, alarming calls were made to reduce the size of the fishing fleet and fishing efforts. In this context, Tamil Nadu is not an exception. With the large influx of giant mechanized fishing crafts and gears over the years, Tamil Nadu has also seen notable progressions in fishing technology.

More than 80% of the world's marine fish stocks are overexploited or almost fully exploited due to their high nutritional value, local market demand and export demand (Kituyi and Thomson, 2018). India's tuna fisheries are in the initial stages of exploitation due to the adoption of advanced fishing gear (Lecomte *et al*., 2017). Tuna landings contribute 2.93% of India's total marine fish landings (CMFRI, 2019). Tamil Nadu holds the second rank in total tuna production in the country, next to Kerala (CMFRI, 2018). Information on the stock assessment of coastal tuna is limited (Silas *et al*., 1985, James *et al*., 1987; Kasim and Mohan, 2009; Sivadas *et al*., 2020) and less information is available on the tuna fishery of Tamil Nadu (Joseph and Jayaprakash, 2003; Abdussamad *et al*., 2008; Kumar *et al*., 2019; Sivadas *et al*., 2019). Frigate tuna *Auxis thazard* (Lacepède, 1800) is one of the most important neritic tuna species in Indian waters. They live closer to the continental shelf and do not undertake transoceanic migrations (Lecomte *et al*., 2017). *Auxis* spp. contributed 11.9 and 13.1 % of total tuna landings in India and Tamil Nadu, respectively (CMFRI, 2019). Ghosh *et al*. (2012), Mudumala *et al*. (2018) and Dan (2021) provided some information on biological reference points (BRPs) of frigate tuna fishery from Indian waters. However, there is no record of BRPs of frigate tuna stock off the Tamil Nadu coast. Hence, the present study made an attempt to investigate the sustainability status of frigate tuna fisheries off Tamil Nadu.

In India, the Department of Animal Husbandry, Dairy and Fisheries (DADF) submits national fish catch statistics to international organizations such as the FAO. The DADF collects information from the state fisheries departments and central institutes, namely the Central Marine Fisheries Research Institute (CMFRI) and the Fishery Survey of India (FSI) (Malhotra and Sinha, 2007). CMFRI publishes group-wise landing data every year, but there is no record of species-wise landing data (CMFRI, 2019). The effort used for the Indian fishery is not available in any public domain. Hence present study attempted to reconstruct the catch and effort data of frigate tuna from 1998 to 2018. This reconstructed catch and effort data from 1998 to 2018 were utilized to understand the dynamics of tuna fishery and the stock status of frigate tuna fisheries off Tamil Nadu.

2. Materials and methods

The catch and effort statistics of frigate tuna from 1998 to 2018 (21 years) were reconstructed using the handbook of Fisheries Statistics (CMFRI, 2006; 2010; 2011; 2012a; 2012b; 2013; 2014; 2015; 2016; 2017; 2018; DADF, 2009; 2012; 2015; 2018) as well as

several other historical fisheries survey reports and State Government reports (GOT, 2004; 2005; 2006; 2010; 2011; 2012; 2013; 2014; 2015; 2016; 2017; 2018; 2019; 2020) following Bhathal (2014). The fishing effort and annual total catch were estimated as million horsepower (HP) days and metric tonnes (MT), respectively. Statewise and species-wise fish landing data was not available for frigate tuna landings of Tamil Nadu during the study period. Catch data of the frigate tuna fishery of Tamil Nadu from 1998 to 2005 was taken from Bhathal (2014). Landing data of frigate tuna from 2006 to 2012 was reconstructed by converting groupwise neritic tuna landing data (DADF, 2012) to species-wise based on the composition of neritic tuna landings (MOA, 2001). Landing data from 2013 to 2018 was collected from CMFRI (2013; 2014; 2015; 2016; 2017; 2018). In Tamil Nadu waters, tunas were harvested with drift gillnets of mesh size of 120-140 mm and net pieces of 40-50 (98.75%), long lines with a hook size of 4 to 8 (0.75%), trawl nets (0.42%) and handlines (0.08%) (Kumar *et al*., 2018; 2019). The first step in the rebuilding of the fishing effort was to collect data (number of boats, fishing days and gear category) from national and state Government documents, research articles, fisheries survey reports, grey literature and databases between 1998 and 2018 (GOT, 2006; 2010; 2011; 2012; 2013; 2014; 2015; 2016; 2017; 2018; 2019; 2020; CMFRI, 2010; 2011; 2012; 2013; 2014; 2015; 2016; 2017; 2018; CMFRI, 2006; 2012a; Bhathal, 2014).

Data were collected and formalized with the essential elements such as vessels with and without engines, including the total number of vessels, total power (HP units), fishing days and crew size (Abinaya and Sajeevan, 2022a). Fishing effort for vessels without engines (HP days) was estimated by integrating the number of vessels, crew size and fishing days. An average crew size of 8 was used to reconstruct the effort of a vessel without an engine (non-mechanized and non-motorized) from 1998 to 2018 following Bhathal (2014). The fishing effort of mechanized and motorized vessels was calculated using the average engine power of vessels with an estimated number of fishing days by each gear sector at a given time. The number of fishing days was calculated assuming that six fishing days were carried out each week of the year. Downtime and spiritual holidays were subtracted from the total number (6 multiplied by the number of weeks in a year) to calculate the actual fishing days. The average number of days spent for fishing with gillnets, liners and hand lines, and trawl nets were 216, 75 and 228 days, respectively. To accommodate variations and differences in fishing power and efficiency, the nominal effort was corrected to a standard type (Bhathal, 2014).

Different approaches have been used to estimate the biological characteristics (*MSY*, B_{MSY} and F_{MSY}) of species. The ordinary least squares method estimated surplus production functions, especially the Schaefer model, the Fox model, the Schnute model, and the Clark, Yoshimoto and Pooley (CY & P) model (Sin and Yew, 2016). The Schnute Model and the CY & P models have limited use in tropical areas as they were developed for long-lived species (Sparre and Venema, 1998; Lindawati *et al*., 2021). Hence, the biological parameters were evaluated in the present study using the Fox (1970), Schaefer (1954), and Pella-Tomlinson (1969) models.

Reconstructed time series of catch and effort data of frigate tuna fishery was analyzed using the fishery-specific computer program Catch and Effort Data Analysis version 3.1 (CEDA) (MRAG, 2016). CEDA is built to carry out the stock assessment in data-deficient fisheries like the frigate tuna fisheries of Tamil Nadu. CEDA used analytical techniques to support and help stock assessments, resulting in a prediction of current population size, either in numbers or biomass and a better estimate of fishing mortality, by correlating catches with the size of the population (Hoggarrth *et al*., 2006). Surplus production models (SPMs) used

in these assessment tools include three types of non-equilibrium models: Fox, Schaefer and Pella-Tomlinson models with three error assumptions (normal, log-normal and gamma). Schaefer (1954) developed the first surplus production model. Here, the logistic population growth model serves as a basis for the Schaefer model:

$$\frac{dB}{dt} = rB(B_\infty - B) \quad (1)$$

Biological reference points can be calculated from the model parameters

$$MSY = {}^{K\,r}/_{4} \quad (2)$$

$$B_{MSY} = K / 2 \quad (3)$$

$$F_{MSY} = r / 2 \quad (4)$$

$$q = CPUEt / B \quad (5)$$

$$K = n\,1/(n-1)\,X\,B_{MSY} \quad (6)$$

$$r = n\,X\,F_{MSY} \quad (7)$$

Following that, Pella-Tomlinson (1969) recognized a generalized production equation:

$$\frac{dB}{dt} = rB\left(B_\infty{}^{n-1} - B^{n-1}\right) \quad (8)$$

And Fox (1970) proposed a Gompertz growth equation:

$$\frac{dB}{dt} = rB(InB_\infty - InB) \quad (9)$$

where *B*, fish stock biomass; *t*, time in the year; *r*, intrinsic rate of population increase; B_∞ and *K*, carrying capacity; *MSY*, maximum sustainable yield; *q*, catchability coefficient; *CPUE*, catch per unit effort; $B_{MSY,}$ biomass corresponding to *MSY*; $F_{MSY,}$ exploitation rate corresponding to *MSY*; *n*, a parameter that controls the shape of the production curve.

Output parameters of CEDA software were *MSY*, *K*, *B*, in MT, catchability coefficient (*q*) (a scaling term) and *r* (per capita change in the population per unit time). CEDA necessitates an initial proportion (IP) input (starting population size over the maximum catch). The fishery began with a virgin population when the initial proportion is set to zero or close to zero, and with an extensively exploited population when it is set to one or close to one. The present study set the initial biomass (B1) as B1=*K* to assure valid results. The carrying capacity (*K*) is the highest population size, density, or biomass that a given area can sustain (Hartvigsen, 2017). Linear regression analysis was carried out to find out the association between catch and effort and the goodness of fit of models (Hanchet *et al*., 1993). Coefficient of determination(R^2) of the goodness of fit model, results of the diagnostic graph and root mean square error (*RMSE*) (Abinaya and Sajeevan, 2022b) were considered for selecting the results of the Fox log-normal model for further investigation on *MSY*, B_{MSY}, and F_{MSY}.

3. Results

The present study reconstructed catch and effort data of frigate tuna of Tamil Nadu from 1998 to 2018 and presented in Table 1. The average annual landings from 1998 to 2018 was 1,732 MT year^{-1} (Standard deviation (*SD*) =565), with the production in 2001 yielding the lowest catch of 555 MT and the production in 2010 yielding the highest catch of 2,523 MT. From 1998 to 2018, the catch of frigate tuna increased with wide fluctuations. The reconstructed effort data for frigate tuna was stable in the initial periods, then registered a decreasing trend since 2006.

Table 1: Total catch, effort and catch per unit effort (*CPUE*) of frigate tuna fishery from the Tamil Nadu coast (1998-2018)

Year	Total catch (in metric tonnes)	Effort (in million HP days)	*CPUE* (in MT/Hp days)
1998	1434	6.17	0.00023
1999	903	6.33	0.00014
2000	1008	6.33	0.00016
2001	555	6.33	0.00009
2002	1004	6.33	0.00016
2003	1832	6.33	0.00029
2004	1582	6.33	0.00025
2005	1415	4.70	0.00030
2006	1415	18.27	0.00008
2007	1865	17.44	0. 00011
2008	2022	16.54	0. 00012
2009	2501	15.64	0. 00016
2010	2523	16.90	0. 00015
2011	2344	16.97	0.00014
2012	2481	14.13	0.00018
2013	1740	13.26	0.00013
2014	1588	14.46	0.00011
2015	1977	14.79	0.00013
2016	1919	14.79	0.00013
2017	1778	14.87	0.00012
2018	2482	14.62	0.00017

The *CPUE* of frigate tuna in Tamil Nadu from 1998 to 2018 is depicted in Table 1. As shown in Table 1, *CPUE* decreased from 1998 to 2001, then increased and peaked during 2005. After that, the *CPUE* showed a decreasing trend with minimum annual fluctuation.

CEDA mandates an initial proportion (IP) that yields trustworthy findings. Employing three-production models (Fox, Schaefer and Pella-Tomlinson model) with a three-error assumption model (normal, log-normal and gamma), a different range of Maximum sustainable yield (*MSY*) was anticipated by using various ranges of initial proportions (0.1 to 0.9). Results are furnished in Table 2. The CEDA package produced different *MSY* results for frigate tuna fishery and was sensitive to input IP values ranging from 0.1 to 0.9 (Table 2). IP value measures the extent of stock exploitation before the investigation. The initial landing

(1,434 MT) surpassed the highest catch (2,523 MT) by a proportion of 50%, hence an IP value of 0.5 was used in the study.

Table 2: Various *MSY* estimated (in metric tonnes) from CEDA software using an initial proportion of 0.1 to 0.9 for frigate tuna fishery from 1998 to 2018

IP	Fox		Schaefer		Pella-Tomlinson	
	normal	log-normal	normal	log-normal	normal	log-normal
0.1	9975	5872	7261	5213	7261	5213
0.2	7932	4877	5216	4251	5216	4251
0.3	7598	3214	4982	3621	4982	3621
0.4	6992	2987	4211	2651	4211	2651
0.5	4008	2582	3635	2086	3635	2086
0.6	3222	2028	2281	1982	2281	1982
0.7	2865	1721	1892	1723	1892	1723
0.8	2423	1466	1526	1532	1526	1532
0.9	1876	1299	1199	1182	1199	1182

The BRPs for three surplus productions with their error assumption models evaluated using CEDA software for frigate tuna fisheries in Tamil Nadu coastal waters using an IP of 0.5 were furnished in Table 3. As shown in Table 3, the Schaefer and Pella-Tomlinson (normal) model projected a greater carrying capacity (*K*) (12,991 MT) than the Fox model. The Fox (log-normal) model, on the other hand, predicted a better catchability coefficient (*q*), as well as the Schaefer (normal) model, which revealed a higher intrinsic population growth rate (*r*) than the other surplus production models. Results of the computed *MSY* value varied from 2,086 MT (Schaefer & Pella-Tomlinson log-normal) to 4,008 MT (Fox-normal). *RMSE* value ranged from 444 MT (Fox-normal) to 524 MT (Schaefer & Pella-Tomlinson- normal). The R^2 values of the Fox model (normal and log-normal) results were 0.09 and 0.15, respectively. The R^2 values for the Schaefer and Pella-Tomlinson models with normal and log-normal error assumptions were 0.10 and 0.06, respectively, but the gamma assumption failed to minimize. The expected high R^2 values of the surplus production models demonstrated a superior fit to the data. The result of the B_{MSY} value varied between 3,723 MT (Fox log-normal) and 6,496 MT (Schaefer & Pella-Tomlinson - normal). The result of the F_{MSY} value varied between 0.45 (Schaefer & Pella-Tomlinson log-normal) and 1.32 (Fox-normal).

Table 3: Biological reference points and intermediate parameters of frigate tuna fisheries in Tamil Nadu from 1998 to 2018

Model	*K*	*q*	*r*	*MSY*	*RMSE*	R^2	*B*	B_{MSY}	F_{MSY}
Fox (normal)	11843	1.42E-08	0.92	4008	523	0.09	8987	4357	1.32
Fox (log-normal)	10121	2.18E-08	0.69	2582	444	0.15	6195	3723	0.69
Schaefer (normal)	12991	1.24E-08	1.12	3635	524	0.10	10426	6496	0.46
Schaefer (log-normal)	10633	2.15E-08	0.78	2086	487	0.06	6082	5317	0.45
Pella-Tomlinson (normal)	12991	1.24E-08	1.12	3635	524	0.10	10426	6496	0.46
Pella-Tomlinson (log-normal)	10633	2.15E-08	0.78	2086	487	0.06	6082	5317	0.45

K, carrying capacity; *q*, catchability coefficient; *r*, intrinsic population growth rate; *MSY*, maximum sustainable yield; *RMSE*, root mean square error; R^2, coefficient of determination; *B*, current biomass; $B_{MSY,}$ biomass giving *MSY* (expressed in metric tonnes); $F_{MSY,}$ fishing mortality giving *MSY*.

Estimated high R^2 and low *RMSE* values of Fox (log-normal) demonstrated an excellent fit to the data (Table 3) in addition to residual plot results. Selected best-fitting Fox log-normal model results are furnished in Table 4. As shown in Table 4, the current biomass (6,195 MT) was more than B_{MSY} (3,723 MT) and fishing mortality (0.28) was less than F_{MSY} (0.69 MT), and the ratio of B/B_{MSY} and F/F_{MSY} values were 1.66 and 0.17, respectively.

Table 4: Biological reference points of frigate tuna fisheries in Tamil Nadu from 1998 to 2018 estimated by fitting Fox log-normal model

B	*F*	*MSY*	B_{MSY}	F_{MSY}	B/B_{MSY}	F/F_{MSY}
6195	0.28	2582	3723	0.69	1.66	0. 40

B, current biomass; *F*, fishing mortality; *MSY*, maximum sustainable yield; $B_{MSY,}$ biomass giving *MSY* (expressed in metric tonnes); $F_{MSY,}$ fishing mortality giving *MSY*; B/B_{MSY}, a ratio of biomass to biomass giving *MSY*; $F/F_{MSY,}$ a ratio of fishing mortality to fishing mortality giving *MSY*.

The equilibrium yield curve for frigate tuna in Tamil Nadu is represented in Figure 1. As illustrated in Figure 1, the estimated B_{MSY} was 3,723 MT, with a maximum yield of 2,582 MT.

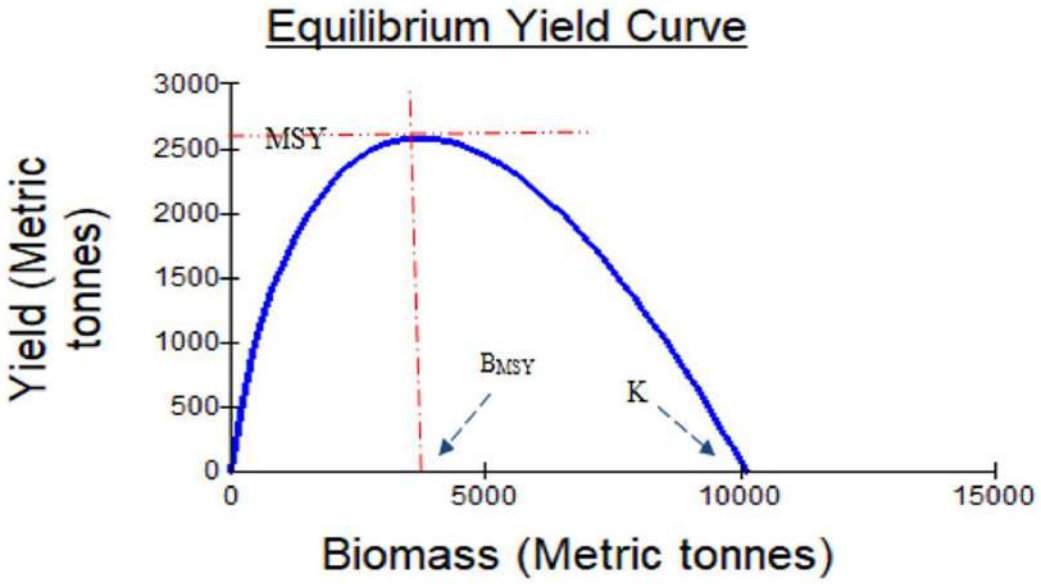


Figure 1: The equilibrium yield curve of the frigate tuna fishery in Tamil Nadu from 1998 to 2018 fitted by the Fox log-normal model

The relationship between expected and observed *CPUE* from 1998 to 2018 is depicted in Figure 2. The expected catch remained stable with slight fluctuation, while observed catches decreased with fluctuation between 1998 and 2018 (Figure 2). The present study used two diagnostic graphs (expected and observed *CPUE* & estimated and observed catches) to show how much the model fits the data. These graphs help to determine the location of a data point on the observed and expected catch graphs on the residual plots. As a result, CEDA can highlight any particular data point as a red square on two diagnostic graphs simultaneously, allowing the user to determine if the point is an outlier or a candidate for exclusions. However, the present study did not exclude any data points from the dataset.

The relationship between estimated and observed catches for all models with an IP value of 0.5 is depicted in Figure 3. Visual inspection demonstrated that the observed catches of normal and log-normal results of the Fox, Schaefer & Pella-Tomlinson models were relatively close to the estimated catch; however, they varied considerably. The estimated and observed catches of the Gamma error model demonstrated a minimization failure to the Fox, Schaefer, and Pella-Tomlinson models (Figure 3).

Linear regression analysis is conducted using catch and effort data from 1998 to 2018 presented in Table 5 and Figure 4. As shown in Table 5, F statistics test the overall significance of the relationship. The relationship between catch and effort data of frigate tuna was statistically significant (p-value < 0.05). A multiple R-value of 0.7 between the two variables indicated that they had a significant and positive association. R^2 and adjusted R^2 were used to determine explained and unexplained variance. According to the results, the regression explained 48% of the total variation in the catch. A histogram of regression analysis over standardized residual is plotted in Figure 4 and a normal P-P plot of regression standardized residuals is illustrated in Figure 5. The residuals of the regression line were normally distributed and confirmed that the regression line satisfies the normality assumption.

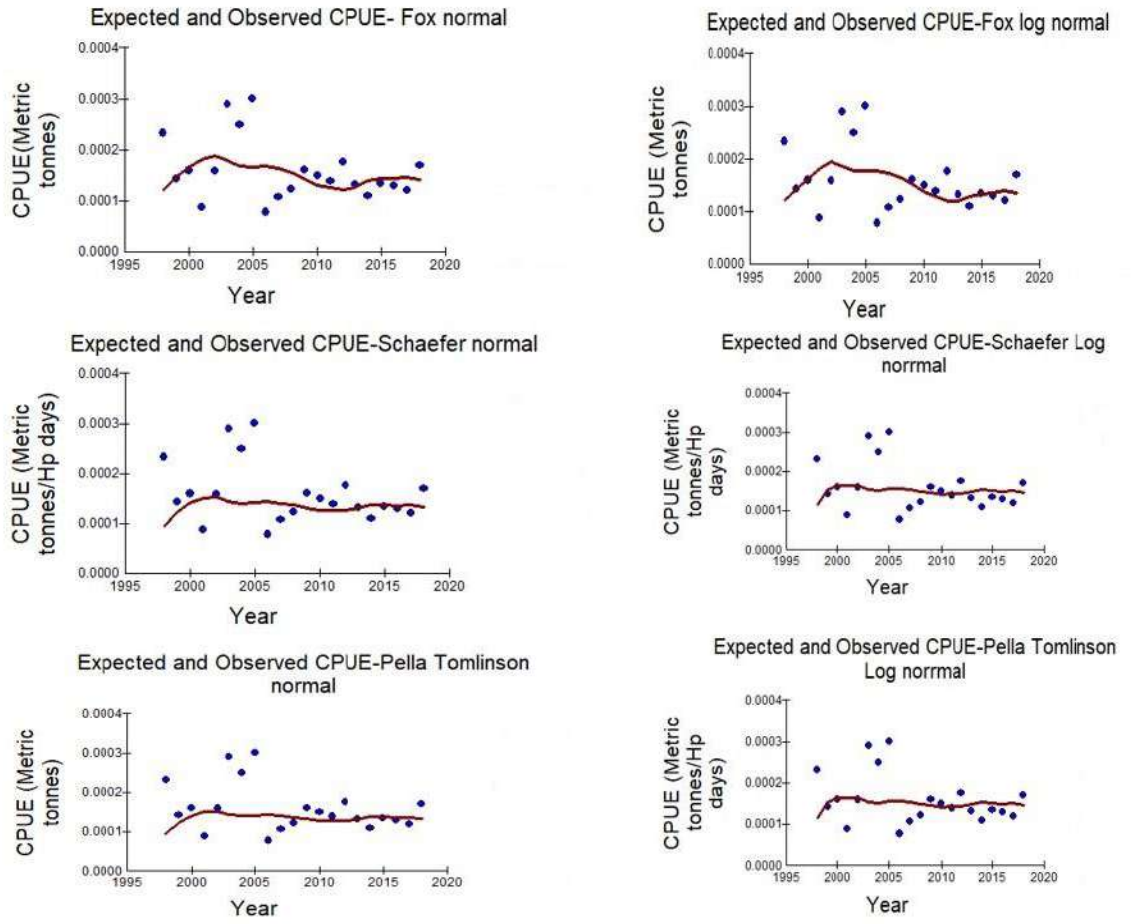


Figure 2: Time series of expected and observed catch per unit effort (*CPUE*) of frigate tuna fishery in Tamil Nadu from 1998 to 2018

4. Discussion

The total landing of frigate tuna in Tamil Nadu in 1998 was 1,434 MT. After a decline of catch to 555 MT in 2001, landings reached their peak of 2,523 MT in 2010. In general, landings registered an increasing trend from 1998 to 2018 with wide fluctuations in some years (Table 1). Kasim and Vivekanandan (2011) observed a decreasing trend in frigate tuna production from 1998 to 2001 and an increasing trend from 2002 to 2010. Increasing trend recorded by the present study concurrent with Kasim and Vivekanandan (2011). Sivadas *et al*. (2019) reported a large-scale increase in fishing efforts after the occurrence of Tsunami. The size of the boat increased from 11-12 meters (m) to 20-23 m overall length, and the fishing net weighing one MT was replaced with more than six MT. The present study recorded a large-scale increase in effort during 2006 and a decrease in fishing effort during subsequent years due to the phasing out of old craft and gears. The sudden increase in fishing efforts resulted in increased landings and *CPUE* during 2007-2012.

In general, the *CPUE* of frigate tuna fisheries showed a declining trend during 1998-2018 (Table 1). The exploitation of the stock close to the *MSY* may be the reason for the reduction in *CPUE* since 2012. Abdussamad *et al*. (2012) reported that frigate tuna in Tamil Nadu waters was very intensively exploited, and production reached very close to the estimated potential. The results of the present study are concurrent with Abdussamad *et al*. (2012) and Sivadas *et al*. (2019). Kirkwood (2001) opined that when fishing and natural

mortality increases the population size decline gradually. Although there were random variation on expected and observed *CPUE*, a specific decreasing trend of CPUE was lacking in observed *CPUE* (Figure 2). Hence it can be assumed that changes in fishing and natural mortality of frigate tuna fisheries doesn't reflected as a decline in population size.

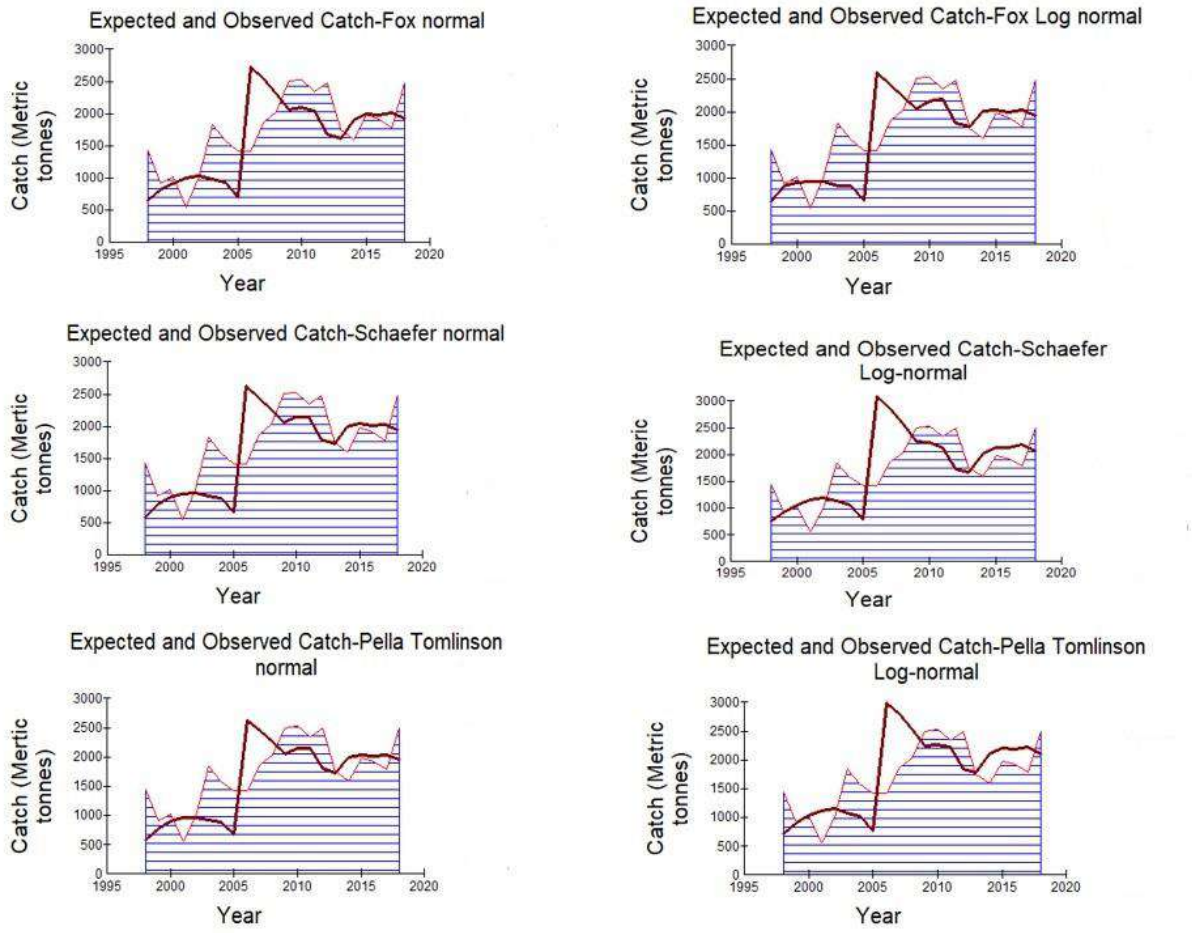


Figure 3: Time series of expected and observed catch of frigate tuna fishery in Tamil Nadu from 1998 to 2018

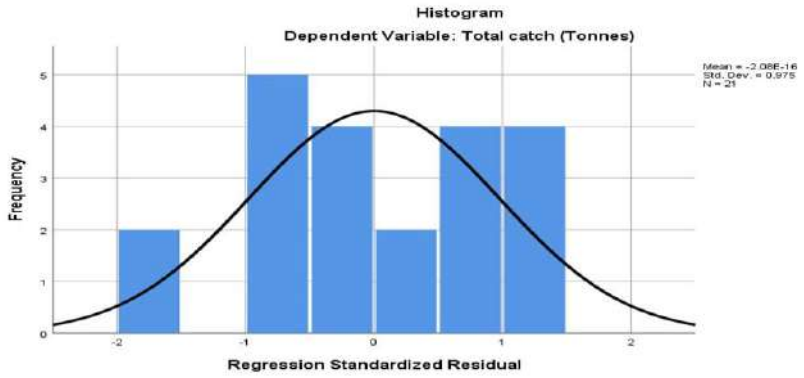


Figure 4: Histogram of regression analysis over standardized residual for frigate tuna fishery from 1998 to 2018

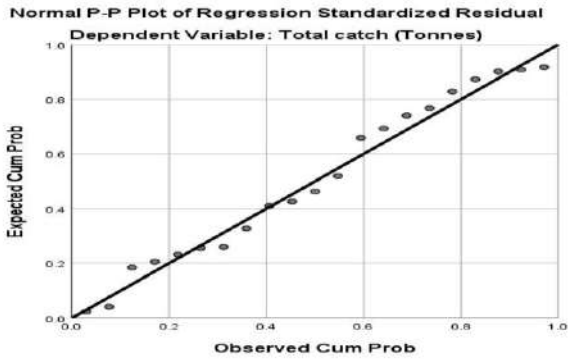


Figure 5: Normal P-P plot of regression standardized residual for frigate tuna fishery from 1998 to 2018

Table 5: Regression output of catch and effort data analysis for frigate tuna fishery from 1998 to 2018

Multiple *R*	R^2	Adjusted R^2	Standard Error	*F*-statistic	*p*-value
0.69447	0.48229	0.45504	35.9723	17.7	0.0004775

Results of the *MSY* estimates indicated that *MSY* values are inversely proportional to IP values (Table 2). Earlier workers reported a similar relationship (Kalhoro *et al*., 2013; Mohsin *et al*., 2018, 2019, 2020, 2021; Talib *et al*., 2017; Abinaya and Sajeevan, 2022a). The estimated BRPs of the Fox, Schaefer and Pella Tomlinson models varied from each other (Table 3). Based on diagnostic plot results, high R^2 and low *RMSE* value, the Fox log-normal model was considered the best-suited model and it made better fits and yielded results near the annual average landings. Hoggarth *et al*. (2006) and Noman *et al*. (2019) recommended that a high R^2 value and strong trend diagnostic plot were considered as a criterion for selecting the best-fit model and Panhwar (2012) suggested that the best-fit model will give results that are close to the annual average landing.

Target reference points (TRPs) and limit reference points (LRPs) are the two categories of reference points in general. TRPs are employed in fisheries management to set desirable fishing limits. *MSY*, F_{MSY}, and B_{MSY} are the three BRPs that have been widely employed in fishery resource management, with *MSY* receiving the most attention (Mohsin *et al*., 2020; Abinaya and Sajeevan, 2022b). Surplus production models are commonly employed in tropical fish stock assessment since they do not estimate cohorts and thus do not necessitate age determination. It can be calculated by using a stock assessment model that incorporates catch and effort statistics and predicts biomass. When the appropriate surplus production model is applied to all species collected by all types of fleets, an immediate *MSY* evaluation for the area is obtained. On the other hand, the challenge of harvesting the same stock by gear of varying effectiveness must be solved by regulating the fishing efforts of all gear active in the fishing (Kuriakose and Kizhakkudan, 2017).

The estimated *MSY* values are compared to the data values. The stock population thrives when the catch quantity is less than the calculated *MSY* value and is much more exploited. Once the stock achieves the *MSY* value, it is stable, and the harvest should be retained at the calculated *MSY* level rather than expanded or diminished. The stock population declines when the catch amount exceeds the actual *MSY* value. The estimated *MSY* of the frigate tuna fishery from Tamil Nadu was 2,543 MT, almost close to the recent catch of 2,482 MT during the 2018 period. BRPs (*MSY*, B_{MSY} and F_{MSY}) estimated by Fox log-normal indicate that the frigate tuna fishery of Tamil Nadu does not come under the status of overfishing and overfished. Estimates of *F* value were less than F_{MSY} and the F/F_{MSY} ratio was on the lower side. This indicates no overfishing sign of frigate tuna in Tamil Nadu waters. Similarly, the B/B_{MSY} value was higher than 0.5, indicating that the stock was not overfished. *MSY* estimates and landing data since 1998, confirm that the average annual landing never exceeded the *MSY* estimates. Moreover, *MSY* estimated by other models also stood above the annual average landing (1,732 MT year^{-1}) during the study period. Similarly, the F_{MSY} estimates of all models were higher than that of the *F* value estimated by the present study.

Coastal tuna stocks in Indian waters were being exploited at near-optimal levels (Silas and Pillai, 1985; James *et al*., 1992, 1993; James and Pillai, 1993; Kasim and Abdussamad, 2005; Pillai *et al*., 2005; Pillai and Ganga, 2008). Abdussamad *et al*. (2005) reported that the frigate tuna stock of Tamil Nadu was underexploited in 2005 and was intensely exploited in 2010 (Abdussamad *et al*., 2012). Ghosh *et al*. (2012) and Mudumala *et al*. (2018) reported that frigate tuna stock occurring on the Northwest coast of India showed signs of overexploitation. Dan (2021) reported that Indian Ocean frigate tuna stock is very close to being fished at *MSY* levels and higher catches may not be sustained. The results of the present study overrule the status of overfishing and overfished stock of frigate tuna. However, the reduction trend of *CPUE* from 2012 against a nominal decrease in the fishing effort is an indication that frigate tuna stock in Tamil Nadu reached the level of optimal exploitation. Any increase in fishing effort and overcapitalization may exert fishing pressure on the stock and lead to overfishing. Therefore, it is suggested that the present level of fishing may be maintained without any replacement for phasing out craft for ensuring sustainable exploitation.

5. Conclusion

The total landing of frigate tuna showed an increasing catch trend from 1998 to 2018. The total effort of frigate tuna registered a large scale increase during 2005 as a post-Tsunami effect and showed a decreasing trend since 2006 due to the phasing out of old craft and gears. The biological reference points (*MSY*, B_{MSY} and F_{MSY}) of frigate tuna rule out designating the frigate tuna stock of Tamil Nadu status as overfishing and overfished. However, an overall reduction trend of *CPUE* since 2012 indicates that stock is exploited very close to *MSY*. Hence, any increase in fishing effort results in heavy fishing pressure on fish stock and may lead to overfishing.

Acknowledgment

We thank Professor (Dr.) Riji John, Hon'ble Vice-Chancellor and Professor (Dr.) Rosalind George, Dean, Kerala University of Fisheries and Ocean Studies, Panangad, for providing all necessary facilities to complete the research.

References

Abinaya, R. and Sajeevan, M. K. (2022a). Fishery appraisal of narrow-barred spanish mackerel *Scomberomorus Commerson* (Lacepède 1800) using surplus production models from tamil nadu, india waters. *Thalassas: An International Journal of Marine Sciences* (online), 1-12. https://doi.org/10.1007/s41208-022-00492-8.

Abinaya, R. and Sajeevan, M. K. (2022b). Stock assessment of sin croaker *Johnius dussumieri* (Cuvier, 1830) fishery by different production model approach from tamil nadu, southeast coast of India. *Turkish Journal of Fisheries and Aquatic Sciences*, **23**, TRJFAS22465. https://doi.org/10.4194/TRJFAS22465.

Abdussamad, E. M., Pillai, P. P., Kasim, M. H., and Balasubramanian, T. S. (2005). Fishery and population characteristics of coastal tunas at Tuticorin. *Journal of the Marine Biological Association of India*, **47**, 50-57.

Abdussamad, E. M., Pillai, N. G. K., and Balasubramanian, T. S. (2008). Population characteristics and fishery of yellow fin tuna, *Thunnus Albacares* landed along the Gulf of Mannar coast, Tamil Nadu, India. *Egyptian Journal of Aquatic Research*, **34**, 330-335.

Abdussamad, E. M., Rao, G .S., Koya, K. P. S., Rohit, P., Joshi, K. K., Sivadas, M., Kuriakose, S., Ghosh S., Jasmine S., Chellappan A., and Koya M. (2012). Indian tuna fishery - production trend during yesteryears and scope for the future. *Indian Journal of Fisheries*, **59**, 1-13.

Ansell, M. (2020). *Marine Fisheries Catches for Mainland India from 1950-2018*. Doctoral dissertation, The University of Western Australia, Australia, 54 pages.

Bhathal, B. (2014). *Government-led Development of India's Mmarine Fisheries Since 1950: Catch and Effort Trends, and Bioeconomic Models for Exploring Alternative Policies*. Doctoral dissertation, University of British, Columbia, 355 pages.

CMFRI (2006). *Marine Fisheries Census* 2005 *Part III* (4). Central Marine Fisheries Research Institute, Kochi, Kerala, 408 pages.

CMFRI (2010). *Annual Report* 2009-2010. Central Marine Fisheries Research Institute, Kochi, Kerala, 173 pages.

CMFRI (2011). *Annual Report* 2010-2011. Central Marine Fisheries Research Institute, Kochi, Kerala, 166 pages.

CMFRI (2012a). *Annual Report* 2011-2012. Central Marine Fisheries Research Institute, Kochi, Kerala, 190 pages.

CMFRI (2012b). *Marine Fisheries Census* 2010 *Part II* (4), Tamil Nadu. Central Marine Fisheries Research Institute, Kochi, Kerala, 110 pages.

CMFRI (2013). *Annual Report* 2012-2013. Central Marine Fisheries Research Institute, Kochi, Kerala, 204 pages.

CMFRI (2014). *Annual Report* 2013-2014. Central Marine Fisheries Research Institute, Kochi, Kerala, 353 pages.

CMFRI (2015). *Annual Report* 2014-2015. Central Marine Fisheries Research Institute, Kochi, Kerala, 291 pages.

CMFRI (2016). *Annual Report* 2015-2016. Central Marine Fisheries Research Institute, Kochi, Kerala, 296 pages.

CMFRI (2017). *Annual Report* 2016-2017. Central Marine Fisheries Research Institute, Kochi, Kerala, 291 pages.

CMFRI (2018). *Annual Report* 2017-2018. Central Marine Fisheries Research Institute, Kochi, Kerala, 304 pages.

CMFRI (2019). *Annual Report* 2019. Central Marine Fisheries Research Institute, Kochi, Kerala, 368 pages.

DADF (2009). *Handbook on Fisheries Statistics* 2008. Department of Animal Husbandry, Dairying & Fisheries, Ministry of Agriculture, New Delhi, 91 pages.

DADF (2012). *Handbook on Fisheries Statistics* 2011. Department of Animal Husbandry, Dairying & Fisheries, Ministry of Agriculture, New Delhi, 38 pages.

DADF (2015). *Handbook on Fisheries Statistics* 2015. Department of Animal Husbandry, Dairying & Fisheries, Ministry of Agriculture, New Delhi, 178 pages.

DADF (2018). *Handbook on Fisheries Statistics* 2018. Department of Animal Husbandry, Dairying & Fisheries, Ministry of Agriculture, New Delhi, 190 pages.

Dan (2021). *Assessment of Indian Ocean Frigate Tuna (Auxis thazard) Using Data-Limited Methods*. Food and Agriculture Organization of the United Nations, Rome, Italy, 14 pages.

Devaraj, M. and Vivekanandan, E. (1999). Marine capture fisheries of India: Challenges and opportunities. *Current Science*, **76**, 314-332.

FAO (1994). *Sustainable Development and the Environment: FAO Policies and Actions*. Stockholm 1972 - Rio 1992. Food and Agriculture Organization of the United Nations Rome, 1992, Reprinted, 1994.

Fox, W. W. (1970). An exponential yield model for optimizing exploited fish populations. *Transactions of the American Fisheries Society*, **99**, 80-88. https://doi.org/10.1577/1548-8659(1970)99<80:AESMFO>2.0.CO;2.

George, G. and Gopalakrishnan, A. (2013). *Status of Marine Fisheries Research in India – Capture Trends, Coastal Vulnerability Issues and Sustainable Production Plans*. In: Pearl - Platinum Jubilee Souvenir of the Department of Aquatic Biology and Fisheries, University of Kerala, Thiruvananthapuram, 33-38.

Ghosh, S., Sivadas, M., Abdussamad, E. M., Rohit, P., Koya, K. P., Joshi, K. K., Chellappan, A., Margaret MuthuRathinam, A., Prakasan, D., and Sebastine, M. (2012). Fishery, population dynamics and stock structure of frigate tuna *Auxis Thazard* (Lacepede, 1800) exploited from Indian waters. *Indian Journal of Geo-Marine Sciences*, **59**, 95-100.

GOT (2004). *Tamil Nadu Policy Note* 2003-04. Animal Husbandry and Fisheries Department, Government of Tamil Nadu, India, 29 pages.

GOT (2005). *Tamil Nadu Policy Note* 2004-05. Animal Husbandry and Fisheries Department, Government of Tamil Nadu, India, 15 pages.

GOT (2006). *Tamil Nadu Policy Note* 2005-06. Animal Husbandry and Fisheries Department, Government of Tamil Nadu. India, 25 pages.

GOT (2010). *Tamil Nadu Policy Note* 2009-10. Animal Husbandry and Fisheries Department, Government of Tamil Nadu, India, 56 pages.

GOT (2011). *Tamil Nadu Policy Note* 2010-11. Animal Husbandry and Fisheries Department, Government of Tamil Nadu, India, 37 pages.

GOT (2012). *Tamil Nadu Policy Note* 2011-12. Animal Husbandry and Fisheries Department, Government of Tamil Nadu, India, 34 pages.

GOT (2013). *Tamil Nadu Policy Note* 2012-13. Animal Husbandry and Fisheries Department, Government of Tamil Nadu, India, 31 pages.

GOT (2014). *Tamil Nadu Policy Note* 2013-14. Animal Husbandry and Fisheries Department, Government of Tamil Nadu, India, 80 pages.

GOT (2015). *Tamil Nadu Policy Note* 2014-15. Animal Husbandry and Fisheries Department, Government of Tamil Nadu, India, 173 pages.

GOT (2016). *Tamil Nadu Policy Note* 2015-16. Animal Husbandry and Fisheries Department, Government of Tamil Nadu, India, 183 pages.

GOT (2017). *Tamil Nadu Policy Note* 2016-17. Animal Husbandry and Fisheries Department, Government of Tamil Nadu, India, 177 pages.

GOT (2018). *Tamil Nadu Policy Note* 2017-18. Animal Husbandry and Fisheries Department, Government of Tamil Nadu, India, 185 pages.

GOT (2019). *Tamil Nadu Policy Note* 2018-19. Animal Husbandry and Fisheries Department, Government of Tamil Nadu, India, 131 pages.

GOT (2020). *Tamil Nadu Policy Note* 2019-20. Animal Husbandry and Fisheries Department, Government of Tamil Nadu, India, 138 pages.

Hanchet, S. M., Blackwell, R. G., and Dunn, A. (2005). Development and evaluation of catch per unit effort indices for southern blue whiting (*Micromesistius Australis*) on the Campbell Island Rise, New Zealand. *ICES Journal of Marine Science*, **62**, 1131-1138. https://doi.org/10.1016/j.icesjms.2005.04.011.

Hoggarrth, D. D., Abeyasekera, S., Arthur, R. I., Beddington, J. R., Burn, R. W., Halls, A. S, Kirkwood, G. P., McAllister, M., Medley, P., Mees, C. C., Parkes, G. B., Pilling, G. M., Wakeford, R. C., and Welcomme, R. L. (2006). *Stock Assessment for Fishery Management-a Framework Guide to the Stock Assessment Tools of the Fisheries Management Science Programme.* FAO Fisheries Technical Paper No. 487, Food and Agriculture Organization of the United Nations, Rome, Italy, 261 pages.

James, P. S. B. R. and Pillai P. P. (1993). *Tuna Resources and Fishery in the Indian EEZ – An Update*. Proceedings of National Tuna Conference, Central Marine Fisheries Research Institute, Kochi, Kerala, 19-43.

James, P. S. B. R., Alagarswami, K., Rao, K. N., Muthu, M. S., Rajagopalan, M. S., Alagaraja, K., and Mukundan, C. (1987). *Potential Marine Fishery Resources of India*. CMFRI Special Publication No. 30, Central Marine Fisheries Research Institute, Cochin, India, 44-74.

James, P. S. B. R., Pillai, P. P., Jayaprakash, A. A., Yohannan, T. M., PonSiraimeetan, Muthiah, C. Gopakumar, G., Pillai, N. G. K., Remban, S., Thiagarajan, R., Said Koya, K. P., Kulkarni, G. M., Somaraju, M. V., Kurup, K. N., and Sathianandan, T. V. (1992). Stock assessment of tunas from Indian Seas. *Indian Journal of Fisheries*, **39**, 260-277.

James, P. S. B. R., Pillai, P. P., Pillai, N. G. K., Jayaprakash, A. A., Gopakumar, G., Kasim, H. M., Sivadas, M., and Said Koya, K. P. (1993). *Fishery, Biology and Stock Assessment of Small Tunas*. In: Sudarsan, D. and John, M. E. (Eds.), Tuna Research in India, FSI, Mumbai, India, 123-148.

Joseph, M. M. and Jayaprakash, A. A. (2003). *Status of Exploited Marine Fishery Resources of India*. Central Marine Fisheries Research Institute, Cochin, 1-34.

Kalhoro, M. A., Liu, Q., Memon, K. H., Chang, M. S., and Jatt, A. N. (2013). Estimation of maximum sustainable yield of Bombay duck, *Harpodon Nehereus* fishery in Pakistan using the CEDA and ASPIC packages. *Pakistan Journal of Zoology,* **45**, 1757-1764.

Kasim, H. M. and Abdussamad, E. M. (2005). *Stock Assessment of Coastal Tunas Along the East Coast of India*. In: Somavanshi, V. S., Varghese, S. and Bhargava, A. K. (Eds.), Proceeding of Tuna Meet, 2003, 42-53.

Kasim, H. M. and Mohan, S. (2009). Tuna fishery and stock assessment of component species off Chennai coast. *Asian Fisheries Science*, **22**, 245-256. https://doi.org/10.33997/j.afs.2009.22.1.023.

Kasim, H. M. and Vivekanandan, V. (2011). *Marine Fish Production in Tamil Nadu & Puducherry*. Central Marine Fisheries Research Institute Data, Technical Report No. FIMSUL/ WP5:AR2. FAO/UTF/IND/180/IND, Central Marine Fisheries Research Institute, Cochin, India, 36 pages.

Kirkwood, G. (2001). *CEDA* [*Catch and Effort Data Analysis*] *and LFDA* [*Length Frequency Distribution Analysis*] *Enhancement. Enhancement and Support of Computer Aids for Fisheries Management*. MRAG Ltd, London, 69 pages. https://assets.publishing.service.gov.uk/media/57a08d59ed915d3cfd00199c/R5050CBa.pdf.

Kituyi, M. and Thomson, P. (2018). 90% of fish stocks are used up-fisheries subsidies must stop. Weforum. Org. https://unctad.org/news/90-fish-stocks-are-used-fisheries-subsidies-must-stop.

Kumar, R., Sundaramoorthy, B., Neethiselvan, N., and Athithan, S. (2018). Tuna fishery along Thoothukudi coast, Tamil Nadu. *Journal of Experimental Zoology*, **21**, 281-285.

Kumar, R., Sundaramoorthy, B., Neethiselvan, N., Athithan, S., Kumar, R., Rahangdale, S., and Sakthivel, M. (2019). Length based population characteristics and fishery of skipjack tuna, *Katsuwonus Pelamis* (Linnaeus, 1758) from Tuticorin waters, Tamil Nadu, India. *Indian Journal of Geo-Marine Sciences*, **48**, 52-59.

Kuriakose, S. and Kizhakkudan, S. J. (2017). *Macro Analytical Models*. In: Course Manual Summer School on Advanced Methods for Fish Stock Assessment and Fisheries Management. Central Marine Fisheries Research Institute, Kochi, 246-251.

Lecomte, M., Rochette, J., Laurans, Y., and Lapeyre, R. (2017). Indian ocean tuna fisheries: between development opportunities and sustainability issues. IDDRI Development Durable and Relations Internationales. https://www.iddri.org/sites/default/files/PDF/Publications/Hors%20catalogue%20Iddri/201811-tuna-indian %20oceanEN.pdf.

Lindawati, L., Mardyani, Y., Yanti, N. D., and Boer, M. (2021). Assessing and managing demersal fisheries in Sunda Strait: Bio-economic modelling. In *IOP Conference Series*: *Earth and Environmental Science*, **860**, 012 - 062. doi:10.1088/1755-1315/860/1/012062.

Malhotra, S. P. and Sinha, V. R. P. (2007). *Indian Fisheries and Aquaculture in a Globalizing Economy Part II.* Narendra Publishing House, Delhi. 85 pages.

MRAG (2016). *Marine Resources Assessment Group. CEDA Version* 3. 0. Available: http://www.mrag.co.uk/resources/fisheries-assessment-software.

MOA (2001). *Report of the Working Group for Revalidating the Potential of Fishery Resources in the Indian EEZ 2000*. Department of Animal Husbandry, Dairying and Fisheries, Ministry of Agriculture, New Delhi, 68 pages.

Mohsin, M., Guilin, D., Zhuo, C., Hengbin, Y., and Noman, M. (2019). Maximum Sustainable Yield Estimates of Carangoides Fishery Resource in Pakistan and its Bioeconomic Implications. *Pakistan Journal of Zoology*, **51**, 279-287. http://dx.doi.org/10.17582/journal.pjz/2019.51.1.279.287.

Mohsin, M., Hengbin, Y., and Luyao, Z. (2021). Application of non-equilibrium SPMs to access overexploitation risk faced by *Scomberomorus Sinensis* in Shandong, China. *Pakistan Journal of Zoology*, **53**, 1-8. https://dx.doi.org/10.17582/journal.pjz/20190901040914.

Mohsin, M., Hengbin, Y., and Nisar, U. (2020). Accessing the risk of overfishing faced by mullet fisheries and its ongoing economics in Pakistan. *Indian Journal of Geo-Marine Sciences*, **49**, 1416-1424.

Mohsin, M., Mu, Y. T., Noman, M., Hengbin, Y., and Mehak, A. (2018). Estimation of maximum sustainable harvest levels and bioeconomic implications of *Babylonia spirata* fisheries in Pakistan by using CEDA and ASPIC. *Oceanography and Fisheries*, **7**, 1-8. http://dx.doi.org/10.19080/OFOAJ.2018.07.555715.

Mudumala, V. K., Farejiya, M. K., Mali, K. S., Karri, R. R., Uikey, D. E., Sawant, P. A., and Siva, A. (2018). Studies on population characteristics of frigate tuna, *Auxis Thazard* (Lacepede, 1800) occurring in the north west coast of India. *International Journal of Life-Sciences Scientific Research,* **4**, 1639-1643. https://doi.org/10.21276/ijlssr.2018.4.2.3.

Noman, M., Mu, Y. T., Mohsin, M., Memon, A. M., and Kalhoro, M. T. (2019). Maximum sustainable yield estimates of *Scomberomorus Spp*. from Balochistan, Pakistan. *Pakistan Journal of Zoology*, **51**, 2199-2207. http://dx.doi.org/10.17582/journal.pjz/2019.51.6.2199.2207.

Panhwar, S. K., Liu, Q., Khan, F., and Siddiqui, P. J. (2012). Maximum sustainable yield estimates of Ladypees, *Sillago Sihama* (Forsskål), fishery in Pakistan using the ASPIC and CEDA packages. *Journal of Ocean University of China,* **11**, 93-98. https://doi.org/10.1007/s11802-012-1880-3.

Pella, J. J. and Tomlinson, P. K. (1969). A generalized stock production model. *Inter American Tropical Tuna Commission Bulletin*, **13**, 416-497.

Pillai, N. G. K. and Ganga, U. (2008). *Fishery and Biology of Tunas in the Indian Seas.* In: Joseph, J., Boopendranath, M. R., Sankar, T. V., Jeeva, J. C. and Kumar, R. (Eds.), Harvest and post-harvest technology for tuna. Society of Fisheries Technologists (India), Cochin, 10-35.

Pillai, N. G. K., Ganga, U., Gopakumar, G., Muthiah, C., and Somy Kuriakose. (2005). *Stock Assessment of Coastal Tunas Along the West Coast of India*. In: Somavanshi, V. S., Varghese, S. and Bhargava, A. K. (Eds.), Proceedings of Tuna Meet, 2003, 54-57.

Satyanarayana, K. V., Reddy, M. N., Balasubramani, N., Pillai, N. G. K., and Ganga, U. (2008). Sustainable marine fisheries development. National Institute of Agricultural Extension Management. http://eprints.cmfri.org.in/3741/1/ganga_study_mat%5B1%5D.pdf.

Schaefer, M. B. (1954). Some aspects of the dynamics of populations important to the management of the commercial marine fisheries. *Inter-American Tropical Tuna Commission Bulletin*, **1**, 23-56. https://doi.org/10.1016/S0092-8240(05)80049-7.

Silas, E. G. and Pillai, P.P. (1985). Indian tuna fishery development perspectives and management plan. CMFRI Bulletin No. 36, Central Marine Fisheries Research Institute, Cochin,193-208.

Silas, E. G., Pillai, P. P., Srinath, M., Jayaprakash, A. A., Muthiah, C., Balan, V., Yohannan, T. M., Siraimeetan, P., Mohan, M., Livingston, P., and Kunhikoya, K. K. (1985). *Population Dynamics of Tunas: Stock Assessment*. CMFRI Bulletin No. 36, Central Marine Fisheries Research Institute, Cochin, 20-27.

Sin, M. S. and Yew, T. S. (2016). Assessing the exploitation status of marine fisheries resources for the west coast of peninsular Malaysia trawl fishery. *World Journal of Fish and Marine Sciences*, **8**, 98-107. http://dx.doi.org/10.5829/idosi.wjfms.2016.8.2.102149.

Sivadas, M., Abdussamad, E. M., Margaret MuthuRathinam, A., Mohan, S., Vasu, P., and Laxmilatha, P. (2019). Tuna drift gillnet fishery at Chennai, Tamil Nadu-an update. *Journal of the Marine Biological Association of India*, **61**, 41-46. http://dx.doi.org/10.6024/jmbai.2019.61.2.2066-06.

Sivadas, M., Margaret Muthu Rathinam, A., Vinothkumar, R., Mini, K. G., and Abdussamad, E. M. (2020). Status and prospects of large pelagics fishery in Tamil Nadu and Puducherry. *Marine Fisheries Information Service, Technical and Extension Series*, **245**, 7-12.

Sparre, P. and Venema, S. C. (1998). *Introduction to Fish Stock Assessment. Part* 1*: Manual*. FAO Fisheries Technical Paper No. 306, Food and Agriculture Organization of the United Nations, Rome, Italy, 1-407.

Tabitha, S. N. and Gunalan, B. (2012). A report on mass landings of economically important fish along the south east coast of India. *Advances in Applied Science Research*, **3**, 3855-3859.

Talib, K. M., Yongtong, M., Hussain, S.S., Ali, K. M., Mahmood, M. A., Muhammad, M., and Ramesh, P. T. (2017). Maximum sustainable yield and economic importance of *Rachycentron Canadum* (Linnaeus, 1766) in Pakistani waters. *Pakistan Journal of Agriculture Science*, **54**, 873-80. http://dx.doi.org/10.21162/PAKJAS/17.5855.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 21, No. 2, 2023 (New Series), pp 193–210
http://www.ssca.org.in/journal

The gLinear Failure Rate Distribution: A New Mixture with Bayesian and Non-Bayesian Analysis

R. M. Mandouh[1]
[1]*Department of Mathematical Statistics*
Faculty of Graduate Studies for Statistical Research, Cairo University, Egypt

Received: 02 September 2022; Revised: 27 November 2022; Accepted: 13 March 2023

Abstract

A mixture of the gamma and linear failure rate distributions is constructed and studied. The model parameters are estimated using maximum likelihood and Bayesian based on real and simulated data.

Key words: Mixture Distribution; Maximum Likelihood Estimation; Laplace Approximation; MCMC; Variational Bayes.

AMS Subject Classifications: 62K05, 05B05

1. Introduction

In statistical analysis, the survival function and hazard function are used to model distribution of data representing lifetime or waiting time. The survival function or reliability function is the probability of survival of an item without failing until time t. Alternatively, we can describe the survival experience in term of hazard failure (instantaneous rate of death) which is the chance of death (failure) as a function of age. The hazard function or the instantaneous failure rate has many types which appeared in practice such as unimodal shaped; bathtub shaped and others. The main aim of this paper is to introduce a new distribution with two parameters. The hazard function of this distribution can be constant, unimodal (upside-down bathtub) or increasing-decreasing-increasing depending on the values of its two parameters. The shapes of the hazard function of the new distribution enables it to be a good model to fit various data sets.

The mixture distribution (Everitt (2013)) is one of the means can be utilized to construct these new distributions. The finite mixture is formed as follow:

$$f(x) = \sum_{i=1}^{c} p_i f_i(x)$$

where $\sum_{i=1}^{c} p_i = 1$ with $c = 2$ in our distribution. Many Papers dealing with two mixture models such as, Lindley (1958) introduced a one parameter distribution, now known as the

Corresponding Author: R. M. Mandouh
Email: rshmndoh@gmail.com, rshmndoh@cu.edu.eg

Lindley distribution. Ghitany *et al.* (2008) studied its properties in details. Shanker and Mishra (2013) added one extra parameter to Lindley distribution and introduced the quasi Lindley distribution. They studied some of its properties. Sen *et al.* (2016) proposed and studied another finite mixture distribution which is called the xgamma distribution.Sen and Chandra (2017) added one extra parameter to the xgamma distribution and introduced the quasi xgamma distribution. Moreover, many Papers dealing with three mixture models such as, Sarhan *et al.* (2014) introduced two lifetime distributions. They referred to these two distributions as $N(\beta)$ and $TN(\alpha, \beta)$ respectively and they discussed some properties of these two distribution such as the behavior of their hazard functions. Mahmoud *et al.* (2017) introduced two distributions based on mixing between different types of distributions.

2. The gLinear failure rate distribution

Now, we introduce a mixture density of two mixture components, one follows gamma (2, β) and the other follows linear failure rate (β, β^2) with mixing weights $\frac{\beta}{\alpha+\beta}$ and $\frac{\alpha}{\alpha+\beta}$. The pdf of the new mixture distribution will be as follows:

$$f(x) = \frac{\beta}{\alpha+\beta}(\beta^2 x + \alpha(1+\beta x)e^{-\frac{\beta^2}{2}x^2})e^{-\beta x}, \quad x > 0,\ \beta, \alpha > 0. \tag{1}$$

We refer to this distribution as glfr (α, β). For $\alpha = 1$, we have the following new distribution as a special case

$$f(x) = \frac{\beta}{1+\beta}(\beta^2 x + (1+\beta x)e^{-\frac{\beta^2}{2}x^2})e^{-\beta x}, \quad x > 0,\ \beta > 0, \tag{2}$$

which is a mixture of gamma (2, β) and the other follows linear failure rate (β, β^2) with mixing weights $\frac{\beta}{1+\beta}$ and $\frac{1}{1+\beta}$ and we refer to this distribution as glfr (β). Figure (1) shows pdf of the glfr distribution for different parameter values. The corresponding cdf of (2.1) takes the following form

$$F(x) = \frac{1}{\alpha+\beta}(\beta + \alpha - e^{-\beta x}(\beta(1+\beta x) + \alpha e^{-\frac{\beta^2}{2}x^2})), \quad x > 0,\ \beta, \alpha > 0. \tag{3}$$

Then the survival function is given by

$$S(x) = \frac{1}{\alpha+\beta}(e^{-\beta x}(\beta(1+\beta x) + \alpha e^{-\frac{\beta^2}{2}x^2})), \quad x > 0,\ \beta, \alpha > 0, \tag{4}$$

and the hazard function is given by

$$h(x) = \frac{\beta(\beta^2 x + \alpha(1+\beta x)e^{-\frac{\beta^2}{2}x^2})}{(\beta(1+\beta x) + \alpha e^{-\frac{\beta^2}{2}x^2})}, \quad x > 0,\ \beta, \alpha > 0, \tag{5}$$

One can note that $h(x)$ is bounded, *i.e.* $\frac{\alpha\beta}{\alpha+\beta} < h(x) < \beta$. The hazard function of glfr distribution is plotted in Figure (2) for four different pairs of choices of α and β.

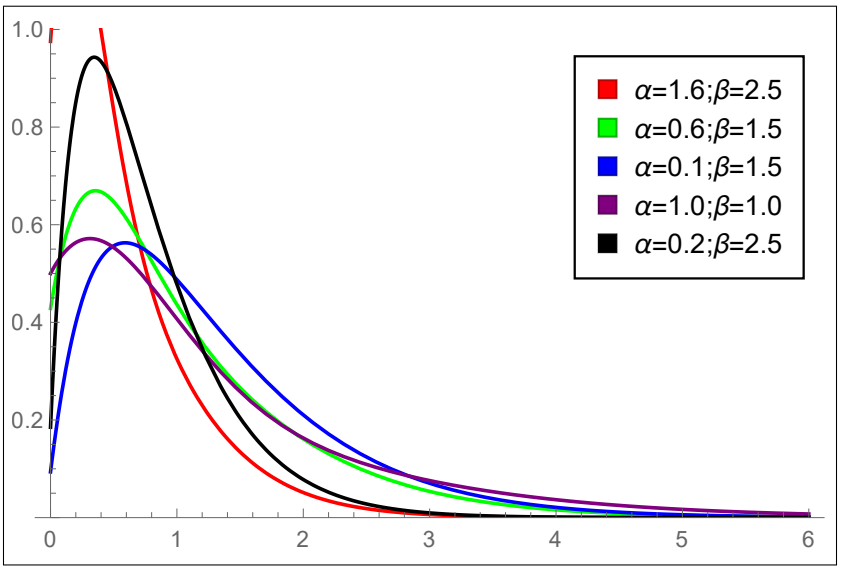


Figure 1: The gLinear failure rate pdfs for some parameter values

The moments and shape measures

Let X follow gLinear failure rate distribution. After some algebra, the rth moment of X is derived as

$$E(X^r) = \frac{\Gamma(r+2)}{\beta^{r-1}(\alpha+\beta)} + \frac{2\alpha\sqrt{e}}{\beta^r(\alpha+\beta)} \int_{1/\sqrt{2}}^{\infty} t(\sqrt{2}t-1)^r e^{-t^2} dt \tag{6}$$

Therefore, the expectation variance of the two parameter glfr distribution in terms of the error function (erf) and its complementary (erfc) are given by

$$E(X) = \frac{4\beta + \alpha\sqrt{2e\pi}(erfc(\frac{1}{\sqrt{2}}))}{2\beta(\alpha+\beta)},$$

and

$$Var(X) = \frac{6\beta + 2\alpha - \alpha\sqrt{2e\pi}(erfc(\frac{1}{\sqrt{2}}))}{\beta^2(\alpha+\beta)} - (\frac{4\beta + \alpha\sqrt{2e\pi}(erfc(\frac{1}{\sqrt{2}}))}{2\beta(\alpha+\beta)})^2,$$

where, $erfc(z) = 1 - erf(z)$ and $erf(z) = \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} dt$.
Also, one can use eq.(6) and the relation between the moments and the central moments to obtain skewness and kurtosis.

The mean residual life

One of special relevance in reliability and survival analysis is the analysis of the lifetime of a device after it has attained age x. Thus, if X is the lifetime with survival function given by (4), the corresponding residual lifetime after age x is the random variable $X_x = (X - x|X > x)$ and the mean residual life of X is defined as $m(x) = E(X - x|X > x)$. It is also called the expected additional lifetime given that a component has survived until

time t is a function of t

$$\begin{aligned} m(x) &= E(X-x|X>x) \\ &= \frac{1}{S(x)}\int_x^\infty S(t)dt \\ &= \frac{\int_x^\infty e^{-\beta t}(\beta(1+\beta t)+\alpha e^{-\frac{\beta^2}{2}t^2})dt}{e^{-\beta x}(\beta(1+\beta x)+\alpha e^{-\frac{\beta^2}{2}x^2})} \\ &= \frac{\beta(2+\beta x)e^{-\beta x}+\sqrt{\frac{e\pi}{2}}erfc(\frac{1+\beta x}{\sqrt{2}})}{\beta(e^{-\beta x}(\beta(1+\beta x)+\alpha e^{-\frac{\beta^2}{2}x^2}))}, \end{aligned}$$

where, $erfc(z) = 1 - erf(z)$ and $erf(z) = \frac{2}{\sqrt{\pi}}\int_0^z e^{-t^2}dt$.

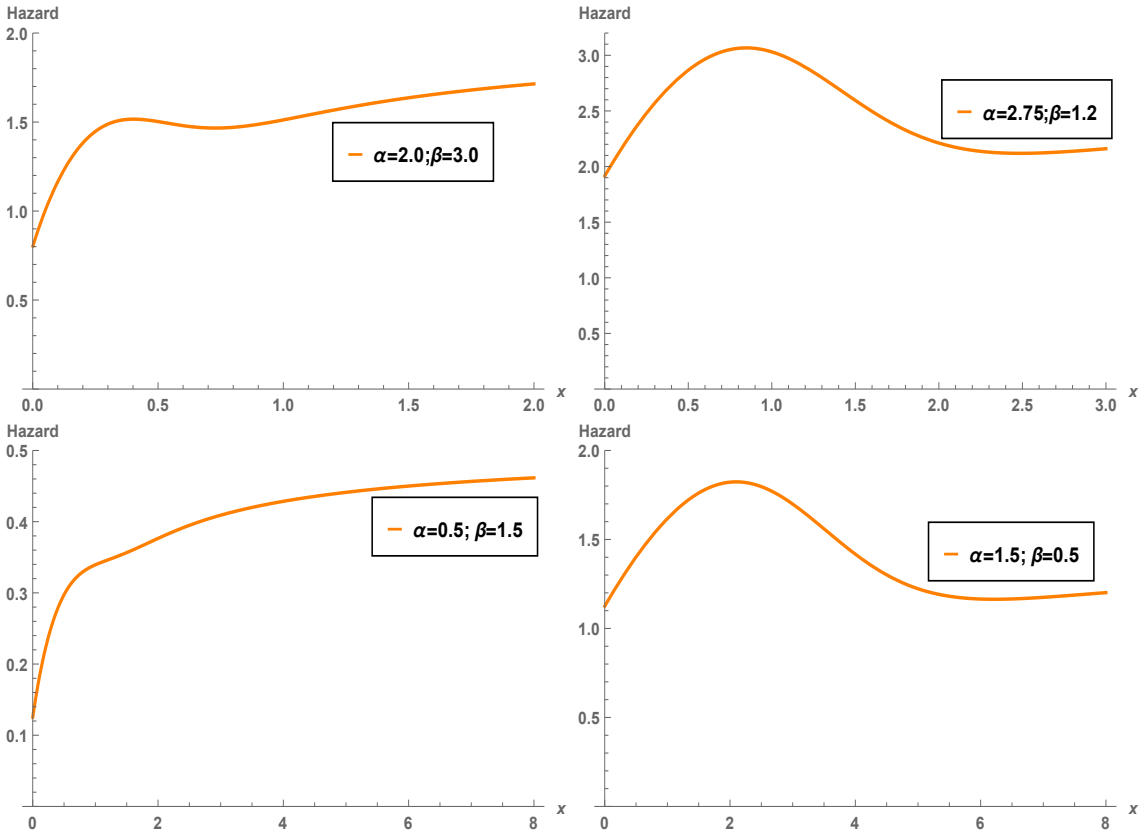


Figure 2: The hazard rate function of the gLinear failure rate for some parameter values

3. Maximum likelihood estimation (MLE)

For different statistical models, MLE is widely utilized to estimate the model parameters. Assume that n independent and identical items are put on a life test simultaneously. The lifetimes of these items are assume to have follow glinear failure rate distribution. Let $\mathbf{x} = (x_1, x_2, ..., x_n)$ be the failure times of the items. The Likelihood function for α, β is given

by

$$L(\alpha,\beta;\mathbf{x}) = \frac{\beta^n}{(\alpha+\beta)^n}\prod_{i=1}^{n}(\beta^2 x_i + \alpha(1+\beta x_i)e^{-\frac{\beta^2}{2}x_i^2})e^{-\beta x_i} \tag{7}$$

The log-likelihood function is

$$\mathcal{L} = \mathcal{L}(\alpha,\beta;\mathbf{x}) = nln\beta - nln(\alpha+\beta) - \beta\sum_{i=1}^{n}x_i + \sum_{i=1}^{n} ln\mathcal{A}_i(\alpha,\beta) \tag{8}$$

where $\mathcal{A}_i(\alpha,\beta) = \beta^2 x_i + \alpha(1+\beta x_i)e^{-\frac{\beta^2}{2}x_i^2}, \quad i = 1,2,...,n.$

Taking partial derivatives of the log-likelihood in (8) w.r.t. α and β, we have

$$\mathcal{L}_\alpha = -\frac{n}{\alpha+\beta} + \sum_{i=1}^{n}\frac{\mathcal{A}_{i,\alpha}(\alpha,\beta)}{\mathcal{A}_i(\alpha,\beta)} \tag{9}$$

$$\mathcal{L}_\beta = \frac{n}{\beta} - \frac{n}{\alpha+\beta} - \sum_{i=1}^{n}x_i + \sum_{i=1}^{n}\frac{\mathcal{A}_{i,\beta}(\alpha,\beta)}{\mathcal{A}_i(\alpha,\beta)} \tag{10}$$

where

$$\begin{aligned}
\mathcal{A}_{i,\alpha}(\alpha,\beta) &= \frac{\partial\mathcal{A}_i(\alpha,\beta)}{\partial\alpha} = (1+\beta x_i)e^{-\frac{\beta^2}{2}x_i^2},\\
\mathcal{A}_{i,\beta}(\alpha,\beta) &= \frac{\partial\mathcal{A}_i(\alpha,\beta)}{\partial\beta} = 2\beta x_i + \alpha x_i e^{-\frac{\beta^2}{2}x_i^2} - \alpha\beta(1+\beta x_i)x_i^2 e^{-\frac{\beta^2}{2}x_i^2}.
\end{aligned}$$

The second derivative of the log-likelihood are

$$\begin{aligned}
\mathcal{L}_{\alpha,\alpha} &= \frac{n}{(\alpha+\beta)^2} + \sum_{i=1}^{n}\frac{\mathcal{A}_i(\alpha,\beta)\mathcal{A}_{i,\alpha^2}(\alpha,\beta) - (\mathcal{A}_{i,\alpha}(\alpha,\beta))^2}{(\mathcal{A}_i(\alpha,\beta))^2}\\
\mathcal{L}_{\alpha,\beta} &= \frac{n}{(\alpha+\beta)^2} + \sum_{i=1}^{n}\frac{\mathcal{A}_i(\alpha,\beta)\mathcal{A}_{i,\alpha\beta}(\alpha,\beta) - (\mathcal{A}_{i,\alpha}(\alpha,\beta))(\mathcal{A}_{i,\beta}(\alpha,\beta))}{(\mathcal{A}_i(\alpha,\beta))^2}\\
\mathcal{L}_{\beta,\beta} &= -\frac{n}{\beta^2} + \frac{n}{(\alpha+\beta)^2} + \sum_{i=1}^{n}\frac{\mathcal{A}_i(\alpha,\beta)\mathcal{A}_{i,\beta^2}(\alpha,\beta) - (\mathcal{A}_{i,\beta}(\alpha,\beta))^2}{(\mathcal{A}_i(\alpha,\beta))^2}
\end{aligned} \tag{11}$$

where

$$\begin{aligned}
\mathcal{A}_{i,\alpha^2}(\alpha,\beta) &= 0,\\
\mathcal{A}_{i,\alpha\beta}(\alpha,\beta) &= x_i e^{-\frac{\beta^2}{2}x_i^2}(1-\beta x_i(1+\beta x_i)),\\
\mathcal{A}_{i,\beta^2}(\alpha,\beta) &= 2x_i + \alpha x_i e^{-\frac{\beta^2}{2}x_i^2}(-\beta x_i - 3\beta x_i^2 + \beta^2 x_i^3(1+\beta x_i)).
\end{aligned}$$

To calculate the information matrix, the expectation of the following matrix is required

$$\mathcal{T}(\alpha,\beta) = -\begin{bmatrix}\mathcal{L}_{\alpha,\alpha} & \mathcal{L}_{\alpha,\beta}\\ \mathcal{L}_{\alpha,\beta} & \mathcal{L}_{\beta,\beta}\end{bmatrix}$$

Equating the derivatives in (9) and (10) to zero and solving them numerically to obtain the mle of α and β, say $\hat{\alpha}$ and $\hat{\beta}$ such that $\mathcal{T}(\hat{\alpha},\hat{\beta})$ is positive definite.

For Interval estimation of (α, β), the mle of parameters α and β are asymptotically normally distributed with means equal the true values of α and β and variances given by the inverse of the observed information matrix,$\mathcal{T}(\hat{\alpha}, \hat{\beta})$, i.e.

$$\begin{bmatrix} \hat{\alpha} \\ \hat{\beta} \end{bmatrix} \sim N_2 \left[\begin{bmatrix} \alpha \\ \beta \end{bmatrix}, \hat{\mathcal{T}}^{-1} \right] \tag{12}$$

where $\hat{\mathcal{T}}^{-1}$ is the inverse of $\mathcal{T}(\hat{\alpha}, \hat{\beta})$. Using (12), large sample $(1-\nu)100\%$ confidence intervals for α and β are $\hat{\alpha} \pm z_{\nu/2}\sqrt{var(\hat{\alpha})}$, $\hat{\beta} \pm z_{\nu/2}\sqrt{var(\hat{\beta})}$, where $z_{\nu/2}$ is the upper $100\nu/2$ quantile of the standard normal distribution and $var(\hat{\alpha})$ and $var(\hat{\beta})$ are the main diagonal of $\hat{\mathcal{T}}^{-1}$.

4. Bayesian estimation

Let $x_1, x_2, ..., x_n$ be a random sample from glinear failure rate distribution. The likelihood of this sample is given by (7). Let the two parameters α and β are independent random variables with prior distributions gamma(a_1, b_1) and gamma(a_2, b_2), respectively. That is, the joint prior density of α and β is

$$g_0(\alpha, \beta) \propto \alpha^{a_1-1}\beta^{a_2-1}e^{-b_1\alpha+-b_2\beta}, \quad \alpha, \beta > 0 \tag{13}$$

where the hyperparameters a_i and $b_i, i = 1, 2.$ are assumed to be positive and known. Using the likelihood function (7) and the joint prior density function (13) and applying Bayes' theorem, we get the joint posterior density function of (α, β), given the data, as

$$g(\alpha, \beta|\boldsymbol{x}) \propto \frac{\alpha^{a_1-1}\beta^{a_2+n-1}}{(\alpha+\beta)^n}e^{-b_1\alpha+-b_2\beta}\prod_{i=1}^{n}(\beta^2 x_i + \alpha(1+\beta x_i)e^{-\frac{\beta^2}{2}x_i^2})e^{-\beta x_i}, \quad \alpha, \beta > 0 \tag{14}$$

Bayes estimators of the unknown parameters of any function of the unknown parameters, say $h(\boldsymbol{\theta})$, can be obtained as follows

$$E(h(\boldsymbol{\theta})|\boldsymbol{x}) = \frac{\int_0^\infty \int_0^\infty h(\boldsymbol{\theta}) g_0(\alpha, \beta) exp(\mathcal{L}) d\alpha d\beta}{\int_0^\infty \int_0^\infty g_0(\alpha, \beta) exp(\mathcal{L}) d\alpha d\beta}, \tag{15}$$

Formula (15) involves a ratio of two multidimentional integrals and does not have analytical solution. Thus, some approximation methods were suggested to approximate these integrals and calculate the ratio of the integrals such as the methods discussed by Lindley (1958) and Tierney and Kadane (1986) . These methods work well for low dimensions. In this paper we will use Tierney and Kadane's approximation method. They approximate (15) by using Laplace method as follow

$$E(h(\boldsymbol{\theta})|\boldsymbol{x}) = \left(\frac{\det\boldsymbol{\Sigma}^*}{\det\boldsymbol{\Sigma}}\right)^{1/2} exp(n(\mathcal{L}(\hat{\boldsymbol{\theta}}^*) - \mathcal{L}(\hat{\boldsymbol{\theta}}))) \tag{16}$$

where $n\mathcal{L}(\hat{\boldsymbol{\theta}}^*) = lnh + lng_0 + \mathcal{L}$, $n\mathcal{L}(\hat{\boldsymbol{\theta}}) = lng_0 + \mathcal{L}$ and $\boldsymbol{\Sigma}^*$ and $\boldsymbol{\Sigma}$ are minus the inverse Hessian of $\mathcal{L}(\hat{\boldsymbol{\theta}}^*)$ and $\mathcal{L}(\hat{\boldsymbol{\theta}})$ evaluated at $\boldsymbol{\theta}^*$ and $\boldsymbol{\theta}$, respectively. For more details about Laplace approximation see Crawford (1994) and Tierney *et al.* (1989).

For many applications, Bayesian inference is performed using Markov Chain Monte Carlo (MCMC), which estimates expectations w.r.t. $g(\boldsymbol{\theta}|\boldsymbol{x})$ by sampling from it. One of MCMC, Metropolis-Hastings (MH) algorithm, is proposed here. MH algorithm requires a proposal distribution and a common choice of it is the multivariate normal distribution. Metropolis-Hastings algorithm steps are

1. Specify the size of the random draws, say m.
2. Choose an initial value of $\boldsymbol{\theta}$, say $\boldsymbol{\theta}^{(0)}$.
3. For $i = 1, 2, \ldots, m$, repeat the following steps:
   (a) Set $\theta^{(i)} = \theta^{(i-1)}$.
   (b) Generate a candidate value θ^* from a proposal distribution $p(\theta^{(*)}|\theta^{(i)})$.
   (c) Compute the ratio $\kappa = min(1, \frac{g(\boldsymbol{\theta}^{(*)}|data)/p(\boldsymbol{\theta}^{(*)}|\boldsymbol{\theta}^{(i)}}{g(\boldsymbol{\theta}^{(i)}|data)/p(\boldsymbol{\theta}^{(i)}|\boldsymbol{\theta}^{(*)})})$.
   (d) Generate a random value u from uniform distribution on (0, 1).
   (e) Put $\boldsymbol{\theta}^{(i)} = \boldsymbol{\theta}^*, if \kappa \geq u,$ otherwise put $\theta^{(i)} = \theta^{(i-1)}$.
4. Return the values $\boldsymbol{\theta}^{(0)}, \boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(m)}$.

For more details about MH algorithm see Puza (2015). For other applications where θ is high dimensional or fast computation is of primary interest, variational Bayesian (VB) is an attractive alternative to MCMC. Yamaguchi *et al.* (2010) developed a VB approach for approximately computing posterior distributions of parameters of mixture of Erlang distribution and they investigated that computation speed of the VB becomes up to 200 times faster than that of MCMC. VB approximates the posterior distribution by a probability distribution with density $q(\boldsymbol{\theta})$ belonging to some tractable family of distributions Q such as Gaussians. The VB method treats an optimization to minimize the Kullback–Leibler (KL) divergence from an approximate posterior distribution to the exact posterior distribution, i.e. The best VB approximation $q^* \in Q$ is

$$q^* = \underset{q \in Q}{argmin} \left\{ KL(q||g(.,\boldsymbol{x})) := \int q(\boldsymbol{\theta}) log \frac{q(\boldsymbol{\theta})}{g(\boldsymbol{\theta}|\boldsymbol{x})} d\boldsymbol{\theta} \right\}.$$

5. Simulation study

A simulation study was carried out to investigate the performance of the accuracy of point and interval estimates of the two parameters of the glfr(α, β) distribution. The following steps are carried out:

1. Specify the values of the parameters α and β.
2. Specify the sample size n.
3. Generate a random sample $(x_1, x_2, x_3, ..., x_n)$ with size n from glfr(α, β) distribution using the following algorithm:

- Generate $U \sim$ uniform$(0, 1)$ with size n.
- Generate $V \sim$ gamma $(2, \beta)$ with size n.
- Generate $W \sim$ linear failure rate (β, β^2) with size n.
- If $u \leq \beta/(\alpha + \beta)$ set $x = u$, otherwise set $u = w$

4. Calculate the mle of the two parameters.

5. Repeat steps 2-4, N times.

6. Calculate the mean squared error (MSE), the average of the confidence interval widths, and the coverage probability for each parameter. The MSE associated with the MLE of the parameter θ, MSE_θ, is

$$\text{MSE}_\theta = \frac{1}{N}\sum_{i=1}^{N}(\hat{\theta} - \theta)^2,$$

where $\hat{\theta}$ is the MLE of θ. Coverage probability is the proportion of the N simulated confidence intervals which include the true parameter θ.

The simulation study is carried out using $N = 1000$. The sample sizes are 50, 75, 100, 150 and 200 and the selected parameter values are (α, β)= (0.8, 0.8), (0.8, 1.0), (1.0, 1.0), (1.0, 1.2), (1.2, 1.2) and (1.6, 2.5). Table 1 presents the MSE, coverage probability (CP_θ) and average width (AW_θ) of 95% confidence intervals of each parameter. This table shows that ,in the most cases, the MSEs and the average widths decrease as the sample size increases and the coverage probability are close to the nominal level of 95%.

6. Applications

In this section, to illustrate the applicability of the two new distributions proposed in this paper, we analyze three data sets. The first data set represents the remission times (in months) of a random sample of 128 bladder cancer patients. Bladder cancer is a disease in which abnormal cells multiply without control in the bladder. The most common type of bladder cancer recapitulates the normal histology of the urothelium and is known as transitional cell carcinoma.This data were studied by Zea *et al.* (2012).The second data represents a complete data with the exact times of failure. This data is considered a data set of the life of fatigue fracture of Kevlar 373/epoxy that are subject to constant pressure at the 90% stress level until all had failed. This data is considered by Ogunde *et al.* (2017). The three data set is provided in Murthy *et al.* (2004), page 278, about time between failures for repairable item.

We will refer to these data sets as data set 1, data set 2 and data set 3, respectively. For each data set, we fit the proposed distributions and other distributions such as the quasi xgamma (qxgamma), xgamma, quasi Lindley (qLinley), Lindley, linear failure rate (lfr) and gamma distributions. Goodness-of-fit tests are applied to verify which distribution better fits these data sets. The tests were carried out at 5% level of significance. We consider the common-known Kolmogorov-Smirnov (K-S) statistic, the Anderson-Darling (A-D), and Cramér-von Mises (C-M) statistics. Moreover, we consider some well-known measures such

as the Akaike information criterion (AIC), the Bayesian information criterion (BIC), the consistent Akaike information criterion (CAIC) and the Hannan-Quinn information criterion (HQIC). These criterions are defined by:

$$\begin{aligned} \text{AIC} &= -2\mathcal{L}(\hat{\boldsymbol{\theta}}) + 2p; \\ \text{BIC} &= -2\mathcal{L}(\hat{\boldsymbol{\theta}}) + plog(n); \\ \text{CAIC} &= -2\mathcal{L}(\hat{\boldsymbol{\theta}}) + \frac{2pn}{n-p-1}; \\ \text{HQIC} &= -2\mathcal{L}(\hat{\boldsymbol{\theta}}) + 2log(log(n)). \end{aligned}$$

where $\mathcal{L}(\hat{\boldsymbol{\theta}})$ denotes the log-likelihood function evaluated at the maximum likelihood estimates for parameters $\boldsymbol{\theta}$, p is the number of parameters and n is the sample size.Table 2 shows the MLE of the parameters of each model, the corresponding maximum log-likelihood value, the AIC, BIC, CAIC and HQIC for the three data sets. Table 3 presents the values of the statistics K-S, (A-D) (A*) and C-M (W*) for the three data sets using each model. The required numerical evaluations are carried out using R software.

For the first two data sets, glfr model has the smallest value of the Kolmogorov-Smirnov (largest P value), Anderson-Darling and the Cramér-von Mises gooness-of-fit tests statistics which indicate that the best fit is provided by glfr model for these data sets. For the third data set, gamma model is a better fit than glfr (α, β) model, see Table 3

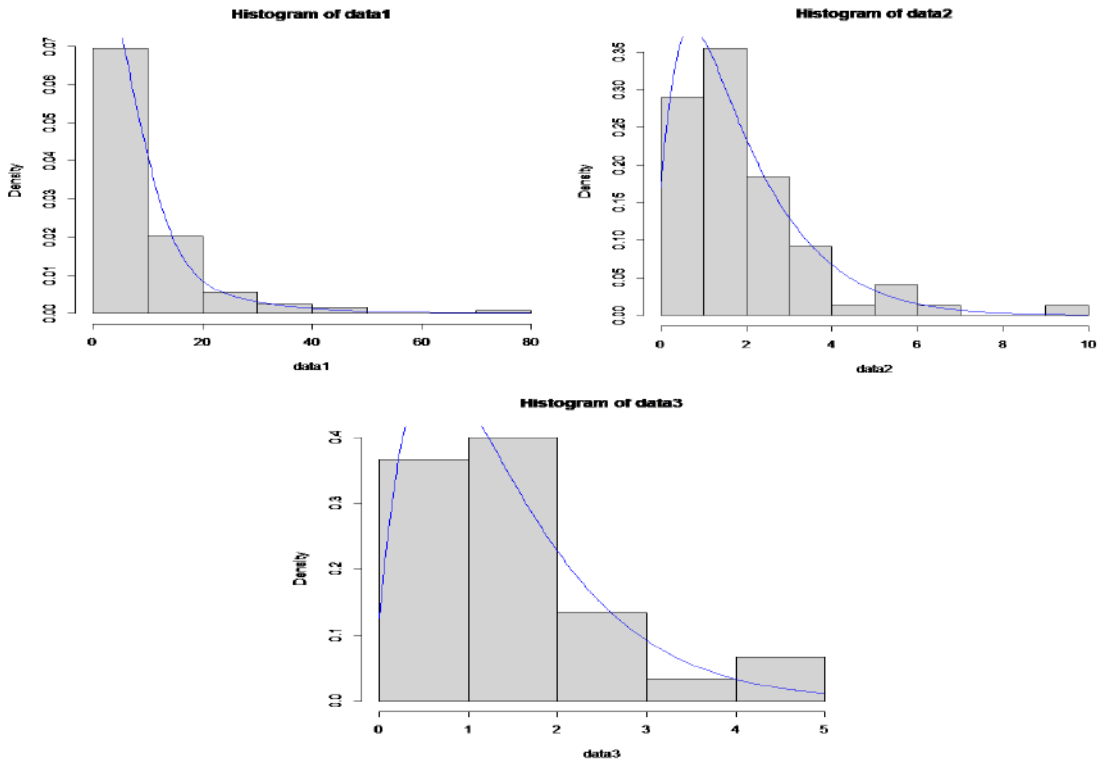


Figure 3: The histogram for the three data sets and fitted pdf of the gLinear failure rate distribution

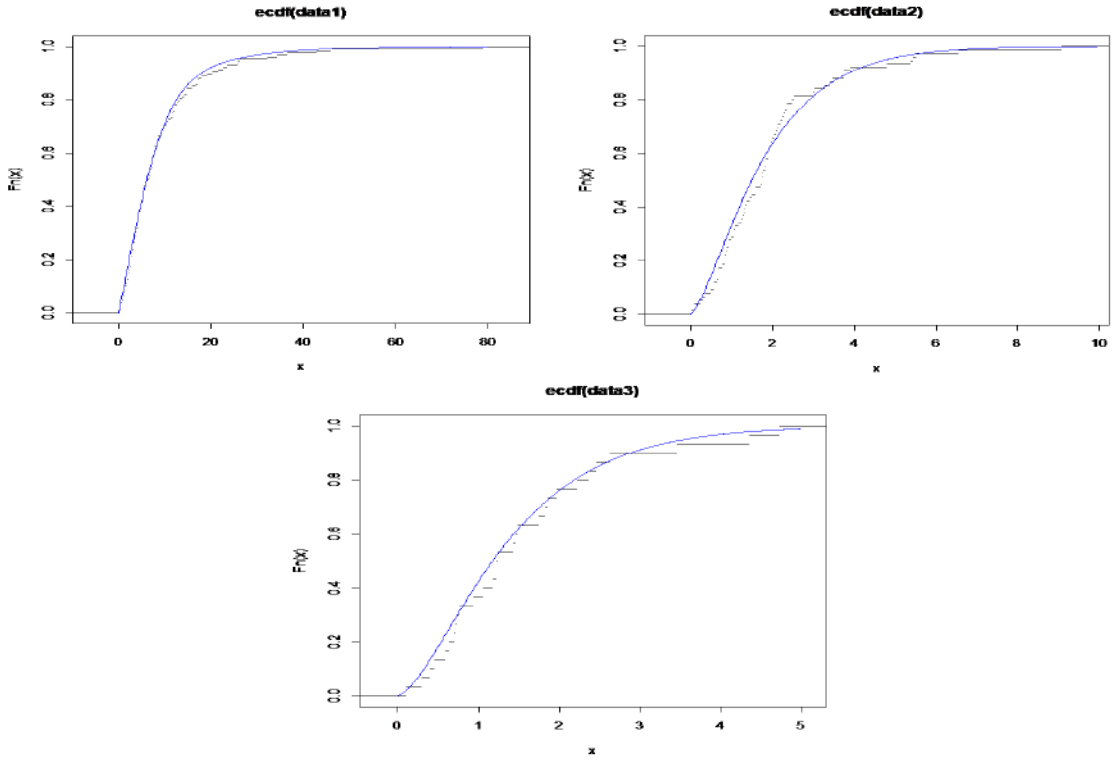


Figure 4: The empirical cdf for the three data sets and fitted cdf of the gLinear failure rate distribution

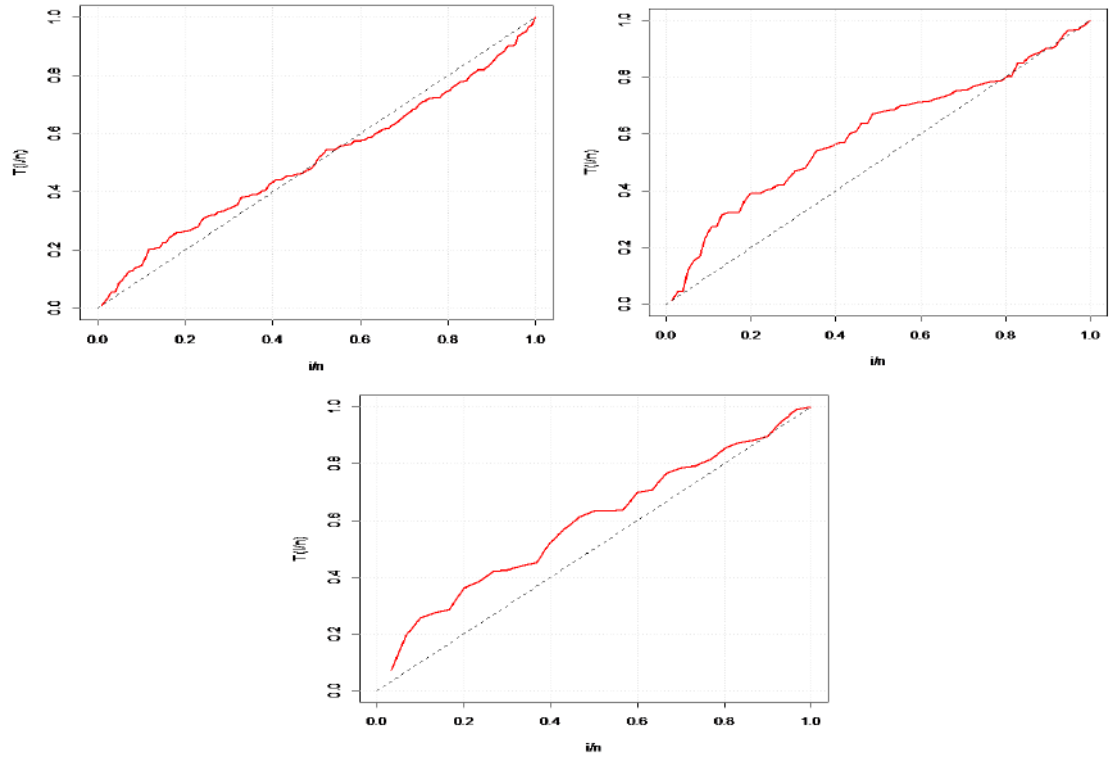


Figure 5: The TTT for the three data sets

From the results in Table 2, one can note that the values of AIC, BIC, CAIC and HQIC are smaller for the glfr distribution compared with those values of the other models, so the new distribution seems to be a very competitive model to the the first two data sets. For the third data set, gamma model has smaller values than glfr model.

Also, we plotted the scaled total time on test transform (TTT) which can help for selecting a model. The empirical scaled TTT transform (Aarset (1987)) can be used to identify the shape of the hazard function. As displayed in Figure 5 The TTT plot shows that the data set 1 has a unimodel hazard, while the rest of data sets have increasing hazards.

For Bayesian computations, we concern with three approaches; the Laplace approximation, MCMC and variational Bayes (VB). We obtain the approximate Bayes estimates of the two unknown parameters of the glfr distribution based on real data sets and simulated samples with true values $\alpha = 1.2, \beta = 1.2$. R package is used to compute these estimates. Using the first two data sets and gamma priors with different values of hyper parameters $((a_i, b_i) = (1, 0.001)$ and $(0.001, 0.001), i = 1, 2)$, the Laplace approximation, MCMC and variational Bayes are carried out and the results are shown in Tables 4-5. From these tables, one can note that the results are close for each other and for simulated data, the results get closer to the true values as the sample size increases. Also, the results are close to each other for different hyper parameters and Tables 4-5 display the results in the case of the gamma priors with $(a_1, b_1) = (1, 0.001)$. For MCMC, Figures 6- 7 show the trace, the approximated posterior density functions and autocorrelation plots of the two parameters of the glfr distribution. These Figures show that as the sample size increases, the chains look stationary, the kernel densities look Gaussian, and the ACF's or autocorrelation function plot show low autocorrelation.

7. Conclusion

A new mixture distribution named glinear failure rate distribution (glfr) is proposed in this paper. The glfr is a mixture of gamma and failure rate distributions. Based on some goodness of fit tests and some criteria for choosing the best fit among several, it is observed that the glfr gives a better fit than some common distributions. The maximum likelihood and Bayesian methods are applied to estimate the two unknown parameters of the glfr distribution. For Bayesian method, we used Laplace approximation, MCMC and Variational Bayes and the results are close to each other and for simulated data, the results get closer to the true values as the sample size increases.

Acknowledgements

I am very grateful to the Editors for their guidance and counsel. I am very grateful to the reviewer for valuable comments.

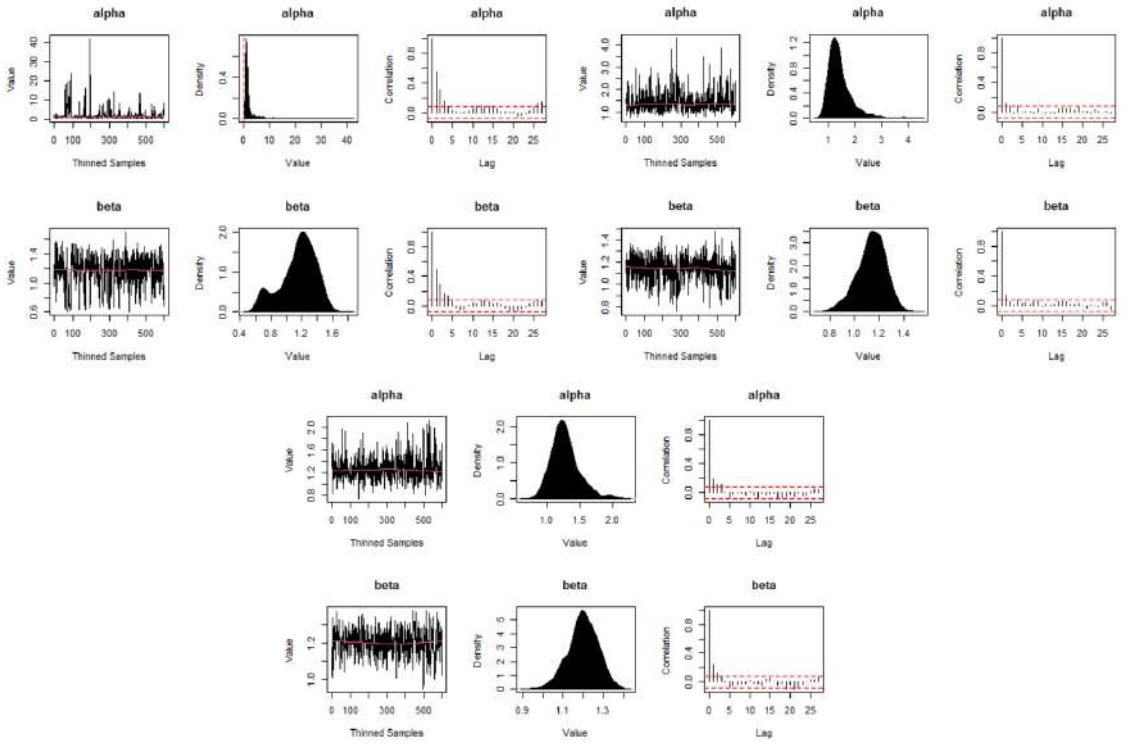


Figure 6: The trace, the approximated posterior density functions and autocorrelation plots of the two parameters for the simulated data

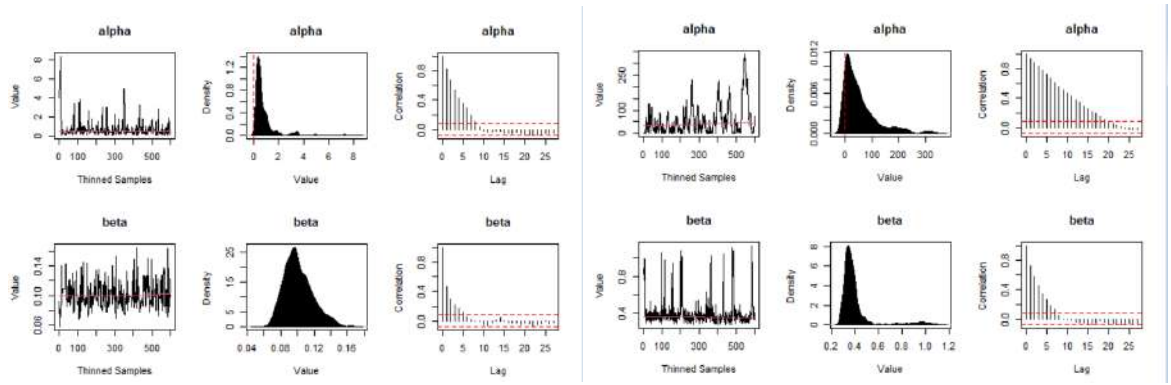


Figure 7: The trace, the approximated posterior density functions and autocorrelation plots of the two parameters for the two data sets

References

Aarset, M. V. (1987). How to identify a bathtub hazard rate. *IEEE Transactions on Reliability*, **36**, 106–108.

Crawford, S. L. (1994). An application of the laplace method to finite mixture distributions. *Journal of the American Statistical Association*, **89**, 259–267.

Everitt, B. (2013). *Finite Mixture Distributions*. Springer Science & Business Media.

Ghitany, M. E., Atieh, B., and Nadarajah, S. (2008). Lindley distribution and its application. *Mathematics and Computers in Simulation*, **78**, 493–506.

Lindley, D. V. (1958). Fiducial distributions and bayes' theorem. *Journal of the Royal Statistical Society, Series B (Methodological)*, **20**, 102–107.

Mahmoud, M. R., Mandouh, R., and Rasheedy, E. (2017). A new lifetime distribution based on finite mixtures. *International Journal of Applied Mathematics & Statistics*, **56**, 13–23.

Murthy, D. P., Xie, M., and Jiang, R. (2004). *Weibull Models*. John Wiley & Sons.

Ogunde, A., Ibraheem, A., and Audu, A. (2017). Performance rating of transmuted nadarajah and haghighi exponential distribution: an analytical approach. *Journal of Statistics: Advances in Theory and Applications*, **17**, 137–151.

Puza, B. (2015). *Bayesian Methods for Statistical Analysis*. ANU press.

Sarhan, A. M., Tadj, L., and Hamilton, D. C. (2014). A new lifetime distribution and its power transformation. *Journal of Probability and Statistics*, **2014**, 1–14.

Sen, S. and Chandra, N. (2017). The quasi xgamma distribution with application in bladder cancer data. *Journal of Data Science*, **15**, 61–76.

Sen, S., Maiti, S. S., and Chandra, N. (2016). The xgamma distribution: statistical properties and application. *Journal of Modern Applied Statistical Methods*, **15**, 38.

Shanker, R. and Mishra, A. (2013). A quasi Lindley distribution. *African Journal of Mathematics and Computer Science Research*, **6**, 64–71.

Tierney, L. and Kadane, J. B. (1986). Accurate approximations for posterior moments and marginal densities. *Journal of the American Statistical Association*, **81**, 82–86.

Tierney, L., Kass, R. E., and Kadane, J. B. (1989). Fully exponential laplace approximations to expectations and variances of nonpositive functions. *Journal of the American Statistical Association*, **84**, 710–716.

Yamaguchi, Y., Okamura, H., and Dohi, T. (2010). A variational bayesian approach for estimating parameters of a mixture of erlang distribution. *Communications in Statistics-Theory and Methods*, **39**, 2333–2350.

Zea, L. M., Silva, R. B., Bourguignon, M., Santos, A. M., and Cordeiro, G. M. (2012). The beta exponentiated pareto distribution with application to bladder cancer susceptibility. *International Journal of Statistics and Probability*, **1**, 8.

ANNEXURE

Table 1: MSE, coverage probability (CP) and average width (AW)

α	β	n	MSE_α	MSE_β	$CP_\alpha\%$	AW_α	$CP_\beta\%$	AW_β
0.8	0.8	50	0.8392	3.5144	99.5	3.2434	99.5	13.7551
		75	0.4296	2.5246	99.6	1.6422	99.7	9.8887
		100	0.3147	1.5077	98.9	1.1966	99.9	5.9097
		150	0.2112	0.1055	93.1	0.7744	94.7	0.3997
		200	0.1907	0.0972	93.1	0.6707	93.6	0.3628
0.8	1.0	50	0.5761	2.0099	99.5	2.2118	99.8	7.8758
		75	0.9416	1.3697	99.9	3.6092	99.9	5.3624
		100	0.3623	0.1921	90.4	1.2151	91.1	0.6762
		150	0.3131	0.1643	86.5	0.9748	90.8	0.5614
		200	0.2879	0.1544	88.1	0.8692	89.8	0.5184
1.0	1.0	50	2.5306	7.2913	98.8	9.8568	98.5	28.4180
		75	1.2069	4.7571	99.3	4.6747	99.3	18.5986
		100	0.7010	3.5606	99.6	2.7089	99.6	13.9390
		150	0.2758	0.1388	94.2	1.0408	95.8	0.5321
		200	0.2635	0.1197	93.7	0.9921	95.3	0.4594
1.0	1.2	50	2.26458	7.1844	99.5	8.8561	99.5	28.1341
		75	0.5315	3.0952	99.4	2.0152	99.8	12.1331
		100	0.3748	2.0887	93.4	1.3431	99.9	8.1863
		150	0.3503	0.1840	89.5	1.1865	92.9	0.6266
		200	0.3500	0.1709	89.6	1.1591	91.1	0.5976
1.2	1.2	50	1.8795	5.0542	99.6	7.2992	99.6	19.7939
		75	0.6501	2.8649	99.8	2.472	99.8	11.2279
		100	0.3362	0.1913	90.5	1.2115	94.8	0.7192
		150	0.3096	0.1483	92.3	1.1403	94.3	0.5555
		200	0.2896	0.1416	92.5	1.0543	93.9	0.5360
1.6	2.5	50	1.4358	0.8157	83.1	4.0621	87.0	2.5023
		75	1.4005	0.7625	76.4	3.3660	94.2	2.5460
		100	1.5233	0.5758	73.0	3.5535	89.8	1.8990
		150	1.3813	0.4837	61.7	2.7944	89.1	1.5026
		200	1.2944	0.4622	55.0	2.4079	85.8	1.3990

Table 2: The MLEs and some measures for the fitted models

Data set	Model	Parameter Estimates	$l(\hat{\boldsymbol{\theta}})$	AIC	BIC	CAIC	HQIC
Data 1	glfr(α,β)	$\hat{\alpha}=0.1911$ and $\hat{\beta}=0.1342$	-411.8	827.6	833.3	827.7	830
	glfr(β)	$\hat{\beta}=0.0888$	-412.6	827.2	830.0	827.2	828.3
	qxgamma	$\hat{\alpha}=2.8289$ and $\hat{\beta}=0.1655$	-416.9	837.9	843.6	838.0	840.2
	xgamma	$\hat{\beta}=0.2689$	-429.4	860.7	863.6	863.8	861.9
	qLindley	$\hat{\alpha}=2.5292$ and $\hat{\beta}=0.1381$	-414.9	833.7	839.4	833.8	836.1
	Lindley	$\hat{\beta}=0.1960$	-419.5	841.1	843.9	841.1	842.2
	lfr	$\hat{\beta}=0.0608$	-427.2	856.6	859.3	856.5	857.6
	gamma	$\hat{\beta}=0.2135$	-426.8	855.6	858.4	855.6	856.8
Data 2	glfr(α,β)	$\hat{\alpha}=0.2086$ and $\hat{\beta}=0.9439$	-122.2	248.4	253.1	248.6	250.3
	glfr(β)	$\hat{\beta}=0.5537$	-124.0	250.0	252.4	250.1	251
	qxgamma	$\hat{\alpha}=0.2009$ and $\hat{\beta}=1.3561$	-122.5	249.0	253.6	249.1	250.8
	xgamma	$\hat{\beta}=1.0330$	-126.3	254.7	257.0	254.7	255.6
	qLindley	$\hat{\alpha}=0.2947$ and $\hat{\beta}=0.8823$	-122.0	248.0	252.7	248.2	249.9
	Lindley	$\hat{\beta}=0.7948$	-123.7	249.4	251.7	294.4	250.3
	lfr	$\hat{\beta}=0.3329$	-124.5	251.0	253.3	251.0	251.9
	gamma	$\hat{\beta}=1.0210$	-123.2	248.4	250.7	248.4	249.3
Data 3	glfr(α,β)	$\hat{\alpha}=0.1394$ and $\hat{\beta}=1.3079$	-39.70	83.40	86.10	83.84	84.30
	glfr(β)	$\hat{\beta}=0.8413$	-41.25	84.50	85.90	84.64	84.94
	qxgamma	$\hat{\alpha}=0.1599$ and $\hat{\beta}=1.156$	-40.52	85.04	87.84	85.48	85.94
	xgamma	$\hat{\beta}=1.2690$	-42.14	86.28	87.69	86.42	86.73
	qLindley	$\hat{\alpha}=0.4026$ and $\hat{\beta}=1.2962$	-40.59	85.18	87.98	85.62	86.07
	Lindley	$\hat{\beta}=0.9947$	-41.09	84.18	85.59	84.33	84.63
	lfr	$\hat{\beta}=0.4408$	-40.73	83.46	84.86	83.60	83.91
	gamma	$\hat{\beta}=1.3250$	-39.52	81.04	82.44	81.19	81.49

Table 3: Statistics K-S (P value), A* and W* for the three data sets

Data set	Model	K-S (P value)	A*	W*
Data 1	glfr(α, β)	0.059 (0.800)	0.3835	0.0607
	glfr(β)	0.055 (0.800)	0.6957	0.1236
	qxgamma	0.100 (0.100)	1.0160	0.1687
	xgamma	0.160(0.003)	2.2250	0.3787
	qLindley	0.074 (0.500)	0.9005	0.1510
	Lindley	0.120 (0.060)	1.0260	0.1717
	lfr	0.180 (6e-04)	2.2630	0.3849
	gamma	0.140 (0.010)	0.7260	0.1211
Data 2	glfr(α, β)	0.091 (0.500)	0.6425	0.1102
	glfr(β)	0.130 (0.200)	0.4919	0.8261
	qxgamma	0.110 (0.300)	0.7579	0.1266
	xgamma	0.150 (0.070)	0.9901	0.1698
	qLindley	0.120 (0.200)	0.6356	0.1082
	Lindley	0.120 (0.200)	0.6907	0.1173
	lfr	0.130 (0.200)	1.0470	0.1807
	gamma	0.098 (0.400)	0.7005	0.1182
Data 3	glfr(α, β)	0.099 (0.900)	0.1520	0.0200
	glfr(β)	0.130 (0.700)	0.1328	0.0188
	qxgamma	0.120 (0.800)	0.2934	0.0401
	xgamma	0.160 (0.400)	0.2481	0.03293
	qLindley	0.120 (0.800)	0.1930	0.0258
	Lindley	0.130 (0.700)	0.1843	0.2436
	lfr	0.110 (0.900)	0.3036	0.0465
	gamma	0.095 (0.900)	0.1496	0.0195

Table 4: Summary results for the posterior parameters in the case of the glfr distribution based on real data sets

Laplace Approximation						
Data set	Parameter	Estimate: Mode	Standard Deviation	LB	UB	Minutes of runtime
Data 1	α	0.3471	0.1829	0.0000	0.7128	0.00
	β	0.1061	0.0172	0.0718	0.1404	
Data 2	α	0.1157	0.1063	0.0000	0.3282	0.00
	β	0.9410	0.1049	0.7312	1.1509	
MCMC						
Data set	Parameter	Estimate: Mode	Standard Deviation	LB	UB	Minutes of runtime
Data 1	α	0.5968	0.5940	0.1383	2.1327	0.09
	β	0.10478	0.0186	0.0784	0.1475	
Data 2	α	56.4094	65.2242	0.0934	230.0292	0.07
	β	0.4108	0.1615	0.2878	0.9777	
Variational Bayesian						
Data set	Parameter	Estimate: Mean	Standard Deviation	LB	UB	Minutes of runtime
Data 1	α	0.4631	0.0996	0.2639	0.6622	0.02
	β	0.1028	0.0114	0.0799	0.1257	
Data 2	α	0.1763	0.9364	0.0000	0.3636	0.02
	β	0.9162	0.1047	0.7068	1.1255	

Table 5: Summary results for the posterior parameters in the case of the glfr distribution based on simulated samples with true values $\alpha = 1.2, \beta = 1.2$

Laplace Approximation						
Samples	Parameter	Estimate: Mode	Standard Deviation	LB	UB	Minutes of run-time
n=300	α	1.2886	0.3798	0.6621	2.1081	0.00
	β	1.2314	0.1419	0.9973	1.5034	
n=500	α	1.2361	0.2786	0.6789	1.7934	0.00
	β	1.1732	0.1003	0.9726	1.3738	
n=1000	α	1.1895	0.1850	0.8196	1.5595	0.01
	β	1.2200	0.0742	1.0792	1.3608	
MCMC						
Samples	Parameter	Estimate: Mode	Standard Deviation	LB	UB	Minutes of run-time
n=300	α	2.3272	3.2792	0.6940	12.1946	0.12
	β	1.1551	0.2263	0.6647	1.4986	
n=500	α	1.4436	0.4626	0.8844	2.6375	0.16
	β	1.1368	0.1152	0.8791	1.3386	
n=1000	α	1.2741	0.2175	0.9189	1.7781	0.26
	β	1.9835	0.0740	1.0441	1.3348	
Variational Bayesian						
Samples	Parameter	Estimate: Mean	Standard Deviation	LB	UB	Minutes of run-time
n=300	α	1.3167	0.3329	0.6508	1.9826	0.02
	β	1.2242	0.13173	0.9607	1.4876	
n=500	α	1.3633	0.2822	0.7990	1.9276	0.04
	β	1.1455	0.0983	0.9490	1.3420	
n=1000	α	1.2414	0.1901	0.8613	1.6215	0.07
	β	1.2086	0.0703	1.0680	1.3492	

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 21, No. 2, 2023 (New Series), pp 211–233
http://www.ssca.org.in/journal

Accelerated Life Test Acceptance Sampling Plans for the Weibull Distribution with Constant Acceleration using EWMA and Modified EWMA Statistics

D. P. Raykudaliya[1], Sanjay Christian[2], and Jyoti Divecha[3]
[1,3]*Department of Statistics, Sardar Patel University, Vallabh Vidyanagar, Gujarat-India*
[2]*JG Institute of Business Administration, Ahmedabad, Gujarat-India*

Received: 06 October 2022; Revised: 17 January 2023; Accepted: 18 March 2023

Abstract

Products or systems with high reliability and durability require more time for inspection as it is difficult to detect failures under normal operating conditions. In literature, accelerated life testing methods are discussed to accelerate the decision, in which products are tested under severe than normal stress conditions with failure mechanisms similar to the one observed in the field. The most commonly applied stressis constant stress. We propose life test sampling plans for establishing a quantile life for the Weibull distribution under a constant acceleration factor. As the accelerated life test plans require more sample size than those without acceleration, the previous history has been used as per EWMA and Modified EWMA schemes to make them somewhat economical. Tables of optimal design parameters are presented for three different acceleration constants for establishing a Weibull median life. Published real-life data sets are used to demonstrate the proposed life test sampling plans.

Key words: Acceleration Factor; Quantile; Weibull Distribution; EWMA; Modified EWMA.

AMS Subject Classifications: 62P30

1. Introduction

The era of a competitive market with rapid developments in technology motivates manufacturers to supply products with high durability and reliability. To achieve this dimension of quality, in literature, online, and offline product control techniques are discussed in which the acceptance sampling or product control is an offline technique for deciding to accept or reject a lot through sample inspection before sending them to market.The most popular acceptance sampling plan is the single sampling plan in which a random sample of size (n) is taken from a lot of size (N) and number of defectives (d) are observed, and the decision to accept the lot if $d \leq c$ (acceptance number) otherwise reject the lot. The concept and usage of acceptance sampling were first narrated by Dodge and Romig (1941) for inspecting bullets.

Corresponding Author: D.P. Raykudaliya
Email: dp_raykundaliya@spuvvn.edu

The acceptance sampling plans are developed based on life tests known as life test sampling plans (LSP). To reduce the cost of testing, a sample of products is put on life test for a time smaller than the specified mean or quantile life, when the mean or quantile life of products exceeds the specified mean or quantile life, the lot of products under inspection is accepted otherwise it is rejected. Several authors have given LSPs which are based on mean lifetimes, to mention are Epstein (1954) gave those for the exponential distribution, Goode and Kao (1960 and 1961) for Weibull distribution, Baklizi and Masri (2004) for Birnbaum Model, Rosaiah and Kantam (2005) for inverse Rayleigh distribution, Tsai and Wu (2006) for Generalized Rayleigh distribution, and, Khan and Alqarni (2020) for inverse Weibull distribution.

Engineers focus more on the product's percentile life during life testing applications than its mean life. The reason is that the mean may not be an appropriate measurement when one perceives a product's quality to be in the low percentile. A small decrease in mean with a simultaneous small increase in variance can result in a downward shift in small percentiles of interest. Gupta (1962), Balakrishnan *et al.* (2007), Singh *et al.* (2015)and Singh and Tripathi (2017) gave time-truncated attribute LSPs for Normal and Log-Normal, generalized Birnbaum-Saunders distribution, generalized inverted exponential distribution, inverse Weibull distribution respectively for median lifetimes. Lio *et al.* (2009), Rao and Kantam (2010), Aslam *et al.* (2011), Rao *et al.* (2012), Rao *et al.* (2013), Rao and Rao (2013), Rao (2013), Malathi and Muthulakshmi (2016), Kaviyarasu and Fawaz (2017), Pradeepaveerakumari and Ponneeswari (2017) and Raykundaliya *et al.* (2022) developed LSPs assuring percentile lifetime of a product when life time distribution are respectively Birnbaum-Saunders, Log-logistic Burr type XII, inverse Rayleigh distribution, Half Normal, generalized Logistic, Marshall-Olkin extended exponential, Gompertz, Modified Weibull, Exponential Rayleigh, and Weibull.

In a competitive environment, to maintain a brand name, one needs to produce products of high reliability and durability. Such products not only require more time for inspection but also fail to detect failures under normal operating conditions. Industries of the above nature require a mechanism that reduces the time required for testing. Accelerated testing reduces testing time and helps to draw quick inferences about the product being tested, that is products under inspection are subjected to higher than usual stress. Nelson (1990) gave the basic idea of accelerated life testing and graphical analysis to estimate product life. Lin and Chiu (1995) constructed a cost model for an accelerated life test sampling plan. Escobar and Meeker (2006) have outlined some of the basic ideas behind accelerated testing. Xiaoyang *et al.* (2015) developed accelerated life testing plans considering lognormal distribution. Aslam et al (2019) gave accelerated LSPs by studying the mean life of Weibull distribution.

The scale family of distributions plays a very important role in lifetime data analysis. Among them the Weibull distribution is widely used in reliability studies because of its flexibility to take various shapes and, IFR - DFR properties. It is popular in warranty analysis, utility services, and industries manufacturing bearings, capacitors, because lifetimes under accelerated testing continue to follow the Weibull distribution with nearly the same shape parameter. This makes development of accelerated LSPs for assuring a specified life of Weibull distribution mathematically tractable. However, the accelerated LSPs require a higher sample size than the usual LSPs because of collecting only restricted lifetime data.

The sample size can be reduced by considering weighted average of the current lot information with the past lot information (see Aslam *et al.*, 2017). Using this approach, Divecha and Raykundaliya (2020) have designed LSPs as per EWMA and Modified EWMA, and showed that they are more economical than those developed based on only current lot information. In this paper, we propose LSPs for assuring quantile Weibull lifetime under constant acceleration using EWMA and Modified EWMA for early decision with lesser sample size. The organization of the paper is as follows:

In Section 2, we briefly discuss the Weibull distribution in the context of the newer term, constant acceleration factor (AF). In Section 3, we discuss the proposed LSPs for the Weibull distribution with acceleration (WDwALSP) followed by economical WDwALSP based on the EWMA and Modified EWMA.In Section 4, we illustrate the LSPs and give hypothetical example. In Section 5, we give two real-life examples to demonstrate the use of proposed plans and compared them with LSPs without acceleration. Concluding remarks are given in Section 6.

2. Weibull distribution and probability of failure under accelerated quantile life ratio

The cumulative distribution function of the Weibull distribution is defined by

$$F(t) = 1 - e^{-(\sigma t)^{\theta}};\ t > 0,\ \sigma > 0,\ \theta > 0, \tag{1}$$

where θ and σ are respectively shape and scale parameters of the distribution.

Let t_U denotes lifetime of a product under normal conditions following Weibull distribution with CDF,

$$F(t_U) = 1 - e^{-(\sigma_U t_U)^{\theta}}. \tag{2}$$

It's 100q$^{\text{th}}$ quantile life time is $t_{qU} = \frac{(-\log(1-q))^{\frac{1}{\theta}}}{\sigma_U} = \frac{b}{\sigma_U}$, where $b = (-log(1-q))^{\frac{1}{\theta}}$.

Therefore,

$$\sigma_U = \frac{b}{t_{qU}}. \tag{3}$$

Let t_A be the lifetime of a product under constant stress (acceleration) having Weibull distribution with CDF,

$$F(t_A) = 1 - e^{-(\sigma_A t_A)^{\theta}} \text{ with } \sigma_A = \frac{b}{t_{qA}}. \tag{4}$$

It's 100q$^{\text{th}}$ quantile life is $t_{qA} = \frac{b(AF)}{\sigma_A}$, where, $AF = \frac{t_U}{t_A}$,
implies

$$t_A = \frac{t_U}{AF} \tag{5}$$

and

$$\frac{\sigma_A}{\sigma_U} = \frac{(AF)t_{qU}}{t_{qA}} \tag{6}$$

Let

$$\frac{\sigma_A}{\sigma_U} = RAR \text{ (ratio of acceleration rate, } RAR > AF) \tag{7}$$

Let, $\tau_A = \frac{t_{U0}}{AF}$, where t_{U0} is truncation time under normal conditions defined as $t_{U0} = \delta_0 t^0_{qU}$, $0 < \delta_0 < 1$ is a constant called termination ratio, and $\frac{t_{qA}}{t^0_{qA}}$, is the ratio of true quantile lifetime to the specified quantile lifetime representingthe quality of a lot. Using (6), $\tau_A = \frac{\delta_0\left(\frac{\sigma_A}{\sigma_U}\frac{1}{AF}t^0_{qA}\right)}{AF}$ for which the CDFin terms of AF and RAR, using (4) and (7) is,

$$F(\tau_A) = 1 - e^{-\left((b)^\theta(\delta_0)^\theta(RAR)^\theta\left(\frac{1}{AF}\right)^\theta\left(\frac{t_{qA}}{t^0_{qA}}\right)^{-\theta}\right)} \tag{8}$$

The specified accelerated quantile life t^0_{qA}, lot quality ratio $\left(\frac{t_{qA}}{t^0_{qA}}\right)$ (> 1), accelerated test termination time truncation ratio δ_0 (< 1), AF (> 1), and RAR $(>$ AF) are decided by the producer.

Further, we know that H_0: $t_{qU} \geq t^0_{qU}$ Vs H_1: $t_{qU} < t^0_{qU}$ and H_0: $t_{qA} \geq t^0_{qA}$ Vs H_1: $t_{qA} < t^0_{qA}$ are equivalent and let truncated time for acceleratedlife tests as τ_A.

3. Procedure of LSPs for the Weibull distribution with acceleration (WD-wALSP)

Suppose that producer submit a lot of units for accelerated testing, whose lifetimes follow the Weibull (θ, σ) and claims that the true quantile life t_{qA} of the lot is better than the specified qaunitle life t^0_{qA}. To support producer's claim, we propose the design of a time truncated LSPs based on accelerated quantile lifetime, with following procedure:

Step1: Suppose 'n' items from a lot are taken and put on life test until accelerated test time τ_A and observe lifetimes X_j $(j = 1, 2, \ldots, n)$.

Step 2: Define an indicator variable say, I_j;$(j = 1, 2, \ldots, n)$

$$I_j = \begin{cases} 1, & \text{if } 0 < X_j \leq \tau_A,\ j = 1, 2, \ldots, n \\ 0, & \text{Otherwise} \end{cases} \tag{9}$$

Let 'd' denote failed items during test time $(0, \tau_A]$. Define $d = \sum_{j=1}^{n} I_j$.

Step 3: If $d \leq c$, accept the lot, otherwise reject the lot.

While inspecting items with acceleration during the time interval $(0, \tau_A]$, the sample under study may fail with probability p where $p = F(\tau_A)$, given by (8), and the count of failed items'd' follows the binomial distribution having parameters n and p, with mean np and variance $np(1-p)$. A lot must be accepted if the data support the null hypothesis H_0 : $t_{qA} \geq t^0_{qA}$ against the alternative hypothesis H_1 : $t_{qA} < t^0_{qA}$ satisfying both producer's and consumer's requirements.

The design parameters (n, c) are to be obtained by solving the following two inequalities simultaneously.

$$\begin{aligned} L(p_1) &= \sum_{d=0}^{c} \binom{n}{d} p_1^d (1-p_1)^{n-d} \geq 1 - \alpha \\ L(p_2) &= \sum_{d=0}^{c} \binom{n}{d} p_2^d (1-p_2)^{n-d} \leq \beta \end{aligned} \tag{10}$$

where, $p_1 = F(\tau_A)$ at $\frac{t_{qA}}{t^0_{qA}} > 1$ and $p_2 = F(\tau_A)$ at $\frac{t_{qA}}{t^0_{qA}} = 1$ denotes the probability of failure for a good lot and for a poor lot during the time $(0, \tau_A]$.

Under normal approximation, binomial probabilities given in equation (10) representing producer's and consumer's requirements are given in terms of standard normal distributions Φ as

$$\begin{aligned} \Phi\left(\frac{c-np_1}{\sqrt{np_1(1-p_1)}}\right) &\geq 1-\alpha \\ \Phi\left(\frac{c-np_2}{\sqrt{np_2(1-p_2)}}\right) &\leq \beta \end{aligned} \tag{11}$$

The normal approximation to the binomial is known to be satisfactory for p_a approximately $\frac{1}{2}$ and $n > 10$. In general, if $\frac{1}{n+1} < p_A < \frac{n}{n+1}$ then the normal approximation is adequate. In numerical form, if $np_A > 10$ and $0.1 \leq p_A \leq 0.9$ then normal approximation is appropriate to the binomial distribution. Referto Montgomery (2001) (Page no. 76-77).

3.1. WDwALSP using EWMA

Accelerated LSP almost requires more than double the sample size, thereby increasing the cost of inspection, than those by the standard LSP that is without accelerated tests. Therefore, there has to be a procedure by which it can be brought down to the sample size of the standard LSPs. Use of EWMA statistics to summarize the previous lot inspection number of failures with the current one brings down the sample size effectively for Weibull (Aslam, 2017), and Generalized Exponential distributions (Divecha and Raykundaliya, 2018). Therefore, if an industry maintains a database of inspection history of lots, step3 in section 3 would be as follows.

Step 3: Calculate the EWMA statistic of Roberts (1959), based on d_l for $l = 1, 2, \ldots$

$$EWMA_l = \lambda d_l + (1-\lambda) EWMA_{l-1};\ EWMA_1 = d_1 \tag{12}$$

where $0 < \lambda < 1$, λ is a smoothing constant chosen because. if $\lambda = 0.6$ means the weight attached to the current sample is 0.6 and the weights attached to the past sample is 0.4, the current number of failures will be affected not by more than the previous 5 lot information.

Then, Accept the lot if $EWMA_l \leq c$, otherwise, reject the lot.

Following Montgomery (2001) the mean and variance of $EWMA$ statistic are given by,

$$\mu_{EWMA} = np$$

and

$$\sigma^2_{EWMA} = np(1-p)\left(\frac{\lambda}{2-\lambda}\right) \tag{13}$$

Therefore, the WDwALSP-based EWMA design consists of estimating (n, c), subject to the producer's and consumer's requirements in terms of probabilities of accepting the lot as those given in (10) are

$$\begin{aligned} L(p_1) &= P(EWMA_l \leq c) \\ L(p_2) &= P(EWMA_l \leq c) \end{aligned} \tag{14}$$

After normalization and using (13), (14) becomes,

$$L(p) = \Phi\left(\frac{c - \mu_{EWMA}}{\sigma_{EWMA}}\right) \tag{15}$$

with $p = p_1$ and p_2 appropriately.

The probability of accepting a good and poor lot using (15) are

$$\begin{aligned} L(p_1) &= \Phi\left(\frac{c-np_1}{\sqrt{np_1(1-p_1)\left(\frac{\lambda}{2-\lambda}\right)}}\right) \geq 1-\alpha \\ L(p_2) &= \Phi\left(\frac{c-np_2}{\sqrt{np_2(1-p_2)\left(\frac{\lambda}{2-\lambda}\right)}}\right) \leq 1-\beta \end{aligned} \tag{16}$$

This pair of equations is used to determine the WDwALSP parameters under EWMA.

3.2. WDwALSP using modified EWMA

A further reduction in sample size for WDwALSP is possible by using Modified EWMA (Khan *et al.*, 2016) as shown in Divecha and Raykundaliya (2020) as the Modified EWMA statistic has a variance smaller than that of EWMA.

Step 3 in Section 3 is taken as follows.

Step 3: Calculate Modified EWMA statistic

$$MEWMA_l = \lambda/2(d_l + d_{l-1}) + (1-\lambda)MEWMA_{l-1};\ MEWMA_1 = d_1 \tag{17}$$

where, $0 < \lambda < 1$; $l > 1, \lambda$ is a smoothing constant chosen just as in the EWMA scheme.

If $MEWMA_l \leq c$, Accept the lot, otherwise, reject the lot.

The mean and variance of Modified EWMA statistic are $\mu_{MEWMA} = np$

and

$$\sigma^2_{MEWMA} = np(1-p)\left(\frac{\lambda}{2}\right) \tag{18}$$

Similar to Section 3.1, the probabilities of accepting a good and poor lot are

$$\begin{aligned} L(p_1) &= \Phi\left(\frac{c-np_1}{\sqrt{np_1(1-p_1)\left(\frac{\lambda}{2}\right)}}\right) \geq 1-\alpha \\ L(p_2) &= \Phi\left(\frac{c-np_2}{\sqrt{np_2(1-p_2)\left(\frac{\lambda}{2}\right)}}\right) \leq \beta \end{aligned} \tag{19}$$

This pair of equations is used to determine the WDwALSP parameters under modified EWMA.

4. Constructions of tables of proposed LSP

Tables 2 – 4 of WDwALSP are constructed by satisfying (11) to establish the 50^{th} quantile life with fixed Producer's risk $\alpha = 0.05$, consumer's risks ($\beta = 0.25, 0.10, 0.05, 0.01$), shape parameters ($\theta = 1.5, 2.0, 2.5$), quality levels $\left(\frac{t_{qA}}{t^0_{qA}} = 1.25, 1.50, 1.75, 2.00, 2.25\right)$, termination ratios ($\delta_0 = 0.6, 0.8$), acceleration factors (AF=1.5,2.0,2.5) and the ratio of acceleration rate(RAR=2.0,2.5,3.0).

Tables 5 – 13 of WDwALSP using EWMA and Tables 14 – 22 of WDwALSP using Modified EWMA are obtained by satisfying equations (16) and (19) with the above parameters and smoothing constant ($\lambda = 0.2, 0.4, 0.6, 0.8, 1.0$) as per EWMA and Modified EWMA statistic. All tables are constructed using R 3.6 language.

From all of the constructed tables, the following observations are made:

1. Keeping fixed producer risk, with the increase in termination ratio (that is increase the truncation time), quality level (that is true process quantile lifetime is much higher than specified quantile lifetime), consumer risk, shape parameter (that is probability of failure of a product is increased) as a result there is a decrease in sample size and acceptance number.

2. With the increase in the smoothing constant, there is an increase in sample size and acceptance number. It means, if information about the quality history of a lot is less to the producer, for testing of a lot, higher sample size is required, and hence cost of inspection of a lot is increased.

Among the comparison of all constructed tables of WDwALSP, WDwALSP with EWMA, and WDwALSPwithModified EWMA lesser sample sizes are required under the last type.

Usually,the question arises in our mind what should be sample size required for inspection when acceleration is given and not given? Which of the LSPs is better? To address these questions, we give tables for optimal design parameters for LSP for Weibull distribution (WDLSP) and LSP for Weibull distribution with acceleration (WDwALSP) with EWMA and modified EWMA respectively in tables 23 - 25 and 2 - 4 for different shape parameters and different process parameters. The Tables 23 - 25 obtains using (11), (16) and (19) after substituting AF=1 and RAR=1 in (8). From both plans' tables, we observed that somewhat higher sample size is required for the inspection of a lot in WDwALSP compared to WDLSP.Lin and Chiu (1995) showed in their paper using an example that, accelerated plans required a higher sample size in comparison to the plans without acceleration. They also showed that the overall cost of the experiment and time of experiments are reduced compared to plans without acceleration. From Tables (2-4), it is observed that the WDwALSP with EWMA or modified EWMA requires a moderately lesser sample size than the WDLSP. Hence under acceleration, plans with EWMA or modified EWMA are economical and preferable provided inspection histories of lots are available.

5. Illustration

We use two data sets, Lawless (1982, Page 228) ball bearing data and Murthy *et al.* (2004) repairable item data to illustrate WDwALSP introduced in this paper.

5.1. Lawless dataset

A million revolutions before failure for each of the 23 ball bearings tested were as follows.

17.88, 28.92, 33.00, 41.52, 42.12, 45.60, 48.80, 51.84, 51.96, 54.12, 55.56, 67.80, 68.44, 68.64, 68.88, 84.12, 93.12, 98.64, 105.12, 105.84, 127.92, 128.04, 173.40.

The maximum likelihood estimators of shape and scale parameters of the well-fitted Weibull distribution ($AIC = 231.37$, KS-test $D = 0.1502(p = 0.62)$) for the number of million revolutions before failure were respectively,2.1014 (considered as 2) and 0.0122.

5.1.1. WDLSP

We illustrate the proposed LSP with and without acceleration using the ball bearing data. We estimate the 50th quantile of lifetimes, which is 67.80. Suppose producer and consumer with specified risks (0.05, 0.1) agrees to accept the lot if the true median life is 1.75 times better than the specified median life, which would be 38.75. Further, they agree on test termination time to be 0.8 of the specified median time, which results in 31. The LSP design parameters for this set of constants are read from the Table 23. So, the parameters (n, c) of chosen LSP is (12,2). During this test time, the total number of failed items in the given sample is 2, which is equal to the acceptance number hence, the lot is accepted.

5.1.2. WDLSP with EWMA and modified EWMA

When the previous history of lot inspection is available, the user can take advantage and inspect the lot with lesser items on the test.

Since the history of data is not available, we have generated four samples (using R Language) of size 6 having shape parameter $\hat{\theta} = 2.1014$ and $\hat{\sigma} = 0.0122$

Under the same set of constants as above the parameter of WDLSP with EWMA from Table 24 are $(n = 6, c = 1)$, requiring only 6 items to be tested instead of 10 items. To illustrate this, we suppose the lifetimes of previous lot inspections are not available, hence we simulate four samples of Weibull distributed lifetimes having shape parameter 2 and scale parameter 0.01 each of size 6, which are:

41.69, 59.15, 39.53, **15.26**, 52.65, **23.28**
124.89, 42.78, 44.61, **24.58**, 139.28, 55.40
139.65, **17.83**, 87.01, 88.02, **23.83**, **26.95**
78.04, 32.49, 81.28, **11.78**, 51.21, 58.62

Real-life data: 17.88, 28.92, 33.00, 41.52, 42.12, 45.60

Their inspection results in the number of failures as $d_1 = 2$, $d_2 = 1$, $d_3 = 3$, $d_4 = 1$,

$d_5 = 2$. As per (12), the corresponding EWMA statistics values are $\boldsymbol{EWMA_1 = 2}$, $\boldsymbol{EWMA_2 = 1.40}$, $\boldsymbol{EWMA_3 = 2.36}$, $\boldsymbol{EWMA_4 = 1.54}$, and $\boldsymbol{EWMA_5 = 1.81}$, respectively. Since $\boldsymbol{EWMA_5 > 1}$, the decision would bea current lot is rejected. Notice that, a bad history may change the decision.

Modified EWMA statistic uses the historical information slightly differently and may further reduce the sample size needed to establish the quality and decision about a lot. Observe that under the same set of constants as above the parameter of WDLSP with Modified EWMA from Table 25 are $(n = 5, c = 1)$, less by one more item than needed by WDLSP under EWMA. The number of failures as $d_1 = 1$, $d_2 = 1$, $d_3 = 2$, $d_4 = 1$, $d_5 = 2$ As far as the decision process is concerned, we calculate as per (17) the Modified EWMA statistics and find them as $\boldsymbol{MEWMA_1 = 1}$, $\boldsymbol{MEWMA_2 = 1}$, $\boldsymbol{MEWMA_3 = 1.3}$, $\boldsymbol{MEWMA_4 = 1.42}$, and $\boldsymbol{MEWMA_5 = 1.468}$, respectively. Since $\boldsymbol{MEWMA_5 = 1.468 > c = 1}$, The decision is once again to reject the lot because of a bad third lot.

5.1.3. WDwALSP

We extend the use of ball bearing data to illustrate the LSPs for Weibull distribution under acceleration factor $AF = 2$. The data set is transformed accordingly (dividing by 2), so they are,

8.94, 14.46, 16.5, 20.76, 21.06, 22.8. 24.4, 25.92, 25.98, 27.06, 27.78, 33.9, 34.22, 34.32, 34.44, 42.06, 46.56, 49.32, 52.56, 52.92, 63.96, 64.02, 86.7.

The accelerated 50^{th} quantile of failure time is $t_{0.5A} = 33.9$. We need to specify the ratio $RAR = \frac{\sigma_A}{\sigma_U} > 2$. ($\because$ Eq. 6 and Eq. 7). The test time to establish a true median life of 1.75 better than the specified, under termination ratio ratio $\delta_0 = 0.8$, and $RAR = 2.5$ would be $\tau_A = 15.49$ units. The WDwLSP(n, c) for consumer risk $\beta = 0.1$, and other constants $(\hat{\theta}, \frac{t_{qA}}{t^0_{qA}}, \delta_0, AF, RAR) = (2, 1.75, 0.8, 2, 2.5)$, are (23,8) as per Table 3. The sample inspection shows that two items have failed. Since $d = 2 < c = 8$, the lot is accepted.

Use of previous lot inspection data and use of EWMA or Modified EWMA statistic to judge a lot is advisable because the sample size increases marginally with an increase in the acceleration factor (See Tables 1-3).

5.1.4. WDwALSP with EWMA and modified EWMA

For consumer risk $\beta = 0.1$, the value of (n, c) for $(\hat{\theta}, \frac{t_{qA}}{t^0_{qA}}, \delta_0, AF, RAR, \lambda) = (2, 1.75, 0.8, 2, 2.5, 0.6)$ from the Table 9 is $(n = 11, c = 4)$

Since the history of data is not available, we have generated four Weibull lifetimes samples of size 11 having shape parameter $\hat{\theta} = 2.1014$ and $\hat{\sigma} = 0.0122$ in R. They are,

20.84, 29.57, 19.76, **7.63**, 26.32, **11.64**, 45.57, 42.43, 66.55, 27.68, 38.695
62.44, 21.39, 22.30, **12.29**, 69.64, 27.7, 16.46, 24.62, 18.40, 39.91, 50.43
69.82, **8.91**, 43.50, 44.01, **11.91**, **13.47**, 38.69, 39.59, 33.165, 42.65, 35.24
39.02, 16.24, 40.64, **5.89**, 25.60, 29.31, 18.97, 43.48, 20.44, 16.01, 23.21

Real-life data: 8.94, 14.46, 16.5, 20.76, 21.06, 22.8. 24.4, 25.92, 25.98, 27.06, 27.78

The inspection of the past 4 lots results in the following number of failures from 5 items using (2) **we have:** $\boldsymbol{d_1 = 2,\ d_2 = 1,\ d_3 = 3,\ d_4 = 1}$ while the current lot is having $\boldsymbol{d_5 = 2}$. As per (12), the corresponding EWMA statistics values are $\boldsymbol{EWMA_1 = 2,\ EWMA_2 = 1.40,\ EWMA_3 = 2.36,\ EWMA_4 = 1.54}$, and $\boldsymbol{EWMA_5 = 1.81}$, respectively. Since $\boldsymbol{EWMA_5 = 1.8176 < c = 4}$, the decision would be a current lot is accepted.

For the same set of constraints, WDwA with Modified EWMA from Table 18 the value of (n, c) is (8,3). Also, the inspection of the past 4 lots results in the following number of failures from 8 items: $\boldsymbol{d_1 = 2,\ d_2 = 1,\ d_3 = 3,\ d_4 = 1}$ while the current lot i.e., the 5^{th} lot is having some failures items as $\boldsymbol{d_5 = 2}$. We calculate as per (17) the Modified EWMA statistics and find them as $\boldsymbol{MEWMA_1 = 2,\ MEWMA_2 = 1.7,\ MEWMA_3 = 1.88,\ MEWMA_4 = 1.952}$, and $\boldsymbol{MEWMA_5 = 1.6808}$ respectively. Since $\boldsymbol{MEWMA_5 = 1.6808 < c = 3}$, the decision is once again accepted. 11.5

Table 1: Summary: Comparison of ball bearing data by considering with and without acceleration

	Acceleration Factor	**50^{th} Quantile**	**Specified Quantile**	**Termination Time**	**Table Values**	**Conclusion**
WDwLSP	$AF = 2$	$t_{qA} = 33.9$	$t^0_{qA} = 19.37$	$\tau_A = 15.496$	(23,8)	Accept the Lot
WDwALSP with EMMA					(11,4)	
WDwALSP with Modified EMMA					(8,3)	
WDLSP	$AF = 1$	$t_{qA} = 67.80$	$t^0_q = 38.7428$	$t_0 = 30.9942$	(12,2)	Accept the Lot
WDLSP with EMMA					(6,1)	Reject the Lot
WDLSP with Modified EMMA					(5,1)	Reject the Lot

Murthy et. al. (2004) represent data from 30 observations about the time between failures for the repairable item. **1.43, 0.11, 0.71, 0.77, 2.63, 1.49, 3.46, 2.46, 0.59, 0.74, 1.23, 0.94, 4.36, 0.40, 1.74, 4.73, 2.23,0.45, 0.70, 1.06, 1.46, 0.30, 1.82, 2.37, 0.63, 1.23, 1.24, 1.97, 1.86, 1.17**

Arranging the data in ascending order

0.11, 0.3, 0.4, 0.45, 0.59, 0.63, 0.7, 0.71, 0.74, 0.77, 0.94, 1.06, 1.17, 1.23, 1.23, 1.24, 1.43, 1.46, 1.49, 1.74, 1.82, 1.86, 1.97, 2.23, 2.37, 2.46, 2.63, 3.46, 4.36, 4.73

The maximum likelihood estimators of shape and scale parameters of the Weibull distribution for the number of million revolutions before failure are respectively $\hat{\theta} = 1.4633 \cong 1.5$ and $\hat{\sigma} = 0.5848$. Further, $AIC = 83.8207$, KS-test value $D = 0.074869$, and the corresponding p-value is 0.996 which is greater than 0.05, suggesting that the Weibull distribution fits well with the data.

Transformed Original Data for the proposed plan by taking $\boldsymbol{AF = 2.5}$ $\left(\because \boldsymbol{AF} = \frac{t_U}{t_A}\right)$

0.57, 0.04, 0.28, 0.31, 1.05, 0.60, 1.38, 0.98, 0.24, 0.30, 0.49, 0.38, 1.74, 0.16, 0.70, 1.89, 0.89, 0.18, 0.28, 0.42, 0.58, 0.12, 0.73, 0.95, 0.25, 0.49, 0.50, 0.79, 0.74, 0.47

Arranging data in ascending order:

0.04, 0.12, 0.16, 0.18, 0.24, 0.25, 0.28, 0.28, 0.30, 0.31, 0.38, 0.42, 0.47, 0.49, 0.49, 0.50, 0.57, 0.58, 0.60, 0.70, 0.73, 0.74, 0.79, 0.89, 0.95, 0.98, 1.05, 1.38, 1.74, 1.89

The accelerated 50^{th} quantile of failure time is $t_{0.5A} = 0.495$ with shape parameter $\hat{\theta} = 1.5$ and let us take the ratio $\frac{\sigma_A}{\sigma_U} = RAR > AF \Rightarrow 3.0 > 2.5$

To establish the ratio $\frac{t_{qA}}{t^0_{qA}} > 1$, we take $\frac{t_{qA}}{t^0_{qA}} = 1.75 \Rightarrow t^0_{qA} = 0.2828$, also we take termination ratio $\delta_0 = 0.8$

$\tau_A = \delta_0 t^0_{qA} = 0.8(0.2828) = 0.2262$

Therefore, the total number of failed items in the given lot is $d = 4$. For consumer risk $\beta = 0.25$, the value of (n, c) for $(\hat{\theta}, \frac{t_{qA}}{(t^0_{qA}}, \delta_0, AF, RAR) = (1.5, 1.75, 0.8, 2.5, 3)$, from Table 2 is (25,10).

Here, $d = 4 < c = 10$ $(d < c)$, Hence, the lot is accepted.

5.1.5. WDwALSP with EWMA and modified EWMA

For consumer risk $\beta = 0.25$,and $(\hat{\theta}, \frac{t_{qA}}{t^0_{qA}}, \delta_0, AF, RAR, \delta) = (1.5, 1.75, 0.8, 2.5, 3, 0.8)$ the value of (n, c) from the Table 7 is (15,6).

To create history, we simulated the previous four samples (using R Language) each of size 15 having shape parameters 1.5 and scale parameter 0.5848 taking AF = 2.5. They are shown below along with the real-life data set.

0.17, 0.17, 0.82, 0.56, 1.16, **0.16**, 0.76, **0.17, 0.08,** 1.55, **0.19**, 0.33, 1.31, 0.51, 0.82
0.21, 0.54, 0.23, **0.04, 0.16,** 1.30, 1.06, 0.34, 0.78, 0.63, **0.18**, 0.65, 1.16, 0.24, 0.41
0.61, 0.27, 0.35, 1.34, 1.67, 0.36, 0.99, 0.84, 1.02, 0.63, 0.34, 0.58, 0.95, 0.62, 0.63
1.50, 0.62, 0.35, 1.67, **0.18**, 0.67, 0.54, **0.13**, 0.39, **0.20**, 0.60, **0.22**, 0.78, **0.17**, **0.04**

0.57, **0.04**, 0.28, 0.31, 1.05, 0.60, 1.38, 0.98, 0.24, 0.30, 0.49, 0.38, 1.74, **0.16**, 0.70

The inspection of the past 4 lots results in the following number of failures from 15 items using (2) we have $\boldsymbol{d_1 = 6,\ d_2 = 4,\ d_3 = 0,\ d_4 = 6}$ while the current lot is having $\boldsymbol{d_5 = 2}$. The EWMA statistics using (12) results in values 6,4.4,0.88, 4.97, and 2.59 respectively, leading to the decision to accept the lot, since $\boldsymbol{EWMA_5 = 2.59 < c = 6}$.

Under the same set of constants, the value of (n, c) for WDwALSP with Modified EWMA from the Table 16 is (10,4). Consequently, an inspection of the first 10 items from the past 4 lots and the current lot gives $\boldsymbol{d_1 = 5,\ d_2 = 3,\ d_3 = 0,\ d_4 = 3}$ and $\boldsymbol{d_5 = 1}$. The resultant Modified EWMA statistics using (17) for each lot are respectively, 5, 4.2, 2.04, 1.61, and 1.92. The lot is accepted as $\boldsymbol{MEWMA_5 = 1.92 < c = 4}$.

6. Conclusion

In this paper, LSPs are proposed based on Weibull distribution for establishing a quantile life of a lot rather quicker than the usual inspection time, which is achieved by testing the items under constant stress calledacceleration factor.More the acceleration, the lesser the inspection times but the marginally higher the sample size.Further, usage of the previous history of the lot reduces the sample size of inspection with acceleration compared to without acceleration and lot history. Hence,under acceleration, plans with EWMA or modified EWMA are more economical. Readymade tables are given for the use of the plan in industries manufacturing durable items. Theplans have straightforward extensions for other lifetime distributions as well as other time-censoring schemes.

Acknowledgments

We thank anonymous referees for their useful suggestions and comments.

References

Aslam, M., Balamurali, S., Jun, C.-H., and Meer, A. (2017). Time-truncated attribute sampling plans using ewma for weibull and burr type x distributions. *Communications in Statistics-Simulation and Computation*, **46**, 4173–4184.

Aslam, M., Balamurali, S., Periyasamypandian, J., and Khan, N. (2019). Designing of an attribute control chart based on modified multiple dependent state sampling using accelerated life test under weibull distribution. *Communications in Statistics-Simulation and Computation*, **50**, 902–916.

Aslam, M., Mahmood, Y., Lio, Y., Tsai, T.-R., and Khan, M. A. (2011). Double acceptance sampling plans for burr type xii distribution percentiles under the truncated life test. *Journal of the Operational Research Society*, **63**, 1010–1017.

Baklizi, A. and El Masri, A. E. Q. (2004). Acceptance sampling based on truncated life tests in the birnbaum saunders model. *Society of Risk Analysis*, **24**, 1453–1457.

Balakrishnan, N., Leiva, V., and Lopez, J. (2007). Acceptance sampling plans from truncated life tests based on the generalized birnbaum–saunders distribution. *Communications in Statistics-Simulation and Computation*, **36**, 643–656.

Divecha, J. and Raykundaliya, D. P. (2018). Cost-effective life test acceptance sampling plans for generalized exponential distribution. *International Journal of Statistics and Reliability Engineering*, **5**, 77–83.

Divecha, J. and Raykundaliya, D. P. (2020). Three economical life test acceptance sampling plans. *Communications in Statistics-Simulation and Computation*, **51**, 3305–3323.

Dodge, H. F. and Romig, H. G. (1941). Single sampling and double sampling inspection tables. *The Bell System Technical Journal*, **20**, 1–61.

Epstein, B. (1954). Truncated life tests in the exponential case. *The Annals of Mathematical Statistics*, **25**, 555–564.

Escobar, L. A. and Meeker, W. Q. (2006). A review of accelerated test models. *Statistical science*, **21**, 552–577.

Goode, H. P. (1961). Sampling procedures and tables for life and reliability testing based on the weibull distribution:(mean life criterion). Technical report.

Goode, H. P. and Kao, J. H. (1960). Sampling plans based on the weibull distribution. Technical report.

Gupta, S. S. (1962). Life test sampling plans for normal and lognormal distributions. *Technometrics*, **4**, 151–175.

Kaviyarasu, V. and Fawaz, P. (2017). Certain studies on acceptance sampling plans for percentiles based on the modified weibull distribution. *International Journal of Statistics and Systems*, **12**, 343–354.

Khan, K. and Alqarni, A. (2020). A group acceptance sampling plan using mean lifetime as a quality parameter for inverse weibull distribution. *Advances and Application in Statistics*, **64**, 237–249.

Khan, N., Aslam, M., and Jun, C.-H. (2016). Design of a control chart using a modified ewma statistic. *Quality and Reliability Engineering International*, **33**, 1095–1104.

Lin, D. and Chiu, W. (1995). An economic accelerated life test acceptance sampling plan. *International Journal of Reliability, Quality and Safety Engineering*, **2**, 49–59.

Lio, Y., Tsai, T.-R., and Wu, S.-J. (2009). Acceptance sampling plans from truncated life tests based on the birnbaum–saunders distribution for percentiles. *Communications in Statistics-Simulation and Computation*, **39**, 119–136.

Malathi, D. and Muthulakshmi, S. (2016). Truncated life test acceptance sampling plans assuring percentile life under gompertz distribution. *IOSR-Journal of Mathematics*, **12**, 27–32.

Montgomery, D. C. (2001). *Introduction To Statistical Quality Control, Ed, IV*. John Wiley and Sons, Inc.

Nelson, W. B. (2009). *Accelerated Testing: Statistical Models, Test Plans, and Data Analysis*. John Wiley & Sons, NewYork.

Patel, A. K. and Divecha, J. (2011). Modified exponentially weighted moving average (ewma) control chart for an analytical process data. *Journal of Chemical Engineering and Materials Science*, **2**, 12–20.

Pradeepa Veerakumari, K. and Ponneeswari, P. (2017). Designing of acceptance double sampling plan for life test based on percentiles of exponentiated rayleigh distribution. *International Journal of Statistics and Systems*, **12**, 475–484.

Rao, B. S., Kumar, C. S., and Rosaiah, K. (2013). Acceptance sampling plans from life tests based on percentiles of half normal distribution. *Journal of Quality and Reliability Engineering*, **2013**, 1–7.

Rao, G. S. (2013). Acceptance sampling plans from truncated life tests based on the marshall–olkin extended exponential distribution for percentiles. *Brazilian Journal of Probability and Statistics*, **27**, 117–132.

Rao, G. S., Kantam, R., Rosaiah, K., and Reddy, J. P. (2012). Acceptance sampling plans for percentiles based on the inverse rayleigh distribution. *Electronic Journal of Applied Statistical Analysis*, **5**, 164–177.

Rao, P. R. and Rao, G. S. (2013). Percentiles based construction of acceptance sampling plans for the truncated type-1 generalized logistic distribution. *American Journal of Mathematics and Statistics*, **3**, 194–203.

Rao, S. G. and Kantam, R. R. L. (2010). Acceptance sampling plans from truncated life tests based on the log-logistic distributions for percentiles. *Economic Quality Control*, **27**, 153–167.

Raykundaliya, D. P., Christian, S., and Divecha, J. (2022). Acceptance sampling plans for export quality manufacturing having weibull lifetimes. *International Journal of Statistics and Reliability Engineering*, **9**, 130–149.

Roberts, S. (1959). Control chart tests based on geometric moving averages. *Technometrics*, **1**, 239–250.

Rosaiah, K. and Kantam, R. R. L. (2005). Acceptance sampling based on the inverse rayleigh distribution. *Economic Quality Control*, **20**, 277–286.

Singh, S., Tripathi, Y. M., and Jun, C. H. (2015). Sampling plans based on truncated life test for a generalized inverted exponential distribution. *Industrial Engineering & Management Systems*, **14**, 183–195.

Singh, S., Tripathi, Y. M., and Jun, C. H. (2017). Acceptance sampling plans for inverse weibull distribution based on truncated life test. *Life Cycle Reliability and Safety Engineering*, **6**, 169–178.

Tsai, T.-R. and Wu, S.-J. (2006). Acceptance sampling based on truncated life tests for generalized rayleigh distribution. *Journal of Applied Statistics*, **33**, 595–600.

Xiaoyang, L., Pengfei, G., and Fuqiang, S. (2015). Acceptance sampling plan of accelerated life testing for lognormal distribution under time-censoring. *Chinese Journal of Aeronautics*, **28**, 814–821.

Table 2: Optimum parameters for WDwA assuring 50$^{\text{th}}$ quantile when shape parameter $\theta = 1.5$, Acceleration Factor ($AF = 1.5, 2, 2.5$) and Acceleration Consumer Ratio ($RA = 2, 2.5, 3$)

		AF = 1.5, RA = 2				**AF = 2, RA = 2.5**				**AF = 2.5, RA = 3**			
β	$\frac{t_{qA}}{t^0_{qA}}$	$\delta_0 = 0.6$		$\delta_0 = 0.8$		$\delta_0 = 0.6$		$\delta_0 = 0.8$		$\delta_0 = 0.6$		$\delta_0 = 0.8$	
		n	c	n	c	n	c	n	c	n	c	n	c
0.25	**1.25**	146	53	106	53	152	51	118	55	160	51	114	51
	1.5	47	16	36	17	51	16	41	18	57	17	38	16
	1.75	28	9	20	9	30	9	24	10	32	9	25	10
	2.0	20	6	14	6	21	6	15	6	22	6	16	6
	2.25	17	5	12	5	15	4	13	5	19	5	13	5
0.1	**1.25**	229	80	174	84	245	79	182	82	261	80	188	81
	1.5	79	25	56	25	82	24	60	25	86	24	63	25
	1.75	44	13	34	14	48	13	34	13	51	13	38	14
	2.0	33	9	23	9	35	9	25	9	37	9	26	9
	2.25	27	7	19	7	29	7	20	7	27	6	21	7
0.05	**1.25**	291	100	214	102	312	99	227	101	325	98	240	102
	1.5	97	30	69	30	109	31	77	31	111	30	78	30
	1.75	57	16	40	16	62	16	43	16	65	16	45	16
	2.0	42	11	29	11	42	10	29	10	44	10	33	11
	2.25	33	8	23	8	35	8	25	8	37	8	26	8
0.01	**1.25**	423	142	308	144	455	141	333	145	482	142	346	144
	1.5	145	43	103	43	158	43	111	43	166	43	116	43
	1.75	83	22	58	22	90	22	65	23	95	22	66	22
	2.0	58	14	40	14	67	15	46	15	67	14	46	14
	2.25	49	11	31	10	53	11	36	11	56	11	38	11

Table 3: Optimum parameters for WDwA assuring 50$^{\text{th}}$ quantile when shape parameter $\theta = 2.0$, Acceleration Factor ($AF = 1.5, 2, 2.5$) and Acceleration Consumer Ratio ($RA = 2, 2.5, 3$)

		AF = 1.5, RA = 2				**AF = 2, RA = 2.5**				**AF = 2.5, RA = 3**			
β	$\frac{t_{qA}}{t^0_{qA}}$	$\delta_0 = 0.6$		$\delta_0 = 0.8$		$\delta_0 = 0.6$		$\delta_0 = 0.8$		$\delta_0 = 0.6$		$\delta_0 = 0.8$	
		n	c	n	c	n	c	n	c	n	c	n	c
0.25	**1.25**	90	29	62	31	100	29	68	31	107	29	70	30
	1.5	34	10	24	11	34	9	24	10	36	9	25	10
	1.75	18	5	14	6	20	5	15	6	22	5	16	6
	2.0	15	4	10	4	13	3	8	3	14	3	11	4
	2.25	12	3	8	3	13	3	8	3	10	2	9	3
0.1	**1.25**	147	45	96	46	160	44	108	47	175	45	112	46
	1.5	52	14	35	15	58	14	36	14	62	14	41	15
	1.75	29	7	20	8	36	8	23	8	35	7	24	8
	2.0	22	5	14	5	25	5	15	5	27	5	17	5
	2.25	19	4	12	4	21	4	13	4	23	4	14	4
0.05	**1.25**	187	56	121	57	208	56	131	56	223	56	142	57
	1.5	66	17	41	17	74	17	48	18	79	17	51	18
	1.75	39	9	24	9	44	9	27	9	47	9	29	9
	2.0	29	6	18	6	33	6	20	6	35	6	21	6
	2.25	22	4	13	4	29	5	17	5	31	5	18	5
0.01	**1.25**	272	79	177	81	307	80	195	81	326	79	205	80
	1.5	98	24	61	24	110	24	70	25	118	24	72	24
	1.75	57	12	37	13	69	13	41	13	74	13	44	13
	2.0	43	8	26	8	49	8	29	8	53	8	31	8
	2.25	36	6	21	6	40	6	24	6	44	6	26	6

Table 4: Optimum parameters for WDwA assuring 50th quantile when shape parameter $\theta = 2.5$, Acceleration Factor ($AF = 1.5, 2, 2.5$) and Acceleration Consumer Ratio ($RA = 2, 2.5, 3$)

		AF = 1.5, RA = 2				**AF = 2, RA = 2.5**				**AF = 2.5, RA = 3**			
β	$\frac{t_{qA}}{t^0_{qA}}$	$\delta_0 = 0.6$		$\delta_0 = 0.8$		$\delta_0 = 0.6$		$\delta_0 = 0.8$		$\delta_0 = 0.6$		$\delta_0 = 0.8$	
		n	c	n	c	n	c	n	c	n	c	n	c
0.25	**1.25**	66	19	40	20	72	18	45	20	83	19	46	19
	1.5	23	6	13	6	27	6	15	6	29	6	16	6
	1.75	13	3	7	3	15	3	8	3	16	3	11	4
	2	10	2	5	2	11	2	6	2	12	2	9	3
	2.25	10	2	5	2	11	2	6	2	12	2	7	2
0.25	**1.25**	108	29	61	29	120	28	71	30	136	29	77	30
	1.5	39	9	22	9	45	9	25	9	50	9	27	9
	1.75	25	5	14	5	29	5	15	5	31	5	17	5
	2	17	3	9	3	20	3	13	4	22	3	14	4
	2.25	13	2	9	3	15	2	11	3	17	2	12	3
0.25	**1.25**	138	36	80	37	159	36	90	37	169	35	95	36
	1.5	51	11	28	11	59	11	32	11	64	11	34	11
	1.75	32	6	17	6	37	6	20	6	41	6	21	6
	2	24	4	13	4	28	4	15	4	31	4	16	4
	2.25	20	3	11	3	23	3	12	3	26	3	13	3
0.25	**1.25**	200	50	114	51	235	51	131	52	257	51	142	52
	1.5	75	15	43	16	87	15	46	15	96	15	50	15
	1.75	48	8	25	8	56	8	29	8	61	8	32	8
	2	35	5	18	5	41	5	21	5	46	5	23	5
	2.25	31	4	16	4	36	4	18	4	40	4	20	4

Table 5: Optimum parameters for WDwAwith EWMA assuring 50th quantile when shape parameter $\theta = 1.5$, Acceleration Factor ($AF = 1.5$) and Acceleration Consumer Ratio ($RA = 2.0$)

β	$\frac{t_{qA}}{t^0_{qA}}$	$\delta_0 = 0.6$										$\delta_0 = 0.8$									
		$\lambda = 0.2$		$\lambda = 0.4$		$\lambda = 0.6$		$\lambda = 0.8$		$\lambda = 1.0$		$\lambda = 0.2$		$\lambda = 0.4$		$\lambda = 0.6$		$\lambda = 0.8$		$\lambda = 1.0$	
		n	c	n	c	n	c	n	c	n	c	n	c	n	c	n	c	n	c	n	c
0.25	**1.25**	22	8	42	15	66	24	99	36	146	53	12	6	28	14	46	23	72	36	106	53
	1.5	6	2	15	5	24	8	35	12	47	16	9	4	13	6	17	8	28	13	36	17
	1.75	6	2	9	3	15	5	19	6	28	9	5	2	7	3	9	4	16	7	20	9
	2	6	2	7	2	10	3	13	4	20	6	5	2	5	2	7	3	12	5	14	6
	2.25	4	1	7	2	7	2	10	3	17	5	3	1	5	2	5	2	7	3	12	5
0.1	**1.25**	32	11	60	21	100	35	152	53	229	80	23	11	50	24	79	38	116	56	174	84
	1.5	13	4	22	7	35	11	53	17	79	25	9	4	16	7	27	12	38	17	56	25
	1.75	7	2	14	4	21	6	31	9	44	13	5	2	10	4	17	7	24	10	34	14
	2	7	2	11	3	15	4	22	6	33	9	5	2	8	3	10	4	18	7	23	9
	2.25	4	1	8	2	12	3	19	5	27	7	3	1	8	3	8	3	13	5	19	7
0.05	**1.25**	35	12	76	26	125	43	195	67	291	100	29	14	59	28	96	46	147	70	214	102
	1.5	13	4	26	8	42	13	65	20	97	30	12	5	21	9	30	13	46	20	69	30
	1.75	7	2	18	5	25	7	39	11	57	16	5	2	10	4	20	8	30	12	40	16
	2	7	2	12	3	19	5	27	7	42	11	5	2	8	3	13	5	19	7	29	11
	2.25	4	1	12	3	16	4	21	5	33	8	3	1	8	3	11	4	17	6	23	8
0.01	**1.25**	54	18	110	37	185	62	283	95	423	142	41	19	77	36	137	64	212	99	308	144
	1.5	17	5	37	11	64	19	98	29	145	43	12	5	29	12	48	20	67	28	103	43
	1.75	11	3	23	6	38	10	57	15	83	22	8	3	16	6	26	10	42	16	58	22
	2	8	2	17	4	25	6	41	10	58	14	6	2	14	5	20	7	29	10	40	14
	2.25	8	2	13	3	22	5	32	7	49	11	6	2	9	3	15	5	22	7	31	10

Table 6: Optimum parameters for WDwAwith EWMA assuring 50th quantile when shape parameter $\theta = 1.5$, Acceleration Factor ($AF = 2.0$) and Acceleration Consumer Ratio ($RA = 2.5$)

β	$\frac{t_{qA}}{t^0_{qA}}$	$\delta_0 = 0.6$										$\delta_0 = 0.8$									
		$\lambda = 0.2$		$\lambda = 0.4$		$\lambda = 0.6$		$\lambda = 0.8$		$\lambda = 1.0$		$\lambda = 0.2$		$\lambda = 0.4$		$\lambda = 0.6$		$\lambda = 0.8$		$\lambda = 1.0$	
		n	c	n	c	n	c	n	c	n	c	n	c	n	c	n	c	n	c	n	c
0.25	1.25	18	6	42	14	69	23	104	35	152	51	15	7	30	14	54	25	77	36	118	55
	1.5	10	3	16	5	25	8	35	11	51	16	7	3	14	6	18	8	25	11	41	18
	1.75	7	2	10	3	17	5	20	6	30	9	5	2	7	3	12	5	17	7	24	10
	2	4	1	7	2	11	3	14	4	21	6	5	2	5	2	8	3	10	4	15	6
	2.25	4	1	7	2	11	3	11	3	15	4	3	1	5	2	5	2	8	3	13	5
0.1	1.25	31	10	62	20	112	36	164	53	245	79	27	12	49	22	80	36	122	55	182	82
	1.5	14	4	24	7	38	11	58	17	82	24	10	4	17	7	29	12	41	17	60	25
	1.75	11	3	15	4	22	6	33	9	48	13	5	2	13	5	18	7	26	10	34	13
	2	4	1	12	3	16	4	24	6	35	9	5	2	8	3	11	4	17	6	25	9
	2.25	4	1	8	2	13	3	17	4	29	7	3	1	6	2	9	3	14	5	20	7
0.05	1.25	38	12	82	26	139	44	208	66	312	99	27	12	61	27	99	44	153	68	227	101
	1.5	14	4	28	8	49	14	74	21	109	31	10	4	20	8	35	14	52	21	77	31
	1.75	11	3	19	5	27	7	42	11	62	16	8	3	13	5	19	7	30	11	43	16
	2	8	2	13	3	21	5	29	7	42	10	6	2	9	3	15	5	23	8	29	10
	2.25	8	2	9	2	17	4	26	6	35	8	3	1	9	3	12	4	18	6	25	8
0.01	1.25	55	17	116	36	200	62	310	96	455	141	39	17	85	37	147	64	223	97	333	145
	1.5	22	6	44	12	70	19	106	29	158	43	18	7	31	12	49	19	75	29	111	43
	1.75	12	3	25	6	41	10	62	15	90	22	11	4	17	6	31	11	43	15	65	23
	2	9	2	18	4	31	7	45	10	67	15	6	2	15	5	21	7	31	10	46	15
	2.25	9	2	15	3	24	5	34	7	53	11	6	2	10	3	17	5	26	8	36	11

Table 7: Optimum parameters for WDwAwith EWMA assuring 50th quantile when shape parameter $\theta = 1.5$, Acceleration Factor ($AF = 2.5$) and Acceleration Consumer Ratio ($RA = 3.0$)

β	$\frac{t_{qA}}{t^0_{qA}}$	$\delta_0 = 0.6$										$\delta_0 = 0.8$									
		$\lambda = 0.2$		$\lambda = 0.4$		$\lambda = 0.6$		$\lambda = 0.8$		$\lambda = 1.0$		$\lambda = 0.2$		$\lambda = 0.4$		$\lambda = 0.6$		$\lambda = 0.8$		$\lambda = 1.0$	
		n	c	n	c	n	c	n	c	n	c	n	c	n	c	n	c	n	c	n	c
0.25	1.25	22	7	44	14	69	22	110	35	160	51	18	8	34	15	54	24	76	34	114	51
	1.5	10	3	17	5	27	8	37	11	57	17	7	3	12	5	19	8	26	11	38	16
	1.75	7	2	11	3	14	4	21	6	32	9	5	2	8	3	10	4	15	6	25	10
	2	4	1	8	2	11	3	15	4	22	6	5	2	5	2	8	3	13	5	16	6
	2.25	4	1	8	2	8	2	12	3	19	5	3	1	5	2	8	3	8	3	13	5
0.1	1.25	36	11	72	22	114	35	173	53	261	80	28	12	51	22	81	35	125	54	188	81
	1.5	11	3	25	7	40	11	61	17	86	24	10	4	18	7	28	11	43	17	63	25
	1.75	8	2	16	4	23	6	35	9	51	13	8	3	11	4	19	7	27	10	38	14
	2	8	2	12	3	17	4	25	6	37	9	6	2	9	3	12	4	20	7	26	9
	2.25	5	1	9	2	13	3	18	4	27	6	3	1	6	2	9	3	15	5	21	7
0.05	1.25	40	12	83	25	146	44	219	66	325	98	33	14	61	26	106	45	160	68	240	102
	1.5	15	4	30	8	48	13	74	20	111	30	13	5	21	8	36	14	52	20	78	30
	1.75	8	2	20	5	32	8	45	11	65	16	8	3	14	5	20	7	31	11	45	16
	2	8	2	13	3	22	5	31	7	44	10	6	2	9	3	15	5	24	8	33	11
	2.25	5	1	13	3	18	4	24	5	37	8	6	2	9	3	13	4	19	6	26	8
0.01	1.25	58	17	122	36	207	61	326	96	482	142	41	17	89	37	149	62	231	96	346	144
	1.5	23	6	46	12	73	19	112	29	166	43	16	6	32	12	54	20	81	30	116	43
	1.75	13	3	26	6	43	10	65	15	95	22	9	3	18	6	30	10	45	15	66	22
	2	9	2	19	4	33	7	47	10	67	14	9	3	13	4	23	7	33	10	46	14
	2.25	9	2	15	3	25	5	36	7	56	11	7	2	11	3	17	5	25	7	38	11

Table 8: Optimum parameters for WDwAwith EWMA assuring 50th quantile when shape parameter $\theta = 1.5$, Acceleration Factor ($AF = 2.5$) and Acceleration Consumer Ratio ($RA = 3.0$)

β	$\frac{t_{qA}}{t^0_{qA}}$	$\delta_0 = 0.6$										$\delta_0 = 0.8$									
		$\lambda = 0.2$		$\lambda = 0.4$		$\lambda = 0.6$		$\lambda = 0.8$		$\lambda = 1.0$		$\lambda = 0.2$		$\lambda = 0.4$		$\lambda = 0.6$		$\lambda = 0.8$		$\lambda = 1.0$	
		n	c	n	c	n	c	n	c	n	c	n	c	n	c	n	c	n	c	n	c
0.25	1.25	16	5	25	8	43	14	62	20	90	29	8	4	16	8	26	13	42	21	62	31
	1.5	7	2	10	3	17	5	24	7	34	10	7	3	9	4	11	5	15	7	24	11
	1.75	4	1	7	2	11	3	14	4	18	5	5	2	5	2	7	3	9	4	14	6
	2	4	1	4	1	8	2	11	3	15	4	3	1	5	2	5	2	7	3	10	4
	2.25	4	1	4	1	4	1	8	2	12	3	3	1	3	1	5	2	5	2	8	3
0.1	1.25	23	7	39	12	62	19	98	30	147	45	15	7	25	12	44	21	67	32	96	46
	1.5	11	3	15	4	26	7	37	10	52	14	7	3	12	5	14	6	23	10	35	15
	1.75	4	1	12	3	16	4	21	5	29	7	5	2	5	2	10	4	13	5	20	8
	2	4	1	9	2	9	2	17	4	22	5	3	1	5	2	8	3	11	4	14	5
	2.25	4	1	5	1	9	2	14	3	19	4	3	1	3	1	6	2	9	3	12	4
0.05	1.25	27	8	50	15	80	24	127	38	187	56	15	7	32	15	53	25	83	39	121	57
	1.5	11	3	19	5	31	8	46	12	66	17	7	3	12	5	19	8	29	12	41	17
	1.75	8	2	13	3	18	4	26	6	39	9	5	2	8	3	11	4	16	6	24	9
	2	5	1	9	2	14	3	19	4	29	6	3	1	6	2	9	3	12	4	18	6
	2.25	5	1	9	2	11	2	16	3	22	4	3	1	6	2	9	3	10	3	13	4
0.01	1.25	31	9	72	21	117	34	186	54	272	79	22	10	46	21	79	36	118	54	177	81
	1.5	12	3	28	7	41	10	66	16	98	24	10	4	18	7	28	11	43	17	61	24
	1.75	9	2	18	4	28	6	38	8	57	12	6	2	11	4	17	6	23	8	37	13
	2	9	2	11	2	21	4	31	6	43	8	6	2	9	3	13	4	19	6	26	8
	2.25	6	1	11	2	17	3	24	4	36	6	4	1	7	2	10	3	14	4	21	6

Table 9: Optimum parameters for WDwAwith EWMA assuring 50th quantile when shape parameter $\theta = 2.0$, Acceleration Factor ($AF = 2.0$) and Acceleration Consumer Ratio ($RA = 2.5$)

β	$\frac{t_{qA}}{t^0_{qA}}$	$\delta_0 = 0.6$										$\delta_0 = 0.8$									
		$\lambda = 0.2$		$\lambda = 0.4$		$\lambda = 0.6$		$\lambda = 0.8$		$\lambda = 1.0$		$\lambda = 0.2$		$\lambda = 0.4$		$\lambda = 0.6$		$\lambda = 0.8$		$\lambda = 1.0$	
		n	c	n	c	n	c	n	c	n	c	n	c	n	c	n	c	n	c	n	c
0.25	1.25	14	4	28	8	45	13	69	20	100	29	11	5	20	9	31	14	46	21	68	31
	1.5	8	2	11	3	15	4	26	7	34	9	5	2	7	3	12	5	17	7	24	10
	1.75	4	1	8	2	12	3	16	4	20	5	3	1	5	2	5	2	10	4	15	6
	2	4	1	5	1	9	2	13	3	13	3	3	1	3	1	5	2	8	3	8	3
	2.25	4	1	5	1	9	2	9	2	13	3	3	1	3	1	5	2	6	2	8	3
0.1	1.25	22	6	40	11	69	19	109	30	160	44	14	6	30	13	46	20	71	31	108	47
	1.5	8	2	17	4	25	6	41	10	58	14	5	2	13	5	18	7	26	10	36	14
	1.75	5	1	9	2	14	3	23	5	36	8	3	1	8	3	11	4	14	5	23	8
	2	5	1	9	2	14	3	19	4	25	5	3	1	6	2	9	3	12	4	15	5
	2.25	5	1	6	1	11	2	16	3	21	4	3	1	6	2	7	2	10	3	13	4
0.05	1.25	26	7	52	14	93	25	141	38	208	56	19	8	35	15	61	26	89	38	131	56
	1.5	13	3	21	5	34	8	52	12	74	17	8	3	13	5	21	8	32	12	48	18
	1.75	5	1	14	3	20	4	30	6	44	9	3	1	9	3	12	4	18	6	27	9
	2	5	1	10	2	16	3	22	4	33	6	3	1	7	2	10	3	13	4	20	6
	2.25	5	1	10	2	12	2	18	3	29	5	3	1	7	2	7	2	11	3	17	5
0.01	1.25	38	10	77	20	134	35	207	54	307	80	24	10	53	22	84	35	130	54	195	81
	1.5	14	3	32	7	46	10	74	16	110	24	11	4	20	7	31	11	45	16	70	25
	1.75	10	2	20	4	31	6	43	8	69	13	6	2	13	4	19	6	29	9	41	13
	2	6	1	17	3	23	4	35	6	49	8	6	2	10	3	14	4	21	6	29	8
	2.25	6	1	13	2	19	3	27	4	40	6	4	1	8	2	12	3	16	4	24	6

Table 10: Optimum parameters for WDwAwith EWMA assuring 50th quantile when shape parameter $\theta = 2.0$, Acceleration Factor ($AF = 2.5$) and Acceleration Consumer Ratio ($RA = 3.0$)

β	$\frac{t_{qA}}{t^0_{qA}}$	$\delta_0 = 0.6$										$\delta_0 = 0.8$									
		$\lambda = 0.2$		$\lambda = 0.4$		$\lambda = 0.6$		$\lambda = 0.8$		$\lambda = 1.0$		$\lambda = 0.2$		$\lambda = 0.4$		$\lambda = 0.6$		$\lambda = 0.8$		$\lambda = 1.0$	
		n	c	n	c	n	c	n	c	n	c	n	c	n	c	n	c	n	c	n	c
0.25	1.25	15	4	30	8	48	13	70	19	107	29	12	5	21	9	33	14	49	21	70	30
	1.5	4	1	12	3	16	4	24	6	36	9	5	2	8	3	10	4	15	6	25	10
	1.75	4	1	9	2	13	3	13	3	22	5	3	1	6	2	8	3	11	4	16	6
	2	4	1	5	1	9	2	13	3	14	3	3	1	3	1	6	2	9	3	11	4
	2.25	4	1	5	1	5	1	10	2	10	2	3	1	3	1	3	1	6	2	9	3
0.1	1.25	23	6	47	12	74	19	113	29	175	45	17	7	32	13	51	21	78	32	112	46
	1.5	9	2	18	4	27	6	40	9	62	14	8	3	11	4	19	7	27	10	41	15
	1.75	5	1	10	2	15	3	25	5	35	7	3	1	6	2	12	4	15	5	24	8
	2	5	1	10	2	11	2	21	4	27	5	3	1	6	2	10	3	10	3	17	5
	2.25	5	1	6	1	11	2	17	3	23	4	3	1	6	2	7	2	10	3	14	4
0.05	1.25	28	7	56	14	96	24	151	38	223	56	20	8	40	16	62	25	92	37	142	57
	1.5	13	3	23	5	37	8	55	12	79	17	8	3	14	5	23	8	34	12	51	18
	1.75	10	2	15	3	21	4	32	6	47	9	6	2	9	3	13	4	19	6	29	9
	2	6	1	11	2	17	3	24	4	35	6	4	1	7	2	10	3	14	4	21	6
	2.25	6	1	11	2	13	2	19	3	31	5	4	1	7	2	8	2	12	3	18	5
0.01	1.25	37	9	86	21	144	35	222	54	326	79	23	9	54	21	92	36	138	54	205	80
	1.5	15	3	34	7	54	11	79	16	118	24	9	3	18	6	33	11	48	16	72	24
	1.75	11	2	22	4	34	6	51	9	74	13	7	2	13	4	20	6	31	9	44	13
	2	11	2	18	3	25	4	38	6	53	8	4	1	11	3	15	4	23	6	31	8
	2.25	7	1	14	2	21	3	29	4	44	6	4	1	8	2	12	3	17	4	26	6

Table 11: Optimum parameters for WDwAwith EWMA assuring 50th quantile when shape parameter $\theta = 2.5$, Acceleration Factor ($AF = 1.5$) and Acceleration Consumer Ratio ($RA = 2.0$)

β	$\frac{t_{qA}}{t^0_{qA}}$	$\delta_0 = 0.6$										$\delta_0 = 0.8$									
		$\lambda = 0.2$		$\lambda = 0.4$		$\lambda = 0.6$		$\lambda = 0.8$		$\lambda = 1.0$		$\lambda = 0.2$		$\lambda = 0.4$		$\lambda = 0.6$		$\lambda = 0.8$		$\lambda = 1.0$	
		n	c	n	c	n	c	n	c	n	c	n	c	n	c	n	c	n	c	n	c
0.25	1.25	11	3	21	6	28	8	45	13	66	19	6	3	10	5	18	9	28	14	40	20
	1.5	4	1	8	2	12	3	16	4	23	6	4	2	7	3	9	4	11	5	13	6
	1.75	4	1	8	2	8	2	12	3	13	3	3	1	5	2	5	2	7	3	7	3
	2	4	1	5	1	5	1	9	2	10	2	3	1	3	1	5	2	5	2	5	2
	2.25	4	1	5	1	5	1	5	1	10	2	3	1	3	1	3	1	5	2	5	2
0.1	1.25	15	4	30	8	49	13	71	19	108	29	11	5	17	8	32	15	42	20	61	29
	1.5	8	2	13	3	18	4	26	6	39	9	5	2	7	3	12	5	17	7	22	9
	1.75	5	1	9	2	10	2	15	3	25	5	3	1	5	2	8	3	11	4	14	5
	2	5	1	6	1	10	2	12	2	17	3	3	1	3	1	6	2	6	2	9	3
	2.25	5	1	6	1	6	1	12	2	13	2	3	1	3	1	6	2	6	2	9	3
0.05	1.25	19	5	38	10	61	16	92	24	138	36	11	5	24	11	37	17	54	25	80	37
	1.5	9	2	14	3	23	5	36	8	51	11	5	2	10	4	15	6	20	8	28	11
	1.75	5	1	10	2	16	3	22	4	32	6	3	1	6	2	9	3	12	4	17	6
	2	5	1	6	1	12	2	18	3	24	4	3	1	6	2	6	2	9	3	13	4
	2.25	5	1	6	1	12	2	14	2	20	3	3	1	4	1	6	2	7	2	11	3
0.01	1.25	24	6	52	13	88	22	136	34	200	50	18	8	29	13	49	22	76	34	114	51
	1.5	10	2	20	4	35	7	50	10	75	15	8	3	13	5	19	7	27	10	43	16
	1.75	6	1	12	2	23	4	35	6	48	8	3	1	9	3	12	4	16	5	25	8
	2	6	1	12	2	19	3	27	4	35	5	3	1	7	2	10	3	14	4	18	5
	2.25	6	1	8	1	15	2	22	3	31	4	3	1	4	1	8	2	12	3	16	4

Table 12: Optimum parameters for WDwAwith EWMA assuring 50th quantile when shape parameter $\theta = 2.5$, Acceleration Factor ($AF = 2.0$) and Acceleration Consumer Ratio ($RA = 2.5$)

β	$\frac{t_{qA}}{t^0_{qA}}$	$\delta_0 = 0.6$										$\delta_0 = 0.8$									
		$\lambda = 0.2$		$\lambda = 0.4$		$\lambda = 0.6$		$\lambda = 0.8$		$\lambda = 1.0$		$\lambda = 0.2$		$\lambda = 0.4$		$\lambda = 0.6$		$\lambda = 0.8$		$\lambda = 1.0$	
		n	c	n	c	n	c	n	c	n	c	n	c	n	c	n	c	n	c	n	c
0.25	1.25	8	2	20	5	32	8	48	12	72	18	7	3	16	7	20	9	29	13	45	20
	1.5	5	1	9	2	13	3	18	4	27	6	5	2	5	2	10	4	10	4	15	6
	1.75	5	1	5	1	10	2	10	2	15	3	3	1	3	1	5	2	8	3	8	3
	2	5	1	5	1	6	1	10	2	11	2	3	1	3	1	3	1	6	2	6	2
	2.25	5	1	5	1	6	1	6	1	11	2	3	1	3	1	3	1	3	1	6	2
0.1	1.25	17	4	30	7	56	13	86	20	120	28	12	5	19	8	31	13	50	21	71	30
	1.5	5	1	15	3	20	4	30	6	45	9	5	2	8	3	11	4	19	7	25	9
	1.75	5	1	11	2	12	2	18	3	29	5	3	1	6	2	9	3	12	4	15	5
	2	5	1	6	1	12	2	13	2	20	3	3	1	4	1	7	2	7	2	13	4
	2.25	5	1	6	1	7	1	13	2	15	2	3	1	4	1	4	1	7	2	11	3
0.05	1.25	22	5	44	10	71	16	106	24	159	36	12	5	27	11	39	16	61	25	90	37
	1.5	10	2	16	3	27	5	42	8	59	11	6	2	11	4	17	6	23	8	32	11
	1.75	6	1	12	2	18	3	25	4	37	6	3	1	7	2	10	3	13	4	20	6
	2	6	1	7	1	14	2	20	3	28	4	3	1	4	1	7	2	11	3	15	4
	2.25	6	1	7	1	14	2	16	2	23	3	3	1	4	1	7	2	8	2	12	3
0.01	1.25	32	7	60	13	101	22	157	34	235	51	18	7	35	14	58	23	88	35	131	52
	1.5	12	2	23	4	40	7	58	10	87	15	6	2	15	5	21	7	33	11	46	15
	1.75	7	1	14	2	27	4	40	6	56	8	6	2	10	3	14	4	21	6	29	8
	2	7	1	14	2	22	3	31	4	41	5	4	1	8	2	12	3	16	4	21	5
	2.25	7	1	9	1	17	2	26	3	36	4	4	1	5	1	9	2	13	3	18	4

Table 13: Optimum parameters for WDwAwith EWMA assuring 50th quantile when shape parameter $\theta = 2.5$, Acceleration Factor ($AF = 2.5$) and Acceleration Consumer Ratio ($RA = 3.0$)

β	$\frac{t_{qA}}{t^0_{qA}}$	$\delta_0 = 0.6$										$\delta_0 = 0.8$									
		$\lambda = 0.2$		$\lambda = 0.4$		$\lambda = 0.6$		$\lambda = 0.8$		$\lambda = 1.0$		$\lambda = 0.2$		$\lambda = 0.4$		$\lambda = 0.6$		$\lambda = 0.8$		$\lambda = 1.0$	
		n	c	n	c	n	c	n	c	n	c	n	c	n	c	n	c	n	c	n	c
0.25	1.25	13	3	22	5	35	8	57	13	83	19	10	4	15	6	22	9	34	14	46	19
	1.5	5	1	10	2	15	3	20	4	29	6	3	1	8	3	8	3	13	5	16	6
	1.75	5	1	6	1	10	2	11	2	16	3	3	1	3	1	6	2	9	3	11	4
	2	5	1	6	1	6	1	11	2	12	2	3	1	3	1	3	1	6	2	9	3
	2.25	5	1	6	1	6	1	7	1	12	2	3	1	3	1	3	1	6	2	7	2
0.1	1.25	19	4	37	8	61	13	89	19	136	29	13	5	23	9	36	14	54	21	77	30
	1.5	10	2	16	3	22	4	33	6	50	9	3	1	9	3	12	4	18	6	27	9
	1.75	6	1	12	2	13	2	24	4	31	5	3	1	6	2	7	2	10	3	17	5
	2	6	1	7	1	13	2	15	2	22	3	3	1	4	1	7	2	10	3	14	4
	2.25	6	1	7	1	8	1	15	2	17	2	3	1	4	1	4	1	8	2	12	3
0.05	1.25	24	5	48	10	77	16	116	24	169	35	13	5	26	10	42	16	66	25	95	36
	1.5	11	2	18	3	29	5	46	8	64	11	6	2	12	4	16	5	25	8	34	11
	1.75	7	1	13	2	20	3	27	4	41	6	4	1	7	2	11	3	14	4	21	6
	2	7	1	8	1	15	2	22	3	31	4	4	1	4	1	8	2	12	3	16	4
	2.25	7	1	8	1	15	2	17	2	26	3	4	1	4	1	8	2	9	2	13	3
0.01	1.25	30	6	70	14	111	22	171	34	257	51	19	7	38	14	60	22	93	34	142	52
	1.5	13	2	25	4	44	7	64	10	96	15	9	3	16	5	23	7	36	11	50	15
	1.75	8	1	21	3	29	4	44	6	61	8	4	1	8	2	15	4	20	5	32	8
	2	8	1	16	2	24	3	34	4	46	5	4	1	8	2	13	3	17	4	23	5
	2.25	8	1	10	1	19	2	29	3	40	4	4	1	5	1	10	2	15	3	20	4

Table 14: Optimum parameters for WDwAwith Modified EWMA assuring 50th quantile when shape parameter $\theta = 1.5$, Acceleration Factor ($AF = 1.5$) and Acceleration Consumer Ratio ($RA = 2$)

β	$\frac{t_{qA}}{t^0_{qA}}$	$\delta_0 = 0.6$								$\delta_0 = 0.8$							
		$\lambda = 0.2$		$\lambda = 0.4$		$\lambda = 0.6$		$\lambda = 0.8$		$\lambda = 0.2$		$\lambda = 0.4$		$\lambda = 0.6$		$\lambda = 0.8$	
		n	c	n	c	n	c	n	c	n	c	n	c	n	c	n	c
0.25	1.25	22	8	33	12	44	16	58	21	12	6	22	11	32	16	44	22
	1.5	6	2	12	4	18	6	21	7	9	4	11	5	13	6	17	8
	1.75	6	2	9	3	12	4	13	4	5	2	7	3	9	4	9	4
	2	6	2	7	2	10	3	10	3	5	2	5	2	7	3	7	3
	2.25	4	1	7	2	7	2	7	2	3	1	5	2	5	2	5	2
0.1	1.25	26	9	49	17	75	26	97	34	23	11	37	18	56	27	72	35
	1.5	12	4	19	6	25	8	35	11	9	4	16	7	18	8	25	11
	1.75	7	2	10	3	17	5	17	5	5	2	10	4	12	5	12	5
	2	7	2	10	3	11	3	15	4	5	2	5	2	8	3	10	4
	2.25	4	1	8	2	8	2	12	3	3	1	5	2	8	3	8	3
0.05	1.25	32	11	61	21	90	31	122	42	23	11	44	21	65	31	86	41
	1.5	13	4	23	7	32	10	39	12	9	4	16	7	23	10	30	13
	1.75	7	2	14	4	18	5	25	7	5	2	10	4	15	6	20	8
	2	7	2	11	3	15	4	19	5	5	2	8	3	8	3	13	5
	2.25	4	1	8	2	12	3	16	4	3	1	6	2	8	3	11	4
0.01	1.25	48	16	89	30	128	43	170	57	32	15	62	29	92	43	124	58
	1.5	17	5	34	10	44	13	61	18	12	5	24	10	31	13	43	18
	1.75	11	3	19	5	30	8	34	9	8	3	13	5	21	8	26	10
	2	8	2	16	4	20	5	25	6	6	2	11	4	14	5	17	6
	2.25	8	2	13	3	17	4	22	5	6	2	9	3	12	4	15	5

Table 15: Optimum parameters for WDwAwith Modified EWMA assuring 50th quantile when shape parameter $\theta = 1.5$, Acceleration Factor ($AF = 2.0$) and Acceleration Consumer Ratio ($RA = 2.5$)

β	$\frac{t_{qA}}{t^0_{qA}}$	$\delta_0 = 0.6$								$\delta_0 = 0.8$							
		$\lambda = 0.2$		$\lambda = 0.4$		$\lambda = 0.6$		$\lambda = 0.8$		$\lambda = 0.2$		$\lambda = 0.4$		$\lambda = 0.6$		$\lambda = 0.8$	
		n	c	n	c	n	c	n	c	n	c	n	c	n	c	n	c
0.25	1.25	18	6	33	11	51	17	66	22	13	6	26	12	39	18	47	22
	1.5	10	3	13	4	16	5	22	7	7	3	9	4	14	6	16	7
	1.75	7	2	7	2	10	3	14	4	5	2	7	3	7	3	12	5
	2	4	1	7	2	7	2	11	3	5	2	5	2	5	2	8	3
	2.25	4	1	4	1	7	2	8	2	3	1	5	2	5	2	5	2
0.1	1.25	28	9	53	17	81	26	99	32	20	9	40	18	60	27	73	33
	1.5	10	3	17	5	27	8	34	10	10	4	12	5	22	9	24	10
	1.75	7	2	11	3	15	4	22	6	5	2	8	3	13	5	16	6
	2	4	1	8	2	12	3	16	4	5	2	8	3	8	3	11	4
	2.25	4	1	8	2	12	3	12	3	3	1	6	2	6	2	9	3
0.05	1.25	38	12	63	20	98	31	126	40	27	12	52	23	72	32	92	41
	1.5	14	4	25	7	35	10	46	13	10	4	20	8	25	10	32	13
	1.75	8	2	15	4	23	6	27	7	8	3	11	4	16	6	19	7
	2	8	2	12	3	16	4	20	5	6	2	9	3	11	4	14	5
	2.25	8	2	9	2	13	3	17	4	3	1	6	2	9	3	12	4
0.01	1.25	52	16	97	30	142	44	184	57	39	17	69	30	101	44	133	58
	1.5	18	5	33	9	51	14	66	18	13	5	23	9	36	14	46	18
	1.75	12	3	20	5	29	7	37	9	11	4	14	5	20	7	28	10
	2	9	2	17	4	22	5	27	6	6	2	12	4	15	5	19	6
	2.25	9	2	14	3	19	4	24	5	6	2	10	3	13	4	16	5

Table 16: Optimum parameters for WDwAwith Modified EWMA assuring 50th quantile when shape parameter $\theta = 1.5$, Acceleration Factor ($AF = 2.5$) and Acceleration Consumer Ratio ($RA = 3.0$)

β	$\frac{t_{qA}}{t^0_{qA}}$	$\delta_0 = 0.6$								$\delta_0 = 0.8$							
		$\lambda = 0.2$		$\lambda = 0.4$		$\lambda = 0.6$		$\lambda = 0.8$		$\lambda = 0.2$		$\lambda = 0.4$		$\lambda = 0.6$		$\lambda = 0.8$	
		n	c	n	c	n	c	n	c	n	c	n	c	n	c	n	c
0.25	1.25	19	6	38	12	50	16	66	21	16	7	27	12	38	17	47	21
	1.5	7	2	14	4	20	6	27	8	7	3	12	5	14	6	19	8
	1.75	7	2	7	2	11	3	14	4	5	2	5	2	10	4	10	4
	2	4	1	7	2	8	2	11	3	3	1	5	2	8	3	8	3
	2.25	4	1	4	1	8	2	8	2	3	1	5	2	6	2	8	3
0.1	1.25	33	10	59	18	85	26	108	33	21	9	42	18	58	25	79	34
	1.5	11	3	18	5	29	8	36	10	10	4	15	6	20	8	28	11
	1.75	8	2	12	3	19	5	23	6	8	3	8	3	11	4	16	6
	2	8	2	12	3	13	3	17	4	3	1	6	2	9	3	12	4
	2.25	5	1	9	2	9	2	13	3	3	1	6	2	6	2	9	3
0.05	1.25	40	12	73	22	103	31	136	41	26	11	52	22	73	31	99	42
	1.5	15	4	26	7	37	10	48	13	13	5	18	7	26	10	34	13
	1.75	8	2	16	4	24	6	28	7	8	3	11	4	14	5	20	7
	2	8	2	13	3	17	4	22	5	6	2	9	3	12	4	15	5
	2.25	5	1	9	2	14	3	18	4	6	2	7	2	10	3	13	4
0.01	1.25	51	15	102	30	146	43	197	58	36	15	72	30	108	45	137	57
	1.5	19	5	35	9	54	14	69	18	16	6	27	10	38	14	46	17
	1.75	13	3	22	5	30	7	39	9	9	3	15	5	21	7	27	9
	2	9	2	18	4	23	5	29	6	7	2	10	3	16	5	20	6
	2.25	9	2	15	3	20	4	25	5	7	2	10	3	14	4	17	5

Table 17: Optimum parameters for WDwAwith Modified EWMA assuring 50th quantile when shape parameter $\theta = 2.0$, Acceleration Factor ($AF = 1.5$) and Acceleration Consumer Ratio ($RA = 2$)

β	$\frac{t_{qA}}{t^0_{qA}}$	$\delta_0 = 0.6$								$\delta_0 = 0.8$							
		$\lambda = 0.2$		$\lambda = 0.4$		$\lambda = 0.6$		$\lambda = 0.8$		$\lambda = 0.2$		$\lambda = 0.4$		$\lambda = 0.6$		$\lambda = 0.8$	
		n	c	n	c	n	c	n	c	n	c	n	c	n	c	n	c
0.25	1.25	16	5	22	7	28	9	37	12	8	4	14	7	20	10	26	13
	1.5	7	2	7	2	10	3	14	4	6	3	7	3	9	4	11	5
	1.75	4	1	7	2	7	2	11	3	5	2	5	2	5	2	7	3
	2	4	1	4	1	4	1	8	2	3	1	5	2	5	2	5	2
	2.25	4	1	4	1	4	1	4	1	3	1	3	1	3	1	5	2
0.1	1.25	20	6	33	10	49	15	62	19	15	7	21	10	35	17	42	20
	1.5	8	2	11	3	18	5	22	6	7	3	7	3	12	5	14	6
	1.75	4	1	8	2	12	3	16	4	5	2	5	2	8	3	10	4
	2	4	1	8	2	9	2	9	2	3	1	5	2	6	2	8	3
	2.25	4	1	5	1	5	1	9	2	3	1	3	1	3	1	6	2
0.05	1.25	20	6	40	12	60	18	77	23	15	7	30	14	42	20	49	23
	1.5	8	2	15	4	23	6	27	7	7	3	12	5	17	7	17	7
	1.75	8	2	9	2	13	3	17	4	5	2	8	3	8	3	11	4
	2	5	1	9	2	10	2	14	3	3	1	6	2	6	2	9	3
	2.25	5	1	5	1	10	2	10	2	3	1	3	1	6	2	6	2
0.01	1.25	31	9	55	16	86	25	110	32	22	10	37	17	57	26	72	33
	1.5	12	3	24	6	32	8	41	10	10	4	15	6	20	8	28	11
	1.75	9	2	14	3	19	4	24	5	6	2	9	3	14	5	17	6
	2	5	1	10	2	15	3	20	4	3	1	6	2	9	3	12	4
	2.25	5	1	10	2	12	2	17	3	3	1	6	2	7	2	10	3

Table 18: Optimum parameters for WDwAwith Modified EWMA assuring 50th quantile when shape parameter $\theta = 2.0$, Acceleration Factor ($AF = 2$) and Acceleration Consumer Ratio ($RA = 2.5$)

β	$\frac{t_{qA}}{t^0_{qA}}$	$\delta_0 = 0.6$								$\delta_0 = 0.8$							
		$\lambda = 0.2$		$\lambda = 0.4$		$\lambda = 0.6$		$\lambda = 0.8$		$\lambda = 0.2$		$\lambda = 0.4$		$\lambda = 0.6$		$\lambda = 0.8$	
		n	c	n	c	n	c	n	c	n	c	n	c	n	c	n	c
0.25	1.25	14	4	24	7	31	9	45	13	9	4	18	8	22	10	31	14
	1.5	8	2	11	3	15	4	15	4	5	2	7	3	7	3	12	5
	1.75	4	1	4	1	8	2	8	2	3	1	5	2	5	2	5	2
	2	4	1	4	1	8	2	8	2	3	1	3	1	5	2	5	2
	2.25	4	1	4	1	5	1	5	1	3	1	3	1	3	1	3	1
0.1	1.25	18	5	36	10	51	14	69	19	14	6	23	10	32	14	46	20
	1.5	8	2	16	4	21	5	25	6	5	2	10	4	13	5	18	7
	1.75	5	1	9	2	13	3	14	3	3	1	6	2	8	3	11	4
	2	5	1	5	1	10	2	10	2	3	1	3	1	6	2	6	2
	2.25	5	1	5	1	10	2	10	2	3	1	3	1	6	2	6	2
0.05	1.25	26	7	45	12	67	18	86	23	14	6	28	12	42	18	56	24
	1.5	12	3	17	4	26	6	30	7	8	3	11	4	16	6	21	8
	1.75	5	1	10	2	15	3	19	4	3	1	6	2	9	3	12	4
	2	5	1	10	2	11	2	16	3	3	1	6	2	7	2	10	3
	2.25	5	1	6	1	11	2	12	2	3	1	6	2	7	2	7	2
0.01	1.25	38	10	65	17	96	25	123	32	24	10	41	17	60	25	77	32
	1.5	14	3	23	5	36	8	46	10	11	4	14	5	22	8	28	10
	1.75	10	2	16	3	21	4	27	5	6	2	10	3	13	4	16	5
	2	6	1	12	2	17	3	23	4	6	2	7	2	11	3	14	4
	2.25	6	1	12	2	13	2	19	3	4	1	7	2	8	2	11	3

Table 19: Optimum parameters for WDwAwith Modified EWMA assuring 50th quantile when shape parameter $\theta = 2.0$, Acceleration Factor ($AF = 2.5$) and Acceleration Consumer Ratio ($RA = 3$)

β	$\frac{t_{qA}}{t^0_{qA}}$	$\delta_0 = 0.6$								$\delta_0 = 0.8$							
		$\lambda = 0.2$		$\lambda = 0.4$		$\lambda = 0.6$		$\lambda = 0.8$		$\lambda = 0.2$		$\lambda = 0.4$		$\lambda = 0.6$		$\lambda = 0.8$	
		n	c	n	c	n	c	n	c	n	c	n	c	n	c	n	c
0.25	1.25	15	4	26	7	37	10	48	13	7	3	14	6	21	9	28	12
	1.5	4	1	8	2	12	3	16	4	5	2	5	2	10	4	10	4
	1.75	4	1	8	2	9	2	9	2	3	1	5	2	6	2	8	3
	2	4	1	5	1	5	1	9	2	3	1	3	1	3	1	6	2
	2.25	4	1	5	1	5	1	5	1	3	1	3	1	3	1	3	1
0.1	1.25	23	6	35	9	55	14	70	18	15	6	27	11	34	14	49	20
	1.5	9	2	17	4	22	5	27	6	8	3	11	4	14	5	19	7
	1.75	5	1	10	2	14	3	15	3	3	1	6	2	9	3	9	3
	2	5	1	10	2	11	2	11	2	3	1	6	2	7	2	9	3
	2.25	5	1	6	1	6	1	11	2	3	1	4	1	7	2	7	2
0.05	1.25	24	6	48	12	68	17	92	23	15	6	30	12	45	18	60	24
	1.5	9	2	18	4	27	6	33	7	8	3	14	5	17	6	20	7
	1.75	9	2	11	2	16	3	21	4	6	2	9	3	10	3	13	4
	2	6	1	11	2	12	2	17	3	4	1	7	2	7	2	10	3
	2.25	6	1	7	1	12	2	13	2	4	1	4	1	7	2	8	2
0.01	1.25	33	8	66	16	99	24	132	32	23	9	41	16	64	25	82	32
	1.5	15	3	25	5	39	8	49	10	9	3	15	5	24	8	30	10
	1.75	11	2	17	3	23	4	29	5	7	2	10	3	14	4	20	6
	2	7	1	13	2	19	3	25	4	4	1	8	2	11	3	15	4
	2.25	7	1	13	2	14	2	20	3	4	1	8	2	9	2	12	3

Table 20: Optimum parameters for WDwAwith Modified EWMA assuring 50th quantile when shape parameter $\theta = 2.5$, Acceleration Factor ($AF = 1.5$) and Acceleration Consumer Ratio ($RA = 2$)

β	$\frac{t_{qA}}{t^0_{qA}}$	$\delta_0 = 0.6$								$\delta_0 = 0.8$							
		$\lambda = 0.2$		$\lambda = 0.4$		$\lambda = 0.6$		$\lambda = 0.8$		$\lambda = 0.2$		$\lambda = 0.4$		$\lambda = 0.6$		$\lambda = 0.8$	
		n	c	n	c	n	c	n	c	n	c	n	c	n	c	n	c
0.25	1.25	7	2	14	4	21	6	28	8	4	2	8	4	12	6	16	8
	1.5	4	1	8	2	8	2	12	3	4	2	5	2	7	3	7	3
	1.75	4	1	4	1	8	2	8	2	3	1	5	2	5	2	5	2
	2	4	1	4	1	5	1	5	1	3	1	3	1	3	1	5	2
	2.25	4	1	4	1	5	1	5	1	3	1	3	1	3	1	3	1
0.1	1.25	15	4	26	7	37	10	45	12	11	5	15	7	19	9	30	14
	1.5	8	2	9	2	13	3	18	4	5	2	5	2	10	4	10	4
	1.75	5	1	5	1	10	2	10	2	3	1	5	2	5	2	8	3
	2	5	1	5	1	6	1	10	2	3	1	3	1	3	1	6	2
	2.25	5	1	5	1	6	1	6	1	3	1	3	1	3	1	6	2
0.05	1.25	19	5	31	8	42	11	58	15	11	5	21	10	26	12	37	17
	1.5	9	2	13	3	18	4	23	5	5	2	8	3	10	4	13	5
	1.75	5	1	10	2	11	2	15	3	3	1	6	2	6	2	9	3
	2	5	1	6	1	11	2	12	2	3	1	3	1	6	2	6	2
	2.25	5	1	6	1	7	1	7	1	3	1	3	1	4	1	6	2
0.01	1.25	20	5	40	10	60	15	80	20	16	7	27	12	38	17	47	21
	1.5	10	2	15	3	25	5	30	6	8	3	11	4	16	6	19	7
	1.75	6	1	12	2	17	3	23	4	3	1	6	2	9	3	12	4
	2	6	1	7	1	13	2	14	2	3	1	6	2	7	2	10	3
	2.25	6	1	7	1	13	2	14	2	3	1	4	1	7	2	8	2

Table 21: Optimum parameters for WDwAwith Modified EWMA assuring 50th quantile when shape parameter $\theta = 2.5$, Acceleration Factor ($AF = 2.0$) and Acceleration Consumer Ratio ($RA = 2.5$)

β	$\frac{t_{qA}}{t^0_{qA}}$	$\delta_0 = 0.6$								$\delta_0 = 0.8$							
		$\lambda = 0.2$		$\lambda = 0.4$		$\lambda = 0.6$		$\lambda = 0.8$		$\lambda = 0.2$		$\lambda = 0.4$		$\lambda = 0.6$		$\lambda = 0.8$	
		n	c	n	c	n	c	n	c	n	c	n	c	n	c	n	c
0.25	1.25	8	2	16	4	24	6	32	8	7	3	9	4	16	7	18	8
	1.5	5	1	9	2	9	2	13	3	5	2	5	2	5	2	10	4
	1.75	5	1	5	1	5	1	9	2	3	1	3	1	5	2	5	2
	2	5	1	5	1	5	1	5	1	3	1	3	1	3	1	3	1
	2.25	5	1	5	1	5	1	5	1	3	1	3	1	3	1	3	1
0.1	1.25	13	3	26	6	39	9	52	12	12	5	17	7	24	10	31	13
	1.5	5	1	10	2	15	3	20	4	5	2	8	3	8	3	11	4
	1.75	5	1	6	1	11	2	12	2	3	1	3	1	6	2	6	2
	2	5	1	6	1	7	1	12	2	3	1	3	1	6	2	6	2
	2.25	5	1	6	1	7	1	7	1	3	1	3	1	4	1	4	1
0.05	1.25	22	5	35	8	53	12	66	15	12	5	22	9	29	12	39	16
	1.5	10	2	15	3	21	4	26	5	6	2	9	3	11	4	14	5
	1.75	6	1	11	2	12	2	18	3	3	1	6	2	7	2	10	3
	2	6	1	7	1	12	2	13	2	3	1	4	1	7	2	7	2
	2.25	6	1	7	1	8	1	13	2	3	1	4	1	4	1	7	2
0.01	1.25	27	6	50	11	74	16	97	21	15	6	28	11	43	17	53	21
	1.5	11	2	22	4	29	5	39	7	6	2	12	4	15	5	21	7
	1.75	7	1	13	2	20	3	26	4	6	2	7	2	11	3	11	3
	2	7	1	13	2	15	2	17	2	4	1	7	2	8	2	11	3
	2.25	7	1	9	1	15	2	17	2	4	1	5	1	8	2	9	2

Table 22: Optimum parameters for WDwAwith Modified EWMA assuring 50th quantile when shape parameter $\theta = 2.5$, Acceleration Factor ($AF = 2.5$) and Acceleration Consumer Ratio ($RA = 3.0$)

β	$\frac{t_{qA}}{t^0_{qA}}$	$\delta_0 = 0.6$								$\delta_0 = 0.8$							
		$\lambda = 0.2$		$\lambda = 0.4$		$\lambda = 0.6$		$\lambda = 0.8$		$\lambda = 0.2$		$\lambda = 0.4$		$\lambda = 0.6$		$\lambda = 0.8$	
		n	c	n	c	n	c	n	c	n	c	n	c	n	c	n	c
0.25	1.25	13	3	22	5	26	6	35	8	10	4	12	5	17	7	22	9
	1.5	5	1	10	2	10	2	15	3	3	1	6	2	8	3	8	3
	1.75	5	1	5	1	10	2	10	2	3	1	3	1	3	1	6	2
	2	5	1	5	1	6	1	6	1	3	1	3	1	3	1	3	1
	2.25	5	1	5	1	6	1	6	1	3	1	3	1	3	1	3	1
0.1	1.25	14	3	28	6	42	9	56	12	10	4	18	7	23	9	31	12
	1.5	10	2	11	2	17	3	22	4	3	1	6	2	9	3	12	4
	1.75	6	1	11	2	12	2	13	2	3	1	6	2	7	2	7	2
	2	6	1	7	1	7	1	13	2	3	1	4	1	4	1	7	2
	2.25	6	1	7	1	7	1	8	1	3	1	4	1	4	1	4	1
0.05	1.25	24	5	39	8	53	11	72	15	13	5	21	8	29	11	42	16
	1.5	11	2	17	3	23	4	29	5	6	2	9	3	12	4	16	5
	1.75	6	1	12	2	14	2	20	3	4	1	7	2	7	2	10	3
	2	6	1	8	1	14	2	15	2	4	1	4	1	7	2	8	2
	2.25	6	1	8	1	9	1	15	2	4	1	4	1	5	1	8	2
0.01	1.25	30	6	55	11	80	16	106	21	19	7	30	11	41	15	60	22
	1.5	12	2	24	4	31	5	43	7	7	2	10	3	17	5	20	6
	1.75	8	1	15	2	22	3	29	4	4	1	8	2	11	3	12	3
	2	8	1	15	2	17	2	24	3	4	1	8	2	9	2	12	3
	2.25	8	1	10	1	17	2	19	2	4	1	5	1	9	2	10	2

Table 23: Optimum parameters for the Weibull Distribution assuring 50th quantile when shape parameter $\theta = 2$

β	$\frac{t_{qA}}{t^0_{qA}}$	$\delta_0 = 0.6$		$\delta_0 = 0.8$	
		n	c	n	c
0.25	1.25	20	3	15	4
	1.5	14	2	9	2
	1.75	14	2	9	2
	2	9	1	5	1
	2.25	9	1	5	1
0.1	1.25	38	5	22	5
	1.5	26	3	16	3
	1.75	20	2	12	2
	2	20	2	12	2
	2.25	14	1	8	1
0.05	1.25	49	6	29	6
	1.5	31	3	18	3
	1.75	31	3	14	2
	2	25	2	14	2
	2.25	25	2	14	2
0.01	1.25	74	8	43	8
	1.5	56	5	32	5
	1.75	42	3	24	3
	2	42	3	24	3
	2.25	35	2	20	2

Table 24: Optimum parameters for the Weibull Distributionwith EWMA assuring 50^{th} quantilewhen shape parameter $\theta = 2$

β	$\frac{t_{qA}}{t^0_{qA}}$	$\delta_0 = 0.6$										$\delta_0 = 0.8$									
		$\lambda = 0.2$		$\lambda = 0.4$		$\lambda = 0.6$		$\lambda = 0.8$		$\lambda = 1.0$		$\lambda = 0.2$		$\lambda = 0.4$		$\lambda = 0.6$		$\lambda = 0.8$		$\lambda = 1.0$	
		n	c	n	c	n	c	n	c	n	c	n	c	n	c	n	c	n	c	n	c
0.25	1.25	6	1	7	1	12	2	13	2	20	3	4	1	4	1	8	2	11	3	15	4
	1.5	6	1	7	1	7	1	13	2	14	2	4	1	4	1	4	1	8	2	9	2
	1.75	6	1	7	1	7	1	8	1	14	2	4	1	4	1	4	1	5	1	9	2
	2	6	1	7	1	7	1	8	1	9	1	4	1	4	1	4	1	5	1	5	1
	2.25	6	1	7	1	7	1	8	1	9	1	4	1	4	1	4	1	5	1	5	1
0.1	1.25	7	1	14	2	16	2	29	4	38	5	4	1	9	2	9	2	17	4	22	5
	1.5	7	1	8	1	16	2	18	2	26	3	4	1	5	1	9	2	11	2	16	3
	1.75	7	1	8	1	10	1	18	2	20	2	4	1	5	1	6	1	11	2	12	2
	2	7	1	8	1	10	1	12	1	20	2	4	1	5	1	6	1	7	1	12	2
	2.25	7	1	8	1	10	1	12	1	14	1	4	1	5	1	6	1	7	1	8	1
0.05	1.25	8	1	16	2	24	3	33	4	49	6	5	1	9	2	14	3	19	4	29	6
	1.5	8	1	10	1	18	2	27	3	31	3	5	1	6	1	11	2	12	2	18	3
	1.75	8	1	10	1	12	1	21	2	31	3	5	1	6	1	7	1	12	2	14	2
	2	8	1	10	1	12	1	14	1	25	2	5	1	6	1	7	1	8	1	14	2
	2.25	8	1	10	1	12	1	14	1	25	2	5	1	6	1	7	1	8	1	14	2
0.01	1.25	9	1	25	3	36	4	54	6	74	8	9	2	11	2	21	4	31	6	43	8
	1.5	9	1	19	2	23	2	35	3	56	5	6	1	11	2	13	2	20	3	32	5
	1.75	9	1	13	1	23	2	28	2	42	3	6	1	7	1	13	2	16	2	24	3
	2	9	1	13	1	16	1	28	2	42	3	6	1	7	1	9	1	16	2	24	3
	2.25	9	1	13	1	16	1	28	2	35	2	6	1	7	1	9	1	16	2	20	2

Table 25: Optimum parameters for the Weibull Distributionwith Modified EWMA assuring 50^{th} quantile when shape parameter $\theta = 2.5$

β	$\frac{t_{qA}}{t^0_{qA}}$	$\delta_0 = 0.6$								$\delta_0 = 0.8$							
		$\lambda = 0.2$		$\lambda = 0.4$		$\lambda = 0.6$		$\lambda = 0.8$		$\lambda = 0.2$		$\lambda = 0.4$		$\lambda = 0.6$		$\lambda = 0.8$	
		n	c	n	c	n	c	n	c	n	c	n	c	n	c	n	c
0.25	1.25	6	1	6	1	7	1	12	2	4	1	4	1	4	1	8	2
	1.5	6	1	6	1	7	1	7	1	4	1	4	1	4	1	4	1
	1.75	6	1	6	1	7	1	7	1	4	1	4	1	4	1	4	1
	2	6	1	6	1	7	1	7	1	4	1	4	1	4	1	4	1
	2.25	6	1	6	1	7	1	7	1	4	1	4	1	4	1	4	1
0.1	1.25	7	1	8	1	14	2	15	2	4	1	8	2	9	2	9	2
	1.5	7	1	8	1	9	1	15	2	4	1	5	1	5	1	9	2
	1.75	7	1	8	1	9	1	10	1	4	1	5	1	5	1	6	1
	2	7	1	8	1	9	1	10	1	4	1	5	1	5	1	6	1
	2.25	7	1	8	1	9	1	10	1	4	1	5	1	5	1	6	1
0.05	1.25	8	1	15	2	16	2	23	3	5	1	9	2	10	2	14	3
	1.5	8	1	9	1	10	1	18	2	5	1	5	1	6	1	10	2
	1.75	8	1	9	1	10	1	12	1	5	1	5	1	6	1	7	1
	2	8	1	9	1	10	1	12	1	5	1	5	1	6	1	7	1
	2.25	8	1	9	1	10	1	12	1	5	1	5	1	6	1	7	1
0.01	1.25	9	1	18	2	26	3	35	4	5	1	10	2	15	3	20	4
	1.5	9	1	12	1	20	2	23	2	5	1	7	1	12	2	13	2
	1.75	9	1	12	1	14	1	23	2	5	1	7	1	8	1	13	2
	2	9	1	12	1	14	1	16	1	5	1	7	1	8	1	9	1
	2.25	9	1	12	1	14	1	16	1	5	1	7	1	8	1	9	1

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 21, No. 2, 2023 (New Series), pp 235–257
http://www.ssca.org.in/journal

Agricultural Price Forecasting Based on Variational Mode Decomposition and Time-Delay Neural Network

Kapil Choudhary[1], Girish K. Jha[2], Ronit Jaiswal[1], P. Venkatesh[2] and Rajender Parsad[1]

[1]*ICAR-Indian Agricultural Statistics Research Institute, PUSA, New Delhi-110012*
[2]*ICAR-Indian Agricultural Research Institute, PUSA, New Delhi-110012*

Received: 24 May 2022; Revised: 16 March 2023; Accepted: 21 March 2023

Abstract

Agricultural commodities prices are very unpredictable and complex, and thus, forecasting these prices is one of the research hotspots. In this paper, we propose a new hybrid VMD-TDNN model combining variational mode decomposition (VMD) and time-delay neural network (TDNN) to improve the accuracy of agricultural price forecasting. Specifically, the VMD decomposes a price series into a set of intrinsic mode functions (IMFs), and the obtained IMFs are modelled and forecasted separately using the TDNN models. Finally, the forecasts of all IMFs are combined to provide an ensemble output for the price series. VMD overcomes the limitation of the mode mixing and end effect problems of the empirical mode decomposition (EMD) based variants. The prediction ability of the proposed model is compared with TDNN, and EMD based variants coupled with TDNN model using international monthly price series of maize, palm oil, and soybean in terms of evaluation criteria like root mean squared error, mean absolute percentage error and, directional prediction statistics. Additionally, Diebold-Mariano test and Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS), a ranking system, are used to evaluate the accuracy of the models. The empirical results confirm that the proposed hybrid model is superior in terms of evaluation criteria and improves the prediction accuracy significantly.

Key words: Agricultural price forecasting; Empirical mode decomposition; Intrinsic mode function; Time-delay neural network; Variational mode decomposition.

1. Introduction

Price forecasting of agricultural commodities is a challenging task as there are several unpredictable factors, both natural and man-made, which influence the production and price of the commodities. Thus, the price series become inherently nonstationary and nonlinear in nature posing a severe threat to food security in developing countries, see FAO (2011). Accurate and reliable agricultural price forecasts are thus very necessary not only

Corresponding Author: Girish K. Jha
Email: girish.stat@gmail.com

for mitigating the threat of food security but also to balance the demand with supply, ensure remunerative prices to farmers, and the welfare of the consumers, see Jaiswal *et al.* (2022). A thorough review of the existing literature confirms that significant efforts have been done to improve price forecasting using various time series models. The various time series models developed for price forecasting can be broadly classified into two categories, *i.e.* statistical models and artificial intelligence (AI) models. Among statistical models, autoregressive integrated moving average (ARIMA), see Box *et al.* (2015), and constituent models, see Hayat and Bhatti (2013); Jadhav *et al.* (2017), are most frequently used as prediction models. However, ARIMA models assume linear relationships among data points, despite real-world agricultural price data being usually nonlinear. As a result, the ARIMA model is unable to capture the hidden patterns in the agricultural price series effectively, leading to unsatisfactory forecasting results.

In recent years, artificial neural network (ANN) in the category of AI models has become the most efficient modelling method in dealing with the complex nature of time series. ANN has been widely utilized to model nonlinear time series with minimal assumptions and high prediction accuracy due to its self-learning capabilities, see Zhang *et al.* (1998). ANN has been effectively employed as a universal function approximator in a wide range of research areas like electricity price forecasting, exchange forecasting, wind speed forecasting, solar energy forecasting, *etc.* In agricultural price forecasting, Jha and Sinha (2014)used the time-delay neural network (TDNN) model to predict monthly wholesale prices of different oilseeds and concluded that the ANN-based forecasting model outperforms the ARIMA model in terms of prediction accuracy. Similarly, Xu and Zhang (2021) investigated both univariate and bivariate neural network modelling for corn cash prices and found that simple neural networks with twenty hidden nodes and two lags provided better forecasting accuracy for short-term forecasting. Despite the better prediction performance of ANN-based models in many areas, their accuracy is still not satisfactory when dealing with nonstationary and nonlinear time series data. However, the accuracy can be further increased using the hybridization technique, *i.e.* combining different models according to their strength and producing a synergetic effect. In a hybrid model class, the decomposition-and-ensemble-methodology is the most promising one, see Qian *et al.* (2019). This methodology follows the principle of "divide and conquer" whereby using some techniques, a complex series is divided into a number of simple subseries such that each subseries now has better characterization and thus can be easily captured, resulting in better forecasting accuracy.

For the decomposition of any nonlinear and nonstationary time series, empirical mode decomposition (EMD), see Huang *et al.* (1998), and its variants like ensemble empirical mode decomposition (EEMD), see Wu and Huang (2009), and complementary ensemble empirical mode decomposition with adaptive noise (CEEMDAN), see Torres *et al.* (2011), are commonly used. The essence of the EMD and its variants is that they decompose a time series into a set of subseries (modes) called intrinsic mode functions (IMFs) and residual. These IMFs and residual are further modelled by any of the forecasting techniques like TDNN. For instance, Yu *et al.* (2008) evaluated EMD based feed-forward neural network (FNN) for crude oil predictions and concluded that decomposition-based hybrid models outperform standalone forecasting models. Choudhary *et al.* (2019) used EEMD for decomposing the daily potato price series of two different markets. Fang *et al.* (2020) applied EEMD to different agricultural commodities for decomposition, whereas ARIMA, neural network (NN) and support vector machine (SVM) models for predicting the decomposed components. Prasad

et al. (2018) demonstrated the superiority of a hybrid model that combines CEEMDAN and an extreme learning machine (ELM) to forecast soil moisture.

However, EMD and its variants have major drawbacks, such as the frequent appearance of mode mixing, noise sensitiveness and end effects, leading to meaningless subseries that negatively impact the precision of decomposition. In order to address these limitations, variational mode decomposition (VMD) is proposed as an adaptive, non-recursive and multiresolution decomposition technique by Dragomiretskiy and Zosso (2014). VMD decomposes original time series into a set of distinct independent IMFs based on their central frequencies. The VMD proved its superiority over EMD based decomposition in the different areas of time series forecasting, see Bisoi *et al.* (2019); Dragomiretskiy and Zosso (2014); Lahmiri (2016); Liu *et al.* (2018). Therefore,in view of the superior performance of VMD as a unique data-adaptive decomposition technique and the advantageous properties of TDNN for forecasting any nonlinear series, a novel hybrid VMD-TDNN model is proposed for agricultural price forecasting.

The main idea of our study is to utilise a new adaptive multiresolution technique in the context of modelling and predicting nonstationary and nonlinear agricultural price series. However, the most significant contributions of this paper are as follows. First, a novel agricultural price forecasting framework is proposed by combining VMD with the TDNN model. VMD is a decomposition technique that breaks a highly complex agricultural price series into several uniform subseries with stable fluctuations. VMD has the advantage of being noise robust as it denoises a time series using simulated harmonic functions. In this context, for an agricultural price series that is known to be very noisy, the VMD is more suitable for its better characterization leading to faster convergence and better predictive accuracy. Second, for empirical evaluation of our proposed model, we use three real agricultural price series to test how well the proposed model can tackle high-frequency events such as fluctuations in fuel prices, strikes, *etc.* and also the low-frequency events such as lower production, higher export, optimum rainfall, *etc.* Third, it is seen that many scientists believe that machine learning is a black box, and the results obtained from any machine learning technique are either not trustworthy or are not able to provide proper interpretations otherwise. Another reason for this belief is that the datasets and the code for machine learning-related forecasting research are often not publicly available, making it difficult for the forecasting community to adapt such research and verify the claimed performance. Thus, considering one of the major goals of our study to make this work replicable by the whole research fraternity for practical forecasting tasks, we use the datasets available in the public domain and for each of the hybrid models used in this study, we develop and publish packages namely, "eemdTDNN" and "vmdTDNN", see Choudhary *et al.* (2021, 2022), in CRAN. Fourth, we compare the prediction accuracy of the VMD-TDNN with different decomposition-based techniques, and the empirical evidence shows that the VMD-TDNN model outperforms EMD and its variants based hybrid models in terms of each evaluation criteria. Finally, for the robust validation and to check the superiority of the forecasting ability of the developed model, we use Diebold-Mariano test for checking the significant improvement achieved by it. Further, we also use Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS) for ranking the models based on overall performance.

The remainder of the paper is organised as follows: Section 2 describes the proposed VMD-TDNN hybrid model for agricultural price forecasting in detail. For empirically eval-

uating the proposed model, three internationally traded agricultural commodities, namely maize, soybean oil and palm oil monthly price series, are described in Section 3. Section 4 concludes the work.

2. Methodology

2.1. Variational mode decomposition

VMD, see Dragomiretskiy and Zosso (2014), is a novel data-adaptive decomposition technique that overcomes the limitations of traditional frequency-based decomposition techniques. This technique effectively improves the end effect, mode mixing, recursive sifting process, sensitivity to noise, a fixed number of modes, and other shortcomings of EMD variants, see Wu and Huang (2009). The algorithm used in VMD is non-recursive as it extracts modes concurrently, assuming limited bandwidth of central frequency for each IMF. Moreover, the modes obtained after VMD have a particular property called sparsity which means each mode is mostly compact around a centre pulsation in the frequency spectrum. The advantages of using VMD over other techniques are that the modes are robust with respect to noise and have faster convergence with better accuracy. These characteristics of VMD make it highly suitable for addressing complex agricultural price data consisting of multi-frequency signals. The bandwidth of a mode is estimated using the following steps:

1. For each mode $c_j(t)$, the Hilbert transformation (HT) is used to obtain a one-sided frequency spectrum.
2. The sifting of the frequency spectrum of the mode is determined using the modulation properties.
3. The bandwidth of $c_j(t)$ is estimated finally using H^1 Gaussian smoothness of the demodulated signal *i.e.* the squared L^2-norm of the gradient.

The following constraints of the variation problem can be used to explain VMD:

$$\min_{\{\omega_j\},\{c_j\}} \left\{ \sum_j \left\| \partial_t \left[\left(\delta(t) + \frac{i}{\pi t} \right) * c_j(t) \right] e^{-i\omega_j t} \right\|_2^2 \right\}$$

such that $\sum_{j=1}^{n} c_j(t) = y(t)$; where $y(t)$ is the original price series, $\{c_j\} := \{c_1, c_2, \cdots c_n\}$ is the set of modes, $\{w_j\} := \{w_1, w_2, \cdots w_n\}$ is the of central frequencies, $\delta(t)$ is the impulse function, $\partial_t(.)$ is the partial derivative of time t, n is the number of modes, $*$ denotes convolution operation, $\|.\|$ denotes norm processing, and $i = \sqrt{-1}$. Lagrangian multipliers $\lambda(t)$ and quadratic penalty terms are used to transform a constraint problem into an unconstrained problem that is easy to solve:

$$\begin{aligned} L\left(\{c_j\}, \{\omega_j\}, \lambda\right) = \alpha \sum_j \left\| \partial_t \left[\left(\delta(t) + \frac{i}{\pi t} \right) c_j(t) \right] e^{-i\omega_j t} \right\|_2^2 \\ + \left\| y(t) - \sum_j c_j(t) \right\|_2^2 + \left[\lambda(t), y(t) - \sum_j c_j(t) \right] \end{aligned}$$

where α is said to be a balance parameter or penalty parameter of data fidelity constraint.

Furthermore, an iterative sequence called the alternate direction method of multipliers (ADMM) is applied to the above equation for updating c_j, ω_j and λ in two directions. The results are obtained as follows:

$$\hat{c}_j^{k+1}(\omega) = \frac{\hat{y}(\omega) - \sum_{l \neq j} \hat{c}_l(\omega) + \frac{\hat{\lambda}(\omega)}{2}}{1 + 2\alpha\left(\omega - \omega_j\right)^2}; \quad \omega_j^{k+1} = \frac{\int_0^\infty \omega \left|\hat{c}_j(\omega)\right|^2 d\omega}{\int_0^\infty \left|\hat{c}_j(\omega)\right|^2 d\omega}$$

$$\text{and } \hat{\lambda}^{k+1}(\omega) \leftarrow \hat{\lambda}^k(\omega) + \tau\left[\hat{y}(\omega) - \sum_j \hat{c}_j^{k+1}(\omega)\right].$$

The stopping criterion for the iterations is $\sum_j \frac{\left\|\hat{c}_j^{k+1} - \hat{c}_j^k\right\|_2^2}{\left\|\hat{c}_j^k\right\|_2^2} < \epsilon$, where $\epsilon > 0, \hat{y}(\omega), \hat{c}_j(\omega), \hat{\lambda}(\omega)$ and $\hat{c}_j^{k+1}(\omega)$ are fourier transformations of $y(t), c_j(t), \lambda(t)$, and c_j^{k+1}(t), respectively, and k is the number of iterations.

2.2. Time-delay neural network (TDNN)

The artificial neural network (ANN) technique, which uses nonlinear units (neurons) to model any complex nonlinear time series, is being frequently utilized in many applications. There are three layers in a standard ANN architecture: input layer, where data is introduced to the network; hidden layer, where data is processed; and output layer, where the results of the given inputs are produced. There are two ways to model time series using neural networks: either using a recurrent neural network or creating short-term memory at the network's input layer, see Haykin (2009). TDNN is an example of the latter, which uses the temporal dimension of a univariate time series to develop a short-term memory, called heteroassociative memory, in its network. The usual TDNN is a feed-forward network with interconnected hidden and output neurons. A TDNN with a single hidden layer has the following generic expression, see Jha and Sinha (2014)

$$\hat{y}(t) = g\left(\alpha_0 + \sum_{j=1}^{q} \alpha_j f\left(\beta_{0j} + \sum_{i=1}^{p} \beta_{ij} y(t-i)\right)\right)$$

where $\hat{y}(t)$ is the predicted value, $y(t-i)$ is the i^{th} input (lag), $\alpha_j (j = 0, 1, 2, \ldots, q)$ and $\beta_{ij} (i = 0, 1, 2, \ldots, p; j = 1, 2, \ldots, q)$ are connection weights, p and q are the numbers of input and hidden nodes, respectively, f and g denote the activation functions at the hidden and output layer of the model.

2.3. VMD-TDNN model for agricultural price series

A nonstationary and nonlinear time series is decomposed into IMFs using VMD as a decomposition tool. Several models of TDNN are built for each IMF separately, varying the hyperparameters of the TDNN, and the best-fitted model is selected for each IMF to predict them separately, followed by ensemble prediction. Thus, a hybrid model, namely, VMD-TDNN, is proposed by integrating VMD and TDNN, and its details are displayed in

Figure 1. The procedure for this model can be separated into three parts:

1. **Data decomposition** – VMD decomposes agricultural price series $y(t)$ into n independent modes (IMFs), which are stationary and nonlinear. These IMFs have different oscillations of agricultural prices from high to low frequencies. These modes have a regular structure and stable fluctuation. Now, the patterns of each IMF can be captured more conveniently and accurately through TDNN.

2. **Individual prediction** – Each IMF is split into training and testing sets to ensure the generalization ability of the forecasting model. The TDNN model is used for modelling each of the IMFs as it is well suited for capturing nonlinear patterns.

3. **Ensemble prediction** - The final forecast of the original price series is obtained by adding the predicted values of all IMFs as:

$$\hat{y}(t) = \sum_{j=1}^{n} \hat{c}_j(t)$$

where, $\sum_{j=1}^{n} \hat{c}_j(t)$ represent the ensemble of predicted values of IMFs.

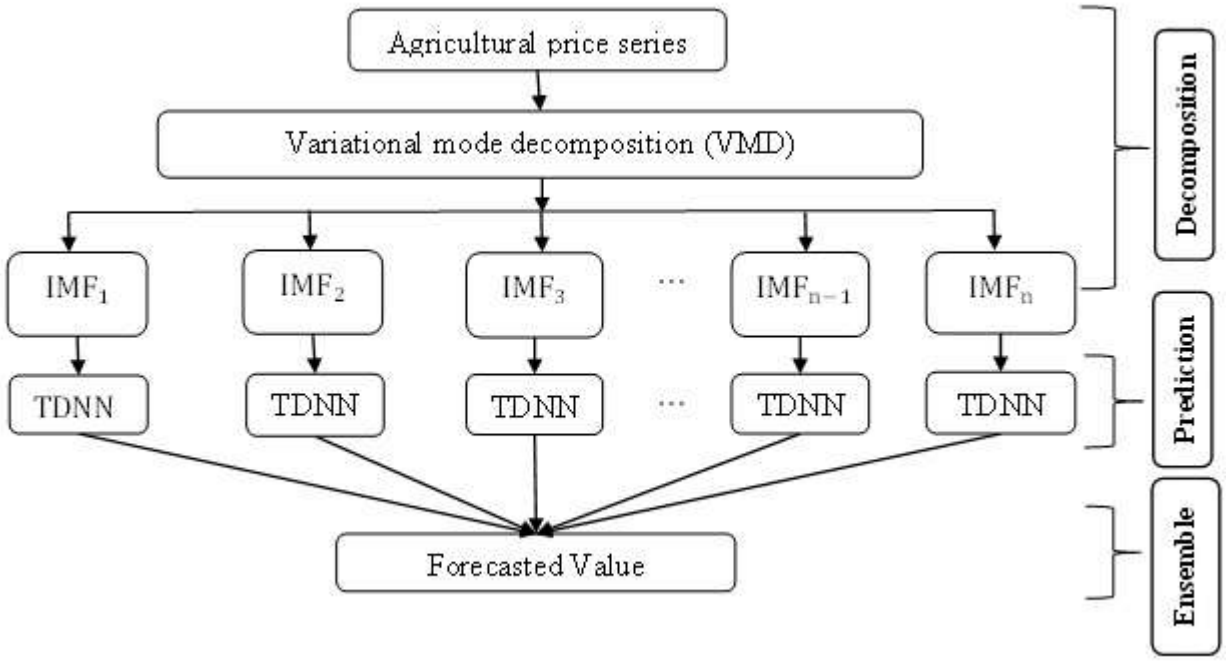


Figure 1: Flowchart of VMD-TDNN model for agricultural price forecasting

2.4. Forecasting evaluation criteria

Each prediction model employed in this paper is evaluated in terms of root mean squared error (RMSE), mean absolute percentage error (MAPE), directional prediction statistics (D_{stat}), and Diebold-Mariano (DM) test, since individual decision criteria is unable

to capture errors completely, see Jaiswal *et al.* (2022). Moreover, for the ranking of each model, TOPSIS method, see Hwang and Yoon (1981), is employed, which ranks each model by giving weights after normalizing the decision matrix of all evaluation criteria and calculating the geometric distance between the different models. The following are the forecasting evaluation criteria for comparing the proposed model with other models:

1. **Root mean squared error (RMSE):**

$$RMSE = \sqrt{\frac{\sum_{t=1}^{h}(y(t) - \hat{y}(t))^2}{h}}$$

2. **Mean absolute percentage error (MAPE):**

$$MAPE = \frac{1}{h}\sum_{t=1}^{h}\left|\frac{y(t) - \hat{y}(t)}{y(t)}\right|$$

3. **Directional prediction statistics** ($D_{\mathbf{stat}}$)**:**

$$D_{\text{stat}} = \frac{1}{h}\sum_{t=1}^{h} a_t \times 100\%$$

where $y(t)$ and $\hat{y}(t)$ are the actual value and predicted value, respectively, h is the size of the testing set and $a_t = \begin{cases} 1, & \text{if } [y(t+1) - y(t)][\hat{y}(t+1) - y(t)] \geq 0 \\ 0, & \text{otherwise} \end{cases}$.

4. **Diebold-Mariano (DM) test:**

For a given time series $y(t)$, the Diebold-Mariano (DM) test statistics is defined as:

$$z_{DM} = \frac{\bar{d}}{\sqrt{\widehat{V}_{\bar{d}}}}$$

where h is the test size, $\{e_{tet}\}_{t=1}^{h}$ and $\{e_{ref}\}_{t=1}^{h}$ are error for test model and reference model respectively, g is the loss function, $\bar{d} = \frac{1}{h}\sum_{t=1}^{h}\left[g\left(e_{\text{tet}}\right) - g\left(e_{ref}\right)\right]$ is the sample mean, $\widehat{V}_{\bar{d}} = \frac{1}{h}\left[\gamma_0 + 2\sum_{j=1}^{l-1}\gamma_j\right]$ is the estimate of variance using l step forecasts and $\gamma_j = \operatorname{cov}\left(d_t, d_{t-j}\right)$ is the estimate of j^{th} autocovariance of $[g\left(e_{tet}\right) - g\left(e_{ref}\right)]$.

5. **TOPSIS:**

For a given decision matrix $\mathbf{X} = (x_{ij})$ and a weight vector $\boldsymbol{W} = [w_1, w_2, \cdots, w_e]$, rank of i^{th} model is defined as:

$$R_i = \frac{d_i^-}{d_i^- + d_i{}^+}$$

where x_{ij} denotes j^{th} evaluation criteria for i^{th} prediction model for $1 \leq i \leq m$ and $1 \leq j \leq e, d_i^+ = \sqrt{\sum_{j=1}^{e} \left(v_{ij} - v_j^+\right)^2}$ and $d_i^- = \sqrt{\sum_{j=1}^{e} \left(v_{ij} - v_j^-\right)^2}$ are measures of separation between the positive and negative ideal solutions, $v_{ij} = w_j * n_{ij}$ is weighted normalized decision matrix where $\sum_{j=1}^{e} w_j = 1, n_{ij} = \frac{x_{ij}}{\sqrt{\sum_i {x_{ij}}^2}}$ is the normalized value of x_{ij}, $v_j^+ = \begin{cases} \max v_{ij}, & \text{if } j \text{ is positive criterion} \\ \min v_{ij}, & \text{if } j \text{ is negative criterion} \end{cases}$ and $v_j^- = \begin{cases} \min v_{ij}, & \text{if } j \text{ is positive criterion} \\ \max v_{ij}, & \text{if } j \text{ is negative criterion} \end{cases}$ are extremely positive and extremely negative performance on each criterion.

3. Empirical results and discussion

Three different agricultural commodity price series are used in this section to empirically evaluate the proposed model's performance. In this study, all the model developments and their statistical analysis are done in R statistical software of version 4.1.2. The detailed R codes are given in the Appendix. In this section, data description, different decomposition techniques, and prediction results of the models are analysed for the price series.

3.1. Data description

This paper examines the efficiency of the proposed hybrid VMD-TDNN model using monthly international Maize, Palm oil, and Soybean oil price (dollar per metric tonne, $\$/MT$) data. Data are obtained from the "World Bank Commodity market" from January 1960 to December 2021(`https://www.worldbank.org/en/research/commodity-markets`). Each price series contains 744 observations divided into training and testing sets to ensure generalization capability. The training set carrying 732 data points is used to train the model, while the remaining 12 data points are used to test the effectiveness of the proposed model. Figure 2 shows time plots and the complex behaviour of each series, which is the characteristic of agricultural price data. Table 1 shows the basic descriptive statistics for each price series.

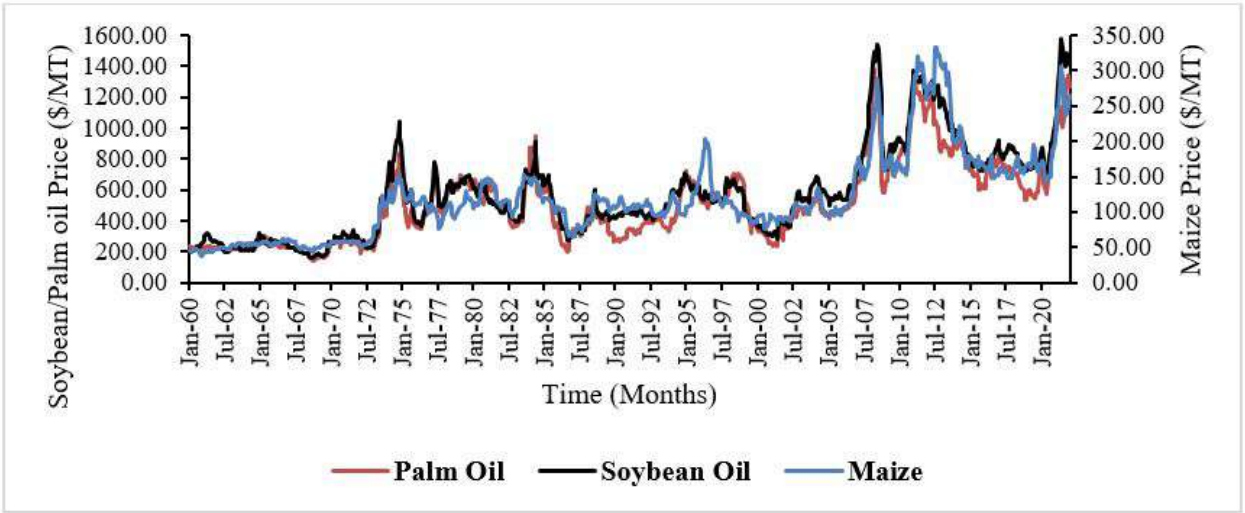


Figure 2: Time plots for monthly international Palm oil, Soybean oil and Maize price ($\$/MT$) series

Table 1: Descriptive statistics of the price ($/MT) series (from January 1960 to December 2021)

Statistics	Maize	Palm oil	Soybean oil
Mean	120.31	512.49	574.07
Maximum	333.05	1377.22	1574.67
Minimum	38.00	141.73	157.00
Standard deviation	59.65	257.75	294.08
Skewness	1.28	0.93	1.01
Kurtosis	1.80	3.58	0.82
Jarque-Bera	299.40	118.26	146.97

Since agricultural price series are complex and may exhibit nonstationarity and nonlinearity properties, it becomes necessary to test these properties, which can be helpful in skilful handling of data while fitting the model. The Augmented Dickey-Fuller (ADF) test, see Kumar *et al.* (2020), and Brock-Dechert-Scheinkman (BDS) test, see Choudhary *et al.* (2019), are used to check the stationarity and linearity characteristics, respectively. Table 2 shows the ADF test results that confirm the nonstationarity of each price series. Table 3 presents the BDS test results, which confirm the nonlinearity nature of each price series.

Table 2: Augmented Dickey-Fuller (ADF) test results

Price Series	ADF Test		Conclusion
	t-statistic	Probability	
Maize	−3.12	0.10	Nonstationary
Pam Oil	−3.14	0.09	Nonstationary
Soybean Oil	−3.01	0.15	Nonstationary

Table 3: Brock-Dechert-Scheinkman (BDS) test results

Price Series	Embedding dimension					Conclusion
	Epsilon	2		3		
		Statistics	Probability	Statistics	Probability	
Maize	0.5 σ	133.23	< 0.001	224.66	< 0.001	Nonlinear
	1.0 σ	67.16	< 0.001	77.15	< 0.001	
	1.5 σ	52.64	< 0.001	53.64	< 0.001	
	2.0 σ	41.41	< 0.001	39.72	< 0.001	
Palm oil	0.5 σ	230.22	< 0.001	399.29	< 0.001	Nonlinear
	1.0 σ	98.24	< 0.001	119.26	< 0.001	
	1.5 σ	64.78	< 0.001	67.36	< 0.001	
	2.0 σ	53.68	< 0.001	51.77	< 0.001	
Soybean oil	0.5 σ	193.52	< 0.001	335.36	< 0.001	Nonlinear
	1.0 σ	85.31	< 0.001	102.54	< 0.001	
	1.5 σ	60.46	< 0.001	62.45	< 0.001	
	2.0 σ	52.53	< 0.001	50.74	< 0.001	

Table 4: VMD parameters for different price series

Price series	α	τ	n	ϵ
Maize	2000	0	9	1×10^{-6}
Palm Oil	2000	0	9	1×10^{-7}
Soybean Oil	2000	1	9	1×10^{-6}

Table 5: Comparison of different decomposition algorithm in terms of θ

Price series	EMD	EEMD	CEEMDAN	VMD
Maize	0.1669	0.0472	0.0435	0.0214
Palm Oil	0.0420	0.0292	0.0190	0.0121
Soybean Oil	0.0608	0.0443	0.0279	0.0140

3.2. Decomposition of the agricultural price series

For the hybrid VMD-TDNN model, VMD is used to decompose the original agricultural price series into a set of IMFs. For decomposition, VMD requires four hyperparameters: (i) balancing parameter of the data-fidelity constrain(α), (ii) tolerance of convergence criterion (τ), (iii) number of modes (n), and (iv) time-step of the dual ascent (ϵ). The values of these hyperparameters are selected through experimentation in order to keep the energy evaluation parameter value (θ) as close to zero as possible to achieve superior decomposition outcomes and are presented in Table 4. While in the case of the number of modes, unlike EMD variants, a VMD technique provides as many modes as it is asked to produce, which significantly affects the accuracy of decomposition results. However, there is no practical or theoretical method to determine the optimum number of modes, see Dragomiretskiy and Zosso (2014); Lahmiri (2016). Therefore, in order to make all models comparable, the number of modes by VMD is chosen the same as that obtained by EMD and its variants. Accordingly, each price series is decomposed into nine different independent IMFs through VMD. Figure 3 shows the decomposed IMFs through VMD of the three price series from high frequency to low frequency. Here, high frequency shows the effect of short term fluctuations of the market, whereas low frequency represents any particularly significant event (like changes in policy, adverse effects of several biotic and abiotic factors, *etc.*) affecting the demand-supply equilibrium at that time. For instance, in our case, the two most significant events are observed in 2008 and 2011, which can be observed in Figure 3 in the form of spikes around 580^{th} and 620^{th} observations, respectively. Reasons behind both the events are the 2007-08's world food crisis and the production of biofuels, see Trostle (2011). For ethanol fuel production, usage of maize increased from 15% (2006) to 40% (2012) of total U.S. maize production. Moreover, the VMD based decomposed IMFs show more independent frequency distribution than the EMD variants, which can be empirically verified through the energy evaluation parameter (θ), see Zhu *et al.* (2016), defined as:

$$\theta = \frac{\left|\sqrt{\sum_{j=1}^{n} E_{j(t)}^2} - E_{y(t)}\right|}{E_{y(t)}}; \quad E_{y(t)} = \sqrt{\frac{\sum_{t=1}^{T} y^2(t)}{T}};$$

where $E_{y(t)}$ and $E_{j(t)}$ are the energy values of the original time series and j^{th}IMF, respectively. Here, θ is used as an evaluation parameter for orthogonality of IMFs such that θ closer to 0 indicates more orthogonality, whereas greater θ indicates the presence of elusive components among IMFs. Table 5 compares different decomposition methods in terms of θ for each price

series, which shows that the value of θ in the case of VMD is the smallest. This motivates us to choose VMD over other techniques to construct the TDNN based hybrid model.

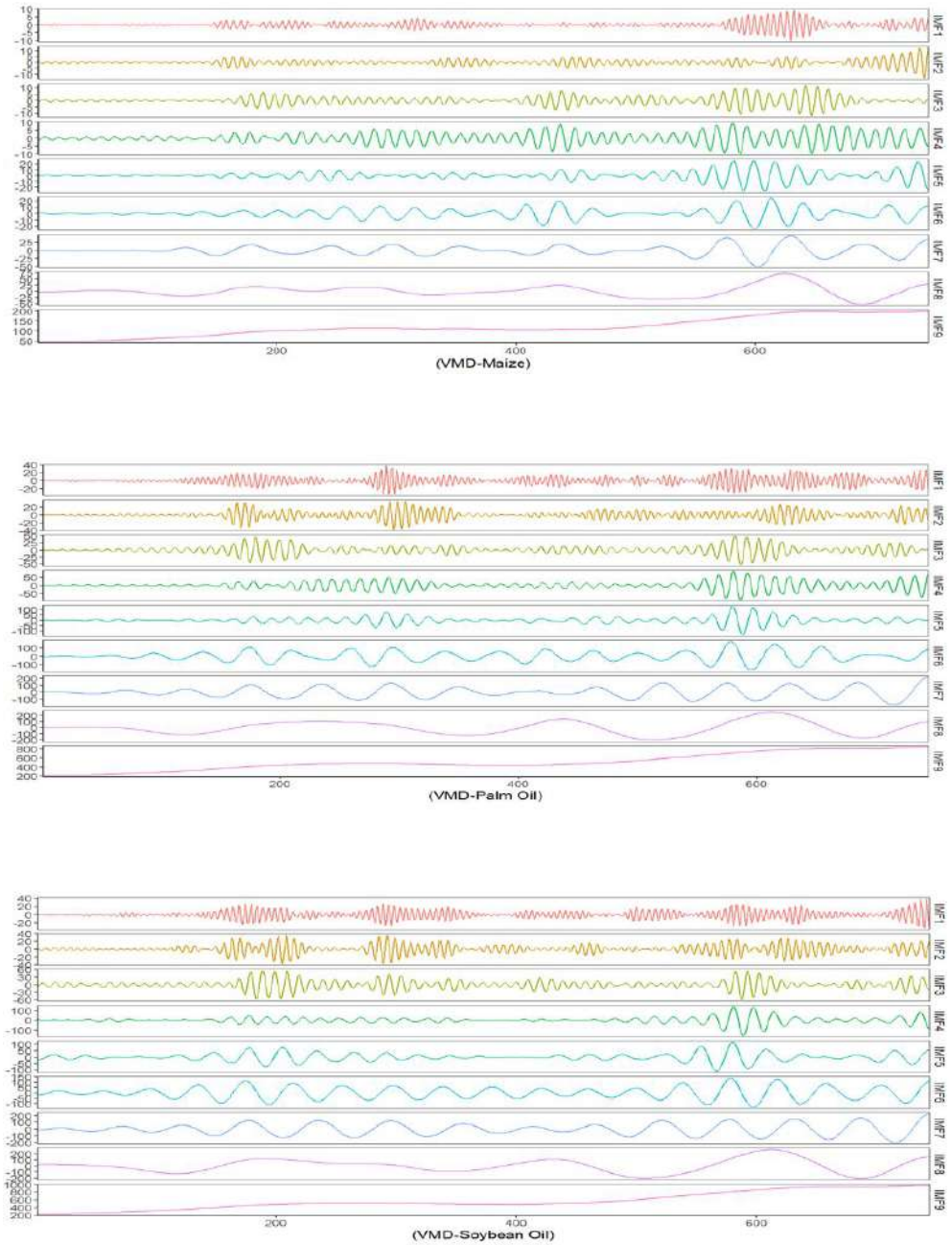


Figure 3: The decomposed IMFs for Maize, Palm oil and Soybean oil price series

3.3. Forecasting results and discussion

The datasets and code for machine learning-related forecasting studies are often not publicly available, making it impossible for the forecasting community to replicate and validate the stated performance. Thus, keeping in mind that one of the primary goals of our study is to make this work replicable by the entire research community for practical forecasting tasks, we develop two R software packages named eemdTDNN, see Choudhary *et al.* (2021) and vmdTDNN, see Choudhary *et al.* (2022), which are published in the comprehensive R archive network (CRAN). Here, the emdTDNN, EEMDTDNN, ceemdanTDNN and VMDTDNN are the functions of the above packages which are used to model and forecast each price series. The forecasting performance of the proposed VMD-TDNN model is compared with the existing individual model, *i.e.* TDNN and different hybrid models like EMD-TDNN, EEMD-TDNN, and CEEMDAN-TDNN for each price series. Figure 4 displays the plots of the predicted series by all models, along with the level series for each price series. The figure clearly shows that the VMD-TDNN model captures price movement patterns and directions better than conventional models. Moreover, the prediction ability of different models is tested in terms of different forecasting evaluation criteria. In this paper, RMSE, MAPE and directional prediction statistics (D_{Stat}) are employed to evaluate the performance of each model. Table 6 shows that all the hybrid models, including EMD-TDNN, EEMD-TDNN, CEEMDAN-TDNN and VMD-TDNN, outperform the single prediction model, *i.e.* TDNN, for each price series in terms of RMSE and MAPE.

It is mainly due to the "decomposition-ensemble principle" where decomposition techniques (EMD, EEMD, CEEMDAN, and VMD) reveal the hidden patterns of agricultural prices series and produce stationary and nonlinear modes which improve the forecasting ability of the TDNN. Among hybrid models, VMD-TDNN outperforms EMD-TDNN, EEMD-TDNN and CEEMDAN-TDNN in terms of both level and directional statistics since VMD is better than EMD variants, as discussed in section 3.2. With regards to D_{Stat} in particular, the results of the proposed VMD-TDNN model show better directional prediction than its competing models by showing 90% direction accuracy for maize series, almost 82% for palm oil, and 100% for soybean oil (Table 6). Though the different evaluation criteria used above show the superiority of the proposed model individually, there is ambiguity in choosing the best among other benchmark models as their results are not consistent. To get a better interpretation and a proper order of all models, we employ a novel technique called TOPSIS, which ranks all the models by combining their performances in both level and directional measures. Table 6 shows the ranks of each model obtained by the TOPSIS method.

Apart from these assessment criteria, the Diebold-Mariano (DM) test is also used to compare the predicting accuracy of various models statistically. Table 7 summarises the results of the DM test for each prediction model, and the following conclusions can be drawn. Firstly, the proposed VMD-TDNN model outperforms all existing models at a 5% significance level for each series. Secondly, all the hybrid models perform better than the TDNN model at the significance level of less than 1% for each series except for EMD-TDNN for palm oil which is significant at 4%.

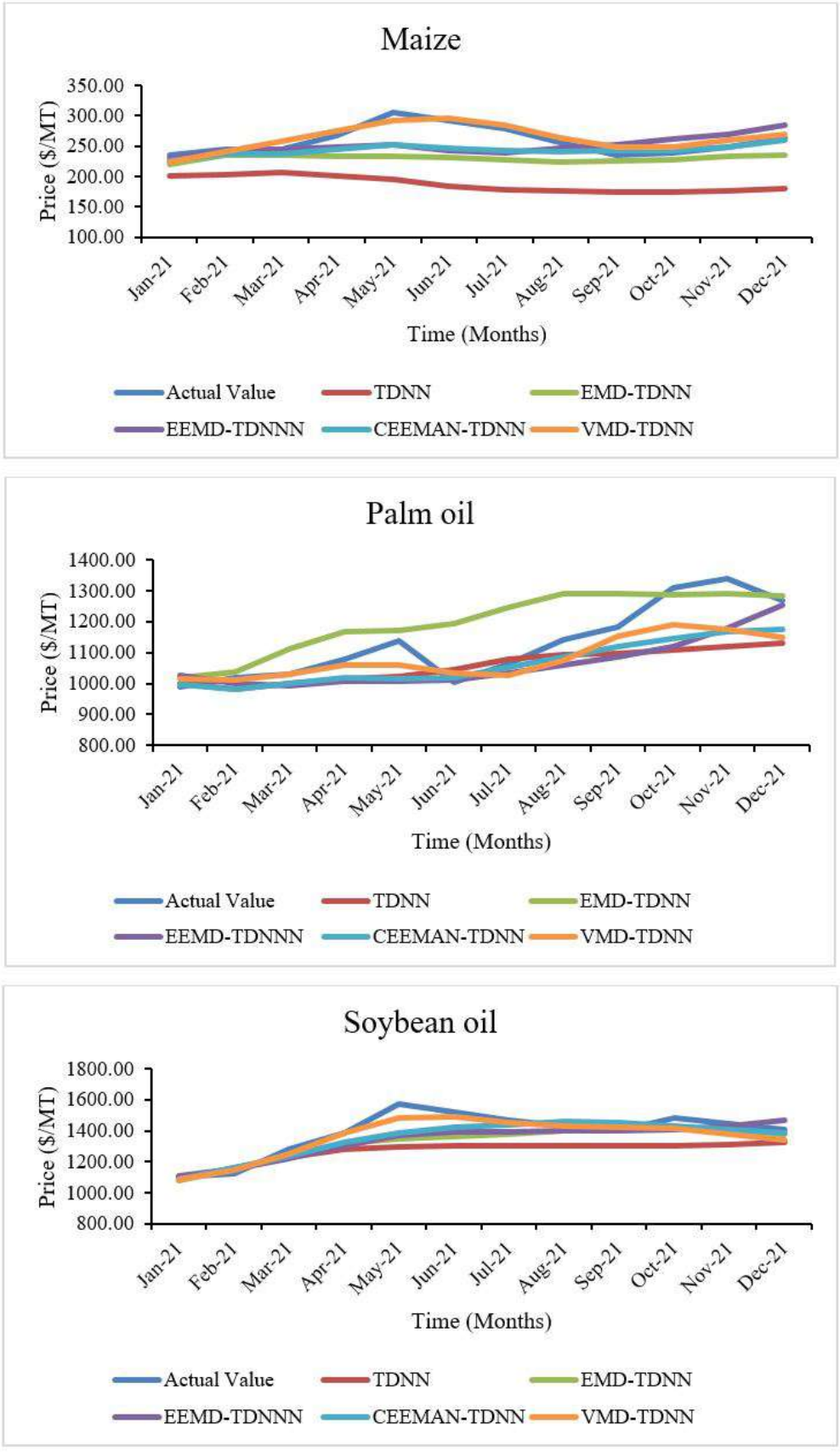


Figure 4: The predicted results of different models for Maize, Palm oil and Soybean oil price series

From the empirical analysis of various models using maize, palm oil, and soybean oil price series, it is clear that the proposed VMD-TDNN model significantly outperforms all other models regarding different forecasting evaluation criteria and thus can be considered as a competitive model for agricultural price forecasting. However, the VMD algorithm requires predetermination of the number of variational modes to be extracted contrary to the EMD and its variants. For EMD variants, the total number of modes is equal to $log_2 T$, where T is the total number of observations in the price series. Further, there is no theoretical or practical approach to determine the number (n) of extracted modes by VMD. For simplicity and to make all models comparable, the number of modes by VMD is chosen the same as that obtained by EMD and its variants, see Bisoi *et al.* (2019); Dragomiretskiy and Zosso (2014); Lahmiri (2016); Liu *et al.* (2018). Indeed, setting a higher number will further reduce the θ even if it is just a little but this will inevitably lead to higher computational burden and processing time during the decomposition process and the training of TDNN. In contrary, setting a lower number may lead to an inefficient representation and characterization of the original time series. The fact that with nine IMFs the VMD–TDNN achieved higher accuracy than others for all price series is very encouraging and promising in itself. However, a formal methodology should be developed in this regard in future works.

Table 6: Forecasting performance of different models for Maize, Palm oil and Soybean oil prices

Forecasting Models	Maize			Palm oil			Soybean oil			TOPSIS Rank
	MAPE	RMSE	Dstat	MAPE	RMSE	Dstat	MAPE	RMSE	Dstat	
TDNN	0.2725	76.29	45.45	0.0690	107.68	36.36	0.0854	143.64	72.73	5
EMD-TDNN	0.1075	36.09	54.54	0.0739	101.38	36.36	0.0477	91.02	90.91	4
EEMD-TDNN	0.0794	27.26	90.90	0.0613	92.57	36.36	0.0449	85.06	81.81	3
CEEMDAN-TDNN	0.0644	24.85	81.82	0.0575	88.86	81.81	0.0377	70.98	90.90	2
VMD-TDNN	0.0345	9.49	90.90	0.0478	76.90	81.81	0.0259	47.28	100.00	1

Table 7: Forecasting performance in terms of DM test of different models for Maize, Palm oil and Soybean oil prices for the one-year forecast horizon

Series	Tested Model	Benchmark Models			
		TDNN	EMD-TDNN	EEMD-TDNN	CEEMDAN-TDNN
Maize	**EMD-TDNN**	12.16(0.000)			
	EEMD-TDNN	13.96(0.000)	2.54(0.013)		
	CEEMDAN-TDNN	11.03(0.000)	5.44(0.000)	1.28(0.111)	
	VMD-TDNN	8.05(0.000)	3.12(0.004)	2.46(0.015)	3.67(0.001)
Palm Oil	**EMD-TDNN**	2.18(0.046)			
	EEMD-TDNN	2.83(0.002)	2.27(0.019)		
	CEEMDAN-TDNN	2.36(0.018)	2.31(0.018)	2.39(0.015)	
	VMD-TDNN	2.76(0.009)	2.81(0.005)	2.08(0.053)	2.56(0.012)
Soybean Oil	**EMD-TDNN**	5.19(0.000)			
	EEMD-TDNN	4.42(0.000)	1.76(0.042)		
	CEEMDAN-TDNN	4.87(0.000)	1.91(0.040)	1.39(0.095)	
	VMD-TDNN	4.66(0.000)	2.05(0.031)	1.90(0.041)	3.44(0.002)

4. Conclusions

Agricultural price series are highly vulnerable to several risks due to biotic and abiotic factors, which account for several characteristics, including nonlinearity and nonstationarity. This paper proposes a new hybrid VMD-TDNN model to improve the prediction accuracy of agricultural price data. The VMD algorithm decomposes a series into a set of subseries

or modes for the proposed model. These obtained modes are forecasted separately using the TDNN model, and their forecasted values are aggregated to give a final forecast for a given price series data. VMD has many advantages over EMD based methods, including a better mathematical foundation, data adaptiveness capability, robustness to noise and faster convergence with better accuracy. For empirical evaluation, an extensive comparative analysis of the forecasting performance of the proposed VMD-TDNN model with the four different models is performed using three monthly international price series. The empirical results show that the VMD-TDNN outperforms the competing models in terms of different forecasting evaluation criteria like MAPE, RMSE and D_{stat}. In addition, to better understand the proper order, we utilise a unique technique called TOPSIS, which ranks all models by combining their performances of both level and directional metrics, and VMD-TDNN stands first among all. Further, the DM test result shows that the VMD-TDNN model significantly improves forecasting accuracy from other models. Overall, we can state that the proposed model provides a valuable decision support tool for every agricultural stakeholder who falls in the domain of agricultural price forecasting.

Acknowledgements

The first author is grateful to the University Grants Commission (UGC) for giving financial assistance and the PG School, ICAR-Indian Agricultural Research Institute, New Delhi, for providing the essential facilities for this research. We are also thankful to the reviewer for many insightful comments which have improved the quality of the paper. Further, the reviewer also suggested different packages in LaTeX for improving the readability of the manuscript.

Conflict of interest

There are no possible conflicts of interest that the authors should disclose that are relevant to the content of the work.

References

Bisoi, R., Dash, P. K., and Parida, A. K. (2019). Hybrid variational mode decomposition and evolutionary robust kernel extreme learning machine for stock price and movement prediction on daily basis. *Applied Soft Computing*, **74**, 652–678.

Box, G., Jenkins, G., Reinsel, G., and Ljung, G. (2015). *Time Series Analysis: Forecasting and Control*. John Wiley and Sons, 4 edition.

Choudhary, K., Jha, G. K., Kumar, R. R., and Jaiswal, R. (2021). eemdTDNN: EEMD and its variant based time-delay neural network model, https://cran.r-project.org/package=eemdTDNN.

Choudhary, K., Jha, G. K., Kumar, R. R., and Mishra, D. C. (2019). Agricultural commodity price analysis using ensemble empirical mode decomposition: A case study of daily potato price series. *Indian Journal of Agricultural Sciences*, **89**, 882–886.

Choudhary, K., Jha, G. K., Parsad, R., and Jaiswal, R. (2022). vmdTDNN :VMD based time-delay neural network model, https://cran.r-project.org/package=vmdTDNN.

Dragomiretskiy, K. and Zosso, D. (2014). Variational mode decomposition. *IEEE Transactions on Signal Processing*, **62**, 531–544.

Fang, Y., Guan, B., Wu, S., and Heravi, S. (2020). Optimal forecast combination based on ensemble empirical mode decomposition for agricultural commodity futures prices. *Journal of Forecasting*, **39**, 877–886.

FAO (2011). Price volatility in food and agricultural markets : policy responses. pages 1–68.

Hayat, A. and Bhatti, M. I. (2013). Masking of volatility by seasonal adjustment methods. *Economic Modelling*, **33**, 676–688.

Haykin, S. (2009). *Neural Networks and Learning Machines, Person Education*, volume 1-3. PHI Learning, INDIA, 3 edition.

Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Snin, H. H., Zheng, Q., Yen, N. C., Tung, C. C., and Liu, H. H. (1998). The empirical mode decomposition and the Hubert spectrum for nonlinear and nonstationary time series analysis. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences*, **454**, 903–995.

Hwang, C. L. and Yoon, K. (1981). *In Multiple attribute decision making*, pages 58–191. Springer.

Jadhav, V., Chinnappa Reddy, B. V., and Gaddi, G. M. (2017). Application of arima model for forecasting agricultural prices. *Journal of Agricultural Science and Technology*, **19**, 981–992.

Jaiswal, R., Jha, G. K., Kumar, R. R., and Choudhary, K. (2022). Deep long short-term memory based model for agricultural price forecasting. *Neural Computing and Applications*, **34**, 4661–4676.

Jha, G. K. and Sinha, K. (2014). Time-delay neural networks for time series prediction: An application to the monthly wholesale price of oilseeds in India. *Neural Computing and Applications*, **24**, 563–571.

Kumar, R. R., Jha, G. K., Choudhary, K., and Mishra, C. (2020). Spatial integration and price transmission among major potato markets in India. *Indian Journal of Agricultural Sciences*, **90**, 581–585.

Lahmiri, S. (2016). Intraday stock price forecasting based on variational mode decomposition. *Journal of Computational Science*, **12**, 23–27.

Liu, H., Mi, X., and Li, Y. (2018). Smart multi-step deep learning model for wind speed forecasting based on variational mode decomposition, singular spectrum analysis, LSTM network and ELM. *Energy Conversion and Management*, **159**, 54–64.

Prasad, R., Deo, R. C., Li, Y., and Maraseni, T. (2018). Soil moisture forecasting by a hybrid machine learning technique: ELM integrated with ensemble empirical mode decomposition. *Geoderma*, **330**, 136–161.

Qian, Z., Pei, Y., Zareipour, H., and Chen, N. (2019). A review and discussion of decomposition-based hybrid models for wind energy forecasting applications. *Applied Energy*, **235**, 939–953.

Torres, M. E., Colominas, M. A., Schlotthauer, G., and Flandrin, P. (2011). A complete ensemble empirical mode decomposition with adaptive noise. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, pages 4144–4147.

Trostle, R. (2011). Why another food commodity price spike? Technical report.

Wu, Z. and Huang, N. E. (2009). Ensemble empirical mode decomposition: A noise-assisted data analysis method. *Advances in Adaptive Data Analysis*, **1**, 1–41.

Xu, X. and Zhang, Y. (2021). Corn cash price forecasting with neural networks. *Computers and Electronics in Agriculture*, **184**, 106–120.

Yu, L., Wang, S., and Lai, K. K. (2008). Forecasting crude oil price with an EMD-based neural network ensemble learning paradigm. *Energy Economics*, **30**, 2623–2635.

Zhang, G., Eddy Patuwo, B., and Y. Hu, M. (1998). Forecasting with artificial neural networks: The state of the art. *International Journal of Forecasting*, **14**, 35–62.

Zhu, B., Shi, X., Chevallier, J., Wang, P., and Wei, Y. M. (2016). An Adaptive Multiscale Ensemble Learning Paradigm for Nonstationary and Nonlinear Energy Price Time Series Forecasting. *Journal of Forecasting*, **35**, 633–651.

Appendix

R codes used for empirical evaluation of the study

```
# To check and install the packages used in this analysis
ipak <- function(pkg){
  new.pkg <- pkg[!(pkg %in% installed.packages()[ ,"Package"])]
  if (length(new.pkg))
    install.packages(new.pkg, dependencies = TRUE)
  sapply(pkg, require, character.only = TRUE)
}
# usage
packages <- c("tseries", "moments", "Rlibeemd", "VMDecomp")
ipak(packages)
library(tseries)
library(moments)
library(Rlibeemd)
library(VMDecomp)


#Importing the actual price data specifying the location of the data file
data=read.csv(file.choose(), header=TRUE)
data
#plotting of the imported data
plot(ts(data))

#transforming the data into numeric vector (1-dimensional)
data=as.matrix(data)
data=as.vector(data)

#Basic descriptive of the data set
library(moments)
summary(data)
sd(data)
skewness(data)
kurtosis(data)
jarque.test(data)

#Stationarity and linearity test
adf.test(data) # Augmented Dickey-fuller test for testing stationarity
bds.test(data) # Brock-Dechert Shienkman test for testing nonlinearity

#Data Decomposition through EMD, EEMD, CEEMDAN, and VMD
library(Rlibeemd)

# Decomposition of price data using EMD technique
EMD=emd(data, num_imfs = 0, S_number = 4L, num_siftings = 50L)
```

```
# Decomposition of price data using EEMD technique
EEMD=eemd(ts(data), num_imfs = 0, ensemble_size = 250L, noise_strength = 0.2,
↪  S_number = 4L, num_siftings = 50L, rng_seed = 0L, threads = 0L)

# Decomposition of price data using CEEMDAN technique
CEEMDAN=ceemdan(ts(data), num_imfs = 0, ensemble_size = 250L, noise_strength =
↪  0.2, S_number = 4L, num_siftings = 50L, rng_seed = 0L, threads = 0L)

# Decomposition of price data using VMD technique
VMD=vmd(data, alpha = 2000, tau = 0, K = 9, DC = FALSE, init = 1, tol = 1e-06)

# Plotting of decomposed series

# PLotting of decomposed series extracted by EMD technique
plot(EMD,xlab="Time (Month)")

# PLotting of decomposed series extracted by EEMD technique
plot(EEMD,xlab="Time (Month)")

# PLotting of decomposed series extracted by CEEMDAN technique
plot(CEEMDAN,xlab="Time (Month)")


# VMdecomp package does not allow for auto-plot of all series,
# so we will extract all the decomposed series one by one done by VMD technique
# and then combine them in a two dimensional matrix and then plot them

#Extraction of all IMFS
AllIMF <- ts(VMD$u)

# VMD decompose price series in reverse order (From low to high frequency)
# in contrary to EMD variants.So IMF1 will be the last column, IMF2 will be
# the second last column,...

# Extraction of each IMF one by one
IMF1=ts(AllIMF[,9])
IMF2=ts(AllIMF[,8])
IMF3=ts(AllIMF[,7])
IMF4=ts(AllIMF[,6])
IMF5=ts(AllIMF[,5])
IMF6=ts(AllIMF[,4])
IMF7=ts(AllIMF[,3])
IMF8=ts(AllIMF[,2])
IMF9=ts(AllIMF[,1])

# Combining of all IMFs
VMD_IMFs <- cbind.data.frame(IMF1, IMF2, IMF3, IMF4, IMF5, IMF6, IMF7, IMF8,IMF9)
VMD_IMFs  <- ts(VMD_IMFs)

# Plotting of all IMFs of VMD together
```

```
plot(VMD_IMFs)

# Modelling and Forecasting results of EMDTDNN model
emd_tdnn=function(data, stepahead = 12, num.IMFs = emd_num_imfs(length(data)),
          s.num = 4L, num.sift = 50L)
{
  n.IMF <- num.IMFs # To find the total number of IMFs
  AllIMF <- emd(data, num_imfs = n.IMF, S_number = s.num, num_siftings =
  ↪  num.sift)
  data_trn <- ts(head(data, round(length(data) - stepahead))) # Extracting the
  ↪  training set
  data_test <- ts(tail(data, stepahead)) # Extracting the testing set
  IMF_trn <- AllIMF[-c(((length(data) - stepahead) + 1):length(data)),
  ]
  Fcast_AllIMF <- NULL
  # Applying For loop to model and forecast each decomposed series using TDNN
  ↪  model
  for (IMF in 1:ncol(IMF_trn)) {
    IndIMF <- NULL
    IndIMF <- IMF_trn[, IMF]
    EMDTDNNFit <- forecast::nnetar(as.ts(IndIMF))
    EMDTDNN_fcast = forecast::forecast(EMDTDNNFit, h = stepahead)
    EMDTDNN_fcast_Mean = EMDTDNN_fcast$mean
    Fcast_AllIMF <- cbind(Fcast_AllIMF, as.matrix(EMDTDNN_fcast_Mean))
  }
  # Combining all the forecasts to get final forecast using EMD-TDNN
  FinalEMDTDNN_fcast <- ts(rowSums(Fcast_AllIMF, na.rm = T))
  # Finding different evaluation criteria based on testing data set
  MAE_EMDTDNN = mean(abs(data_test - FinalEMDTDNN_fcast))
  MAPE_EMDTDNN = mean(abs(data_test - FinalEMDTDNN_fcast)/data_test)
  rmse_EMDTDNN = sqrt(mean((data_test - FinalEMDTDNN_fcast)^2))
  Plot_IMFs <- AllIMF
  AllIMF_plots <- plot(Plot_IMFs)
  return(list(TotalIMF = n.IMF, AllIMF = AllIMF, data_test = data_test,
              AllIMF_forecast = Fcast_AllIMF, FinalEMDTDNN_forecast =
              ↪  FinalEMDTDNN_fcast,
              MAE_EMDTDNN = MAE_EMDTDNN, MAPE_EMDTDNN = MAPE_EMDTDNN,
              rmse_EMDTDNN = rmse_EMDTDNN, AllIMF_plots = AllIMF_plots))
}
EMDTDNN=emd_tdnn(data, stepahead = 12, num.IMFs = emd_num_imfs(length(data)),
                 s.num = 4L, num.sift = 50L)
EMDTDNN

# Forecasting result of EEMDTDNN model
EEMD_TDNN=function (data, stepahead = 12, num.IMFs = emd_num_imfs(length(data)),
          s.num = 4L, num.sift = 50L, ensem.size = 250L, noise.st = 0.2)
{
  n.IMF <- num.IMFs # To find the total number of IMFs
  AllIMF <- eemd(ts(data), num_imfs = n.IMF, ensemble_size = ensem.size,
```

```
                noise_strength = noise.st, S_number = s.num, num_siftings =
                ↪  num.sift,
                rng_seed = 0L, threads = 0L)
  data_trn <- ts(head(data, round(length(data) - stepahead)))  # Extracting the
  ↪  training set
  data_test <- ts(tail(data, stepahead))  # Extracting the testing set
  IMF_trn <- AllIMF[-c(((length(data) - stepahead) + 1):length(data)),
  ]
  Fcast_AllIMF <- NULL
  # Applying For loop to model and forecast each decomposed series using TDNN
  ↪  model
  for (IMF in 1:ncol(IMF_trn)) {
    IndIMF <- NULL
    IndIMF <- IMF_trn[, IMF]
    EEMDTDNNFit <- forecast::nnetar(as.ts(IndIMF))
    EEMDTDNN_fcast = forecast::forecast(EEMDTDNNFit, h = stepahead)
    EEMDTDNN_fcast_Mean = EEMDTDNN_fcast$mean
    Fcast_AllIMF <- cbind(Fcast_AllIMF, as.matrix(EEMDTDNN_fcast_Mean))
  }
  # Combining all the forecasts to get final forecast using EMD-TDNN
  FinalEEMDTDNN_fcast <- ts(rowSums(Fcast_AllIMF, na.rm = T))
  # Finding different evaluation criteria based on testing data set
  MAE_EEMDTDNN = mean(abs(data_test - FinalEEMDTDNN_fcast))
  MAPE_EEMDTDNN = mean(abs(data_test - FinalEEMDTDNN_fcast)/data_test)
  rmse_EEMDTDNN = sqrt(mean((data_test - FinalEEMDTDNN_fcast)^2))
  Plot_IMFs <- AllIMF
  AllIMF_plots <- plot(Plot_IMFs)
  return(list(TotalIMF = n.IMF, data_test = data_test, AllIMF_forecast =
  ↪  Fcast_AllIMF,
              FinalEEMDTDNN_forecast = FinalEEMDTDNN_fcast, MAE_EEMDTDNN =
              ↪  MAE_EEMDTDNN,
              MAPE_EEMDTDNN = MAPE_EEMDTDNN, rmse_EEMDTDNN = rmse_EEMDTDNN,
              AllIMF_plots = AllIMF_plots))
}
EEMDTDNN=EEMD_TDNN(data, stepahead = 12, num.IMFs = emd_num_imfs(length(data)),
                   s.num = 4L, num.sift = 50L, ensem.size = 250L, noise.st = 0.2)
EEMDTDNN

#Forecasting Result of CEEMDANTDNN model
ceemdan_TDNN=function (data, stepahead = 12, num.IMFs =
↪  emd_num_imfs(length(data)),
          s.num = 4L, num.sift = 50L, ensem.size = 250L, noise.st = 0.2)
{
  n.IMF <- num.IMFs # To find the total number of IMFs
  AllIMF <- ceemdan(ts(data), num_imfs = n.IMF, ensemble_size = ensem.size,
                    noise_strength = noise.st, S_number = s.num, num_siftings =
                    ↪  num.sift,
                    rng_seed = 0L, threads = 0L)
  data_trn <- ts(head(data, round(length(data) - stepahead)))  # Extracting the
  ↪  training set
```

```
  data_test <- ts(tail(data, stepahead))  # Extracting the testing set
  IMF_trn <- AllIMF[-c(((length(data) - stepahead) + 1):length(data)),
  ]
  Fcast_AllIMF <- NULL
  # Applying For loop to model and forecast each decomposed series using TDNN
  ↪ model
  for (IMF in 1:ncol(IMF_trn)) {
    IndIMF <- NULL
    IndIMF <- IMF_trn[, IMF]
    CEEMDANTDNNFit <- forecast::nnetar(as.ts(IndIMF))
    CEEMDANTDNN_fcast = forecast::forecast(CEEMDANTDNNFit,
                                           h = stepahead)
    CEEMDANTDNN_fcast_Mean = CEEMDANTDNN_fcast$mean
    Fcast_AllIMF <- cbind(Fcast_AllIMF, as.matrix(CEEMDANTDNN_fcast_Mean))
  }
  # Combining all the forecasts to get final forecast using EMD-TDNN
  FinalCEEMDANTDNN_fcast <- ts(rowSums(Fcast_AllIMF, na.rm = T))
  # Finding different evaluation criteria based on testing data set
  MAE_CEEMDANTDNN = mean(abs(data_test - FinalCEEMDANTDNN_fcast))
  MAPE_CEEMDANTDNN = mean(abs(data_test - FinalCEEMDANTDNN_fcast)/data_test)
  rmse_CEEMDANTDNN = sqrt(mean((data_test - FinalCEEMDANTDNN_fcast)^2))
  Plot_IMFs <- AllIMF
  AllIMF_plots <- plot(Plot_IMFs)
  return(list(TotalIMF = n.IMF, data_test = data_test, AllIMF_forecast =
  ↪ Fcast_AllIMF,
              FinalCEEMDANTDNN_forecast = FinalCEEMDANTDNN_fcast, MAE_CEEMDANTDNN
              ↪ = MAE_CEEMDANTDNN,
              MAPE_CEEMDANTDNN = MAPE_CEEMDANTDNN, rmse_CEEMDANTDNN =
              ↪ rmse_CEEMDANTDNN,
              AllIMF_plots = AllIMF_plots))
}
CEEMDANTDNN=ceemdan_TDNN(data, stepahead = 12, num.IMFs =
↪ emd_num_imfs(length(data)),
                         s.num = 4L, num.sift = 50L, ensem.size = 250L, noise.st =
                         ↪ 0.2)
CEEMDANTDNN

# Forecasting Result of VMDTDNN model
VMD_TDNN=function (data, stepahead = 12, nIMF = 9, alpha = 2000, tau = 0,
          D = FALSE)
{
  data <- ts(data)
  data <- as.vector(data)
  v <- vmd(data, alpha = 2000, tau = 0, K = nIMF, DC = D, init = 1,
           tol = 1e-06)
  AllIMF <- v$u
  data_trn <- ts(head(data, round(length(data) - stepahead))) # Extracting the
  ↪ training set
  data_test <- ts(tail(data, stepahead)) # Extracting the testing set
  IMF_trn <- AllIMF[-c(((length(data) - stepahead) + 1):length(data)),
```

```
  ]
  Fcast_AllIMF <- NULL
  # Applying For loop to model and forecast each decomposed series using TDNN
  ↪  model
  for (AllIMF in 1:(ncol(IMF_trn))) {
    IndIMF <- NULL
    IndIMF <- IMF_trn[, AllIMF]
    VMDTDNNFit <- forecast::nnetar(as.ts(IndIMF))
    VMDTDNN_fcast = forecast::forecast(VMDTDNNFit, h = stepahead)
    VMDTDNN_fcast_Mean = VMDTDNN_fcast$mean
    Fcast_AllIMF <- cbind(Fcast_AllIMF, as.matrix(VMDTDNN_fcast_Mean))
  }
  # Combining all the forecasts to get final forecast using EMD-TDNN
  FinalVMDTDNN_fcast <- ts(rowSums(Fcast_AllIMF, na.rm = T))
  # Finding different evaluation criteria based on testing data set
  MAE_VMDTDNN = mean(abs(data_test - FinalVMDTDNN_fcast))
  MAPE_VMDTDNN = mean(abs(data_test - FinalVMDTDNN_fcast)/data_test)
  RMSE_VMDTDNN = sqrt(mean((data_test - FinalVMDTDNN_fcast)^2))
  return(list(AllIMF = AllIMF, data_test = data_test, AllIMF_forecast =
  ↪  Fcast_AllIMF,
              FinalVMDTDNN_forecast = FinalVMDTDNN_fcast, MAE_VMDTDNN =
              ↪  MAE_VMDTDNN,
              MAPE_VMDTDNN = MAPE_VMDTDNN, RMSE_VMDTDNN = RMSE_VMDTDNN))
}
VMDTDNN=VMD_TDNN(data, stepahead = 12, nIMF = 9, alpha = 2000, tau = 0,
                 D = FALSE)
VMDTDNN
```

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 21, No. 2, 2023 (New Series), pp 259–278
http://www.ssca.org.in/journal

Number of Overlapping Runs Until a Stopping Time for Higher Order Markov Chain

Anuradha
Department of Statistics, Lady Shri Ram College for Women
University of Delhi, Lajpat Nagar-IV, New Delhi – 110024, India

Received: 02 November 2022; Revised: 23 December 2022; Accepted: 05 April 2023

Abstract

Let $\{X_j : j \geq -m+1\}$ be a homogeneous Markov chain of order m taking values in $\{\mathbf{0}, \mathbf{1}\}$. For $j = 0, -1, \ldots, -l+1$, we will set $R_j = 0$ and we define $R_j = \prod_{i=j-1}^{j-l}(1 - R_i) \prod_{i=j}^{j+k-1} X_i$. Now $R_j = 1$ implies that an l-look-back run of length k has occurred starting at j. Here R_j is defined inductively as a run of $\mathbf{1}$'s starting at j, provided that no l-look-back run of length k occurs, starting at time $j-1, j-2, \ldots, j-l$ respectively. We study the conditional distribution of the number of overlapping runs of length k_1 until the stopping time *i.e.* the r^{th} occurrence of the l-look-back run of length k where $k_1 \leq k$ and obtain it's probability generating function. The number of overlapping runs of length k_1 until the stopping time has been expressed as the sum of r independent random variables with the first random variable having a slightly different distribution. We introduce a new discrete distribution, namely *generalized Binomial type* distribution, which plays a central role in our study. The conditional distributions are identified using this and other known distributions, such as extended negative binomial distribution of order k. Our results also generalize the known results for the number of successes until a stopping time.

Key words: Overlapping runs; Stopping time; Markov chain; Strong Markov property; Probability generating functions.

AMS Subject Classifications: 60C05, 60E05, 60F05

1. Introduction

Since Feller (1968) introduced runs of successes as an example of a renewal event, the theory of distributions of runs has been explored widely by the researchers. The application of powerful techniques such as Markov embedding technique (see Fu and Koutras (1994)), method of conditional p.g.f.s (see Ebneshahrashoob and Sobel (1990)) *etc.* has paved way to develop and study the distributions of various run statistics and their properties extensively.

Two schemes of counting runs, namely non-overlapping counting and the overlapping counting, have been extensively studied in the literature. As the name suggests, in the

Corresponding Author: Anuradha
Email: sarkar.anuradha@gmail.com

non-overlapping counting, runs are not allowed to overlap while in the overlapping counting scheme the runs may overlap as much as possible. Philippou and Makri (1986) studied the distribution of number of non-overlapping runs of successes of length k for *i.i.d.* Bernoulli trials and introduced the Binomial distribution of order k. Ling (1988) derived the distribution of the number of overlapping runs of successes of length k for a sequence of *i.i.d.* Bernoulli trials. This distribution is referred as the Type II Binomial distribution of order k. Aki and Hirano (1994) obtained the marginal distributions of number of failures, successes and success-runs of length less than k until the first occurrence of consecutive k successes when the underlying random variables were either *i.i.d.* or Markov dependent or binary sequence of order k. Aki and Hirano (1995) derived the joint distributions of number of failures, successes and runs of success under the same set up. Under different types of counting schemes like runs of length k_1, non overlapping runs of length k_1, overlapping runs of length k_1 *etc.*, Hirano *et. al.* (1997) gave interesting results on the distributions of number of success runs of length l until the first occurrence of the success run of length k for an m^{th} order homogeneous Markov chain. The joint distributions of the waiting time and the number of outcomes such as failures, successes and success runs of length less than k under the set up of an m^{th} order homogeneous Markov chain was developed by Uchida (1998) for various enumeration schemes of runs. Chadjiconstantindis and Koutras (2001) also obtained the distribution of failures and successes in a waiting time problem.

Another scheme of μ-overlapping counting was introduced by Aki and Hirano (2000) where an overlap of at most μ successes was allowed between two consecutive runs of length k where $0 \leq \mu \leq k-1$. They also introduced the generalized Binomial distribution of order k and investigated some of it's properties. It is easy to observe that when $\mu = 0$, the counting scheme matches with the non-overlapping counting while $\mu = k - 1$ yields the overlapping counting scheme. Han and Aki (2000) have extended this counting scheme for the negative values of μ in which there should be at least $|\mu|$ trials between any two consecutive success runs of length k. They have derived recurrence relations for the probability generating function (*pgf*) of the number of μ-overlapping success runs of length k. Inoue and Aki (2003) derived exact formulae for the *pgf* of the above-mentioned random variable in the case of two-state Markov dependent trials. They also derived explicitly, in the same case, the *pgf* of the waiting time until the r^{th} occurrence of the μ-overlapping success run of length k. Makri and Philippou (2005) obtained the exact formulas for the probability distribution function of the number of μ-overlapping success runs of length k in n trials. Makri *et. al.* (2007) considered the concept of μ-overlapping success runs in the Polya-Eggenberger sampling scheme and obtained the distribution of the number of drawings according to the Polya-Eggenberger sampling scheme until the r^{th} occurrence of the μ-overlapping success run of length k. They have also introduced Polya, inverse Polya, and circular Polya distributions of order k for μ-overlapping success runs of length k.

Anuradha (2023) introduced the l-look-back counting scheme for runs of successes. In this scheme, if a run has been counted starting at time i, *i.e.*, $\{X_i = X_{i+1} = \cdots = X_{i+k-1} = \mathbf{1}\}$, then no runs can be counted till the time point $i + l$ and the next counting of runs can start only from the time point $i + l + 1$, where $X_i = \mathbf{1}$ represents a success at time i and l is a non-negative integer. This process is repeated every time a run is counted. In other words, if a run is counted starting at time i, then there are k-consecutive successes starting from the time point i and no runs of length k has been counted starting at time points $i-1, i-2, \ldots, i-l$. The mathematical definition has been provided in the section 3.

The look-back counting scheme generalizes the concept of run counting and encompasses both the definitions of overlapping counting as well as the non-overlapping counting thereby giving rise to new objects for further study. Indeed, if $l = 0$, this matches exactly with the counting of overlapping runs of length k, and if $l = k - 1$, this counting scheme results in the counting of non-overlapping runs of length k. It should further be noted that μ-overlapping scheme, for positive values of μ, can also be identified as l-look-back counting where $\mu = k - l - 1$. However, for negative values of μ, the definitions do not match with the corresponding value of $l = k - \mu - 1$. We illustrate this difference with the same example as cited in Han and Aki (2000). Consider the following sequence of successes and failures:

$$\mathbf{11110110001111111110000111}.$$

In this sequence, for $k = 3$ and $l = 3$, we have four 3-look-back runs of length 3 starting at trials $1, 11, 15$ and 23, while there are only three (-1)-overlapping runs of length 3, starting at $1, 11$ and 15. Therefore l-look-back counting scheme is an entirely new scheme of counting which has not yet been studied in detail.

Under the set up of m^{th} order homogeneous Markov chain, Anuradha (2023) proved that the waiting time distribution of the n^{th} occurrence of the l-look-back run of length k converges to an extended Poisson distribution when the system exhibits strong propensity towards success. Further central limit theorem was established for the number of l-look-back runs of length k till the n^{th} trial.

Aki and Hirano (2000) established that the number of $(l-1)$-overlapping runs of length k $(l < k)$ until the n^{th} overlapping occurrence of success run of length l follows a generalized Binomial distribution of order $(k - l)$ for the *i.i.d.* as well as the m^{th} order homogeneous Markov chain. In this paper, we pose a different problem from the counting perspective. We fix two positive integers $k_1 \leq k$ and another integer $l \geq 0$ and count the number of overlapping runs of length k_1 until the n^{th} occurrence of l-look-back run of length k. The stopping time originates from the l-look-back counting scheme which encompasses non-overlapping, overlapping as well as μ-overlapping (for positive μ) counting schemes. We should also note that there is no restriction on l, which may equal or exceed k. Our focus is on counting of runs of smaller lengths (k_1) until a stopping rule which involves occurrences of runs of larger length (k). We obtain a decomposition of the number of runs until the stopping time into a sum of independent random variables. This, in turn, brings out a new discrete distribution of order k and also establishes new connections with the other known discrete distributions.

Koutras (1997) defined a Markov Negative Binomial distribution of order k where he studied the waiting time distributions associated with the runs of length k for a two-state Markov chain. In this paper, we introduce a new distribution which is different from the above. We denote it by *generalized Binomial type distribution*. The probability generating function of this distribution has been derived, which also shows how it generalizes the classical Binomial distribution and the classical negative Binomial distribution (refer to Definition 1). In our study, the generalized Binomial type distribution will play a central role, along with the extended negative binomial distribution of order k with parameters n and $(p_1, p_2, \ldots, p_k)$ which was introduced by Aki (1985).

Our results show that the number of overlapping runs of length k_1 up to the r^{th} occurrence of the l-look-back run of length k $(k_1 \leq k)$ can be split into a sum of r independent

random variables. We further establish that except the first one, all the other random variables are identically distributed. The result has a number of interesting corollaries. For example, the results of Aki and Hirano (1994), on the number of successes until the first occurrence of the k consecutive successes for the *i.i.d.* as well as the Markov chain set-up can be derived as a corollary from our result (see Corollary 3). We also show that under the assumption of strong tendency towards failure after k consecutive successes, the number of overlapping success runs of length k_1 can be approximated by Poisson random variable translated by r (see Corollary 2).

We employ a new technique to prove our results. First we convert the m^{th} order Markov chain to a first order Markov chain which takes values in a finite set and recast our problem into this new set-up, *i.e.*, define the success / failure in the original chain in terms of the new chain and convert all relevant definitions in terms of the new chain. Thereafter, the main tool that we employ is the method of generating functions. We use the strong Markov property on this first order Markov chain to derive a recurrence relations between the probabilities. This, in turn, yields recurrence relations between the probability generating functions (*pgfs*). Finally we consider the generating function of the *pgfs*. Using the recurrence relations between the *pgfs* we obtain a linear equation involving the generating function of the *pgfs* which is used to establish its expression. Expanding this generating function of the *pgfs*, we obtain the expression for the individual *pgf*.

In the next section, we introduce the new discrete distribution, namely *generalized Binomial type distribution* and provide the probability generating function of the distribution. In section 3, we give all the definitions and state the main result and the corollaries. Section 4 is devoted to setting up the new Markov chain and recasting of the problem in terms of the new Markov chain. In Section 5, the proof of the main theorem has been established. In the final section, we provide the conclusion of the paper.

2. A new discrete distribution

In this section we introduce a discrete distribution which will be important for our work.

Definition 1: We say that a random variable W follows a *generalized Binomial type* distribution with parameters $0 < p < 1$, $n \geq 0$ and $t \geq 1$ (denoted by $GB(p, n, t)$) if

$$W = \sum_{i=1}^{n} W_i$$

where each $\{W_i : i = 1, \ldots, n\}$ is *i.i.d.* geometric random variable truncated at t with parameter p. In case $n = 0$, the sum should be understood as 0. In other words, for $i = 1, \ldots, n$,

$$P(W_i = u) = \begin{cases} qp^u & \text{if } 0 \leq u < t \\ p^t & \text{if } u = t \\ 0 & \text{otherwise.} \end{cases}$$

If $n = 1$, we will refer a $GB(p, n, t)$ random variable as a *generalized Bernoulli type* and we will denote it by a $GBer(p, t)$.

The probability generating function $\chi_{(p,n,t)}$ of the $GB(p,n,t)$ is given by

$$\chi_{(p,n,t)}(s) = \Big(q + qps + \cdots + qp^{t-1}s^{t-1} + p^t s^t\Big)^n. \tag{1}$$

Thus, the generating function of a $GBer(p,t)$ is given by

$$\chi_{(p,t)}(s) = \Big(q + qps + \cdots + qp^{t-1}s^{t-1} + p^t s^t\Big).$$

It should be noted that if $t = 1$, W follows the binomial distribution with parameters n and p. In this sense, this can be thought of as a generalization of the binomial distribution. Further, if n is fixed and $t \uparrow \infty$, then W follows the usual negative binomial distribution with parameters n and p. Also note that, if we set $p = \lambda/n$, then

$$\chi_{(p,n,t)}(s) = \Big[1 - \frac{\lambda}{n}\Big(1-s\Big) + o\Big(\frac{1}{n}\Big)\Big]^n \to \exp\Big(-\lambda(1-s)\Big) \tag{2}$$

as $n \to \infty$. The limit is the probability generating function of a Poisson random variable with parameter λ. Hence, when p and n are related in such a way as above, then $GB(p,n,t)$ converges to a Poisson random variable as $n \to \infty$.

Another discrete distribution will be important for our results. Aki (1985) had defined an *extended negative binomial distribution of order t with parameters* n and $(p_1, p_2, \ldots, p_t)$ and gave the probability generating function as

$$\varphi(s; n, (p_1, p_2, \ldots, p_t)) = \left[\frac{p_1 p_2 \ldots p_t s^t}{1 - \sum_{j=1}^{t} p_1 p_2 \cdots p_{j-1} q_j s^j}\right]^n. \tag{3}$$

We will mostly consider the case when $p_1 = p_2 = \cdots = p_t = p$. Indeed, when $t = 1$, this is the usual negative binomial distribution with parameters $0 < p < 1$ and $n \geq 1$. When $n = 1$ and $p_1 = p_2 = \cdots = p_t = p$, we will call this distribution as *extended geometric distribution of order t* with parameter p.

3. Definitions and statement of results

Let $X_{-m+1}, \ldots, X_0, X_1, \ldots$ be a sequence of stationary m-order $\{\mathbf{0}, \mathbf{1}\}$ valued Markov chain. Assume that the states of $X_{-m+1}, \ldots, X_0$ are known *i.e.*, $x_0, x_{-1}, \ldots, x_{-m+1}$ are known and we take the initial state as $X_0 = x_0, X_{-1} = x_{-1}, \ldots, X_{-m+1} = x_{-m+1}$.

Define the set $S_i = \{0, 1, \ldots, 2^i - 1\}$ for any $i \geq 0$. It is clear that S_i and $\{\mathbf{0}, \mathbf{1}\}^i$ can be connected by the one-to-one and onto mapping $x = (x_0, x_1, \ldots, x_{i-1}) \longrightarrow \sum_{j=0}^{i-1} 2^j x_j$. Since $\{X_n : n \geq -m+1\}$ is the m^{th} order Markov chain, we have the transition probabilities

$$p_x = \mathbb{P}(X_{n+1} = 1 | X_n = x_0, X_{n-1} = x_1, \ldots, X_{n-m+1} = x_{m-1}) \tag{4}$$

where $x = \sum_{j=0}^{m-1} 2^j x_j \in S_m$, for any $n \geq 0$. Therefore, we have $q_x = \mathbb{P}(X_{n+1} = 0 | X_n = x_0, X_{n-1} = x_1, \ldots, X_{n-m+1} = x_{m-1}) = 1 - p_x$. We assume that $0 < p_x < 1$ for all $x \in S_m$.

Definition 2: (**l-look-back run**) (Anuradha (2023)) Fix two integers $k \geq 1$ and $l \geq 0$. We set $R_i(k,l) = 0$ for $i = 0, -1, \ldots, -l+1$ and for any $i \geq 1$, define inductively,

$$R_i(k,l) = \prod_{j=i-l}^{i-l} (1 - R_j(k,l)) \prod_{j=i}^{i+k-1} X_j \tag{5}$$

where the first product is to be taken as 1 when $l = 0$. If $R_i(k,l) = 1$, we say that a *l-look-back run of length k* has been recorded which started at time i.

It should be noted that for a l-look-back run to start at the time point i, we need to look back at the preceding l many time points, *i.e.*, $i-1$ to $i-l$, none of which can be the starting point of a l-look-back run of length k.

Next we define the stopping times where the r^{th} l-look-back run of length k is completed.

Definition 3: For $r \geq 1$, the stopping time $\tau_r(k,l)$ be the (random) time point at which the r^{th} l-look-back run of length k is completed. In other words,

$$\tau_r(k,l) = k - 1 + \inf\{n : \sum_{i=1}^{n} R_i(k,l) = r\}. \tag{6}$$

Note that r^{th} l-look-back run of length k is completed at time point $\tau_r(k,l)$. Next we define the overlapping runs of length k.

Definition 4: (**Overlapping runs of length *k***) When $k(\geq 1)$ consecutive successes occur, we call it an overlapping run of length k.

We may represent this mathematically as follows:

$$R_i^{(k)} = \prod_{j=1}^{k} X_{i+j-1}.$$

Note here that $R_i^{(k)} = 1$ if and only if an overlapping run of length k starts at time point i. Here a trial can contribute to more than one runs. Indeed, if $k+1$ successes appear consecutively, starting from time i, two overlapping runs will be counted with first one starting at i and the next one starting at $i+1$. Clearly all successes between time $i+1$ to $i+k-1$ will contribute to two overlapping runs.

Let $N_n(k)$ be the number of occurrences of overlapping runs of length k until time n. In other words,

$$N_n(k) = \sum_{i=1}^{n-k+1} R_i^{(k)}.$$

In this paper, we study the number of overlapping runs of length k_1 till the stopping time $\tau_r(k,l)$ (see Definition (3)). Fix any constant $k_1 \leq k$. For each $r \geq 1$, we define the random variable

$$N_r(k_1) := N_{\tau_r(k,l)}(k_1) = \sum_{i=1}^{\tau_r(k,l)} R_i^{(k_1)} \tag{7}$$

as the number of overlapping runs of length k_1 until the stopping time $\tau_r(k,l)$.

Let us consider the following example to facilitate understanding: Consider the following sequence of **1**'s and **0**'s of length 25

$$\mathbf{1110111010110111111101011}.$$

Let $k=3$ and $l=1$. Now using the definition we have $R_1(3,1)=R_5(3,1)=R_{14}(3,1)=R_{16}(3,1)=R_{18}(3,1)=1$, while for other values of i, $R_i(3,1)=0$. Thus, stopping times become $\tau_1(3,1)=3, \tau_2(3,1)=7, \tau_3(3,1)=16, \tau_4(3,1)=18$ and $\tau_5(3,1)=20$. For $k_1=2$, the number of the overlapping runs of length 2 till the stopping times are given by $N_1(2)=2, N_2(2)=4, N_3(2)=7$ and $N_4(2)=9$ and $N_5(2)=11$.

Let us define the probability generating function of $N_r(k_1)$ as follows

$$\zeta_r(s;k_1)=\sum_{n=0}^{\infty}\mathbb{P}(N_r(k_1)=n)s^n=\sum_{n=0}^{\infty}g_r(n;k_1)s^n. \tag{8}$$

Now we state our main result which we prove in Section 5.

Theorem 1: For any initial condition $x\in S_m$ and $k_2=k-k_1$ and $k_1\geq m$, the probability generating function of $N_r(k_1)$ is given by,

$$\zeta_r(s;k_1)=\frac{s\Big(p_{2^m-1}s\Big)^{k_2}}{1-\sum_{j=0}^{k_2-1}q_{2^m-1}\Big(p_{2^m-1}\Big)^j s^{j+1}}\Bigg[\Big(p_{2^m-1}s\Big)^{l+1} + \frac{s\Big(p_{2^m-1}s\Big)^{k_2}}{1-\sum_{j=0}^{k_2-1}q_{2^m-1}\Big(p_{2^m-1}\Big)^j s^{j+1}}\sum_{j=0}^{l}q_{2^m-1}\Big(p_{2^m-1}s\Big)^j\Bigg]^{r-1}.$$

In the above and subsequently, we have used the convention that the sum is taken to be 0 if the starting index of the sum is bigger than the ending index of the sum (which happens in the above expression when we take $k_2=0$).

Now the result of theorem 1 provides a powerful representation of $N_r(k_1)$ through the extended geometric random variables and generalized Bernoulli type distribution.

Let us define the indicator function as follows:

$$\mathbb{I}_{\{u\}}(v)=\begin{cases}1 & \text{if } u=v\\ 0 & \text{otherwise.}\end{cases} \tag{9}$$

Corollary 1: Suppose that $\{G_i^{(E)}: i=1,\ldots,r\}$ and $\{B_i^{(G)}: i=1,\ldots,r\}$ are independent families of *i.i.d.* random variables where each $G_i^{(E)}$ is having an extended geometric distribution of order k_2 with parameter p_{2^m-1} and each $B_i^{(G)}$ is a generalized Bernoulli type random variable $\mathrm{GBer}(p_{2^m-1}, l+1)$. Then

$$N_r(k_1)\stackrel{d}{=}\Big(1+G_1^{(E)}\Big)+\sum_{i=2}^{r}\Big[B_i^{(G)}+\Big(1+G_i^{(E)}\Big)\Big(1-\mathbb{I}_{\{l+1\}}\Big(B_i^{(G)}\Big)\Big)\Big].$$

Indeed, we have that the generating function of any $G_i^{(E)}$ is given by the equation (3). Also, the generating function of $B_i^{(G)} + \left(1 + G_i^{(E)}\right)\left(1 - \mathbb{I}_{\{l+1\}}\left(B_i^{(G)}\right)\right)$ is given by

$$\begin{aligned}
&\sum_{i=0}^{\infty}\sum_{j=0}^{l+1} s^{j+(1+i)\left(1-\mathbb{I}_{\{l+1\}}(j)\right)}\mathbb{P}(G_i^{(E)}=i)\mathbb{P}(B_i^{(G)}=j)\\
&=\sum_{i=0}^{\infty}\sum_{j=0}^{l} s^{j+(1+i)}\mathbb{P}(G_i^{(E)}=i)\mathbb{P}(B_i^{(G)}=j)+\sum_{i=0}^{\infty} s^{l+1}\mathbb{P}(G_i^{(E)}=i)\mathbb{P}(B_i^{(G)}=l+1)\\
&=s\sum_{i=0}^{\infty} s^i\mathbb{P}(G_i^{(E)}=i)\sum_{j=0}^{l} s^j\mathbb{P}(B_i^{(G)}=j)+s^{l+1}\left(p_{2^m-1}\right)^{l+1}\sum_{i=0}^{\infty}\mathbb{P}(G_i^{(E)}=i)\\
&=\frac{s\left(p_{2^m-1}s\right)^{k_2}}{1-\sum_{j=0}^{k_2-1} q_{2^m-1}\left(p_{2^m-1}\right)^j s^{j+1}}\sum_{j=0}^{l} q_{2^m-1}\left(p_{2^m-1}s\right)^j+\left(p_{2^m-1}s\right)^{l+1}.
\end{aligned}$$

Thus using the independence of the random variables, we now conclude that the generating functions of the random variables of both sides of the corollary 1 are same. This proves the corollary.

If $r = 1$, the distribution of $N_1(k_1)$ is actually an extended geometric distribution of order k_2 and parameter p_{2^m-1} translated by 1. If $k_2 = 0$, *i.e.*, $k = k_1$, we have that $G_i^{(E)} = 0$ and hence we have

$$\begin{aligned}
N_r(k_1) &\stackrel{d}{=} 1+\sum_{i=2}^{r}\left[B_i^{(G)}+1-\mathbb{I}_{\{l+1\}}\left(B_i^{(G)}\right)\right]\\
&=r+\sum_{i=2}^{r}\left[B_i^{(G)}-\mathbb{I}_{\{l+1\}}\left(B_i^{(G)}\right)\right]=r+\sum_{i=1}^{r-1} D_i^{(G)}
\end{aligned}$$

where $D_i^{(G)} = B_{i+1}^{(G)} - \mathbb{I}_{\{l+1\}}\left(B_{i+1}^{(G)}\right)$ for $i = 1, 2, \ldots, r-1$. Now, we observe that $D_i^{(G)}$ has a geometric distribution truncated at l. Indeed, for $j < l$, it is easy to see that $\mathbb{P}(D_i^{(G)} = j) = \mathbb{P}(B_{i+1}^{(G)} = j) = q_{2^m-1}\left(p_{2^m-1}\right)^j$ and for $j = l$, we have $\mathbb{P}(D_i^{(G)} = l) = \mathbb{P}(B_{i+1}^{(G)} = l) + \mathbb{P}(B_{i+1}^{(G)} = l+1) = q_{2^m-1}\left(p_{2^m-1}\right)^l + \left(p_{2^m-1}\right)^{l+1} = \left(p_{2^m-1}\right)^l$. Thus, $N_r(k_1) - r$ has generalized Binomial type distribution with parameters $p_{2^m-1}, r-1$ and l.

Under the assumption that the system has a strong tendency towards failure when m consecutive successes are observed, *i.e.*, p_{2^m-1} as a function of r converges to 0 in such a way that

$$rp_{2^m-1} \to \lambda > 0 \text{ as } r \to \infty, \tag{10}$$

using the equation (2) and the subsequent discussion, we can easily obtain the following corollary.

Corollary 2: For any initial condition $x \in S_m$, if the condition (10) holds and if $k_2 = 0$, we have

$$N_r(k_1) - r \quad \Rightarrow \text{Poi}(\lambda)$$

where $\text{Poi}(\lambda)$ is the Poisson distribution with parameter λ.

If we set $k_1 = 1$, $N_r(k_1)$ represents the number of successes till the r^{th} occurrence of the l-look-back run of length k. Thus, we have the following corollary:

Corollary 3: For the *i.i.d.* case or the Markov dependent case, the probability generating function of the number of successes till the r^{th} occurrence of the l-look-back run of length k, *i.e.*, $N_r(1)$ is given by,

$$\zeta_r(s;1) = \frac{s\Big(p_{2^m-1}s\Big)^{k-1}}{1-\sum_{j=0}^{k-2} q_{2^m-1}\Big(p_{2^m-1}\Big)^j s^{j+1}}\Bigg[\Big(p_{2^m-1}s\Big)^{l+1} + \frac{s\Big(p_{2^m-1}s\Big)^{k-1}}{1-\sum_{j=0}^{k-2} q_{2^m-1}\Big(p_{2^m-1}\Big)^j s^{j+1}}\sum_{j=0}^{l} q_{2^m-1}\Big(p_{2^m-1}s\Big)^j\Bigg]^{r-1}.$$

For $r = 1$, the expression reduces to

$$\zeta_1(s;1) = \frac{\Big(p_{2^m-1}\Big)^{k-1}(1-p_{2^m-1}s)s^k}{1-s+q_{2^m-1}\Big(p_{2^m-1}\Big)^{k-1}s^k}$$

which is the probability generating function of the number of successes until the first occurrence of k consecutive successes. For the *i.i.d.* case, we have $p_{2^m-1} = p$ and for the Markov dependent case, we have $p_{2^m-1} = p_{11}$. Putting these values, we observe that we may obtain the results (Proposition 3.4 and Theorem 3.2) of Aki and Hirano (1994). Therefore our result provides a generalized version of *pgf* for all values of r.

4. A new Markov chain

Now we outline the underlying set up which will be used in the subsequent sections to establish the results. Let us define two functions $f_0, f_1 : S_{k_1} \to S_{k_1}$ by

$$f_1(x) = 2x + 1 \pmod{2^{k_1}} \text{ and } f_0(x) = 2x \pmod{2^{k_1}}.$$

Further define a projection $\theta_m : S_{k_1} \to S_m$ by $\theta_m(x) = x \pmod{2^m}$. Now, set $X_{-m} = X_{-m-1} = \cdots = X_{-k_1+1} = 0$. Define a sequence of random variables $\{Y_n : n \geq 0\}$ as follows:

$$Y_n = \sum_{j=0}^{k_1-1} 2^j X_{n-j}.$$

Since $X_i \in \{\mathbf{0}, \mathbf{1}\}$ for all i, Y_n assumes values in the set S_{k_1}. The random variables X_n's are stationary and forms an m^{th} order Markov chain, hence we have that $\{Y_n : n \geq 0\}$ is a homogeneous Markov chain with transition matrix given by

$$\mathbb{P}(Y_{n+1} = y | Y_n = x) = \begin{cases} p_{\theta_m(x)} & \text{if } y = f_1(x) \\ 1 - p_{\theta_m(x)} & \text{if } y = f_0(x) \\ 0 & \text{otherwise.} \end{cases}$$

Note that Y_n is even if and only if $X_n = 0$. This motivates us to define the function $\kappa : S_{k_1} \to \{0, 1\}$ by

$$\kappa(x) = \begin{cases} 1 & \text{if } x \text{ is odd} \\ 0 & \text{if } x \text{ is even.} \end{cases}$$

Therefore, $\kappa(Y_n) = 1$ if and only if $X_n = 1$. Hence, the definition of l-look-back run can be described in terms of Y_n's as

$$R_i(k, l) = \prod_{j=i-l}^{i-1} (1 - R_j(k, l)) \prod_{j=i}^{i+k-1} \kappa(Y_j).$$

Let us fix any initial condition $x \in S_m$. We denote the probability measure governing the distribution of $\{Y_n : n \geq 1\}$ with $Y_0 = x \in S_k$ by $\mathbb{P}_x$. Since we have set $X_{-m} = X_{-m-1} = \cdots = X_{-k+1} = 0$, we have $Y_0 = x$.

In order to obtain the recurrence relation for the probabilities, we will condition the process after the first occurrence of the run of length k_1. Therefore, we consider the stopping time T when the first occurrence of a run of length k_1 ends, *i.e.*, when we observe k_1 successes consecutively for the first time. More precisely, define

$$T := \inf\{i \geq k_1 : \prod_{j=i-k_1+1}^{i} X_j = 1\}. \tag{11}$$

We would like to translate the above definition to Y_i's. It must be the case that when T occurs, last k_1 trials have resulted in success, which may be described by $\kappa(Y_j) = 1$ for $j = i - k_1 + 1$ to i. Therefore, Y_T must equal $2^{k_1} - 1$. Since this is the first occurrence, this has not happened earlier. So, T can be better described as

$$T = \inf\{i \geq k_1 : Y_i = 2^{k_1} - 1\},$$

i.e., the first visit of the chain to the state $2^{k_1} - 1$ after time $k_1 - 1$. Now, we note that $\{Y_n : n \geq 0\}$ is a Markov chain with finite state space. Further, since $0 < p_u < 1$ for $u \in S_m$, this is an irreducible chain; hence, it is positive recurrent. So we must have $\mathbb{P}_x(T < \infty) = 1$. We observe that when the first occurrence of k consecutive successes happen, we must have the occurrence of k_1 successes previously since $k_1 \leq k$. Therefore, we have $\mathbb{P}_x(T \leq \tau_1(k, l)) = 1$.

5. Overlapping runs till the stopping time

In this section, we study the distribution of overlapping runs of length k_1. We will employ the method of generating functions to derive these results.We obtain a recurrence relation between the probabilities in order to derive the generating functions.

Let us define the probability, for $x \in S_m, n \in \mathbb{Z}$,

$$g_r^{(x)}(n; k_1) = \mathbb{P}_x\Big(N_r(k_1) = n\Big). \tag{12}$$

We note that since $N_r(k_1) \geq 1$, $\mathbb{P}_x\Big(N_r(k_1) = n\Big) = 0$ for $n \leq 0$. Also, if $r = 1$ and $k_2 = k - k_1 = 0$, i.e, $k = k_1$, we have that $N_1(k_1) = 1$.

We will show that these probabilities $g_r^{(x)}(n;k_1)$ is actually independent of the initial condition x. First we consider the case when $r=1$. As we have already observed, if $k_2=0$,

$$g_1^{(x)}(1;k_1)=\mathbb{I}_{\{1\}}(n)$$

where $\mathbb{I}_{\{u\}}(v)$ is the indicator function defined in (9). Clearly we have $g_1^{(x)}(n;k_1)$ is independent of x.

Now, we concentrate on the case when $r=1$ and $k_2=k-k_1>0$, *i.e.*, $k>k_1$. We note that $N_1(k_1)\geq(k_2+1)$ and hence $\mathbb{P}_x(N_1(k_1)=n)=g_1^{(x)}(n;k_1)=0$ for $n\leq k_2$.

Theorem 2: For $n>k_2$ and $k_2=k-k_1>0$, we have

$$g_1^{(x)}(n;k_1)=\sum_{t=0}^{k_2-1}q_{2^m-1}\Big(p_{2^m-1}\Big)^t g_1^{(2^m-2)}(n-t-1;k_1)+\Big(p_{2^m-1}\Big)^{k_2}\mathbb{I}_{\{k_2+1\}}(n) \tag{13}$$

where $\mathbb{I}_{\{u_1\}}(u_2)$ is the indicator function defined in (9).

Proof: When $k_2=k-k_1>0$ and $r=1$, using the fact that $Y_T=2^{k_1}-1$ with probability 1, we have

$$\begin{aligned}
&g_1^{(x)}(n;k_1)=\mathbb{P}_x(N_1(k_1)=n)=\mathbb{P}_x(N_1(k_1)=n,Y_T=2^{k_1}-1)\\
&=\sum_{t=0}^{k_2-1}\mathbb{P}_x(N_1(k_1)=n,Y_T=2^{k_1}-1,Y_{T+1}=2^{k_1}-1,\ldots,\\
&\qquad Y_{T+t}=2^{k_1}-1,Y_{T+t+1}=2^{k_1}-2)\\
&+\mathbb{P}_x(N_1(k_1)=n,Y_T=2^{k_1}-1Y_{T+1}=2^{k_1}-1,Y_{T+2}=2^{k_1}-1,\ldots,\\
&\qquad Y_{T+k_2-1}=2^{k_1}-1,Y_{T+k_2}=2^{k_1}-1).
\end{aligned}\tag{14}$$

We look at the terms in the summation first. For any $0\leq t\leq k_2-1$, we have,

$$\begin{aligned}
&\mathbb{P}_x(N_1(k_1)=n,Y_T=2^{k_1}-1,Y_{T+1}=2^{k_1}-1,\ldots,Y_{T+t}=2^{k_1}-1,Y_{T+t+1}=2^{k_1}-2)\\
&=\mathbb{P}_x(N_1(k_1)=n\mid Y_T=2^{k_1}-1,Y_{T+1}=2^{k_1}-1,\ldots,Y_{T+t}=2^{k_1}-1,Y_{T+t+1}=2^{k_1}-2)\\
&\times\mathbb{P}_x(Y_T=2^{k_1}-1,Y_{T+1}=2^{k_1}-1,\ldots,Y_{T+t}=2^{k_1}-1,Y_{T+t+1}=2^{k_1}-2).
\end{aligned}\tag{15}$$

The second term in (15) can be written as

$$\begin{aligned}
&\mathbb{P}_x(Y_T=2^{k_1}-1,Y_{T+1}=2^{k_1}-1,\ldots,Y_{T+t}=2^{k_1}-1,Y_{T+t+1}=2^{k_1}-2)\\
&=\mathbb{P}_x(Y_{T+t+1}=2^{k_1}-2\mid Y_T=2^{k_1}-1,Y_{T+1}=2^{k_1}-1,\ldots,Y_{T+t}=2^{k_1}-1)\\
&\times\prod_{j=1}^{t}\mathbb{P}_x(Y_{T+j}=2^{k_1}-1\mid Y_T=2^{k_1}-1,Y_{T+1}=2^{k_1}-1,\ldots,Y_{T+j-1}=2^{k_1}-1).
\end{aligned}$$

Now, $T+j-1$ is also a stopping time for any $1\leq j\leq t$. We denote by $\mathcal{F}_{T+j-1}$, the σ-algebra generated by the process Y_n up to the stopping time $T+j-1$, and by $\mathcal{F}_{(T+j-1)+}$, the σ-algebra generated by the process after the stopping time $T+j-1$. Clearly,

$\{Y_{T+1} = 2^{k_1} - 1, \ldots, Y_{T+j-1} = 2^{k_1} - 1\} \in \mathcal{F}_{T+j-1}$ and $\{Y_{T+j} = 2^{k_1} - 1\} \in \mathcal{F}_{(T+j-1)+}$. Thus, by strong Markov property, we can write

$$\begin{aligned}&\mathbb{P}_x(Y_{T+j} = 2^{k_1} - 1 \mid Y_T = 2^{k_1} - 1, Y_{T+1} = 2^{k_1} - 1, \ldots, Y_{T+j-1} = 2^{k_1} - 1)\\ &= \mathbb{P}_{Y_{T+j-1}}(Y_{T+j} = 2^{k_1} - 1) = \mathbb{P}_{2^{k_1}-1}(Y_1 = 2^{k_1} - 1) = p_{2^m-1}. \end{aligned} \tag{16}$$

A similar argument shows that

$$\mathbb{P}_x(Y_{T+t+1} = 2^{k_1} - 2 \mid Y_T = 2^{k_1} - 1, Y_{T+1} = 2^{k_1} - 1, \ldots, Y_{T+t} = 2^{k_1} - 1) = q_{2^m-1}. \tag{17}$$

For the first term in (15), we note that $T + t + 1$ is also a stopping time and $\{Y_{T+1} = 2^{k_1} - 1, \ldots, Y_{T+t} = 2^{k_1} - 1, Y_{T+t+1} = 2^{k_1} - 2\} \in \mathcal{F}_{T+t+1}$. Since $Y_{\tau_1(k_1)} = 2^{k_1} - 1$, we must have $X_{T-k_1} = 0$ and $X_{T-j} = 1$ for $j = 0, 1, \ldots, k_1 - 1$. Further, since $Y_{\tau_1(k_1)+j} = 2^{k_1} - 1$ for $j = 1, \ldots, t$ and $Y_{T+t+1} = 2^{k_1} - 2$, we also have $X_{T+j} = 1$ for $j = 0, 1, \ldots, t$ and $X_{T+t+1} = 0$. Therefore, we have a sequence of $\mathbf{1}'$s of length $k_1 + t$ with $t > 0$ which contributes to $t + 1$ overlapping runs of length k_1 and since there are no runs of length k_1 before T, by the very definition of T, we have that the number of overlapping runs of length k_1 up to time $T+t+1$ is $1 + t$. Since $t \leq k_2 - 1$, we have that $T + t + 1 < \tau_1(k, l)$. Let us define $Y_i' = Y_{i+T+t+1}$ for $i \geq 0$. Now, using the strong Markov property, we have that $\{Y_i' : i \geq 0\}$ is a homogeneous Markov chain with same transition matrix as that of $\{Y_i : i \geq 0\}$ with $Y_0' = 2^{k_1} - 2$. Now, define $\tau_1'(k, l)$ as the stopping time for the process $\{Y_i' : i \geq 0\}$. From the above discussion, we have that $\tau_1(k, l) = T + t + 1 + \tau_1'(k, l)$. Further, if we define, $N_1'(k_1)$ as the number of overlapping runs of length k_1 up to time $\tau_1'(k, l)$ for the process $\{Y_i' : i \geq 0\}$, we must have that $N_1'(k_1) = n - t - 1$. Therefore, we have,

$$\begin{aligned}&\mathbb{P}_x(N_1(k_1) = n \mid Y_T = 2^{k_1} - 1, Y_{T+1} = 2^{k_1} - 1, \ldots, Y_{T+t} = 2^{k_1} - 1, Y_{T+t+1} = 2^{k_1} - 2)\\ &= \mathbb{P}_{(2^m-2)}(N_1'(k_1) = n - t - 1) = g_1^{(2^m-2)}(n - t - 1; k_1). \end{aligned} \tag{18}$$

The last term in (14) can be similarly written as

$$\begin{aligned}&\mathbb{P}_x(N_1(k_1) = n, Y_T = 2^{k_1} - 1, Y_{T+1} = 2^{k_1} - 1, \ldots, Y_{T+k_2} = 2^{k_1} - 1)\\ &= \prod_{j=1}^{k_2} \mathbb{P}_x(Y_{T+j} = 2^{k_1} - 1 \mid Y_T = 2^{k_1} - 1, Y_{T+1} = 2^{k_1} - 1, \ldots, Y_{T+j-1} = 2^{k_1} - 1)\\ &\quad \times \mathbb{P}_x(N_1(k_1) = n \mid Y_T = 2^{k_1} - 1, Y_{T+1} = 2^{k_1} - 1, \ldots,\\ &\qquad Y_{T+k_2-1} = 2^{k_1} - 1, Y_{T+k_2} = 2^{k_1} - 1)\\ &= \left(p_{2^m-1}\right)^{k_2} \mathbb{P}_x(N_1(k_1) = n \mid Y_T = 2^{k_1} - 1, Y_{T+1} = 2^{k_1} - 1, \ldots,\\ &\qquad Y_{T+k_2-1} = 2^{k_1} - 1, Y_{T+k_2} = 2^{k_1} - 1). \end{aligned}$$

Note that given $\{Y_T = 2^{k_1} - 1, Y_{T+1} = 2^{k_1} - 1, \ldots, Y_{T+k_2-1} = 2^{k_1} - 1, Y_{T+k_2} = 2^{k_1} - 1\}$, we have $\tau_1(k, l) = T + k_2$. Further, in such a case we have exactly $k_2 + 1$ many overlapping runs of length k_1 until time $T + k_2$. Therefore, $N_1(k_1) = n$ if and only if $n = k_2 + 1$. In other words, $\mathbb{P}_x(N_1(k_1) = n \mid Y_T = 2^{k_1} - 1, Y_{T+1} = 2^{k_1} - 1, \ldots, Y_{T+k_2-1} = 2^{k_1} - 1, Y_{T+k_2} = 2^{k_1} - 1) = 1_{\{k_2+1\}}(n)$ where $\mathbb{I}$ is the indicator function as defined in (9).

Thus combining the above equation with equations (14) - (18), we can express

$$g_1^{(x)}(n;k_1) = \sum_{t=0}^{k_2-1} q_{2^m-1}\Big(p_{2^m-1}\Big)^t g_1^{(2^m-2)}(n-t-1;k_1) + \Big(p_{2^m-1}\Big)^{k_2}\mathbb{I}_{\{k_2+1\}}(n).$$

This completes the proof. □

We note that the right hand side of (13) does not involve the initial condition $x \in S_m$. Therefore $g_1^{(x)}(n;k_1)$ must be independent of x. So, we will drop x and denote the above probability by $g_1(n;k_1)$. Thus, we have the following corollary from theorem 2.

Corollary 4: For $n \geq k_2+1$ and $k_2 > 0$, we have

$$g_1(n;k_1) = \sum_{t=0}^{k_2-1} q_{2^m-1}\Big(p_{2^m-1}\Big)^t g_1(n-t-1;k_1) + \Big(p_{2^m-1}\Big)^{k_2}\mathbb{I}_{\{k_2+1\}}(n). \tag{19}$$

Let us recall that

$$\zeta_r(s;k_1) = \sum_{n=0}^{\infty}\mathbb{P}(N_r(k_1)=n)s^n = \sum_{n=0}^{\infty} g_r(n;k_1)s^n.$$

When $k_2 = 0$, we have

$$\zeta_1(s;k_1) = s.$$

For $k_2 > 0$, we may use the equation (19) to derive its generating function. We have

$$\begin{aligned}
\zeta_1(s;k_1) &= \sum_{n=k_2+1}^{\infty} g_r(n;k_1)s^n \\
&= \sum_{n=k_2+1}^{\infty}\Big[\sum_{t=0}^{k_2-1} q_{2^m-1}\Big(p_{2^m-1}\Big)^t g_1(n-t-1;k_1) + \Big(p_{2^m-1}\Big)^{k_2}\mathbb{I}_{\{k_2+1\}}(n)\Big]s^n \\
&= \Big(p_{2^m-1}\Big)^{k_2}s^{k_2+1} + \sum_{t=0}^{k_2-1} q_{2^m-1}\Big(p_{2^m-1}\Big)^t s^{t+1}\sum_{n=k_2+1}^{\infty} g_1(n-t-1;k_1)s^{n-t-1} \\
&= \Big(p_{2^m-1}\Big)^{k_2}s^{k_2+1} + \zeta_1(s;k_1)\sum_{t=0}^{k_2-1} q_{2^m-1}\Big(p_{2^m-1}\Big)^t s^{t+1}.
\end{aligned}$$

This linear equation can now be solved to yield the following corollary.

Corollary 5: For $r = 1$, we have

$$\zeta_1(s;k_1) = \frac{s\Big(p_{2^m-1}s\Big)^{k_2}}{1-\sum_{j=0}^{k_2-1} q_{2^m-1}\Big(p_{2^m-1}\Big)^j s^{j+1}}. \tag{20}$$

Now we consider the case when $r > 1$. In this case also, we note that $N_r(k_1) \geq (k_2+1)+(r-1)(l+1)$. Hence $g_r^{(x)}(n;k_1) = \mathbb{P}_x(N_r(k_1)=n) = 0$ for $n \leq (r-1)(l+1)+k_2$. Now, we derive the recurrence relation.

Theorem 3: For $n \geq (k_2+1)+(r-1)(l+1)$ and $x \in S_m$, we have

$$\begin{aligned} g_r^{(x)}(n;k_1) &= \Big(p_{2^m-1}\Big)^{k_2+(r-1)(l+1)} \mathbb{I}_{\{n\}}(k_2+(r-1)(l+1)+1) \\ &+ \sum_{j=0}^{k_2-1} q_{2^m-1}\Big(p_{2^m-1}\Big)^{j} g_r^{(2^m-2)}(n-j-1;k_1) \\ &+ \sum_{j_1=0}^{r-2}\sum_{j_2=0}^{l} q_{2^m-1}\Big(p_{2^m-1}\Big)^{k_2+j_1(l+1)+j_2} g_{r-1-j_1}^{(2^m-2)}(n-1-k_2-j_1(l+1)-j_2;k_1). \end{aligned} \tag{21}$$

where $\mathbb{I}_{v_1}(v_2)$ is the indicator function, as defined in the previous theorem.

Proof: We proceed in the same way as in the previous theorem. Conditioning on the first occurrence of k_1 many successes, *i.e.*, T, we have, for any $n \geq (k_2+1)+(r-1)(l+1)$,

$$\begin{aligned} g_r^{(x)}(n;k_1) = \sum_{t=0}^{k_2+(r-1)(l+1)-1} & \mathbb{P}_x(N_r(k_1)=n, Y_T = 2^{k_1}-1, Y_{T+1}=2^{k_1}-1,\ldots, \\ & Y_{T+t}=2^{k_1}-1, Y_{T+t+1}=2^{k_1}-2) \\ + \mathbb{P}_x(N_r(k_1)=n, Y_T = 2^{k_1}-1, Y_{T+1} &= 2^{k_1}-1, Y_{T+2}=2^{k_1}-1,\ldots, \\ & Y_{T+(r-1)(l+1)-1}=2^{k_1}-1, Y_{T+(r-1)(l+1)}=2^{k_1}-1). \end{aligned} \tag{22}$$

The last term in (22) is similar to the last term in equation (14) in the previous theorem. Thus this term can be simplified in the similar way. Indeed using the same arguments, as done after equation (18), we get

$$\begin{aligned} &\mathbb{P}_x(N_r(k_1)=n, Y_T = 2^{k_1}-1, Y_{T+1}=2^{k_1}-1, Y_{T+2}=2^{k_1}-1,\ldots, \\ &\qquad Y_{T+(r-1)(l+1)-1}=2^{k_1}-1, Y_{T+(r-1)(l+1)}=2^{k_1}-1) \\ &= \Big(p_{2^m-1}\Big)^{k_2+(r-1)(l+1)} \mathbb{I}_{\{n\}}(k_2+(r-1)(l+1)+1). \end{aligned} \tag{23}$$

The terms in the summation in (22) can also be handled in the similar way as done in the previous theorem. Fix any j with $0 \leq t \leq k_2+(r-1)(l+1)-1$ and we obtain that

$$\begin{aligned} &\mathbb{P}_x(N_r(k_1)=n, Y_T = 2^{k_1}-1, Y_{T+1}=2^{k_1}-1,\ldots, Y_{T+t}=2^{k_1}-1, Y_{T+t+1}=2^{k_1}-2) \\ &= \mathbb{P}_x(N_r(k_1)=n \mid Y_T = 2^{k_1}-1, Y_{T+1}=2^{k_1}-1,\ldots, Y_{T+t}=2^{k_1}-1, Y_{T+t+1}=2^{k_1}-2) \\ &\times \mathbb{P}(Y_T = 2^{k_1}-1, Y_{T+1}=2^{k_1}-1,\ldots, Y_{T+t}=2^{k_1}-1, Y_{T+t+1}=2^{k_1}-2). \end{aligned} \tag{24}$$

The last term in the product above is again simplified using the product of conditional

terms and the strong Markov property. Since $Y_T = 2^{k_1} - 1$ with probability 1, we have

$$\begin{aligned}
&\mathbb{P}(Y_T = 2^{k_1}-1, Y_{T+1} = 2^{k_1}-1, \ldots, Y_{T+t} = 2^{k_1}-1, Y_{T+t+1} = 2^{k_1}-2)\\
&= \mathbb{P}_x(Y_{T+t+1} = 2^{k_1}-2 \mid Y_T = 2^{k_1}-1, Y_{T+1} = 2^{k_1}-1, \ldots, Y_{T+t} = 2^{k_1}-1)\\
&\times \prod_{j=1}^{t} \mathbb{P}_x(Y_{T+j} = 2^{k_1}-1 \mid Y_T = 2^{k_1}-1, Y_{T+1} = 2^{k_1}-1, \ldots, Y_{T+j-1} = 2^{k_1}-1)\\
&= \mathbb{P}_x(Y_{T+t+1} = 2^{k_1}-2 \mid Y_{T+t} = 2^{k_1}-1) \times \prod_{j=1}^{t} \mathbb{P}_x(Y_{T+j} = 2^{k_1}-1 \mid Y_{T+j-1} = 2^{k_1}-1)\\
&= q_{2^m-1}\Big(p_{2^m-1}\Big)^t. \qquad (25)
\end{aligned}$$

For the first term, we note that the event $\{Y_T = 2^{k_1}-1, Y_{T+1} = 2^{k_1}-1, \ldots, Y_{T+t} = 2^{k_1}-1, Y_{T+t+1} = 2^{k_1}-2\}$ implies that at time $T+t+1$, we have just observed k_1+t many successes followed by a failure. This string of k_1+t many successes, will contribute $t+1$ many overlapping runs of successes. Since T is the first time when we observe first k_1 many consecutive successes, we have $t+1$ overlapping success runs completed at time $T+t+1$. Thus, we are left with $n-t-1$ many runs for the remaining part, *i.e.*, after time $T+t+1$.

At time $T+t+1$, we have the information that k_1+t many successes followed by a failure has just been observed. Using the strong Markov property, we can think that the process restarts with this information. In other words, considering the converted Y process, we are restarting the process with the initial condition $Y_{T+t+1} = 2^{k_1}-2$.

Now, we examine the two cases namely $k_2 = k - k_1 = 0$, *i.e.*, $k = k_1$ and $k_2 > 0$, *i.e.*, $k > k_1$ separately. When $k_2 = 0$ and k_1+t many successes followed by a failure has just been observed, then we have already completed $1 + \lfloor t/(l+1) \rfloor$ many l-look-back runs of length k where $\lfloor a \rfloor$ is the largest integer smaller or equal to a. Thus, we are left with $r - \lfloor t/(l+1) \rfloor - 1$ many l-look-back runs of length k, which is to be completed by the process after the time $T+t+1$. Hence, we obtain,

$$\begin{aligned}
&\mathbb{P}_x(N_r(k_1) = n \mid Y_T = 2^{k_1}-1, Y_{T+1} = 2^{k_1}-1, \ldots, Y_{T+t} = 2^{k_1}-1, Y_{T+t+1} = 2^{k_1}-2)\\
&= \mathbb{P}_{2^m-2}(N_{r-\lfloor t/(l+1) \rfloor - 1}(k_1) = (n-t-1)) = g^{(2^m-2)}_{r-\lfloor t/(l+1) \rfloor - 1}(n-t-1; k_1). \qquad (26)
\end{aligned}$$

For $k_2 > 0$, the argument is essentially the same, except for one part. When $t \le k_2 - 1$, we would have $k_1 + t \le k_1 + k - k_1 - 1 = k-1$ many successes followed by a failure. This will not contribute to any run of l-look-back run of length k. But for $t \ge k_2$, we will have $1 + \lfloor (t-k_2)/(l+1) \rfloor$ many l-look-back runs of length k which have been completed. Thus, we have

$$\begin{aligned}
&\mathbb{P}_x(N_r(k_1) = n \mid Y_T = 2^{k_1}-1, Y_{T+1} = 2^{k_1}-1, \ldots, Y_{T+t} = 2^{k_1}-1, Y_{T+t+1} = 2^{k_1}-2)\\
&= \begin{cases} g^{(2^m-2)}_r(n-t-1; k_1) & \text{if } t \le k_2 - 1\\ g^{(2^m-2)}_{r-\lfloor (t-k_2)/(l+1) \rfloor - 1}(n-t-1; k_1) & \text{if } t \ge k_2. \end{cases} \qquad (27)
\end{aligned}$$

Therefore, combining all the terms above from equations (24), (25), (26) and (27), we have

$$\begin{aligned}
&\sum_{t=0}^{k_2+(r-1)(l+1)-1} \mathbb{P}_x(N_r(k_1)=n, Y_T=2^{k_1}-1, Y_{T+1}=2^{k_1}-1, Y_{T+2}=2^{k_1}-1,\ldots,\\
&\qquad\qquad Y_{T+t}=2^{k_1}-1, Y_{T+t+1}=2^{k_1}-2)\\
&=\sum_{j=0}^{k_2-1} q_{2^m-1}\Big(p_{2^m-1}\Big)^j g_r^{(2^m-2)}(n-j-1;k_1)\\
&\quad+\sum_{j_1=0}^{r-2}\sum_{j_2=0}^{l} q_{2^m-1}\Big(p_{2^m-1}\Big)^{k_2+j_1(l+1)+j_2} g_{r-1-j_1}^{(2^m-2)}(n-1-k_2-j_1(l+1)-j_2;k_1). \qquad (28)
\end{aligned}$$

Now combining the equations (22), (23) and (28), the proof of the theorem is completed. □

If $r=1$, we have that $g_r^{(x)}(\cdot;k_1)$ is independent of $x\in S_m$ (see Corollary 4). By induction, assume that $g_r^{(x)}(\cdot)$ is independent of $x\in S_m$. Clearly, from the above relation, we have that $g_{r+1}^{(x)}(\cdot;k_1)$ can be expressed as weighted sums of $g_i^{(x)}(\cdot;k_1)$ for $i=1,2,\ldots,r$. Since the right hand side of the above relation does not involve any $x\in S_m$, $g_{r+1}^{(x)}(\cdot;k_1)$ must be independent of x. Therefore, from now on we will drop the superscript x from $g_r^{(x)}(\cdot;k_1)$. Hence we have the following corollary.

Corollary 6: For any $x\in S_m$, the probability $g_r^{(x)}(n;k_1)=\mathbb{P}_x(N_r(k_1)=n)$ is independent of x and satisfies the recurrence relation

$$\begin{aligned}
&g_r(n;k_1)\\
&=\sum_{j=0}^{k_2-1} q_{2^m-1}\Big(p_{2^m-1}\Big)^j g_r(n-j-1;k_1)+\Big(p_{2^m-1}\Big)^{k_2+(r-1)(l+1)} 1_n(k_2+(r-1)(l+1)+1)\\
&\quad+\sum_{j_1=0}^{r-2}\sum_{j_2=0}^{l} q_{2^m-1}\Big(p_{2^m-1}\Big)^{k_2+j_1(l+1)+j_2} g_{r-1-j_1}(n-1-k_2-j_1(l+1)-j_2;k_1). \qquad (29)
\end{aligned}$$

We now derive the generating function $\zeta_r(s;k_1)$ of $N_r(k_1)$ using the recurrence relation. For $r=1$, we have already obtained the expression of $\zeta_1(s;k_1)$ (see Corollary 5). For $r\geq 2$, we can't directly obtain the expression of $\zeta_r(s;k_1)$. Instead, we will obtain a recurrence relation in terms of the generating functions. Indeed, for $r\geq 2$, we have

$$\begin{aligned}
\zeta_r(s;k_1) &= s\Big(p_{2^m-1}s\Big)^{k_2+(r-1)(l+1)}+\sum_{n=0}^{\infty}\sum_{j=0}^{k_2-1} q_{2^m-1}\Big(p_{2^m-1}\Big)^j g_r(n-1-j;k_1)s^n\\
&\quad+\sum_{n=0}^{\infty}\sum_{j_1=0}^{r-2}\sum_{j_2=0}^{l} q_{2^m-1}\Big(p_{2^m-1}\Big)^{k_2+j_1(l+1)+j_2}\\
&\qquad\times g_{r-1-j_1}(n-1-k_2-j_1(l+1)-j_2;k_1)s^n\\
&= s\Big(p_{2^m-1}s\Big)^{k_2+(r-1)(l+1)}+\sum_{j=0}^{k_2-1} q_{2^m-1}\Big(p_{2^m-1}\Big)^j s^{j+1}\zeta_r(s;k_1)\\
&\quad+\sum_{j_1=0}^{r-2}\sum_{j_2=0}^{l} q_{2^m-1}s\Big(p_{2^m-1}s\Big)^{k_2+j_1(l+1)+j_2}\zeta_{r-1-j_1}(s;k_1). \qquad (30)
\end{aligned}$$

Simplifying equation (30), we obtain a recurrence relation involving $\zeta_r(s;k_1)$. This is given in the following lemma.

Lemma 1: For $r \geq 2$, the sequence of the probability generating functions satisfies the following recurrence relation

$$\begin{aligned}
&\Big(1-\sum_{j=0}^{k_2-1} q_{2^m-1}\Big(p_{2^m-1}\Big)^j s^{j+1}\Big)\zeta_r(s;k_1)\\
&=\Big(p_{2^m-1}s\Big)^{k_2}\Big(\sum_{j=0}^{l} q_{2^m-1}s\Big(p_{2^m-1}s\Big)^j\Big)\sum_{j_1=0}^{r-2}\Big(p_{2^m-1}s\Big)^{j_1(l+1)}\zeta_{r-1-j_1}(s;k_1)\\
&\quad+s\Big(p_{2^m-1}s\Big)^{k_2+(r-1)(l+1)}. \qquad (31)
\end{aligned}$$

Now, we are ready to prove the main theorem, namely Theorem 1.

Proof: The generating function of the sequence $\{\zeta_r(s;k_1) : r \geq 1\}$ is denoted by $\Xi(z;k_1)$, *i.e.*,

$$\Xi(z;k_1)=\sum_{r=1}^{\infty}\zeta_r(s;k_1)z^r.$$

Now, using (31) we obtain the generating function $\Xi(z;k_1)$ as follows:

$$\begin{aligned}
&\Big(1-\sum_{j=0}^{k_2-1} q_{2^m-1}\Big(p_{2^m-1}\Big)^j s^{j+1}\Big)\Xi(z;k_1)\\
&=\sum_{r=1}^{\infty}\Big(1-\sum_{j=0}^{k_2-1} q_{2^m-1}\Big(p_{2^m-1}\Big)^j s^{j+1}\Big)\zeta_r(s;k_1)z^r\\
&=\Big(1-\sum_{j=0}^{k_2-1} q_{2^m-1}\Big(p_{2^m-1}\Big)^j s^{j+1}\Big)\zeta_1(s;k_1)z+\sum_{r=2}^{\infty}s\Big(p_{2^m-1}s\Big)^{k_2+(r-1)(l+1)}z^r\\
&\quad+\sum_{r=2}^{\infty}\Big(p_{2^m-1}s\Big)^{k_2}\Big(\sum_{j=0}^{l} q_{2^m-1}s\Big(p_{2^m-1}s\Big)^j\Big)\sum_{j_1=0}^{r-2}\Big(p_{2^m-1}s\Big)^{j_1(l+1)}\zeta_{r-1-j_1}(s;k_1)z^r\\
&=sz\Big(p_{2^m-1}s\Big)^{k_2}+sz\Big(p_{2^m-1}s\Big)^{k_2}\sum_{r=1}^{\infty}\Big(p_{2^m-1}s\Big)^{r(l+1)}z^r\\
&\quad+\Big(p_{2^m-1}s\Big)^{k_2}\Big(\sum_{j=0}^{l} q_{2^m-1}s\Big(p_{2^m-1}s\Big)^j\Big)\sum_{j_1=0}^{\infty}\Big(p_{2^m-1}s\Big)^{(l+1)j_1}\sum_{r=j_1}^{\infty}\zeta_{r-j_1+1}(s;k_1)z^{r+2}\\
&=\frac{sz\Big(p_{2^m-1}s\Big)^{k_2}}{1-\Big(p_{2^m-1}s\Big)^{(l+1)}z}+\frac{\Big(p_{2^m-1}s\Big)^{k_2}\Big(\sum_{j=0}^{l} q_{2^m-1}s\Big(p_{2^m-1}s\Big)^j\Big)z\Xi(z;k_1)}{1-\Big(p_{2^m-1}s\Big)^{(l+1)}z}. \qquad (32)
\end{aligned}$$

Now, from the above equation (32), we can easily solve $\Xi(z;k_1)$ to obtain

$$\begin{aligned}\Xi(z;k_1) &= \left[sz\left(p_{2^m-1}s\right)^{k_2}\right]\left[\left(1-\sum_{j=0}^{k_2-1} q_{2^m-1}\left(p_{2^m-1}\right)^j s^{j+1}\right)\right.\\ &\quad\left.\times\left(1-\left(p_{2^m-1}s\right)^{(l+1)}z\right)-z\left(p_{2^m-1}s\right)^{k_2}\left(\sum_{j=0}^{l} q_{2^m-1}s\left(p_{2^m-1}s\right)^j\right)\right]^{-1}\\ &= \frac{zs\left(p_{2^m-1}s\right)^{k_2}}{1-\sum_{j=0}^{k_2-1} q_{2^m-1}\left(p_{2^m-1}\right)^j s^{j+1}}\left[1-\left(p_{2^m-1}s\right)^{(l+1)}z\right.\\ &\quad\left.-\frac{z\left(p_{2^m-1}s\right)^{k_2}\left(\sum_{j=0}^{l} q_{2^m-1}s\left(p_{2^m-1}s\right)^j\right)}{1-\sum_{j=0}^{k_2-1} q_{2^m-1}\left(p_{2^m-1}\right)^j s^{j+1}}\right]^{-1}\\ &= \frac{zs\left(p_{2^m-1}s\right)^{k_2}}{1-\sum_{j=0}^{k_2-1} q_{2^m-1}\left(p_{2^m-1}\right)^j s^{j+1}}\\ &\quad\times\left[1-z\left(\left(p_{2^m-1}s\right)^{(l+1)}+\frac{\left(p_{2^m-1}s\right)^{k_2}\left(\sum_{j=0}^{l} q_{2^m-1}s\left(p_{2^m-1}s\right)^j\right)}{1-\sum_{j=0}^{k_2-1} q_{2^m-1}\left(p_{2^m-1}\right)^j s^{j+1}}\right)\right]^{-1}. \qquad (33)\end{aligned}$$

Now, we obtain $\zeta_r(s;k_1)$ by calculating the coefficient of z^r in the equation (33). Observe that coefficient of z^r is obtained by multiplying the coefficient of z^{r-1} in the expression in the last line in (33) by $s\left(p_{2^m-1}s\right)^{k_2}/\left(1-\sum_{j=0}^{k_2-1} q_{2^m-1}\left(p_{2^m-1}\right)^j s^{j+1}\right)$. Using the expansion $(1-az)^{-1}=\sum_{t=0}^{\infty} a^t z^t$, we have

$$\begin{aligned}\zeta_r(s;k_1) &= \frac{s\left(p_{2^m-1}s\right)^{k_2}}{1-\sum_{j=0}^{k_2-1} q_{2^m-1}s\left(p_{2^m-1}s\right)^j}\left[\left(p_{2^m-1}s\right)^{l+1}\right.\\ &\quad\left.+\frac{s\left(p_{2^m-1}s\right)^{k_2}}{1-\sum_{j=0}^{k_2-1} q_{2^m-1}\left(p_{2^m-1}\right)^j s^{j+1}}\sum_{j=0}^{l} q_{2^m-1}\left(p_{2^m-1}s\right)^j\right]^{r-1}.\end{aligned}$$

This completes the proof of theorem 1. □

6. Conclusion

In this article we have defined a new discrete distribution, called *generalized Binomial type distribution*. The probability generating function of the distribution has been given along with its connections to the classical Binomial distribution as well as the negative Binomial distribution. We have studied the number of overlapping runs of length k_1 until the r^{th} occurrence of l-look-back run of length k ($k_1 \le k$) for the m^{th} order Markov chain and obtained the explicit expression of its probability generating function. Further, we have shown that our result generalizes the results of Aki and Hirano (1994) when we consider $r = 1$ for both *i.i.d.* as well as Markov dependent case. Since our stopping time is quite

general, our theorem will also provide similar results when we apply it to different cases such as the r^{th} occurrence of non-overlapping runs or r^{th} occurrence of overlapping runs or r^{th} occurrence of μ-overlapping runs (for positive μ).

Our result shows that the conditional distribution, that we have considered, has a renewal structure (see Feller (1968)) in the sense that it splits into independent sums of random variables, which may be interpreted as arrival times in a renewal process. Further, it is also seen that the arrival times are identical except the first arrival time. In other words, it admits a delayed renewal structure. We are able to identify the arrival times through the newly defined generalized Binomial type distribution and extended geometric discrete distribution. This renewal structure, in turn, can also be used to obtain approximate distribution of number of runs when the value of r is large. For instance, we may obtain the strong law of large numbers for the number of overlapping runs of length k_1.

We also provide a versatile method of proving the result where we convert our problem from the m^{th} order Markov chain into a simple Markov chain by combining the states. This allows us to use the Markov chain machinery, namely the strong Markov property, to derive the recurrence relation and use the method of generating functions effectively to obtain our results. Our method is quite powerful and can be used to prove similar results for other run statistic. We expect that, in future, there will be more applications of our method.

Acknowledgement

The author wishes to thank the referees for their helpful comments which have improved the presentation of the paper.

References

Aki, S. (1985). Discrete distributions of order k on a binary sequence. *Annals of the Institute of Statistical Mathematics*, **37**, 205–224.

Aki, S. and Hirano, K. (1994). Distributions of numbers of failures and successes until the first consecutive k successes. *Annals of the Institute of Statistical Mathematics*, **46**, 193–202.

Aki, S. and Hirano, K. (1995). Joint distributions of numbers of success-runs and failures until the first k consecutive successes. *Annals of the Institute of Statistical Mathematics*, **47**, 225–235.

Aki, S. and Hirano, K. (2000). Number of success-runs of specified length until certain stopping time rules and generalized binomial distributions of order k. *Annals of the Institute of Statistical Mathematics*, **52**, 767–777.

Anuradha. (2023). Asymptotic results for generalized runs in higher order Markov Chains. *Statistics and Applications*, **21**, 189–207.

Chadjiconstantinidis, S. and Koutras, M. V. (2001). Distributions of the numbers of failures and successes in a waiting time problem. *Annals of the Institute of Statistical Mathematics*, **53**, 576–598.

Ebneshahrashoob, M. and Sobel, M. (1990). Sooner or Later waiting time problems for Bernoulli trials: Frequency and run quotas. *Statistics and Probability Letters*, **9**, 5–11.

Feller, W. (1968). *An Introduction to Probability Theory and its Applications. Vol - I.* John Wiley, New York, 3rd ed.

Fu, J. C. and Koutras, M. V. (1994). Distribution Theory of runs: a Markov Chain approach. *Journal of the American Statistical Association*, **89**, 1050–1058.

Han, Q. and Aki, S. (2000). A unified approach to binomial type distributions of order k. *Communications in Statistics - Theory and Methods*, **29**, 1929–1943.

Hirano, K., Aki, S., and Uchida, M. (1997). Distributions of numbers of success-runs until the first consecutive k successes in higher order Markov dependent trials. In: Balakrishnan, N. (eds) *Advances in Combinatorial Methods and Applications to Probability and Statistics*, Birkhäuser Boston, Boston, 401–410.

Inoue, K. and Aki, S. (2002). Generalized binomial and negative binomial distributions of order k by the l-overlapping enumeration scheme. *Annals of the Institute of Statistical Mathematics*, **55**, 153–167.

Koutras, M. V. (1997). Waiting time distributions associated with runs of fixed length in two-state Markov chains. *Annals of the Institute of Statistical Mathematics*, **49**, 123–139.

Makri, F. S. and Philippou, A. N. (2005). On binomial and circular binomial distributions of order k for l- overlapping success runs of length k. *Statistical Papers*, **46**, 411–432.

Makri, F. S., Philippou, A. N., and Psillakis, Z. M. (2007). Polya, inverse Polya, and circular Polya distributions of order k for l- overlapping success runs. *Communications in Statistics - Theory and Methods*, **36**, 657–668.

Philippou, A. N. and Makri, F. S. (1986). Successes, runs and longest runs. *Statistics and Probability Letters*, **4**, 211–215.

Uchida, M. (1998). On number of occurrences of success runs of specified length in a higher order two-state Markov chain. *Annals of the Institute of Statistical Mathematics*, **50**, 587–601.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 21, No. 2, 2023 (New Series), pp 279–297
http://www.ssca.org.in/journal

Topp-Leone Generated q-Weibull Distribution and its Applications

Nicy Sebastian, Jeena Joseph and Sona Santhosh
Department of Statistics, St Thomas College, Thrissur, University of Calicut, Kerala, India-680 001

Received: 24 September 2022; Revised:14 January 2023; Accepted: Accepted: 15 April 2023

Abstract

In this paper, we introduce a new generated distribution called the Topp-Leone Generated q-Weibull(TLqW) Distribution. The described distribution's many distributional attributes and reliability traits are covered. Some well-known special cases of the mentioned model are also listed. When the lifetimes follow this distribution, it is better to establish a new reliability test plan, which aids in picking the best choices. The maximum likelihood method is investigated for parameter estimation in models. Using actual data sets, we used empirical evidence to demonstrate the value and adaptability of the new model in the model building process. The new test plan is then used to demonstrate how it may be used for creating dependable software in commercial settings.

Key words: Topp-Leone generated q-Weibull distribution; Quantile function; Reliability test plan; Maximum likelihood estimation.

AMS Subject Classifications: 60E05, 33B20, 62G05, 62N05, 62G07

1. Introduction

Numerous statistical distributions, including exponential, Weibull, logistic, and others, are significant in modelling survival and life-time data. The support for almost all of these distributions is unbounded. However, there are instances in real life where observations can only represent values in a small range, such percentages, proportions, or fractions. According to Papke and Wooldridge (1996), the variable is limited between zero and one in many economic scenarios, including the percentage of total weekly hours spent working, pension plan participation rates, industry market shares, percentage of land area given to agriculture, *etc.* As a result, for models to produce results that make sense, the unit interval must be used as the definition. Additionally, some writers use continuous models with finite support to characterise lifetime data while conducting reliability analysis. The most prevalent distribution for modelling continuous variables in the unit interval is the beta distribution, as is widely known. Due to the excellent flexibility of its density function, this distribution is well-liked in the fields of engineering, economics, biology, and ecology,

Corresponding Author: Nicy Sebastian
Email: nicycms@gmail.com

among others. However, the mathematical formulation is found to be challenging because its distribution function cannot be written in closed form and it incorporates the incomplete beta function. In contrast, a number of scholars have suggested alternatives to the beta distribution by reviving the one Kumaraswamy suggested in 1980.

Topp and Leone's new distribution, known as the Topp Leone (TL) distribution, defined on finite support, was introduced in 1955. Several authors researched this distribution. The Topp Leone distribution offers closed variants of the probability density function (pdf) and cumulative density function (cdf), and it describes empirical data with a J-shaped histogram, such as powered tool band failures and automatic adding machine failure. Prior to being identified by Nadarajah and Kotz (2003), the Topp Leone distribution has gotten little attention. They examined various aspects of TL distribution and supplied its moments, central moments, and characteristic functions. Some reliability metrics of the TL distribution, including a hazard function, mean residual life, reversed hazard rate, predicted inactivity time, and its stochastic orderings were presented by Ghitany *et al.* (2005). Kotz and Seier (2002) reported a discussion on the TL distribution's kurtosis.

If a random variable X belongs to the TL distribution, it can have either finite ($0 < x < b$) or infinite ($0 < x < b < \infty$) support. To avoid using any additional functions for creating a new family of produced distributions, we here largely concentrate on the TL distribution with $b = 1$ (see Zografos and Balakrishnan (2009), Alzaatreh *et al.* (2013), Lee *et al.* (2013)).

Topp and Leone concentrated on creating J-shaped histogram distributions for empirical data. A random variable X is distributed as the TL, bounded on (0,1) with cdf

$$F_{TL}(x) = x^{\alpha}(2-x)^{\alpha}; 0 < x < 1, \tag{1}$$

where $\alpha > 0$. Its pdf associated with equation (1) is

$$f_{TL}(x) = 2\alpha x^{\alpha-1}(1-x)(2-x)^{\alpha-1}. \tag{2}$$

This distribution can alternatively be seen as one in which the failure rate is proportional to a power of time, assuming the random variable X represents the failure times. The survival and hazard functions are the other crucial traits. They are respectively

$$s(x) = 1 - x^{\alpha}(2-x)^{\alpha},$$

and

$$h(x) = \frac{2\alpha x^{\alpha-1}(1-x)(2-x)^{\alpha-1}}{1-x^{\alpha}(2-x)^{\alpha}}.$$

Life time distributions' hazard rate functions can be monotonically increasing, monotonically decreasing, or U-shaped (bath tub shaped). Each example has applications in the real world. In the case of the TL distribution, the failure rate decreases over time if the shape parameter's value is less than one, remains constant over time if it is equal to one, and rises over time if it is more than one. Additionally, Nadarajah and Kotz noted that the bathtub shape of the hazard function is provided by the TL distribution when $0 < \alpha < 1$.

To get the Topp-Leone generated (TLG) family of distribution, use the TL distribution as the generating distribution. Then relation of a random variable X having the TLG distribution and a random variable T having TL distribution is $X = G^{-1}(T)$, with $T \sim TL(\alpha)$. This relationship depicts how the function $G(\cdot)$ transforms the TL distribution's pdf (2) into a new probability function.

$$F_{TLG}(x) = 2\alpha \int_0^{G(x)} t^{\alpha-1}(1-t)(2-t)^{\alpha-1}dt = G(x)^{\alpha}(2-G(x))^{\alpha}. \tag{3}$$

By differentiating, we get the corresponding pdf,

$$f_{TLG}(x) = 2\alpha g(x)(1-G(x))G(x)^{\alpha-1}(2-G(x))^{\alpha-1}. \tag{4}$$

The Topp-Leone generated exponential (TLE) distribution was introduced by Sangsanit and Bodhisuwan (2016) as an illustration of the Topp-Leone generated distribution. Even though exponential distribution is frequently used in reliability analysis, its constant hazard rate still remains a limitation of this distribution. The two-parameter Weibull distribution is one of the most well-known generalisations of the exponential distribution. Weibull distribution has many applications in real data analysis. Aryal *et al.* (2017) discussed characterizations and applications of Topp-Leone generated Weibull distribution. We can generalize TLE distribution into TLW distribution using a transformation. If X follows TLE distribution then the distribution of $Y = X^{\frac{1}{\gamma}}, \gamma > 0$ follows TLW distribution. Hence, a random variable X is said to follow TLW distribution if it has the cdf and the pdf as in the form

$$F_{TLW}(x) = 1 - exp(-2(\nu x)^{\gamma})^{\alpha}, \gamma, \alpha, \nu > 0, \tag{5}$$

$$f_{TLW}(x) = 2\alpha\gamma\nu^{\gamma}x^{\gamma-1}exp(-2(\nu x)^{\gamma})(1-exp(-2(\nu x)^{\gamma})^{\alpha-1}, \tag{6}$$

where α, γ are shape parameter and ν is the scale parameter.

Authors have recently examined a variety of q-type distributions, including q-exponential, q-Weibull, q-logistic, etc. Since the exponential form can be attained gradually as $q \to 1$, the q-exponential distribution can be seen as a stretched model of the exponential distribution (see Beck (2006), Beck and Cohen (2003), Mathai(2005)). According to Tsallis statistics and many research based on q-type distributions, including q-Weibull, Wilk and Wlodarczyk (2000, 2001) and Tsallis (1988). Costa *et al.* (2006) described a research of dielectric breakdown in electronic device oxides and demonstrated that a q-Weibull distribution provides a satisfactory fit for the data. For $x > 0$ and for $q > 1$ the distribution function and the density function of the q-Weibull distribution is,

$$F_1(x) = 1 - [1 + (q-1)(\lambda x)^{\gamma}]^{\frac{q-2}{q-1}}, \tag{7}$$

$$f_1(x) = \gamma\lambda^{\gamma}(2-q)x^{\gamma-1}[1 + (q-1)(\lambda x)^{\gamma}]^{\frac{-1}{q-1}}. \tag{8}$$

where, $\gamma, \lambda > 0, 1 < q < 2$. For $x > 0$ and $q < 1$,the cdf and the density function of q - Weibull distribution becomes

$$F_2(x) = 1 - [1 - (1-q)(\lambda x)^{\gamma}]^{\frac{2-q}{1-q}}, \quad 0 \le x \le \frac{1}{\lambda(1-q)^{\frac{1}{\gamma}}}, \tag{9}$$

$$f_2(x) = \gamma\lambda^{\gamma}(2-q)x^{\gamma-1}[1-(1-q)(\lambda x)^{\gamma}]^{\frac{1}{1-q}}. \tag{10}$$

Clearly, as q tends to 1 $f_1(x)$ and $f_2(x)$ tend to the usual Weibull distribution with two parameters γ, λ.

The rest of the paper is organized as follows. In section 2 we will introduce the Topp-Leone q-Weibull Distribution (TLqW) and further properties. In section 3, we described a new reliability test plan for TLqW distribution. In section 4, we study the estimation of parameters of the TLqW distribution, using the method of maximum likelihood. Simulation studies, real data illustrations, and reliability test applications of TLqW distribution are also discussed in section 5. Concluding remarks are addressed in section 6.

2. Topp Leone q - Weibull distribution

The applications of the q-weibull distribution have recently been studied by a number of researchers in the contexts of information theory, statistical mechanics, reliability modelling, *etc.* In terms of reliability, the TL distribution is a fairly adaptable distribution. We therefore use the origin of the TLG distribution to merge these two distributions inspired by this. As a result, we present the TLqW distribution.

A random variable X possessing the TLqW distribution with $q > 1$ has the cdf and pdf respectively are

$$F_{1TLqW} = \left(1-(1+(q-1)(\lambda x)^{\gamma})^{\frac{2q-4}{q-1}}\right)^{\alpha}, x > 0, \lambda, \alpha, \gamma > 0, 1 < q < 2. \tag{11}$$

and

$$f_{1TLqW}(x) = 2\alpha\gamma\lambda^{\gamma}(2-q)x^{\gamma-1}[1+(q-1)(\lambda x)^{\gamma}]^{\frac{q-3}{q-1}}\{1-[1+(q-1)(\lambda x)^{\gamma}]^{\frac{2q-4}{q-1}}\}^{\alpha-1}, \tag{12}$$

where $x > 0, \lambda, \alpha, \gamma > 0, 1 < q < 2$.

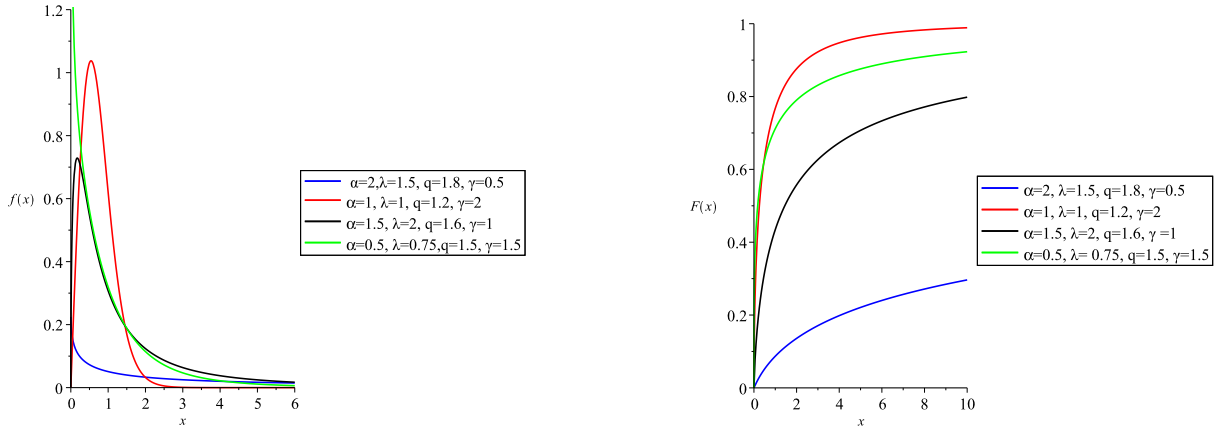


Figure 1: Plots of pdf and cdf of TLqW distribution

The plots of pdf and cdf of TLqW for various values of the shape parameters α, γ and q are shown in Figure 1. Survival function is the probability that a system will survive beyond a

given time. The survival function S(x) for TLqW distribution is

$$\begin{aligned} S(x) &= 1 - F_{TLqW}(x) \\ &= 1 - \{1 - (1 + (q-1)(\lambda x)^{\gamma})^{2\frac{q-2}{q-1}}\}^{\alpha}. \end{aligned} \tag{13}$$

The TLqW distribution's hazard function is

$$\begin{aligned} h(x) &= \frac{f_{TLqW}(x)}{1 - F_{TLqW}(x)} \\ &= \frac{2\alpha\gamma\lambda^{\gamma}(2-q)x^{\gamma-1}[1 + (q-1)(\lambda x)^{\gamma}]^{\frac{q-3}{q-1}}\{1 - [1 + (q-1)(\lambda x)^{\gamma}]2\frac{q-2}{q-1}\}^{\alpha-1}}{1 - \{1 - (1 - [1 + (q-1)(\lambda x)^{\gamma}]^{2\frac{q-2}{q-1}})\}^{\alpha}}. \end{aligned} \tag{14}$$

One can see the behaviour of hazard function using Figure 2. The Cumulative hazard

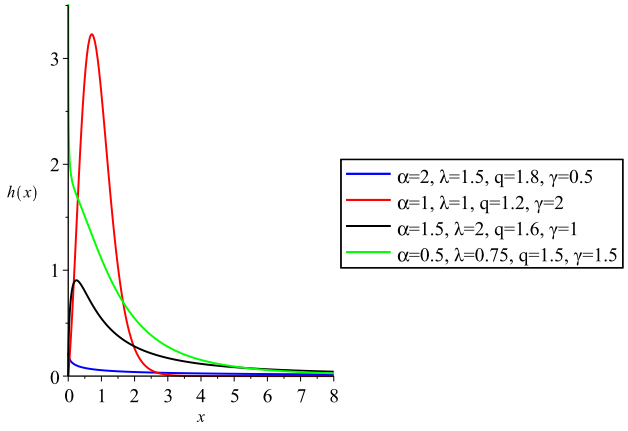


Figure 2: Plot of h(x) of TLqW distribution

function H(x)is defined as

$$\begin{aligned} H(x) &= \int_0^t h(t)dt \\ &= -\ln\{1 - \{1 - (1 + (q-1)(\lambda x)^{\gamma})^{2\frac{q-2}{q-1}}\}^{\alpha}\}. \end{aligned} \tag{15}$$

There are several new as well as well known distributions that can be obtained from the TLqW distributions. The sub-models include the following distributions:

1. When $q \to 1$, we obtain Topp Leone Weibull (TPW) distribution
2. When γ =1, we obtain Topp Leone q Exponential (TPqE) distribution
3. If $q \to 1$ and γ =1, we have the Topp Leone Exponential (TLE) distribution

2.1. L - Class property

The class L distributions are a significant class of distributions utilised in queuing theory and risk theory.

A distribution F belongs to the class L if

$$\lim_{x\to\infty} \frac{1-F(x-y)}{1-F(x)} = 1, \forall y \in R.$$

2.2. Quantile function

Probability distributions can be defined in terms of distribution functions or quantile functions when modelling and analysing statistical data. Quantile functions are more practical for analysis since they possess a number of intriguing characteristics that distributions do not share. The quantile function $Q(u)$ is defined as for a non-negative random variable X with distribution function $F(x)$ (see Nair *et al.* (2013)),

$$Q(u) = F^{-1}(u) = inf\{x : F(x) \geq u\}, \quad 0 \leq u \leq 1.$$

For every -$\infty < x < \infty$ and $0 < u < 1$, we have

$$F(x) \geq u \text{ if and only if } Q(u) \leq x.$$

As a result, $Q(u)$ is the smallest value of x satisfying $F(x) = u$ and $F(Q(u)) = u$ if there is an x such that $F(x) = u$. By solving the equation $F(x) = u$, we may get x in terms of u, which is the quantile function of X, if $F(x)$ is continuous and strictly growing. Moreover, if $Q(u)$ is the only value of x such that $F(x) = u$, then $F(x) = u$. The quantile function of TLqW distribution when $1 < q < 2$ is obtained as,

$$Q(u) = \left(\frac{\left(\sqrt{1-u^{\frac{1}{\alpha}}}\right)^{(\frac{q-1}{q-2})} - 1}{(q-1)\lambda^{\gamma}} \right)^{1/\gamma}, 1 < q < 2.$$

where u is chosen at random from the uniform distribution throughout the range $(0, 1)$. By matching population features with comparable sample characteristics, quantile-based measures of distributional properties such as location, dispersion, skewness, and kurtosis can be used to estimate model parameters. We can obtain the median as

$$M = Q(\frac{1}{2}) = \left(\frac{1-(1-0.50^{\frac{1}{\alpha}})^{\frac{1-q}{4-2q}}}{(1-q)\lambda^{\gamma}} \right)^{\frac{1}{\gamma}}.$$

The inter-quartile-range (IQR) of the TLqW model is,

$$IQR = Q(\frac{3}{4}) - Q(\frac{1}{4}) = \left(\frac{1-(1-0.75^{\frac{1}{\alpha}})^{\frac{1-q}{4-2q}}}{(1-q)\lambda^{\gamma}} \right)^{\frac{1}{\gamma}} - \left(\frac{1-(1-0.25^{\frac{1}{\alpha}})^{\frac{1-q}{4-2q}}}{(1-q)\lambda^{\gamma}} \right)^{\frac{1}{\gamma}}.$$

The Galton's coefficient of skewness (S) of the TLqW model is,

$$S = \frac{Q(\frac{3}{4}) + Q(\frac{1}{4}) - 2Median}{IQR}$$
$$= \frac{\left(\frac{1-(1-0.75^{\frac{1}{\alpha}})^{\frac{1-q}{4-2q}}}{(1-q)\lambda^{\gamma}}\right)^{\frac{1}{r}} - \left(\frac{1-(1-0.25^{\frac{1}{\alpha}})^{\frac{1-q}{4-2q}}}{(1-q)\lambda^{\gamma}}\right)^{\frac{1}{\gamma}} - 2\left(\frac{1-(1-0.50^{\frac{1}{\alpha}})^{\frac{1-q}{4-2q}}}{(1-q)\lambda^{\gamma}}\right)^{\frac{1}{\gamma}}}{\left(\frac{1-(1-0.75^{\frac{1}{\alpha}})^{\frac{1-q}{4-2q}}}{(1-q)\lambda^{\gamma}}\right)^{\frac{1}{\gamma}} - \left(\frac{1-(1-0.25^{\frac{1}{\alpha}})^{\frac{1-q}{4-2q}}}{(1-q)\lambda^{\gamma}}\right)^{\frac{1}{\gamma}}}.$$

Moor's coefficient of kurtosis (T) of the TLqW model is,

$$T = \frac{Q(\frac{7}{8}) - Q(\frac{5}{8}) + Q(\frac{3}{8}) - Q(\frac{1}{8})}{IQR}$$
$$= \frac{\left(\frac{1-(1-\frac{7}{8}^{\frac{1}{\alpha}})^{\frac{1-q}{4-2q}}}{(1-q)\lambda^{\gamma}}\right)^{\frac{1}{\gamma}} - \left(\frac{1-(1-\frac{5}{8}^{\frac{1}{\alpha}})^{\frac{1-q}{4-2q}}}{(1-q)\lambda^{\gamma}}\right)^{\frac{1}{\gamma}} - \left(\frac{1-(1-\frac{3}{8}^{\frac{1}{\alpha}})^{\frac{1-q}{4-2q}}}{(1-q)\lambda^{\gamma}}\right)^{\frac{1}{\gamma}} - \left(\frac{1-(1-\frac{1}{8}^{\frac{1}{\alpha}})^{\frac{1-q}{4-2q}}}{(1-q)\lambda^{\gamma}}\right)^{\frac{1}{\gamma}}}{\left(\frac{1-(1-\frac{3}{4}^{\frac{1}{\alpha}})^{\frac{1-q}{4-2q}}}{(1-q)\lambda^{\gamma}}\right)^{\frac{1}{r}} - \left(\frac{1-(1-\frac{1}{4}^{\frac{1}{\alpha}})^{\frac{1-q}{4-2q}}}{(1-q)\lambda^{\gamma}}\right)^{\frac{1}{\gamma}}}.$$

2.3. Simulation

A random variable Y having TLqW distribution can be simulated, for $1 < q < 2$ as,

$$Y = \left\{ \frac{[1 - U^{\frac{1}{\alpha}}]^{\frac{q-1}{2(q-2)}} - 1}{(q-1)(\lambda)^{\gamma}} \right\}^{\frac{1}{\gamma}},$$

where $U \sim U(0, 1)$.

3. Reliability test plan

The acceptance sampling plan inspection method, which is used to decide whether to accept or reject a specific quantity of material (see Kantam *et al.* (2001), Rao *et al.* (2011), Jose and Joseph (2018), *etc.*), is prescribed. If it is applied to a series of lots, the method will give a specific probability of accepting lots of a given quality. Here we establish the reliability test, with its operating characteristic function plan for accepting or rejecting a lot where the lifetime of the product follows Topp-Leone generated q-Weibull distribution. The process in a life testing experiment is to call the test off at a predetermined time t and record the number of failures. We accept the lot with a specified probability of at least p if the number of failures at the end of time t does not exceed a predetermined number c, known as the acceptance number. However, we reject the lot if the failure rate reaches c before time t. We are interested in obtaining the smallest sample size possible in order to arrive at a decision for such a truncated life test and the accompanying decision rule.

Although many distributions from the Topp-Leone produced family have been created with a variety of uses, none of them have been used in acceptance sampling to create reliability test plans. Assume that the lifetime of a product T follows the Topp-Leone generated

q -Weibull distribution with cdf

$$F(t) = \{1 - [1 + (q-1)(\frac{t}{\lambda})^{\gamma}]^{2(\frac{q-2}{q-1})}\}^{\alpha}, t > 0, \lambda, \alpha, \gamma > 0, 1 < q < 2. \tag{16}$$

Let λ_0 be the required minimum average life time and the shape parameters α,γ and q are known. Then

$$F_{TLqW}(t; \alpha, q, \gamma, \lambda) \leq G_{TLqW}(t; \alpha, q, \gamma, \lambda_0) \Leftrightarrow \lambda \geq \lambda_0. \tag{17}$$

The number of units under test n, the acceptance number c, the maximum test time t, and the minimum average lifetime λ_0 are used to define a sampling plan. The consumer's risk (chance of accepting a bad lot) shouldn't be higher than the value $1 - p^*$, where p^* is a lower bound on the likelihood that the sampling plan will reject a lot with a true value of λ below λ_0 . The sampling plan is defined by $(n, c, t/\lambda_0)$ for fixed p^* . For sufficiently large lots, the acceptance probability can be determined using the binomial distribution. For given values of c and t/λ_0, the goal is to find the smallest positive integer n such that

$$L(p_0) = \sum_{i=0}^{c} \binom{n}{i} {p_0}^i (1 - p_0)^{n-i} \leq 1 - p^*. \tag{18}$$

The operational characteristic function is increasing in λ, as indicated by the fact that the product's average lifespan increases with λ and the failure probability $p(\lambda)$ decreases. Where $p_0 = F_{TLqW}(t; \alpha, q, \gamma, \lambda_0)$ is given in (16) and denotes the failure probability before time t, which solely depends on the ratio t/λ_0. For $\alpha = 2, q = 1.1$,$\gamma = 1.2$ and p^*=0.75 and $t/\lambda_0 = 0.248, 0.361, 0.482, 0.602, 0.903, 1.204, 1.505$ and 1.806, the minimal values of n fulfilling (18) are obtained. Table 1 presents the findings.

The binomial probability can be approximated by the Poisson probability with the parameter $\theta = np_0$ if $p_0 = F_{TLqW}(t; \alpha, q, \gamma, \lambda_0)$ is small and n is very large. As a result, (18) becomes true.

$$L_1(p_0) = \sum_{i=0}^{c} \frac{\theta^i}{i!} e^{-\theta} \leq 1 - p^*. \tag{19}$$

For the same set of values for α,γ, q, p^* and t/λ_0, the minimum values of n satisfying (19) are obtained and shown in Table 2. In the beginning equation,

$$L(p_0) = \sum_{i=0}^{c} \binom{n}{i} {p_0}^i (1 - p_0)^{n-i}, \tag{20}$$

and in the end equation, $p = F(t, \lambda)$, where λ is taken into consideration, the probability $L(p_0)$ of accepting the lot is given by the operating characteristic function of the sampling plan $(n, c, t/\lambda_0)$. When p^* and t/λ_0 are given values, the values of n and c are calculated using the operating characteristics (OC) function, taking into account the fact that$p = F(\frac{t}{\lambda_0}/\frac{\lambda}{\lambda_0})$, and the results are displayed in Table 3.

The probability of rejecting a lot when $\lambda > \lambda_0$ is the producer's risk. By first determining that $p = F(t; \lambda)$ and then employing the binomial distribution function, we may

calculate the producer's risk. For illustration, we generate p from the sample plan provided in Table 1 for the given value of producer's risk, say 0.05, under the constraint that

$$\sum_{i=0}^{c} \binom{n}{i} {p_0}^i (1-p_0)^{n-i} \geq 0.95. \tag{21}$$

The minimal value of meeting (21) λ/λ_0 for the sampling plan $(n, c, t/\lambda_0)$ and for the specified p^* are reported in Table 6.

3.1. Explanation of the tables

Assume that q=1.1 and α=2 correspond to the TLqW distribution throughout the lifespan. Let us say the experimenter wants to confirm that the true unknown average life is at least 1000 hours with a $p^* = 0.75$ level of confidence. At $t = 903$ hours, the experiment should come to an end. The required n is hence 9 for an acceptance number $c = 4$ (Table 1). With a confidence level of 0.75, the experimenter can claim that the average life is at least 1000 hours if, during the course of 903 hours, no more than 4 failures out of 9 are detected. The value of n is 11 if the Poisson approximation to binomial probability is utilized (Table 2). The operational characteristic values from Table 3 are reported in Table 4 for this sample plan $(n = 9, c = 4, t/\lambda_0 = 0.903)$ under the TLqW distribution. The operational characteristic values from Table 3 are reported in Table 5 for the sample plan $(n = 7, c = 4, t/\lambda_0 = 1.806)$ with the consumer's risk of 0.05 under the TLqW distribution. This demonstrates that producers' risk is 0.05 when $\lambda/\lambda_0 = 3$ and insignificant when $\lambda/\lambda_0 = 4$. According to Table 3 for this plan, the minimum value of λ/λ_0, which represents the producer's risk as 0.05, is 3. When the consumer's risk is 0.25 or p^* =0.75, $c = 4$ and $t/\lambda_0 = 0.903$, the minimum ratio, $\lambda/\lambda_0 = 1.9619$ (from Table 6) which indicates that if $\lambda \geq 1.9619 \times (t/0.903) = 2.1726t = 1961.9$ hours, then, with sample size $n = 9$ and $c = 4$, the lot will be rejected with probability less than or equal to 0.05.

4. Maximum likelihood estimation

Let $x_1, x_2, ..., x_n$ be an observed random sample from TLqW distribution with $1 < q < 2$ unknown parameter vector $\theta = (\alpha, \gamma, \lambda, q)^T$. The likelihood function is then expressed as

$$L(\theta) = \prod_{i=1}^{n} 2\alpha\gamma\lambda^{\gamma}(2-q)x_i^{\gamma-1}(1+(q-1)(\lambda x_i)^{\gamma})^{\frac{q-3}{q-1}}\{1-(1+(q-1)(\lambda x_i)^{\gamma})^{2\frac{q-2}{q-1}}\}^{\alpha-1}.$$

The log-likelihood function is given by,

$$l(\theta) = \ln L(\theta) = n\ln 2 + n\ln\alpha + n\ln\gamma + n\gamma\ln\lambda + n\ln(2-q) + (\gamma-1)\sum_{i=1}^{n}\ln x_i + \frac{q-3}{q-1}\sum_{i=1}^{n}\ln(1+(q-1)(\lambda x_i)^{\gamma}) + (\alpha-1)\sum_{i=1}^{n}\ln\{1-(1+(q-1)(\lambda x_i)^{\gamma})^{2\frac{2-q}{1-q}}\}.$$

Let $z_i(x) = 1 + (q-1)(\lambda x_i)^{\gamma}$and $k = 2\frac{(q-2)}{(q-1)}$, then $l(\theta)$ can be written as,

Table 1: Using the Binomial approximation, minimum sample size

p^*	c	t/λ_0							
		0.248	0.361	0.482	0.602	0.903	1.204	1.505	1.806
0.75	0	18	8	5	3	2	1	1	1
	1	35	16	10	7	4	3	2	2
	2	51	21	14	10	6	4	4	3
	3	66	30	18	13	8	6	5	5
	4	82	37	22	16	9	7	6	6
	5	60	33	22	17	11	9	8	7
	6	70	38	25	19	13	10	9	8
	7	79	43	29	22	14	12	10	9
	8	88	48	32	24	16	13	11	11
	9	97	52	35	27	18	14	13	12
	10	106	57	39	30	20	16	14	13
0.90	0	29	13	8	5	3	2	2	1
	1	50	23	13	9	5	4	3	3
	2	69	31	18	13	7	5	4	4
	3	86	39	23	16	9	7	6	5
	4	103	47	28	19	11	8	7	6
	5	120	54	32	23	13	10	8	7
	6	136	62	37	26	15	11	9	9
	7	152	69	41	29	17	13	11	10
	8	168	76	46	32	19	14	12	11
	9	184	84	50	35	21	16	13	12
	10	200	91	55	38	22	17	14	13
0.95	0	39	17	10	7	4	3	2	2
	1	61	27	16	11	6	4	3	3
	2	81	36	21	15	8	6	5	4
	3	100	45	27	18	10	8	6	5
	4	118	53	32	22	12	9	7	7
	5	135	61	36	25	14	11	9	8
	6	153	69	41	29	16	12	10	9
	7	170	77	46	32	18	14	11	10
	8	186	84	50	35	20	15	13	11
	9	203	92	55	38	22	16	14	12
	10	219	99	59	42	24	18	15	14
0.99	0	58	26	15	10	5	4	3	2
	1	85	38	22	15	8	6	4	4
	2	107	48	28	19	10	7	6	5
	3	128	58	34	23	13	9	7	6
	4	149	67	39	27	15	11	9	7
	5	168	75	45	31	17	12	10	9
	6	187	84	50	34	19	14	11	10
	7	205	92	55	38	21	15	13	11
	8	224	101	60	41	23	17	14	12
	9	242	109	65	45	25	18	15	14
	10	259	117	70	47	27	20	17	15

Table 2: Using the Poisson approximation, the minimal sample size

p^*	c	t/λ_0							
		0.248	0.361	0.482	0.602	0.903	1.204	1.505	1.806
0.75	0	19	9	6	4	3	2	2	2
	1	36	17	10	7	5	4	4	3
	2	52	24	15	11	7	6	5	5
	3	67	31	19	14	9	7	6	6
	4	83	38	24	17	11	9	8	7
	5	98	45	28	20	13	10	9	8
	6	113	52	32	23	14	12	10	10
	7	127	59	36	26	16	13	12	11
	8	142	66	40	29	18	14	13	12
	9	157	72	45	32	20	16	14	13
	10	171	79	49	35	22	17	15	15
0.90	0	31	14	9	7	4	3	3	4
	1	51	24	15	11	7	6	5	6
	2	70	33	20	15	9	7	7	7
	3	88	41	25	18	11	9	8	9
	4	105	49	30	22	14	11	10	10
	5	122	57	35	25	16	12	11	12
	6	138	64	39	28	18	14	13	13
	7	155	72	44	32	20	16	14	15
	8	171	79	49	35	22	17	15	16
	9	187	86	53	38	24	19	17	17
	10	202	94	57	41	26	20	18	18
0.95	0	40	19	12	8	5	4	4	4
	1	63	29	18	13	8	7	6	6
	2	83	39	24	17	11	9	8	7
	3	102	47	29	21	13	11	9	9
	4	120	56	34	25	15	12	11	10
	5	138	64	39	28	18	14	13	12
	6	156	72	44	32	20	16	14	13
	7	173	80	49	35	22	17	16	15
	8	190	88	54	39	24	19	17	16
	9	206	95	59	42	26	21	18	17
	10	223	103	63	45	28	22	20	19
0.99	0	61	28	18	13	8	6	6	5
	1	87	41	25	18	11	9	8	8
	2	111	51	32	23	14	11	10	10
	3	132	61	38	27	17	13	12	11
	4	152	71	43	31	19	15	14	13
	5	172	80	49	35	22	17	16	15
	6	191	89	54	39	24	19	17	16
	7	210	97	60	43	27	21	19	18
	8	228	106	65	47	29	23	20	19
	9	246	114	70	50	31	25	22	21
	10	264	122	75	54	33	27	24	22

Table 3: Values for the sample plan's operating characteristic function $(n, c, t/\lambda_0)$

				λ/λ_0						
p^*	n	c	t/λ_0	2	2.5	3	3.5	4	4.5	5
0.75	82	4	0.241	0.9997	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
	37	4	0.361	0.9991	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
	22	4	0.482	0.9980	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999
	16	4	0.602	0.9952	0.9996	0.9999	0.9999	0.9999	0.9999	0.9999
	9	4	0.903	0.9856	0.9984	0.9997	0.9999	0.9999	0.9999	0.9999
	7	4	1.204	0.9587	0.9938	0.9989	0.9998	0.9999	0.9999	0.9999
	6	4	1.505	0.9162	0.9838	0.9968	0.9993	0.9998	0.9999	0.9999
	6	4	1.806	0.7663	0.9360	0.9838	0.9958	0.9988	0.9996	0.9998
0.90	103	4	0.241	0.9993	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
	47	4	0.361	0.9975	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999
	28	4	0.482	0.9940	0.9995	0.9999	0.9999	0.9999	0.9999	0.9999
	19	4	0.602	0.9894	0.9990	0.9998	0.9999	0.9999	0.9999	0.9999
	11	4	0.903	0.9621	0.9951	0.9993	0.9998	0.9999	0.9999	0.9999
	8	4	1.204	0.9192	0.9864	0.9976	0.9995	0.9998	0.9999	0.9999
	7	4	1.505	0.8195	0.9587	0.9910	0.9979	0.9994	0.9998	0.9999
	6	4	1.806	0.7663	0.9360	0.9838	0.9958	0.9988	0.9996	0.9998
0.95	118	4	0.241	0.9987	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
	53	4	0.361	0.9958	0.9997	0.9999	0.9999	0.9999	0.9999	0.9999
	32	4	0.482	0.9894	0.9991	0.9999	0.9999	0.9999	0.9999	0.9999
	22	4	0.602	0.9801	0.9980	0.9997	0.9999	0.9999	0.9999	0.9999
	12	4	0.903	0.9448	0.9925	0.9988	0.9998	0.9999	0.9999	0.9999
	9	4	1.204	0.8659	0.9747	0.9953	0.9990	0.9997	0.9999	0.9999
	7	4	1.505	0.8195	0.9587	0.9910	0.9979	0.9994	0.9998	0.9999
	7	4	1.806	0.5795	0.8572	0.9587	0.9884	0.9966	0.9989	0.9996
0.99	149	4	0.241	0.9965	0.9997	0.9999	0.9999	0.9999	0.9999	0.9999
	67	4	0.361	0.9889	0.9991	0.9999	0.9999	0.9999	0.9999	0.9999
	39	4	0.482	0.9763	0.9977	0.9997	0.9999	0.9999	0.9999	0.9999
	27	4	0.602	0.9551	0.9950	0.9993	0.9999	0.9999	0.9999	0.9999
	15	4	0.903	0.8705	0.9788	0.9965	0.9993	0.9998	0.9999	0.9999
	11	4	1.204	0.7287	0.9366	0.9866	0.9970	0.9993	0.9998	0.9999
	9	4	1.505	0.5732	0.8659	0.9645	0.9908	0.9975	0.9993	0.9997
	7	4	1.806	0.5795	0.8572	0.9587	0.9884	0.9966	0.9989	0.9996

Table 4: Values of the OC function for values of λ/λ_0 **at** $(n = 9, c = 4, t/\lambda_0 = 0.903)$

λ/λ_0	2	2.5	3	3.5	4
L(p)	0.9856	0.9984	0.9997	0.9999	0.9999

Table 5: Values of the OC function for values of λ/λ_0 **at** $(n = 7, c = 4, t/\lambda_0 = 1.806)$

λ/λ_0	2	2.5	3	3.5	4	4.5	5
L(p)	0.5795	0.8572	0.9587	0.9884	0.9966	0.9989	0.9996

Table 6: Minimum of λ/λ_0 for the acceptability of a lot with producer's risk of 0.05

p^*	c	t/λ_0							
		0.241	0.361	0.482	0.602	0.903	1.204	1.505	1.806
0.75	0	4.4202	4.6647	5.0961	5.4557	6.3239	6.7439	7.9027	9.2078
	1	2.6177	2.6953	2.9073	3.1296	3.4508	3.7272	3.7349	4.2172
	2	2.0644	2.1350	2.2313	2.3367	2.6114	2.7601	3.2916	3.2154
	3	1.8393	1.8984	1.9580	2.0957	2.3727	2.5925	2.7851	3.2154
	4	1.7188	1.7616	1.8008	1.8853	1.9619	2.1962	2.3978	2.7845
	5	1.6234	1.6725	1.7327	1.7773	1.8712	2.1316	2.1554	2.5061
	6	1.5622	1.6067	1.6543	1.6999	1.8052	1.9422	2.1922	2.2987
	7	1.5160	1.5582	1.5945	1.6415	1.7548	1.8168	2.0356	2.1467
	8	1.4796	1.5078	1.5479	1.5962	1.6361	1.8256	1.9156	2.0420
	9	1.4459	1.4784	1.5107	1.5293	1.6141	1.7264	1.8159	1.9256
	10	1.4182	1.4546	1.4804	1.5026	1.5949	1.6536	1.7332	1.8456
0.90	0	5.5043	5.8420	6.3102	6.4500	7.6763	8.4810	10.5436	10.1158
	1	2.9925	3.1559	3.2234	3.3993	3.7805	4.4528	4.6590	5.5908
	2	2.3788	2.4760	2.5309	2.7041	2.8758	3.1049	3.3046	3.9655
	3	2.0771	2.1603	2.2278	2.3005	2.4810	2.7966	3.1275	3.2283
	4	1.9092	1.9837	2.0451	2.0857	2.2539	2.4017	2.6903	2.7896
	5	1.8031	1.8494	1.9003	1.9837	2.1086	2.3241	2.3978	2.5061
	6	1.7165	1.7739	1.8216	1.8785	2.0026	2.1140	2.1922	2.6307
	7	1.6558	1.6997	1.7458	1.7932	1.9211	2.0932	2.2509	2.4552
	8	1.6064	1.6464	1.6942	1.7292	1.8635	1.9541	2.1100	2.2987
	9	1.5685	1.6095	1.6466	1.6791	1.8133	1.9541	1.9979	2.1791
	10	1.5355	1.5748	1.6167	1.6377	1.7172	1.8530	1.8996	2.0798
0.95	0	6.1698	6.4672	6.8688	7.3394	8.6576	10.1799	10.6012	12.7215
	1	3.2479	3.3789	3.5622	3.7318	4.1507	4.4294	4.6590	5.5908
	2	2.5480	2.6414	2.7313	2.9068	3.1138	3.4832	3.8811	3.9558
	3	2.2262	2.3104	2.4118	2.4561	2.6408	3.0678	3.1275	3.2348
	4	2.0322	2.0980	2.1983	2.2504	2.3868	2.6242	2.6903	3.2283
	5	1.9003	1.9664	2.0179	2.0759	2.2092	2.4962	2.6688	2.8774
	6	1.8144	1.8693	1.9220	1.9924	2.0919	2.2655	2.4336	2.6307
	7	1.7428	1.7964	1.8492	1.9015	2.0026	2.2189	2.2509	2.4552
	8	1.6853	1.7283	1.7741	1.8205	1.9304	2.0729	2.2832	2.2987
	9	1.6416	1.6859	1.7281	1.7605	1.8721	1.9541	2.1580	2.2269
	10	1.6012	1.6414	1.6779	1.7354	1.8256	1.9541	2.0530	2.2796
0.99	0	7.3513	7.8632	8.1828	8.6528	9.5893	11.4865	12.6403	12.7215
	1	3.7743	3.9580	4.1543	4.3619	4.7903	5.5126	5.5958	6.6232
	2	2.8869	3.0280	3.1469	3.2614	3.5104	3.8505	4.3540	4.6433
	3	2.4840	2.6006	2.6998	2.7758	3.0848	3.3080	3.5050	3.7417
	4	2.2557	2.3496	2.4185	2.5203	2.7413	3.0228	3.2723	3.2283
	5	2.1068	2.1774	2.2695	2.3402	2.5059	2.6634	2.9051	3.2026
	6	1.9953	2.0701	2.1332	2.1804	2.3501	2.5615	2.6635	2.9203
	7	1.9227	1.9558	2.0314	2.0857	2.2155	2.3494	2.6372	2.7011
	8	1.8377	1.8936	1.9542	1.9924	2.1256	2.2896	2.4426	2.5942
	9	1.7811	1.8370	1.8958	1.9414	2.0488	2.1522	2.3121	2.5897
	10	1.7328	1.7899	1.8419	1.8636	1.9827	2.1266	2.3142	2.4587

$$l(\theta) = n\ln 2 + n\ln\alpha + n\ln\gamma + n\gamma\ln\lambda + n\ln(2-q) + (\gamma-1)\sum_{i=1}^{n}\ln x_i + \frac{q-3}{q-1}\sum_{i=1}^{n}\ln z_i(x) + (\alpha-1)\sum_{i=1}^{n}\ln\{1-z_i(x)^k\}.$$

Differentiating $l(\theta)$ with respect to α, γ, λ,and q,we have

$$\frac{\partial l(\theta)}{\partial\alpha} = \frac{n}{\alpha} + \sum_{i=1}^{n}\ln\{1-z_i(x)^k\}.$$

$$\frac{\partial l(\theta)}{\partial\gamma} = \frac{n}{\gamma} + n\ln(\lambda) + \sum_{i=1}^{n}\ln(x_i) + (q-3)\sum_{i=1}^{n}\ln(\lambda x_i)(\lambda x_i)^{\gamma}\frac{1}{z_i(x)} - (\alpha-1)(q-1)k\sum_{i=1}^{n}\ln(\lambda x_i)(\lambda x_i)^{\gamma}\frac{z_i(x)^{k-1}}{1-z_i(x)^k}.$$

$$\frac{\partial l(\theta)}{\partial\lambda} = \frac{n\gamma}{\lambda} + \left[(q-1)\gamma\lambda^{\gamma-1}\right]\left\{\frac{q-3}{q-2}\sum_{i=1}^{n}\frac{x_i^{\gamma}}{z_i(x)} - (\alpha-1)k\sum_{i=1}^{n}{x_i}^{\gamma}\frac{z_i(x)^{k-1}}{1-z_i(x)^k}\right\}.$$

$$\frac{\partial l(\theta)}{\partial q} = -\frac{n}{2-q} + \frac{2}{(q-1)^2}\sum_{i=1}^{n}\ln z_i(x) + \frac{q-3}{q-1}\sum_{i=1}^{n}\frac{(\lambda x_i)^{\gamma}}{z_i(x)} - (\alpha-1)\sum_{i=1}^{n}\frac{1}{1-z_i(x)^k}\frac{d}{dq}(z_i(x))^k,$$

where

$$\frac{d}{dq}(z_i(x))^k = z_i(x)^k\{k\frac{(\lambda x_i)^{\gamma}}{z_i(x)} + \frac{k+2}{q-1}\ln z_i(x)\}.$$

Now, setting the non-linear system of equations,$\frac{\partial l(\theta)}{\partial\alpha} = 0$, $\frac{\partial l(\theta)}{\partial\gamma} = 0$, $\frac{\partial l(\theta)}{\partial\lambda} = 0$, $\frac{\partial l(\theta)}{\partial q} = 0$ and solving them simultaneously we obtain the maximum likelihood estimate, $\hat{\theta} = (\hat{\alpha}, \hat{\lambda}, \hat{\gamma}, \hat{q})^T$. One can utilise iterative techniques like the Newton-Raphson type algorithm to calculate the estimate when solving non-linear equations numerically.

5. Numerical illustration

5.1. Simulation study

In this section, we do simulation tests to assess how well the MLEs of the TLqW distribution parameters perform over the long term. Numerous finite sample sizes are taken into account and to be more specific, we create samples from the TLqW distribution with n = 50, 75, 100 and 110 for the parameter values $\alpha = 1.275$, $\lambda = 1.5$, $\gamma = 7.8$ and $q = 1.7$. Also, the iteration is conducted 1000 times. The mean values of the biases, root mean squared errors (RMSEs), 95% (asymptotic) coverage probabilities (CPs), and average lengths (ALs) of the 95% (asymptotic) CIs corresponding to each of the parameter estimates for every replication are calculated with respect to the corresponding sample sizes. From Table 7 it can be seen that the RMSEs and ALs corresponding to each estimate decrease as the sample size increases.

Table 7: Simulation results

Sample Size	Parameter	MLE	Bias	RMSE	CP	AL
n=50	α	1.415	0.140	0.141	1.000	0.425
	λ	1.469	-0.031	0.131	0.931	0.437
	γ	9.557	1.758	10.14	0.971	11.469
	q	0.024	-0.023	0.197	0.906	0.534
n=75	α	1.415	0.140	0.140	1.000	0.347
	λ	1.477	-0.023	0.100	0.937	0.351
	γ	8.727	0.927	3.177	0.967	7.500
	q	1.695	-0.005	0.111	0.917	0.371
n=100	α	1.415	0.139	0.140	0.999	0.300
	λ	1.481	-0.019	0.084	0.936	0.299
	γ	8.562	0.762	1.838	0.969	6.146
	q	1.703	0.003	0.083	0.925	0.302
n=110	α	1.415	0.139	0.139	0.878	0.286
	λ	1.482	-0.017	0.077	0.944	0.282
	γ	8.539	0.738	1.674	0.964	5.779
	q	1.706	0.006	0.076	0.916	0.283

5.2. Data illustration for failure time data

In the reliability tests described in this section, lifetime data from engineering equipment are used to show one use of the TLqW distribution. The example uses data from a set measuring how long it took 500 MW generators to fail for the first time (see Jia *et al.* (2020)). The data are (thousands of hours) 0.058, 0.070, 0.090, 0.105, 0.113, 0.121, 0.153, 0.159, 0.224, 0.421, 0.570, 0.596, 0.618, 0.834, 1.019, 1.104, 1.497, 2.027, 2.234, 2.372, 2.433, 2.505, 2.690, 2.877, 2.879, 3.166, 3.455, 3.551, 4.378, 4.872, 5.085, 5.272, 5.341, 8.952, 9.188 and 11.399. The use of the TLqW illustrates the ability of this distribution in dealing with the non-monotonic hazard rate function, which includes a set of problems with relevant applications in the reliability context; for more information, see Jiang *et al.* (2003). Commonly used distributions like Weibull are barely suitable to fit the mentioned failure data. The

Table 8: Goodness of fit for different distributions on the failure time data

Model	Estimates(SE)	lnL	K-S	p value	AIC
Weibull	$\hat{\lambda}$=2.3118(0.256) $\hat{\gamma}$= 0.8156(0.058)	-68.6906	0.1219	0.1880	141.3812
MWE	$\hat{\lambda}$=0.2130(0.133) $\hat{\theta}$= 10.0923(0.003) $\hat{\gamma}$= 0.6920(0.001)	-68.2628	0.1046	0.2900	142.5276
ENH	$\hat{\lambda}$=0.1430(1.934) $\hat{\eta}$= 1.6347 (0.248) $\hat{\gamma}$= 0.6415(0.181)	-68.3560	0.1021	0.3330	142.712
TPW	$\hat{\lambda}$=0.4754(0.424) $\hat{\alpha}$= 1.3378(0.859) $\hat{\gamma}$= 0.1337(0.046)	-68.4044	0.2483	0.0192	142.8089
TLqW	$\hat{\lambda}$=0.1816(0.029) $\hat{\alpha}$= 0.0794(0.019) $\hat{\gamma}$= 6.4944(0.154) $\hat{q}$=1.8799(0.013)	-46.9633	0.0852	0.9361	101.9265

$P-P$ plot of the failure time data is given in Figure 3. The estimated standard error values

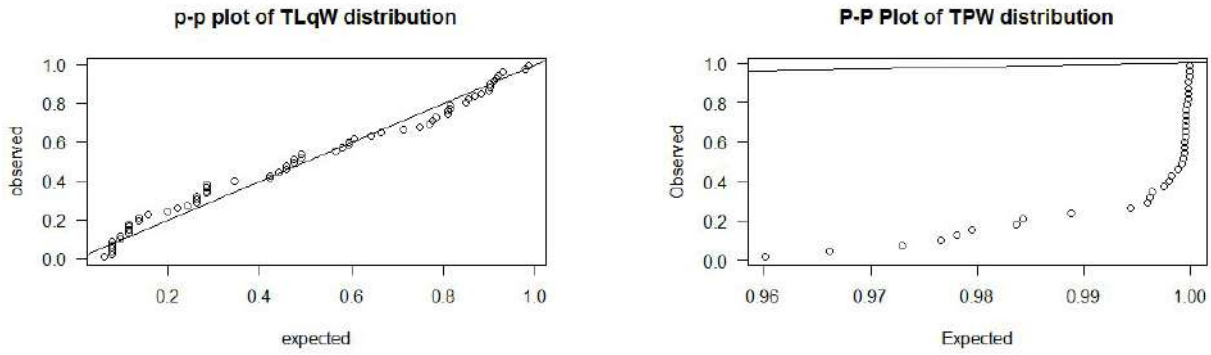


Figure 3: *P*-*P* **plot of failure time data**

are given in parentheses. It can be easily seen from the Table 8 that the TLqW distribution is a good alternative to the other lifetime models, namely the Weibull, modified Weibull extension (MWE), exponentiated Nadarajah-Haghighi (ENH), and TPW distributions.

5.3. Data illustration for fibre strength data

We use the original uncensored observations of the 1.5 cm glass fibre strengths made by employees of the UK National Physical Laboratory (see Merovci *et al.* (2016)). The fibre strength data are 0.55, 0.74, 0.77, 0.81, 0.84, 1.24, 0.93, 1.04, 1.11, 1.13, 1.30, 1.25, 1.27, 1.28, 1.29, 1.48, 1.36, 1.39, 1.42, 1.48, 1.51, 1.49, 1.49, 1.50, 1.50, 1.55, 1.52, 1.53, 1.54, 1.55, 1.61, 1.58, 1.59, 1.60, 1.61, 1.63, 1.61, 1.61, 1.62, 1.62, 1.67, 1.64, 1.66, 1.66, 1.66, 1.70, 1.68, 1.68, 1.69, 1.70, 1.78, 1.73, 1.76, 1.78, 1.73, 1.76, 1.76, 1.77, 1.89, 1.81, 1.82, 1.84, 1.84, 2.00, 2.01 and 2.24.

Table 9: Goodness of fit for different distributions on fibre strength data

Model	Estimates(SE)	lnL	K-S	p value	AIC
qW	$\hat{\lambda}$=0.0357(0.028) $\hat{\gamma}$= 1.2934(0.830) $\hat{q}$= 1.2934(0.142)	-296.15	0.1113	0.4053	598.31
TLqW	$\hat{\lambda}$=0.0360(0.007) $\hat{\gamma}$= 0.8304(0.206) $\hat{\alpha}$= 3.0546(0.024) $\hat{q}$=1.1747(0.012)	-294.74	0.1062	0.4654	594.49

Figure 4 gives the $P-P$ plot of the fibre strength data. It can be easily seen from the Table 9 that TLqW distribution gives better fit than q-Weibull (qW) distribution.

5.4. Reliability test comparison with Marshall-Olkin extended exponential distribution

Comparing Reliability Test Plans for Marshall-Olkin Extended Exponential distribution (see Rao *et al.* (2011)) with TLqW distribution, the minimal sample size is 49 using binomial approximation, whereas for α=2, acceptance number c=9, for the stated ratio t/λ_0=0.482 and confidence level p^*=0.75, whereas for TLqW distribution it is 35. The

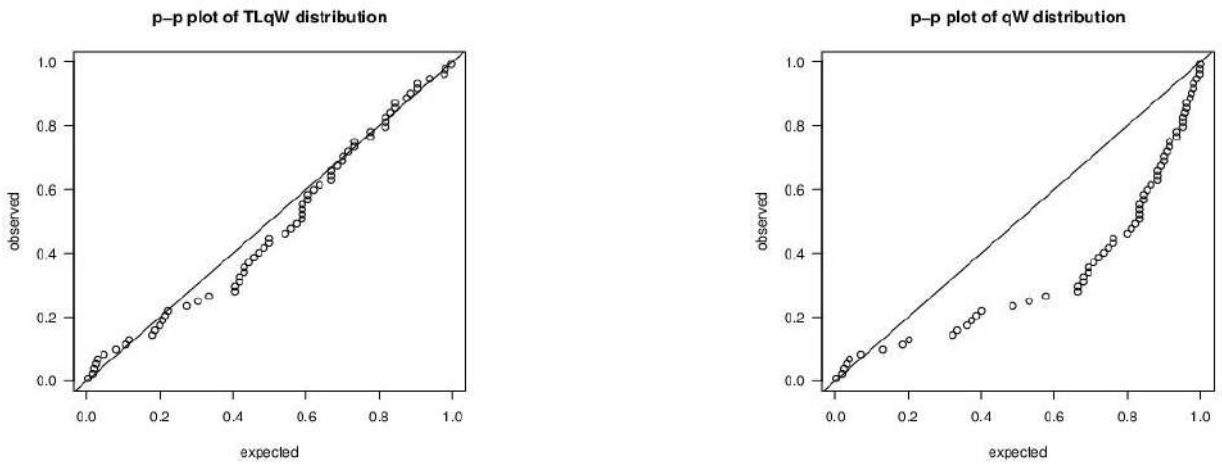


Figure 4: P-P plot of fibre strength data

scaled termination time is uniformly less than that for the current reliability test plans if we take into account each value of c and each value of t/λ_0. The new test plan is now more advantageous due to this modification, which also aids in selecting the best possible decisions.

5.5. Real life application of the new test plan

Take into account the following software release failure times, which are ordered and expressed in hours from the moment the software begins to run until a failure occurs (see, Wood(1996)). The observations 254, 788, 1054, 1393, 2216, 2880, 3593, 4281, and 5180 make up an ordered sample of this data with a size of $n = 9$.

Let's assume that the desired average lifetime is 1000 hours and that the testing time is 903 hours. This results in a ratio of $t/\lambda_0 = 0.903$, with a corresponding sample size of $n = 9$ and an acceptance number of $c = 4$, which is determined from Table 1 for $p^* = 0.75$. As a result, the sampling strategy for the sample data presented above is $(n = 9, c = 4, t/\lambda_0 = 0.903)$. We must choose whether to accept or reject the product in light of the observations. Only products with fewer than or equal to four failures prior to 903 hours are accepted. The sampling plan, however, only ensures the confidence level if the given life times follow the TLqW distribution. We compared the sample quantiles and the corresponding population quantiles and discovered a reasonable agreement, proving that the given sample is produced by lifetimes following the TLqW distribution. As a result, it would seem appropriate to adopt the sampling plan's decision rule. There are just two failures in the sample of 9 units, occurring 254 and 788 hours before $t = 903$ hours. Consequently, we approve the product.

6. Conclusion

The TLqW distribution is introduced in this paper as a generalization of the Weibull distribution. Class L is where the new distribution fits in. Additionally, the generation of random variates using the new model is straightforward. The Weibull distribution is shown to be a competitor of the new model, and the model's adaptability is demonstrated by fitting it

to two sets of real-world data. Additionally, we determine the minimal sample size required for a lot to be accepted or rejected using percentiles. The test strategy was established using some helpful tables that were provided. Therefore, we draw the conclusion that the Topp Leone q-Weibull distribution is the most appropriate model among those taken into consideration, as well as a model that is particularly capable of explaining lifetime scenarios. We anticipate that the new model will grab researchers' attention as a serious threat to the Weibull distribution.

Acknowledgements

The authors would like to thank the referees for their valuable comments, which enabled the authors to improve the presentation of the material in the paper. The authors declare no conflict of interest. This research is not supported by any grant from any granting agency.

References

Aryal, G. R., Ortega, E. M., Hamedani, G., and Yousof, H. M. (2017). The Topp-Leone generated Weibull distribution: regression model characterizations and applications. *International Journal of Statistics and Probability*, **6**, 126–141.

Alzaatreh, A., Lee, C., and Famoye, F. (2013). A new method for generating families of continuous distributions. *Metron*, **71**, 63–79.

Beck, C. (2006). Stretched exponentials from superstatistics. *Physica A*, **365**, 96–101.

Beck, C. and Cohen, E. C. D. (2003). Superstatistics. *Physica A*, **322**, 267–275.

Costa, U. M. S., Freire, V. N., Malacarne, L. C., Mendes, R. S., Picoli, S. Jr., Vasconcelos, E. A., and da Silva E. F. Jr. (2006). An improved description of the dielectric breakdown in oxides based on a generalized Weibull distribution. *Physica A*, **361**, 209–215.

Ghitany M. E., Kotz, S., and Xie, M. (2005). On some reliability measures and their stochastic orderings for the Topp-Leone distribution. *Journal of Applied Statistics*, **32**, 715–722.

Jia, X., Nadarajah, S., and Guo, B. (2020). Inference on q-Weibull parameters. *Statistical Papers*, **61**, 575-593.

Jiang, R., Ji, P., and Xiao, X. (2003). Aging property of unimodal failure rate models. *Reliability Engineering & System Safety*, **79**, 113–116.

Jose, K. K. and Joseph, J. (2018). Reliability test plan for the Gumbel-uniform distribution. *Stochastics and Quality Control*, **33**, 71–81.

Kantam, R. R. L., Rosaiah, K., and Rao, G. S. (2001). Acceptance sampling based on life tests: log-logistic model. *Journal of Applied Statistics*, **28**, 121–128.

Kotz, S. and Seier, E. (2007). Kurtosis of the Topp-Leone distributions. *Interstat*,**1**, 1–15.

Kumaraswamy, P. (1980). Generalized probability density-function for double-bounded random processes. *Journal of Hydrology*, **46**, 79–88.

Lee, C., Famoye, F., and Olumolade, O. (2007). Beta-Weibull distribution: some properties and applications to censored data. *Journal of Modern Applied Statistical Methods*, **6**, 173–186.

Mathai, A. M. (2005). A pathway to matrix-variate Gamma and Normal densities. *Linear Algebra and its Applications*, **396**, 317–328.

Merovci, F., Khaleel, M. A., Ibrahim, N. A., and Shitan, M. (2016). The beta type X distribution: properties with applications. *Springer Plus*, **5**, 697.

Nadarajah, S. and Kotz, S. (2003). Moments of some J-shaped distributions. *Journal of Applied Statistics*. **30**, 311–317.

Nair, N. U., Sankaran, P. G., and Balakrishnan, N. (2013). *Quantile-Based Reliability Analysis: Statistics for Industry and Technology*. Springer, New York.

Papke, L. and Wooldridge, J. (1996). Econometric methods for fractional response variables with an application to 401(K) plan participation rates. *Journal of Applied Econometrics*, **11**, 619–632.

Rao, S. G., Ghitany, M. E., and Kantam, R. R. L. (2009). Reliability test plans for Marshall-Olkin extended Exponential distribution. *Applied Mathematical Sciences*, **3**, 2745–2755.

Sangsanit, Y. and Bodhisuwan, W. (2016). The Topp-Leone generator of distributions: properties and inferences. *Songklanakarin Journal of Science and Technology*, **38**, 537–548.

Tsallis, C. (1988). Possible generalizations of Boltzmann-Gibbs statistics. *Journal of Statistical Physics*, **52**, 479–487.

Topp, C. W. and Leone, F. C. (1955). A family of J-shaped frequency function. *Journal of the American Statistical Association*, **50**, 209–219.

Wilk, G. and Wlodarczyk, Z. (2000). Interpretation of the nonextensivity parameter q in some applications of Tsallis statistics and Lévy distributions. *Physical Review Letters*, **84**, 2770–2773.

Wilk, G. and Wlodarczyk, Z. (2001). Non-exponential decays and nonextensivity. *Physica A*, **290**, 55–58.

Wood, A. (1996). Predicting software reliability. *IEEE Transactions on Software Engineering*, **22**, 69–77.

Zografos, K. and Balakrishnan, N. (2009). On families of beta and generalized gamma-generated distributions and associated inference. *Statistical Methodology*, **6**, 344–362.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 21, No. 2, 2023 (New Series), pp 299–316
http://www.ssca.org.in/journal

Modeling Bivariate Survival Data By Compound Frailty Distributions

David D. Hanagal[1] and Alok D. Dabade[2]
[1]*Department of Statistics, Savitribai Phule Pune University, Pune, India*
[2]*Department of Statistics, University of Mumbai, Mumbai, India*

Received: 14 March 2023; Revised: 08 April 2023; Accepted: 17 April 2023

Abstract

Share frailty models are often used to model heterogeneity in survival analysis. In these models, it is assumed that each individual from a group shares common frailty, but sometimes it may be possible that some individuals will have zero susceptibility to an event. In such cases, compound distributions are more proper to model shared frailty than usually preferred distributions, gamma, lognormal etc. In this paper we have considered compound Poisson and compound negative binomial frailty distributions with IDB as baseline distribution. Since it has increasing, decreasing, constant and bathtub shaped hazard function. MCMC approach have been used to estimate the parameters involved in the models. A real life data analysis is also considered by applying the proposed models.. . .

Key words: Bayesian model comparison; Compound negative binomial distribution; Compound Poisson distribution; IDB distribution; MCMC; Shared frailty.

AMS Subject Classifications: 62F15, 62N01, 62P10

1. Introduction

In survival data, researchers are interested to study effect of covariates on life times of individuals from a group. For example, medical practitioner in case of lung cancer patients, may be interested to study how the factors such as age, health condition of the patient and the type of tumor may affect the survival times. In experiments on the time to failure of electrical insulation, engineer is interested to find the effect of the voltage, the insulation is subject to. Also in clinical trials, the experimenter is interested to study effect of the treatment assigned to a patient on the survival time. Unfortunately, many of the times it is impossible to include all relevant covariates. May be because, we have little or no information on the individual level. For example, it is known that excretion of small amounts of albumin in the urine is a diagnostic marker for increased mortality, however we are unable to include this variable, unless we actually obtain urine and analyze samples for each individual under study. Furthermore, we may not aware the relevance of the risk factor or even that the factor we ought to include in the analysis. For example, a genetic factor as we do not know all

Corresponding Author: David D. Hanagal
Email: david.hanagal@gmail.com

possible genes having influence on survival. In other cases, it may be impossible to measure the risk factor without great financial cost or time effort. In such cases, the usual practice is to ignore such covariates. The neglect of such covariates leads to heterogeneity into the data. This heterogeneity is named as frailty by Vapuel *et al.* (1979). To address the frailty, it is necessary to include random effect term into the model. Such models are well known as frailty models.

Sometimes individuals from a group share a common frailty, for example, if we consider data on twins then for monozygotic twins, sex, any other genetically based covariates, date of birth and pre-birth environment is common. For the timings of failures of several paired human organs like kidneys, lungs, eyes, ears etc. shares common frailty because they are of same individual. In case of sequences of times of asthmatic attacks of asthma patients or in tumor diagnosis, tumor recurrence times in individual patients also has common frailty because occurrence time of an event is on same individual. In industrial applications, if we consider the breakdown times of dual generator in a power plant or failure times of two engines in a two engines airplane then common environment is shared by both the engines and generators. In such situations, shared frailty models are suggested in the literature (see Clayton (1978)).

Hanagal (2005) proposed a positive stable frailty model with bivariate exponential of Marshall-Olkin (1967) as baseline distribution. Hanagal (2006) discussed the gamma frailty regression model in the bivariate survival data and Hanagal (2007) also presented the gamma frailty regression models in the mixture distributions. Hanagal and Sharma (2013, 2015a, 2015b, 2015c) analyzed diabetic retinopathy data, acute leukaemia data and kidney infection data using shared gamma and inverse Gaussian frailty models.

In shared frailty models, it is assumed that, each individual from a group experiences an event of interest but sometimes it may be possible that some individuals are immune to a particular event *i.e.*, they are non-susceptible or they have zero susceptibility. For example, some cancer patients survive their cancer. In medicine, there are several examples of diseases primarily attacking people with particular susceptibility, for instance, a genetic kind, other people having virtually zero susceptibility of getting the disease. Another example is fertility, some couples are unable to conceive children so that the time to have first child birth for them have zero susceptibility. In case of marriages, some people never marry, some marriages are not prone to dissolve so that time to divorce for such couples have zero susceptibility. In such type of data, compound distribution having some positive mass at zero value can be a suitable choice. For example, compound Poisson distribution or compound negative binomial distribution.

Aalen (1992) considered a compound Poisson distribution as a mixture distribution in survival analysis. Also, Moger and Aalen (2005), Hanagal (2010a), Hanagal (2010b), Hanagal and Dabade (2012) and Hanagal and Kamble (2015) have considered compound Poisson frailty models. Hanagal and Dabade (2013) and Hanagal and Kamble (2016) have introduced compound negative binomial shared frailty model. Recently Hanagal (2023a, 2024a, 2024b) introduced compound Poisson frailty models based on additive hazard, correlated compound Poisson frailty models based on the hazard rate and reversed hazard rates to analyze kidney infection data and Australian twin data. Hanagal (2023b) proposed correlated compound geometric frailty models to analyze kidney infection data. More details on compound Poisson

frailty models are available in Hanagal (2011, 2019).

A random variable Z following a compound distribution is defined as,

$$Z = \begin{cases} Y_1 + Y_2 + \cdots + Y_N & ; \ N > 0 \\ 0 & ; \ N = 0. \end{cases} \tag{1}$$

where N is also random variable with some statistical distribution and $Y_1, Y_2, \cdots, Y_N$ are independent, identically gamma distributed random variables with scale parameter ν and shape parameter γ having density function,

$$f(y) = \begin{cases} \frac{\nu^\gamma}{\Gamma(\gamma)} y^{\gamma-1} e^{-\nu y} & ; \ y > 0, \nu > 0, \gamma > 0 \\ 0 & ; \ \text{otherwise.} \end{cases}$$

Here, variable Y_i represents length of i^{th} failure. If $N = 0$ frailty is not at all affecting the life times of an individual from a group and if $N > 0$ then frailty is cumulative effect of heterogeneity due to N failures.

Aalen and Tretli (1999) modelled testis cancer data using compound Poisson frailty model. A man receives damages during a critical period of their fetal development which may develop testis cancer after the hormonal process of puberty has started. The damage may be a result of the mother's exposure to environmental factors, for example an excessive estrogenic burden, and may also interact with genetic factors. Aalen and Tretli (1999) represented Y_i as size of the damage at i^{th} occasion and N be the number of damages occurred. Thus Z is now cumulative effect of damages occurred. Some other examples can be given as, in case of marriage data, Z may represents cumulative heterogeneity for not getting a perfect partner due to different unknown difficulties like, medical issues of an individual, hereditary problems etc. In case of fertility, Z may be cumulative effect due to different unknown reasons such as, effect of miss-carriages on health, male infertility, age related issues *etc.* However, Aalen and Tretli (1999) says, this point of view should not be taken too literally as a description of biological reality. The main reason for using compound frailty random variables is statistical convenience. Compound Poisson and compound negative binomial distribution both have simple and closed from expression of Laplace transform, which a requirement of any frailty model.

To complete the parametric form of the model we now make assumption on baseline distribution. Weibull distribution is one of the most widely used baseline distribution. Hazard function for Weibull distribution is a monotone function, which increases with time to infinity when shape parameter α is greater than one and it decreases up to the value zero for $\alpha < 1$. At time zero, it has a zero-failure rate implies that almost no failure will occur which is hardly feasible in real life. Also, other usually preferred baseline distributions such as, gamma, lognormal etc. has monotone hazard function. So, there is a need to have another baseline distribution which is feasible to model increasing, decreasing and bathtub shape hazard function. Hjort (1980) introduced Increasing, Decreasing, Constant and Bathtub-shaped failure rate distribution (IDB) which has all the above shapes. Also at time zero, failure rate is positive. So, we thought IDB distribution can be better than Weibull to model as baseline distribution.

For estimation of parameters of the model, we have considered MCMC technique. To check the performance of the model we have considered simulation study. Also, we have

applied the proposed models to a bivariate survival data set of McGrilchrist and Aisbett (1991) related to kidney infection and suggested the best model by using Bayesian model comparison techniques. The remainder of the paper is organized as follows, in Section 2, we provide introduction to general bivariate shared frailty model. In Section three, baseline distribution IDB is discussed. Section four and five respectively considers compound Poison and compound negative binomial shared frailty models. Section 6 is contributing to proposed models. In section 7 estimation procedure is discussed followed by simulation study and data analysis of kidney infection data in Section 8 and 9 respectively. Finally, paper concluded with Conclusion.

2. General bivariate shared frailty model

Suppose n individuals are observed for the study and let a bivariate random variable (T_{1j}, T_{2j}) be represent first and second survival time of j^{th} individual $(j = 1, 2, 3, \ldots, n)$. Also suppose that there are k observed covariates collected in a vector $\underline{X}_j = (X_{1j}, \ldots, X_{kj})$ for j^{th} individual where X_{aj} $(a = 1, 2, 3, \ldots, k)$ represent the value of a^{th} observed covariate for j^{th} individual. Here we assume that both the survival times for each individual share the same value of the covariates.

Let Z_j be represent shared frailty variable for j^{th} individual. Assuming that the frailties are acting multiplicatively on the baseline hazard function and both the survival times of individuals are conditionally independent for given frailty, the conditional hazard function and hence conditional survival function for j^{th} individual at i^{th} $(i = 1, 2)$ survival time $t_{ij} > 0$ for given frailty $Z_j = z_j$ has the form respectively,

$$h(t_{ij} \mid z_j, \underline{X}_j) = z_j h_0(t_{ij}) \eta_j \tag{2}$$

$$S(t_{ij} \mid z_j, \underline{X}_j) = e^{-z_j H_0(t_{ij}) \eta_j} \tag{3}$$

where $h_0(t_{ij})$ and $H_0(t_{ij})$ are respectively baseline hazard and cumulative baseline hazard functions at time $t_{ij} > 0$; $\eta_j = e^{\underline{X}_j \underline{\beta}}$ and $\underline{\beta}$ is a vector of order k, of regression coefficients. Under the assumption of independence, bivariate conditional survival function for given frailty $Z_j = z_j$ at time $t_{1j} > 0$ and $t_{2j} > 0$ is,

$$S(t_{1j}, t_{2j} \mid z_j, \underline{X}_j) = e^{-z_j (H_{01}(t_{1j}) + H_{02}(t_{2j})) \eta_j} \tag{4}$$

Unconditional bivariate survival function at time $t_{1j} > 0$ and $t_{2j} > 0$ is obtained by integrating over frailty variable Z_j having the probability function f(z_j), for j^{th} individual.

$$S(t_{1j}, t_{2j} \mid \underline{X}_j) = \int_{Z_j} S(t_{1j}, t_{2j} \mid z_j) f(z_j) dz_j = L_{Z_j}[(H_{01}(t_{1j}) + H_{02}(t_{2j}))\eta_j]$$

where $L_{Z_j}(.)$ is Laplace transform of frailty variable of Z_j for j^{th} individual. Thus, unconditional bivariate survival function for j^{th} individual at time $t_{1j} > 0$ and $t_{2j} > 0$ is,

$$S(t_{1j}, t_{2j} \mid \underline{X}_j) = L_{Z_j}[(H_{01}(t_{1j}) + H_{02}(t_{2j}))\eta_j] \tag{5}$$

Here onwards we represent $S(t_{1j}, t_{2j} \mid \underline{X}_j)$ as $S(t_{1j}, t_{2j})$.

Once we have unconditional survival function of bivariate random variable (T_{1j}, T_{2j}) we can obtain likelihood function and estimate the parameters of the model.

3. Baseline distribution

A continuous random variable T is said to follow three parameters Increasing, Decreasing, Constant and Bathtub-shaped (IDB) distribution if its survival function is given by,

$$S_0(t) = \begin{cases} \dfrac{e^{-\frac{\lambda t^2}{2}}}{(1+\alpha t)^{\frac{\theta}{\alpha}}} & ; \quad t>0, \alpha>0, \lambda>0, \theta>0 \\ 1 & ; \qquad \text{otherwise.} \end{cases} \tag{6}$$

Corresponding density function, hazard function and cumulative hazard function are respectively;

$$f_0(t) = \begin{cases} \dfrac{\theta+\lambda t\,(1+\alpha t)}{(1+\alpha t)^{1+\theta/\alpha}} \exp\left(-\frac{\lambda t^2}{2}\right) & ; \quad t>0, \alpha>0, \lambda>0, \theta>0 \\ 0 & ; \qquad \text{otherwise.} \end{cases} \tag{7}$$

$$h_0(t) = \begin{cases} \lambda t + \dfrac{\theta}{1+\alpha t} & ; \quad t>0, \alpha>0, \lambda>0, \theta>0 \\ 0 & ; \qquad \text{otherwise.} \end{cases} \tag{8}$$

$$H_0(t) = \begin{cases} \dfrac{\lambda t^2}{2} + \dfrac{\theta}{\alpha}\log(1+\alpha t) & ; \quad t>0, \alpha>0, \lambda>0, \theta>0 \\ 0 & ; \qquad \text{otherwise.} \end{cases} \tag{9}$$

It is easy to observe that, first term of hazard function increases and second term decreases with increase in time. So, if λ is 0 then hazard function is decreasing function and for $\theta = 0$ it is increasing in nature. From the difference between hazard function for two different time points $0 < t_1 < t_2$, $h_0(t_1) - h_0(t_2) = (t_2 - t_1)\left[\dfrac{\alpha\theta}{(1+\alpha t_1)(1+\alpha t_2)} - \lambda\right]$, we can observe that, for $\lambda \geq \alpha\theta$ hazard function is increasing function and for $0 < \lambda < \alpha\theta$ hazard function will have bathtub shape. For $\lambda = 0 = \alpha$ it has a constant hazard function.

4. Compound Poison shared frailty model

A random variable defined in (1) is said to follow compound Poisson distribution if N is Poisson distributed with mean ρ. The distribution of Z consists of two parts; a discrete part which corresponds to the probability of zero susceptibility, and a continuous part on the positive real line. The discrete part is, $P(Z=0) = e^{-\rho}$, which decreases as ρ increases and the distribution of the continuous part can be found by conditioning N and using the fact that the Y_i's are gamma distributed. It can be written as

$$f(z;\gamma,\nu,\rho) = \begin{cases} \frac{1}{z}e^{-(\rho+\nu z)} \sum\limits_{n=1}^{\infty} \dfrac{\rho^n(\nu z)^{n\gamma}}{\Gamma(n\gamma)n!} & ; \; z>0, \rho>0, \nu>0, \gamma>0 \\ 0 & ; \; \text{otherwise} \end{cases}$$

The parameter set for the compound Poisson distribution is $\rho > 0, \nu > 0, \gamma > 0$. The moments; mean, variance and Laplace transform of compound Poisson distribution are given by,

$$L_Z(s) = exp\left\{-\rho\left[1-\left(\frac{\nu}{\nu+s}\right)^{\gamma}\right]\right\} \tag{10}$$

$$E(Z) = \frac{\rho\gamma}{\nu}; \quad Var(Z) = \frac{\rho\gamma(\gamma+1)}{\nu^2}. \tag{11}$$

The shared frailty models are suffering from non-identifiability. To resolve the issue, as usual, we assume Z has expected value equal, which imposes the restriction $\nu = \rho\gamma$ on the parameters of compound Poisson distribution. Under the restriction Laplace transformation of compound Poisson distribution reduces to,

$$L_Z(s) = exp\left\{-\rho\left[1-\left(1+\frac{s}{\rho\gamma}\right)^{-\gamma}\right]\right\} \tag{12}$$

with variance $\dfrac{\gamma+1}{\rho\gamma}$. Replacing Laplace transformation in equation (5), we get the unconditional bivariate survival function for j^{th} individual at time $t_{1j} > 0$ and $t_{2j} > 0$ as,

$$S(t_{1j}, t_{2j}) = exp\left\{-\rho\left[1-\left(1+\frac{(H_{01}(t_{1j})+H_{02}(t_{2j}))\eta_j}{\rho\gamma}\right)^{-\gamma}\right]\right\} \tag{13}$$

Clayton (1978) defined a cross-ratio function given by,

$$\theta^*(t_1,t_2) = \frac{\lambda_1(t_1 \mid T_2 = t_2)}{\lambda_1(t_1 \mid T_2 > t_2)} = \frac{\lambda_2(t_2 \mid T_1 = t_1)}{\lambda_2(t_2 \mid T_1 > t_1)} = \frac{S(t_1,t_2)\dfrac{\partial^2 S(t_1,t_2)}{\partial t_1 \partial t_2}}{\dfrac{\partial S(t_1,t_2)}{\partial t_1}\dfrac{\partial S(t_1,t_2)}{\partial t_2}}$$

where $\lambda_1(.)$ and $\lambda_2(.)$ are conditional hazard functions of T_1 and T_2. It is an association function such that,

$$\theta^*(t_1,t_2)\begin{cases} > 1 & ;\text{positive association} \\ = 1 & ;\text{no association} \\ < 1 & ;\text{negative association} \end{cases}$$

For compound Poisson shared frailty model cross-ratio function is given by,

$$\theta^*(t_1,t_2) = 1 + \sigma^2\left[1+\frac{lnS(t_1,t_2)}{\rho}\right]^{-1} \tag{14}$$

It is easy to observe that, cross ratio function is greater than one and is a function of t_1, t_2. This implies there is always positive association between the survival times t_1 and t_2. Also, it is decreasing function of $t_1 > 0$, $t_2 > 0$ and decreases from $1 + \sigma^2$ to 1.

5. Compound Negative Binomial shared frailty model

A random variable of (1) is said to follow compound negative binomial distribution if N is negative binomial variate with parameters; the number of successes, r and the probability of success, p. The probability function of N is given by,

$$P(x) = \begin{cases} \begin{pmatrix} x+r-1 \\ x \end{pmatrix} p^r q^x & ; \ x = 0, 1, \cdots; 0 < p < 1; q = 1-p, r = 1, 2, \cdots \\ 0 & ; \text{ otherwise} \end{cases}$$

Discrete part of probability function of Z is, $P(Z = 0) = p^r$ and the continuous part is given by,

$$f(z) = \begin{cases} p^r \frac{1}{z} e^{-\nu z} \sum\limits_{N=1}^{\infty} \begin{pmatrix} N+r-1 \\ N \end{pmatrix} q^N \ \frac{(\nu z)^{N\gamma}}{\Gamma(N\gamma)} & ; \ z > 0, \nu > 0, \gamma > 0, 0 < p < 1; \\ & \quad q = 1-p, r = 1, 2, \cdots \\ 0 & ; \text{ otherwise.} \end{cases}$$

The parameter set for the compound negative binomial distribution is, $r = 1, 2, \cdots$; $0 < p < 1$; $\nu > 0$ and $\gamma > 0$. The Laplace transform, mean and variance of compound negative binomial variate are respectively given by,

$$L_Z(s) = \left\{ \frac{p}{1 - q\left[1 + \frac{s}{\nu}\right]^{-\gamma}} \right\}^r \tag{15}$$

$$E(Z) = \frac{rq\gamma}{p\nu}; Var(Z) = \frac{rq\gamma(p+\gamma)}{p^2\nu^2} \tag{16}$$

Under the identifiability condition, $EZ = 1$, the restriction on parameters is $\nu = \frac{rq\gamma}{p}$. Under this restriction, Laplace transform of compound negative binomial distribution reduces to,

$$L_Z(s) = \left\{ \frac{p}{1 - q\left[1 + d\frac{ps}{rq\gamma}\right]^{-\gamma}} \right\}^r \tag{17}$$

with variance $\sigma^2 = \frac{p+\gamma}{rq\gamma}$. Replacing Laplace transform in equation (5), we get the unconditional bivariate survival function for j^{th} individual at time $t_{1j} > 0$ and $t_{2j} > 0$ as,

$$S(t_{1j}, t_{2j}) = \left\{ \frac{p}{1 - q\left[1 + \frac{p(H_{01}(t_{1j}) + H_{02}(t_{2j}))\eta}{rq\gamma}\right]^{-\gamma}} \right\}^r \tag{18}$$

For negative binomial shared frailty model cross-ratio function is given by,

$$\theta^*(t_1,t_2) = 1 - \frac{1-(\gamma+1)\left[1-pS(t_1,t_2)^{-\frac{1}{r}}\right]^{-1}}{r\gamma}$$

We can easily observe that, cross-ratio function is always positive and decreasing function of t_1, t_2. It decreases between $1-\frac{1}{r\gamma}+\frac{\gamma+1}{rq\gamma}$ to $1-\frac{1}{r\gamma}$. This implies that there is always positive association between the survival times t_1 and t_2 and it decreases as time t_1, t_2 increases.

6. Proposed models

The unconditional bivariate survival functions for compound Poisson and compound negative binomial models at time $t_{1j} > 0$ and $t_{2j} > 0$ after substituting cumulative hazard function for IDB distribution in equations (13) and (18) are,

$$S(t_{1j},t_{2j}) = exp\left\{-\rho\left[1-\left(1+\frac{\phi(t_{1j},t_{2j})\eta_j}{\rho\gamma}\right)^{-\gamma}\right]\right\} \tag{19}$$

$$S(t_{1j},t_{2j}) = p^r\left[1-q\left(1+\frac{p\phi(t_{1j},t_{2j})\eta_j}{rq\gamma}\right)^{-\gamma}\right]^{-r} \tag{20}$$

where $\phi(t_{1j},t_{2j}) = \frac{\lambda_1 t_{1j}^2}{2}+\frac{\lambda_2 t_{2j}^2}{2}+\frac{\theta_1}{\alpha_1}\log(1+\alpha_1 t_{1j})+\frac{\theta_2}{\alpha_2}\log(1+\alpha_2 t_{2j})$. Here onwards we call equation (19) and (20) as model CP and CNB respectively.

7. Likelihood specification and bayesian estimation of parameters

Suppose there are n individuals under study, whose first and second observed failure times are represented by (t_{1j},t_{2j}). Let c_{1j} and c_{2j} be the observed censoring times for j^{th} individual $(j = 1,2,3,...,n)$ for first and second recurrence times respectively. Here we assume the independence between censoring scheme and life times of individuals.

The contribution of bivariate life time random variable of j^{th} individual in likelihood function is given by,

$$L_j(t_{1j},t_{2j}) = \begin{cases} f_1(t_{1j},t_{2j}), & ;\ t_{1j} < c_{1j}, t_{2j} < c_{2j}, \\ f_2(t_{1j},c_{2j}), & ;\ t_{1j} < c_{1j}, t_{2j} > c_{2j}, \\ f_3(c_{1j},t_{2j}), & ;\ t_{1j} > c_{1j}, t_{2j} < c_{2j}, \\ f_4(c_{1j},c_{2j}), & ;\ t_{1j} > c_{1j}, t_{2j} > c_{2j}. \end{cases}$$

and likelihood function is,

$$L(\underline{\theta},\underline{\beta},\underline{\tau}) = \prod_{j=1}^{n_1} f_1(t_{1j},t_{2j})\prod_{j=1}^{n_2} f_2(t_{1j},c_{2j})\prod_{j=1}^{n_3} f_3(c_{1j},t_{2j})\prod_{j=1}^{n_4} f_4(c_{1j},c_{2j}) \tag{21}$$

where $\underline{\tau}$, $\underline{\theta} = (\alpha_1, \lambda_1, \theta_1, \alpha_2, \lambda_2, \theta_2)$ and $\underline{\beta}$ are respectively vector of frailty parameters, vector of baseline parameters and vector of regression coefficients. In compound Poisson model $\underline{\tau} = (\rho, \gamma)$ and in compound negative binomial model $\underline{\tau} = (r, p, \gamma)$. Let n_1, n_2, n_3 and n_4 be the number of pairs for which first and second failure times (t_{1j}, t_{2j}) lie in the ranges $t_{1j} < c_{1j}, t_{2j} < c_{2j}$; $t_{1j} < c_{1j}, t_{2j} > c_{2j}$; $t_{1j} > c_{1j}, t_{2j} < c_{2j}$ and $t_{1j} > c_{1j}, t_{2j} > c_{2j}$ respectively and

$$\begin{aligned} f_1(t_{1j}, t_{2j}) &= \frac{\partial^2 S(t_{1j}, t_{2j})}{\partial t_{1j} \partial t_{2j}} \quad , \quad f_2(t_{1j}, c_{2j}) = -\frac{\partial^2 S(t_{1j}, c_{2j})}{\partial t_{1j}} \\ f_3(c_{1j}, t_{2j}) &= -\frac{\partial^2 S(c_{1j}, t_{2j})}{\partial t_{2j}} \quad , \quad f_4(c_{1j}, c_{2j}) = S(c_{1j}, c_{2j}) \end{aligned}$$

These functions for CP and CNB model respectively are given by,

CP model:

$$\begin{aligned} f_1(t_{1j}, t_{2j}) &= \left[\lambda_1 t_{1j} + \frac{\theta_1}{1+\alpha_1 t_{1j}}\right]\left[\lambda_2 t_{2j} + \frac{\theta_2}{1+\alpha_2 t_{2j}}\right]\left[1 + \frac{\phi(t_{1j}, t_{2j})\eta_j}{\rho\gamma}\right]^{-(\gamma+2)} \\ &\quad \left\{\frac{\gamma+1}{\rho\gamma} + \left[1 + \frac{\phi(t_{1j}, t_{2j})\eta_j}{\rho\gamma}\right]^{-\gamma}\right\} S(t_{1j}, t_{2j})\eta_j^2 \\ f_2(t_{1j}, c_{2j}) &= \left[\lambda_1 t_{1j} + \frac{\theta_1}{1+\alpha_1 t_{1j}}\right]\left[1 + \frac{\phi(t_{1j}, c_{2j})\eta_j}{\rho\gamma}\right]^{-(\gamma+1)} S(t_{1j}, t_{2j})\eta_j \\ f_3(c_{1j}, t_{2j}) &= \left[\lambda_2 t_{2j} + \frac{\theta_2}{1+\alpha_2 t_{2j}}\right]\left[1 + \frac{\phi(c_{1j}, t_{2j})\eta_j}{\rho\gamma}\right]^{-(\gamma+1)} S(t_{1j}, t_{2j})\eta_j \\ f_4(c_{1j}, c_{2j}) &= S(t_{1j}, t_{2j}) \end{aligned}$$

CNB model:

$$\begin{aligned} f_1(t_{1j}, t_{2j}) &= \frac{p^{r+2}\eta_j^2}{rq\gamma} \frac{\left[\lambda_1 t_{1j} + \frac{\theta_1}{1+\alpha_1 t_{1j}}\right]\left[\lambda_2 t_{2j} + \frac{\theta_2}{1+\alpha_2 t_{2j}}\right]\Phi_1(t_{1j}, t_{2j})}{\left[1 + \frac{p\phi(t_{1j}, t_{2j})\eta_j}{rq\gamma}\right]^{2(\gamma+1)}\left\{1 - q\left[1 + \frac{p\phi(t_{1j}, t_{2j})\eta_j}{rq\gamma}\right]^{-\gamma}\right\}^{r+2}} \\ f_2(t_{1j}, c_{2j}) &= p^{r+1}\eta_j \frac{\lambda_1 t_{1j} + \frac{\theta_1}{1+\alpha_1 t_{1j}}}{\left[1 + \frac{p\phi(t_{1j}, c_{2j})\eta_j}{rq\gamma}\right]^{(\gamma+1)}\left\{1 - q\left[1 + \frac{p\phi(t_{1j}, c_{2j})\eta_j}{rq\gamma}\right]^{-\gamma}\right\}^{r+1}} \\ f_3(c_{1j}, t_{2j}) &= p^{r+1}\eta_j \frac{\lambda_2 t_{2j} + \frac{\theta_2}{1+\alpha_2 t_{2j}}}{\left[1 + \frac{p\phi(c_{1j}, t_{2j})\eta_j}{rq\gamma}\right]^{(\gamma+1)}\left\{1 - q\left[1 + \frac{p\phi(c_{1j}, t_{2j})\eta_j}{rq\gamma}\right]^{-\gamma}\right\}^{r+1}} \\ f_4(c_{1j}, c_{2j}) &= p^r\left[1 - q\left(1 + \frac{p\phi(c_{1j}, c_{2j})\eta_j}{rq\gamma}\right)^{-\gamma}\right]^{-r} \end{aligned}$$

where $\Phi_1(t_{1j}, t_{2j}) = q\gamma(r+1) + (\gamma+1)\left[1 + \frac{p\phi(t_{1j}, t_{2j})\eta_j}{rq\gamma}\right]^{\gamma}\left\{1 - q\left[1 + \frac{p\phi(t_{1j}, t_{2j})\eta_j}{rq\gamma}\right]\right\}$

In our study, the likelihood function (21), due to censoring, is not in a simple form and so the first order derivatives. Hence, to estimate the parameters we have to use Newton-Raphson iterative procedure, but may be due to large number of parameters MLE's are not converging. So, we moved to computational Bayesian approach which does not suffer from these difficulties.

The joint posterior density function of parameters for given failure times is given by,

$$\pi(\alpha_1, \lambda_1, \theta_1, \alpha_2, \lambda_2, \theta_2, \underline{\tau}, \underline{\beta}) \propto L(\alpha_1, \lambda_1, \theta_1, \alpha_2, \lambda_2, \theta_2, \underline{\tau}, \underline{\beta}) * g_1(\alpha_1)g_2(\lambda_1)g_3(\theta_1)g_4(\alpha_2) g_5(\lambda_2)g_6(\theta_2)\prod_{i=1}^{f} h_i(\tau_i)\prod_{i=1}^{k} p_i(\beta_i)$$

where $g_i(.)$ $(i = 1, 2, \cdots, 6)$, $h_i(.)$ $(i = 1, 2, \cdots, f)$ and $p_i(.)$ $(i = 1, 2, \cdots, k)$ are prior density functions with known hyper parameters of corresponding arguments for baseline, frailty parameters and regression coefficients. Likelihood function $L(.)$ is given by equation (21). Here we assume that all the parameters are independently distributed.

A widely used prior for frailty parameter is the gamma distribution with mean one and large variance, $G(\phi, \phi)$, say with a small choice of ϕ and the prior for regression coefficient is the normal with mean zero and large variance say ϵ^2. Similar types of prior distributions were used in Ibrahim et al. (2001), Sahu *etal*. (1997) and Santos *et al*. (2010). So, in our study also we have used same noninformative prior for frailty parameters and regression coefficients. We have considered two different noninformative prior distributions for baseline parameters, one is $G(a_1, a_2)$ and another is $U(b_1, b_2)$. All the hyper-parameters $\phi, \epsilon^2, a_1, a_2, b_1$ and b_2 are known. Here $G(a_1, a_2)$ is gamma distribution with shape parameter a_1 and scale parameter a_2 and $U(b_1, b_2)$ represents uniform distribution over the interval b_1 to b_2. We set hyper-parameters $\phi = 0.0001, \epsilon^2 = 1000, a_1 = 1, a_2 = 0.0001, b_1 = 0$ and $b_2 = 100$.

We have fitted the Bayesian model with the above prior density functions and likelihood function (21) using Metropolis-Hastings algorithm. We have monitored convergence of Markov chain to a stationary distribution by Gelman-Rubin convergence statistic and Geweke test. Trace plots, coupling from the past plots and sample autocorrelation function plots have been used, to check the behaviour of the chain, to decide burn-in period and sample autocorrelation lag respectively.

In order to compare the proposed models, we have used Akaike Information criteria (AIC), Bayesian Information Criterion (BIC), Deviance Information Criteria (DIC) and Conditional Predictive Ordinate (CPO) plot (see Gelfand (1996)). Also, we have used the Bayes factor B_{uv} for comparison of the models M_u against M_v. To compute Bayes factor, we have considered MCMC approach given in Kass and Raftery (1995).

8. Simulation study

To evaluate the performance of the Bayesian estimation procedure we have carried out a simulation study. For the simulation purpose we have considered only one covariate X_1. It is assumed to follow normal distribution. As the Bayesian methods are time consuming, we

have generated only fifty pairs of life times. According to the assumption, for given frailty Z, life times of individuals are independent. So, the conditional survival function for an individual for given frailty $Z = z$ and a covariate X_1 at time $t > 0$ is,

$$S(t \mid z, X_1) = e^{-zH_0(t)\eta}$$

Equating $S(t \mid z, X_1)$ to a random number say R $(0 < R < 1)$ over $t > 0$ we get,

$$\psi(t) = \frac{\lambda t^2}{2} + \frac{\theta}{\alpha}\log(1 + \alpha t) + \frac{\log(R)}{z\eta} \tag{22}$$

It is not possible to express explicitly as function of t, so to generate life times we have used bisection method. Exact step-wise procedure to generate sample is:

1. Generate a random sample of size 50 from frailty distribution as shared frailty for j^{th} $(j = 1, 2, \cdots, 50)$ individual. Firstly, generate a random observation $N = n$ from Poisson distribution for CP model and from negative binomial for CNB model. If $n = 0$ then assign frailty $Z = 0$ and if $n > 0$ then generate n gamma variables X_i and assign $Z = \sum_{i=1}^{n} X_i$.

2. Generate 50 covariate values for X_1 from normal distribution and compute $\eta_j = e^{X_{1j}\beta_1}$ for j^{th} individual.

3. Generate 50 pairs of life times (t_{1j}, t_{2j}) for given frailty z_j obtained in step 1 by solving equation (22) using bisection method.

4. Generate censoring times c_{1j} and c_{2j} from exponential distribution and observe survival time for i^{th} time $t^*_{ij} = min(t_{ij}, c_{ij})$ and censoring indicator δ_{ij} for j^{th} individual $(i = 1, 2$ and $j = 1, 2, \ldots, 100)$, where

$$\delta_{ij} = \begin{cases} 1, & ; \ t_{ij} \leq c_{ij} \\ 0, & ; \ t_{ij} > c_{ij} \end{cases}$$

To estimate parameters of the model using simulated data, we have generated two parallel chains for both the models using two sets of prior distributions with the different starting points using Metropolis-Hastings algorithm based on normal transition kernels. We have iterated both the chains for 10000 times. There is no effect of prior distribution on posterior summaries because estimates of parameters are nearly same and convergence rate of chains for both the prior sets is also not greatly different. Also, for both the chains the results are somewhat similar, so we present here the analysis for only one chain with $G(a_1, a_2)$ as prior for baseline parameters, for both the models.

To check the effect of sample size of chain on the posterior summary, we have generated different samples and obtained posterior summary with small, moderate and large sample sizes. We have considered sample of size 7 as small, 16 as moderate and maximum possible sample size allowed by number of iterations and autocorrelation lag as large sample size. Gelman-Rubin convergence statistic values are nearly equal to one and Geweke test values are quite small and corresponding p-values are large enough to say the chain attains stationary

Table 1: Posterior summary for simulation study of CP model

Parameter	α_1	λ_1	θ_1	α_2	λ_2	θ_2	ρ	γ	β_1	$\|Bias\|$
True values	2.2	4.5	0.5	2.2	4.5	0.5	5	0.5	0.5	-
Sample size = 7;										
Estimates	2.0370	4.3470	0.5149	2.2999	4.4598	0.3534	4.5035	0.2865	0.6043	
Standard error	0.4637	0.3185	0.2442	0.4083	0.2336	0.2108	0.2601	0.1310	0.1080	
Bias	0.1630	0.1530	0.0149	0.0999	0.0402	0.1466	0.4965	0.2135	0.1043	0.6215
Sample size = 16;										
Estimates	2.2739	4.3950	0.4292	2.2676	4.5015	0.3621	4.5087	0.3225	0.6281	
Standard error	0.4414	0.3094	0.1980	0.4130	0.2098	0.2310	0.2846	0.1591	0.1283	
Bias	0.0739	0.1050	0.0708	0.0676	0.0015	0.1379	0.4913	0.1775	0.1281	0.5783
Sample size = 85;										
Estimates	2.2268	4.4535	0.4643	2.1980	4.5278	0.4838	4.7066	0.4804	0.5494	
Standard error	0.4192	0.2878	0.2458	0.3284	0.2035	0.2277	0.2659	0.2124	0.1494	
Bias	0.0268	0.0465	0.0357	0.0020	0.0278	0.0162	0.2934	0.0196	0.0494	0.3068

Table 2: Posterior summary for simulation study of CNB model

Parameter	α_1	λ_1	θ_1	α_2	λ_2	θ_2	p	γ	β_1	$\|Bias\|$
True values	2.2	4.5	0.5	2.2	4.5	0.5	0.5	0.5	0.5	-
Sample size = 7;										
Estimates	2.0312	4.6471	0.5990	2.1994	4.4922	0.4522	0.4821	0.8097	0.4954	
Standard error	0.4799	0.1739	0.2175	0.3917	0.2917	0.1978	0.0081	0.1558	0.0569	
Bias	0.1688	0.1471	0.0990	0.0006	0.0078	0.0478	0.0179	0.3097	0.0046	0.3982
Sample size = 16;										
Estimates	2.3001	4.5286	0.5976	2.2886	4.6000	0.4931	4.4826	0.7890	0.4963	
Standard error	0.4100	0.2626	0.1993	0.4110	0.3012	0.1229	0.0112	0.1358	0.0786	
Bias	0.1001	0.0286	0.0976	0.0886	0.1000	0.0069	0.0174	0.2890	0.0037	0.3494
Sample size = 85;										
Estimates	2.2169	4.4515	0.5234	2.1933	4.5115	0.4980	0.4827	0.7582	0.4923	
Standard error	0.4866	0.2707	0.1999	0.3694	0.2571	0.2114	0.0111	0.1457	0.0797	
Bias	0.0169	0.0484	0.0234	0.0067	0.0115	0.0020	0.0173	0.2582	0.0077	0.2653

distribution. Simulated values of parameters have autocorrelation of lag k, so every k^{th} iteration is selected as a sample from posterior distribution. The posterior mean and standard error with absolute bias for different sample sizes are reported in Table 1 and Table 2 for model CP and model CNB respectively. Last column of these Tables gives norm of bias which is calculated as $\sqrt{\sum_{i=1}^{n}(\text{true parameter}_i - \text{estimated value}_i)^2}$. From these Tables, it can be observed that the estimates become closer and closer to true values as sample size increases. Also, the standard error reduces as sample size increases.

9. Analysis of kidney infection data

We fit the proposed models to kidney infection data of McGrilchrist and Aisbett (1991). The data is related to recurrence times to infection at point of insertion of the catheter for 38 kidney patients using portable dialysis equipment. For each patient, first and second recurrence times (in days) of infection from the time of insertion of the catheter until it has to be removed owing to infection is recorded. The catheter may have to be removed for reasons other than kidney infection and this regard as censoring. So, survival time for a patient given

may be first or second infection time or censoring time. After the occurrence or censoring of the first infection sufficient (ten weeks interval) time was allowed for the infection to be cured before the second time the catheter was inserted. So, the first and second recurrence times are taken to be independent apart from the common frailty component. The data consists of three risk variables age, sex and disease type GN, AN and PKD where GN, AN and PKD are short forms of Glomerulo Nephritis, Acute Nephritis and Polycystic Kidney Disease. Let T_1 and T_2 be represents first and second recurrence time to infection. Five covariates age, sex and presence or absence of disease type GN, AN and PKD are represented by X_1, X_2, X_3, X_4, and X_5. To analyze kidney infection data, success is defined as getting infection first time, so we set $r = 1$.

First, we check goodness of fit of the data for both baseline distributions and then apply the Bayesian estimation procedure. To check goodness of fit for kidney data set, we have considered Kolmogorove-Smirnov test, we have applied the test to T_1 and T_2 separately. The p-values for CP and CNB models for T_1 are $0.9996, 0.4935$ and for T_2 are $0.5111, 0.3225$ respectively.

Table 3: Posterior summary for kidney infection data set for CP model

Parameter	Estimates	S.E.	L.C.L	U.C.L
n = 250, B = 1400, k = 390				
α_1	0.721066	0.125690	0.555642	0.962101
λ_1	0.000386	0.000361	0.000068	0.001358
θ_1	0.091341	0.047561	0.026447	0.213807
α_2	0.759945	0.116958	0.570522	0.982734
λ_2	0.000329	0.000300	0.000054	0.001263
θ_2	0.050038	0.028556	0.012666	0.128006
ρ	3.455383	0.805696	2.012911	4.910539
γ	2.440900	1.195852	1.032547	5.256461
β_1	0.007370	0.116010	-0.013778	0.029776
β_2	-1.885846	0.639941	-3.153641	-0.677762
β_3	0.168584	0.547786	-0.898598	1.244601
β_4	0.786868	0.544851	-0.298998	1.820400
β_5	-0.499750	0.980033	-2.549072	1.433960

Table 4: Posterior summary for kidney infection data set for CNB model

Parameter	Estimates	S.E.	L.C.L	U.C.L
n = 242, B = 2000, k = 390				
α_1	0.748220	0.130669	0.555915	0.982400
λ_1	0.000875	0.000742	0.000134	0.002706
θ_1	0.075700	0.053959	0.018796	0.217971
α_2	0.767938	0.124498	0.562309	0.979487
λ_2	0.000658	0.000502	0.000128	0.001859
θ_2	0.041461	0.024286	0.011405	0.099163
p	0.065639	0.021011	0.041000	0.118163
γ	0.496327	0.049441	0.406921	0.591675
β_1	0.009677	0.014753	-0.016473	0.040842
β_2	-2.368412	0.662620	-3.736769	-1.133544
β_3	0.221596	0.681551	-1.067240	1.394427
β_4	0.829265	0.645573	-0.526464	1.853952
β_5	-0.426339	1.044975	-2.423081	1.464310

As in case of simulation, here also we have got same conclusion. So, we present the analysis for only one chain with $G(a_1, a_2)$ as prior for baseline parameters, for both the models. In this case we iterated chains for 99000 times. The posterior summaries for CP and CNB models are presented in Table 3 and Table 4 respectively. In these Tables, second and third column represents estimate (posterior mean) and standard error whereas last two columns represent 95% lower and upper credible limits. The notations n, B and k respectively represent sample size, burn in period and auto-correlation lag.

Table 5: AIC, BIC and DIC values for kidney infection data set

Model	WOF	CP	CNB
AIC	712.3857	711.7692	709.3664
BIC	733.6743	732.7827	730.6550
DIC	708.4433	702.7835	698.9031

Table 5 provides AIC, BIC and DIC values for three models, CP, CNB and the model with ignoring frailty, which we call as without frailty (WOF) model. AIC and BIC values for CP and WOF models are nearly same, so cannot be used for comparing models, these values for CNB model are definitely smaller amongst other models. Further, if we rank DIC values from smallest to largest then CNB model will get first rank then CP and finally WOF model. This suggest that, CP and CNB models both are better than WOF model and CNB is better than CP.

Now consider comparison criteria $D_{uv} = 2\log(B_{uv})$ for comparing u^{th} numerator model against v^{th} denominator model, where B_{uv} is Bayes factor. Negative value of D_{uv} favours denominator model. These values are provided in Table 6.

Table 6: D_{uv} values for comparing CP and CNB models

		Numerator Model	
		WOF	CP
Denominator Model	CP	-0.7609	-
	CNB	-1.9194	-2.6804

From the Table 6 we can observe that, D_{uv} values for CP against WOF and CNB against WOF models are negative indicating CP and CNB models are better than WOF model. This is also confirmed with CPO plot presented in Figures 1 and 2. Large number of positive points in plot favour CP and CNB models. This implies if we ignore frailty then we may lose more informative model.

Thus, all the comparison criteria indicate that CNB model is better than CP model. We are now in a position to say that, both the proposed models, CP and CNB are more informative than ignoring frailty and CNB model is the best model then CP for modelling frailty in kidney infection data.

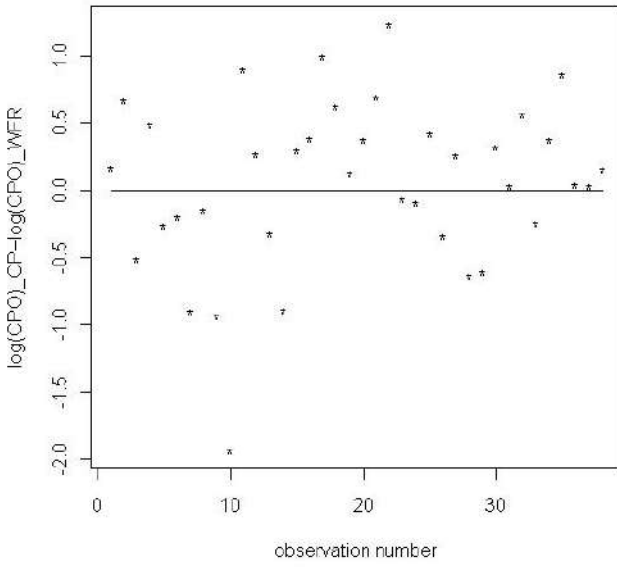


Figure 1: CPO plot for CP against WOF model

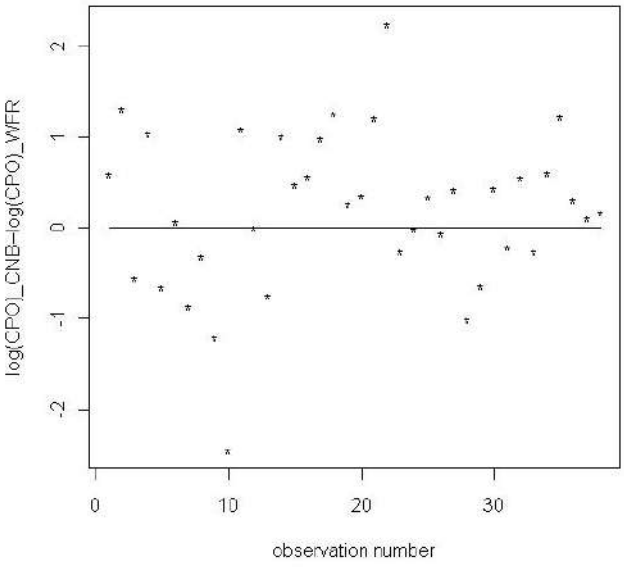


Figure 2: CPO plot for CNB against WOF model

10. Discussions

In the present paper, we have discussed compound Poisson and compound negative binomial shared frailty models. The main advantages of these models in comparison with other share frailty models is that they deal with the zero susceptibility. Further, the cross-ratio function is decreasing function of time unlike the other share frailty models, gamma and inverse Gaussian. Here we have considered IDB as baseline distribution. Even though it is an old distribution but it is more useful to model life times as it has increasing, decreasing, constant and bathtub shaped hazard function.

We have used Metropolis-Hastings algorithm to fit all the models. We analysed kidney infection data using our proposed models and the best model is suggested. We have used self-written programs in R statistical Environment to perform analysis.

The estimated frailty variances (0.4080) and (1.2118) for compound Poisson and compound negative binomial models respectively indicate that there is heterogeneity in the population of patients. Some patients are expected to be very prone to infection compared to others with the same covariate values. In continuation to this, all the model comparison criteria suggested that compound Poisson and compound negative binomial models are better than without frailty model. This indicates importance of frailty component in modelling of kidney infection data. Further comparing compound Poisson and compound negative binomial models, compound negative binomial shared frailty model is performing well for modelling of kidney infection data than compound Poisson model.

In compound negative binomial share frailty model, only one regression coefficient, β_2 is having larger ratio of its estimate to standard error and the value zero is not a credible value for the credible interval. This means, only covariate X_2 *i.e.*, Gender is significantly affecting

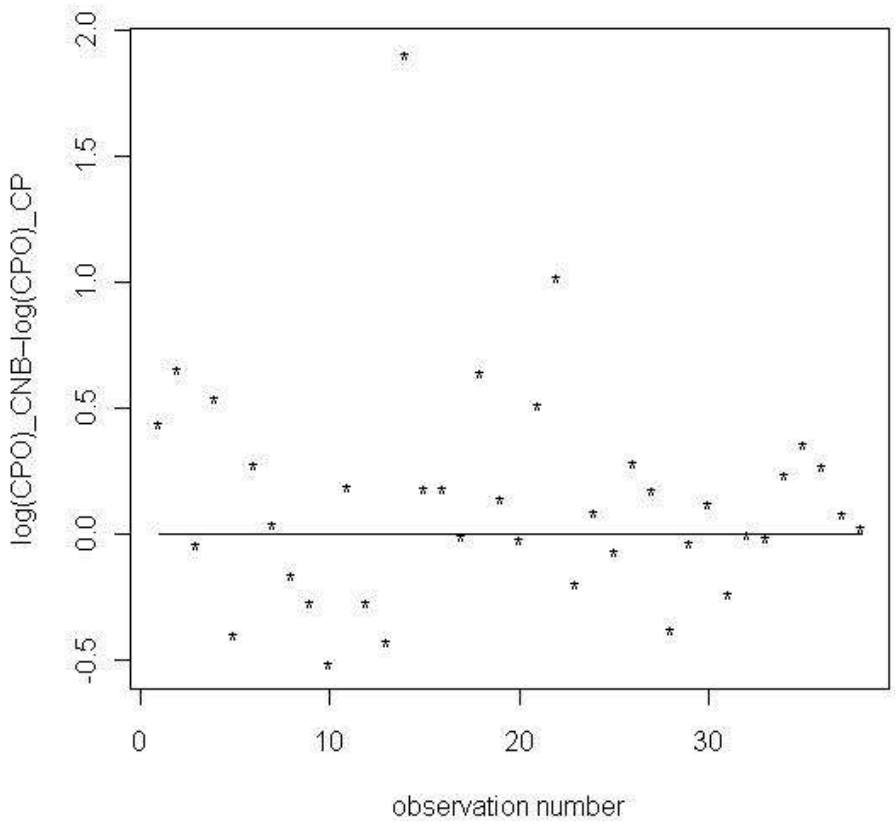


Figure 3: CPO plot for CNB against CP model

on infection rate. Negative value of β_2 indicate that the female patients have a lower risk for infection as compared to male patients. Same conclusion holds for compound Poisson share frailty models also. The estimated probability of non-susceptibility for compound negative binomial shared frailty model is 0.0656 indicating almost 6% of patients in the population are non-susceptible for kidney infection. In case of compound Poisson share frailty model, it is 3%.

In summary, this paper discussed modelling of survival times using compound frailty distributions when population consists of non-susceptible individuals.

Acknowledgements

We thank the referee for the valuable suggestions and comments which improved the earlier version the manuscript.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

Aalen, O. O. (1992). Modelling heterogeneity in survival analysis by the compound Poisson distribution. *Annals of Applied Probability*, **2**, 951-72.

Aalen, O. O. and Tretli, S. (1999). Analyzing incidence of tests cancer by means of a frailty model. *Cancer Causes Control*, **10**, 285-92.

Clayton, D. G. (1978). A model for association in bivariate life tables and its applications to epidemiological studies of familial tendency in chronic disease incidence. *Biometrika*,**65**,141-151.

Gelfand, A. E. (1996). Model determination using sampling based methods. *Markov Chain Monte Carlo in Practice*, Eds. Gilks, W.R., Richardson, S. and Spiegelhalter, D.J., Chapman and Hall, London, 145-161.

Hanagal, D. D. (2005). A positive stable frailty regression model in bivariate survival data. *Journal of the Indian Society for Probability and Statistics*, **9**, 35-44.

Hanagal, D. D. (2006). A gamma frailty regression model in bivariate survival data. *IAPQR Transactions*, **31**, 73-83.

Hanagal, D. D. (2007). Gamma frailty regression models in mixture distributions. *Economic Quality Control*, **22**, 295-302.

Hanagal, D. D. (2010a). Modelling heterogeneity for bivariate survival data by compound Poisson distribution. *Model Assisted Statistics and Applications*, **5**, 01-09.

Hanagal, D. D. (2010b). Modeling heterogeneity for bivariate survival data by the compound Poisson distribution with random scale. *Statistics and Probability Letters*, **80**, 1781-90.

Hanagal, D. D. (2011). *Modeling Survival Data Using Frailty Models, First Edition*, CRC Press, New York.

Hanagal, D. D. and Dabade, A. D. (2012). Modeling heterogeneity in bivariate survival data by compound Poisson distribution using Bayesian approach. *International Journal of Statistics and Management systems*, **7**, 36-84.

Hanagal, D. D. and Dabade, A. D. (2013). Compound negative binomial shared frailty models for bivariate survival data. *Statistics and Probability Letters*, **83**, 2507 -2515.

Hanagal, D. D. and Kamble, A. T. (2015). Bayesian estimation in shared compound Poisson frailty models. *Journal of Reliability and Statistical Studies*, **8**, 159-180.

Hanagal, D. D. and Kamble, A. T. (2016). Bayesian Estimation in Shared Compound Negative Binomial Frailty Models. *Research & Reviews: Journal of Statistics and Mathematical Sciences*, **2**, 53-67.

Hanagal, D. D. and Sharma, R. (2013). Modeling heterogeneity for bivariate survival data by shared gamma frailty regression model. *Model Assisted Statistics and Applications*, **8**, 85-102.

Hanagal, D. D. and Sharma, R. (2015a). Bayesian inference in Marshall-Olkin bivariate exponential shared gamma frailty regression model under random censoring. *Communications in Statistics, Theory & Methods*, **44**, 24-47.

Hanagal, D. D. and Sharma, R. (2015b). Comparison of frailty models for acute leukaemia data under Gompertz baseline distribution. *Communications in Statistics, Theory & Methods*, **44**, 1338-1350.

Hanagal, D. D. and Sharma, R. (2015c). Analysis of bivariate survival data using shared inverse Gaussian frailty model. *Communications in Statistics, Theory & Methods*, **44**, 1351-1380.

Hanagal, D. D. (2019). *Modeling Survival Data Using Frailty Models, Second Edition*, Springer Nature, Singapore.

Hanagal, D. D. (2023a). Compound Poisson shared frailty models based on additive hazards. *Communications in Statistics, Theory and Methods*, **52**, 6287-6309.

Hanagal, D. D. (2023b). Correlated compound geometric frailty models based on reversed hazard rate. *Model Assisted Statistics and Applications*, **18**, 149-164.

Hanagal, D. D. (2024a). Correlated compound Poisson frailty models based on reversed hazard rate. *Communications in Statistics, Theory and Methods*, (to appear). https://doi.org/10.1080/03610926.2022.2098336.

Hanagal, D. D. (2024b). Analysis of kidney infection data using correlated compound Poisson frailty models. *Model Assisted Statistics and Applications*, **19**, (to appear).

Hjorth, U. (1980). A reliability distribution with increasing, decreasing, constant and bathtub-shaped failure rates. *Technometrics*, **22**, 99-107.

Ibrahim, J. G, Ming-Hui Chen, and Sinha, D.(2001). *Bayesian Survival Analysis*. Springer Verlag.

Kass, R. E. and Raftery, A. E. (1995). Bayes Factor.*Journal of the American Statistical Association*, **90**, 773-795.

McGilchrist, C. A. and Aisbett, C. W. (1991). Regression with frailty in survival analysis. *Biometrics*, **47**, 461-466.

Moger, T. A. and Aalen, O. O. (2005). A distribution of multivariate frailty based on the compound Poisson distribution with random scale. *Lifetime Data analysis*, **11**, 41-59.

Sahu, S. K., Dey, D. K., Aslanidou, H., and Sinha, D.(1997). A Weibull regression model with gamma frailties for multivariate survival data. *Lifetime Data Analysis*, **3**, 123-137.

Santos, C. A. and Achcar, J. A.(2010). A Bayesian analysis for multivariate survival data in the presence of covariates. *Journal of Statistical Theory and Applications*, **9**, 233-253.

Vaupel, J. W., Manton, K. G., and Stallard, E. (1979). The impact of heterogeneity in individual frailty on the dynamics of mortality. *Demography*, **16**, 439-454.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 21, No. 2, 2023 (New Series), pp 317–325
http://www.ssca.org.in/journal

A Hybrid Regression Model for Cashew Nuts Price Prediction

Satyanarayana[1] **and Ismail B.**[2]
[1]*Department of Statistics, Mangalore University, Mangalagangothri, India*
[2]*Department of Statistics, Yenepoya (Deemed to be University), Mangalore, India*

Received: 24 June 2022; Revised: 14 January 2023; Accepted: 17 April 2023

Abstract

This paper proposes a hybrid regression model based on the regression tree and multiple linear regression model for improving prediction accuracy and to overcome one of the main disadvantages of the Regression Tree. The performance of the proposed model is compared with regression tree, K-nearest neighbor regression, multiple linear regression, and support vector regression through a Monte-Carlo simulation study. The simulation result indicates that the hybrid model outperforms all other regression models irrespective of sample size when the observations are from a normal distribution and uniform distribution. As an application, the proposed hybrid model is used to solve a problem faced by cashew nuts farmers and buyers to decide the most appropriate prices for the cashew nuts. The results from the hybrid model can be used as a guide by the farmers for fetching better prices in the market and by buyers for getting a lot of ascertained quality.

Key words: Cashew nuts price; Hybrid model; Regression tree; Support vector regression.

AMS Subject Classifications: 62J05

1. Introduction

Improving efficiency the prediction accuracy of regression model is still an interesting topic for researchers due to the natural variations in the systems themselves, which may drastically affect the model performance. In a traditional regression model, one must make assumptions about the functional form that connects the response variable with explanatory variable(s) which may not be valid. Most of the non-parametric regression techniques depend on the appropriate kernel or bandwidth selection and do not perform well in the case of high dimensional. CART (Classification and Regression Tree) is the most popular, efficient and widely used method for constructing decision trees introduced by Breiman *et al.* (1984). Shih (1997, 2004) observe that the splitting procedure in Regression Tree (RT) is biased as it searches for all possible splits and suggests that a proper normalization method will overcome this difficulty.

Corresponding Author: Satyanarayana
Email: sathya1301@gmail.com

The main disadvantages of RT are

a. RT assigns the same predicted value, average value, for all the tuples in a branch that satisfies the same corresponding splitting criterion.

b. Sometimes RT over fit the datasets, i.e., model completely fits the train data but fails to generalize for the test data.

To overcome the problem of over fitting, a sequence of values for threshold parameters are considered. According to cross-validation technique, the final threshold value is selected based on minimum prediction error criterion. Alternatively, one can also select the final threshold value by using the 1-standard error rule, which yields a prediction error of one standard deviation larger than the minimum error estimated by the cross-validation method. CART has several advantages over the traditional regression model.

The review paper of Domor *et al.* (2019) highlights the prediction performance of various decision tree algorithms. They also carried out an in-depth review of various methods used to improve the performance of the algorithms. Many papers have appeared in the direction of a hybrid modelling approach to improve prediction accuracy. Bennett *et al.* (1998) proposed a Support Vector Machine (SVM) approach to a decision tree to build a hybrid model. Kumar and Gopal (2010) hybrid SVM model-based decision tree and Chang and Liu's (2012) decision tree as an accelerator for SVM are the noticeable works in this direction. Muhamad Safiih Lola *et al.* (2016) proposed a hybrid model based on Artificial Neural Network and Multiple Linear Regression model (MLR). They showed that hybrid approach could improve the performance of Multiple Linear Regression model. Tanujit Chakraborty (2019) proposed a hybrid regression model based on Regression Tree and support vector regression for boiler water quality prediction. Regression Tree can model the arbitrary decision boundaries and found to be more robust algorithm. It has a built-in variable selection method and also it can handle missing values.

The proposed approach is similar to local linear regression using the bandwidth method. Here instead of computing bandwidth to fit regression line locally, the linear regression model is fit to each branch separately after arranging the observations according to splitting criterion. Since observations in each branch show high intra class similarity, the fitted model is expected to perform better than the linear model because the linear regression line is fitted globally. In the hybrid model, the strength of Regression Tree is used to improve the strength of the Multiple Linear Regression model. The proposed model can be used to select the best subset of regressors and for the prediction task. It has the advantages of significant accuracy and easy interpretability.

This work is motivated by a problem faced by cashew nuts buyers and the sellers to decide the most appropriate price for the cashew nuts. The price of cashew nuts can be decided from several quality measurements on the raw and kernel of the cashew nuts. The quality of cashew nuts brought to the market by the farmers varies considerably from lot to lot. In the case of farmers, if the quality of grown cashew nuts is good but due to lack of proper assessment about their quality, they may sell their whole lot for a lesser price. From the point of buyers, after offering a reasonable price for the raw cashew nuts, if the buyers do not get good quality kernels after de-shelling raw cashew nuts, it leads to massive losses because raw cashew nuts are purchased in a large number of lots. Also, the

process of producing kernels ready for marketing involves a large amount of human resources. Therefore, it is essential to develop a model which accesses the quality of the cashew nuts with minimal effort and decides optimal remunerative prices for the lot. The cashew nut plays a vital role in economic activities because the cultivation and marketing of cashew nuts involve a considerable amount of manpower in India. India is the largest producer of cashew nuts in the world. The problems associated with its cultivation, trading and marketing are that the growers do not reap optimal return and traders do not get reasonable profit.

2. Methodology

a. MLR method

Consider the multiple regression model $Y = \beta_0 X_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon$, where Y is an $n \times 1$ vector of the response variable, X_0 is a unit vector of size $n \times 1$ and $X_0, X_1, \ldots, X_k$ are regressors, $\beta_0, \beta_1, \ldots, \beta_k$ are unknown parameters and ε is an $n \times 1$ vector of error terms. The OLS estimator of β, the model parameter is given by $\hat{\beta} = (X^T X)^{-1} X^T Y$, where $X = [X_0, X_1, \ldots, X_k]$.

b. KNN method

This algorithm searches the pattern space for the K- training tuples closest to the unknown tuple. The closeness is defined in terms of distance between the tuples. It is better to normalize the values of each attribute before computing the closeness. For KNN, the unknown tuple is assigned the average value of its K- nearest neighbours as the predicted value.

c. SVR method

Support Vector Regression (Smola, 2002) is based on Statistical learning theory (Vapnik, 1995). Consider a linear regression model: $f(x) = w^T X + b$, where w is the weight of vector, b is the bias and X is the input feature vector. Then a solution that minimizes the error function is

$$f(x) = \sum_{i=1}^{n} (a_i^* - a_i) X^T X + b$$

where a_i^* and a_i are lagrange multipliers. The non-zero lagrange multipliers based on training vectors are called support vectors. The model for nonlinear case based on kernel function can be represented as:

$$f(x) = \sum_{i=1}^{n} (a_i^* - a_i) k(X^T X) + b$$

The Gaussian kernel is commonly used kernel function in $k(.)$ in SVR.

d. RT method

Even though the RT is an efficient method to produce outcomes, the main disadvantage of the RT is that it assigns the same average value of the response variable belonging to a particular group as the predicted value (constant) for all the observations in a group. Since

the original values of the response variable are not the same, even though RMSE is minimum, the predicted value of the response value is of great interest to the decision-making process. In RT, two branches are grown from each node N corresponding to the condition $X_i <$ splitting point and $X_i \geq$ splitting point respectively. The splitting variables and splitting point define the rectangles D_j as

$$D_1 = \{X \mid X_i < \text{splitting point}\} \quad i = 1 \text{ to } k$$
$$D_2 = \{X \mid X_i \geq \text{splitting point}\}$$

where X_i is the i-th predictor variable, k denotes number of predictor variables. Then predicted value at p-th node is given by

$$\hat{m}_p = \frac{\sum_h Y_h I\{h \in D_i\}}{|D_j|}$$

where $p = 1$ to m and $|D_j|$ denotes number of observations in p-th node. Thus, an estimate of $m(x)$ in D_j is simply the average response of the Y observations with predictor vector in D_j. The goal is to find that combination of splitting variables and splitting point, which leads to minimum residual sum of squares (RSS). In each node, fitted value of the response variable is constant, $\hat{m}_p$.

Proposed hybrid regression model (RT-MLR)

The formulation of proposed hybrid model is as follows: initially dataset splits into several branches based on the RT algorithm. Branching is depends on the splitting variables (significant variables) and best split point, which produces the minimum error. Using RT, the best subset of variables is selected and redundant features are eliminated. The dataset in each leaf node is arranged based on the position of tuples that satisfy the corresponding splitting criterion. Further, a MLR is built for each leaf node with significant variables. The model parameters are estimated using the least-squares method. Since observations within each group show high intra class similarities, the application of MLR in each group separately ensures that the estimated regression function fits well with the data. This hybrid model is easy, flexible and simplifies the work of selecting the best set of variables separately.

The workflow of the proposed model is as follows:

- Apply RT algorithm to train dataset to construct a RT which holds the split point, leaf node and significant variables.
- In each leaf node, datasets are arranged according to the positions of the observations, which satisfies the corresponding splitting criterion.
- Fit MLR model separately, obtain the fitted values and repeat this for all the leaf nodes.

Therefore, predicted values at p-th node is given by

$$\hat{Y}_{hp} = \hat{\alpha} + \hat{\beta} X_h \qquad h = 1 \text{ to } n_p$$

This model is comprises of two steps: significant variables selection using RT and applying MLR to each leaf node separately to get improved prediction results. Observe that in each

node, fitted values of the response variable are not constant. Since outputs of RT will be used in MLR, the proposed model performs better irrespective of problems such as missing values, noise and outliers. The proposed model can be used to identify the significant variables and causal parameters.

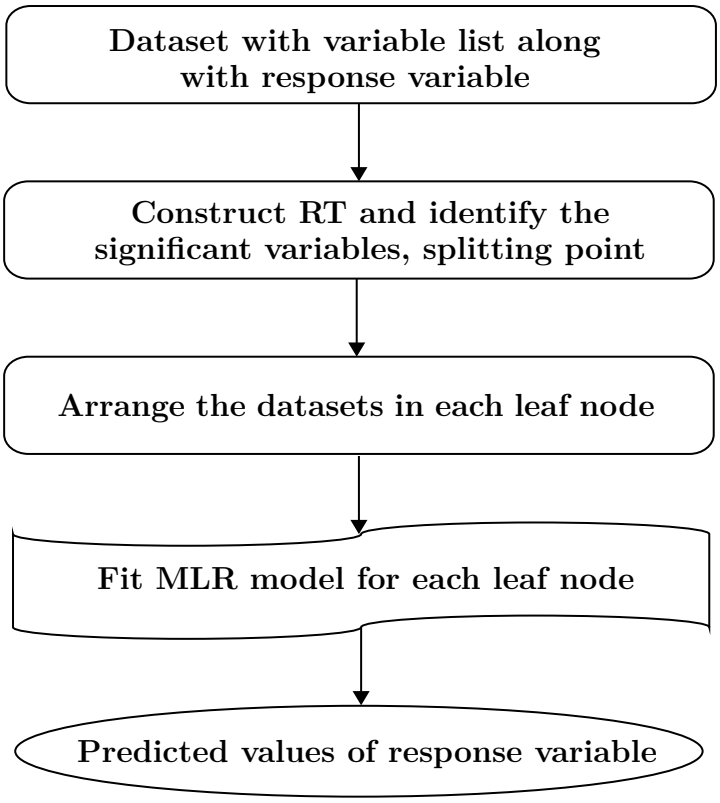


Figure 1: Flow chart of the proposed model

Performance measures

The model performance measure used in the simulation study and data analysis are

$$\begin{aligned}
\text{Root mean square error (RMSE)} &= \sqrt{\tfrac{1}{n}\textstyle\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2} \\
\text{Mean Absolute Error (MSE)} &= \tfrac{1}{n}\textstyle\sum_{i=1}^{n}|Y_i-\hat{Y}_i| \\
\text{Coefficient of determination } (R^2) &= 1-\left[\frac{\sum_{i=1}^{n}(Y_i-\hat{Y}_i)^2}{\sum_{i=1}^{n}(Y_i-\bar{Y}_i)^2}\right]
\end{aligned}$$

3. Simulation study

In this section, a simulation study is performed to highlight the distinction between proposed hybrid RT-MLR model, RT, KNN regression, MLR model and SVR model. The predictive performance of these models is compared in terms of RMSE and MAE. The simulation design is as follows:

1. Considered the linear regression model $Y = \beta_0 X_0 + \beta_1 X_1 + \cdots + \beta_k X_k + \varepsilon$, where all β's are set to 1 and X_1, X_2, X_3 are generated randomly from a standard normal distribution.

2. X_1, X_2, X_3 are also generated randomly from a uniform distribution $(0, 1)$ to check the robustness of the proposed model.

3. The error ϵ is generated from normal distribution with mean = 0 and variance = 5. The samples size used are $20, 50, 80, 100, 200, 500, 800, 1000, 2000, 3000$. The tree was grown to consist of three leaf nodes. The threshold stopping parameter for the RT is chosen as 0.01. For each scenario, 5000 repetitions were performed. In each simulation, the model is constructed using a train set and performance is evaluated using independently generated test data.

Table 1: Performance of different regression models for different sample size when the observations are from a standard normal distribution

Sample size	Method	RT	KNN	MLR	RT-MLR	SVR
20	RMSE	19.6	15.69	4.59	3.74	10.47
	MAE	15.65	12.19	3.75	3.15	8.3
50	RMSE	15.82	12.24	4.78	3.85	9.36
	MAE	12.4	9.38	3.83	3.15	7.45
80	RMSE	13.77	11.01	4.87	3.81	9.08
	MAE	10.72	8.44	3.9	3.12	7.26
100	RMSE	12.92	10.39	4.87	3.82	8.64
	MAE	10.05	7.98	3.9	3.12	6.9
200	RMSE	11.84	9.08	4.94	3.9	8.12
	MAE	9.27	7.02	3.94	3.18	6.47
500	RMSE	12.74	7.94	4.97	4.26	7.35
	MAE	10.04	6.18	3.97	3.45	5.62
800	RMSE	13.29	7.49	4.97	4.6	7.02
	MAE	10.47	5.86	3.97	3.71	5.45
1000	RMSE	13.59	7.33	5	4.71	6.78
	MAE	10.71	5.75	3.98	3.76	5.18
3000	RMSE	14.61	6.58	4.92	4.94	6.23
	MAE	11.52	5.24	3.91	4.04	4.84

From Table 1, the proposed hybrid RT-MLR model outperforms all other with a significant margin irrespective of sample size. The proposed model, along with overcoming the disadvantage of the regression tree, also performs better than all other models.

Robustness of the proposed hybrid RT-MLR model

To check the property of robustness of the proposed model about distributions, observations are generated from uniform distribution and results are summarised in Table 2.

Observe that proposed hybrid model outperforms all other models, irrespective of sample size.

Table 2: Performance of different regression model for different sample size when the observations are from uniform distribution

Sample size	Method	RT	KNN	MLR	RT-MLR	SVR
20	RMSE	7.92	9.22	6.14	4.97	7.76
	MAE	6.26	7.32	4.94	4.03	5.34
50	RMSE	7.26	8.83	6.67	5.04	8.74
	MAE	5.64	6.96	5.35	4.08	6.27
80	RMSE	7.02	8.74	6.8	5.06	9.12
	MAE	5.43	6.9	5.45	4.09	6.74
100	RMSE	6.88	8.65	6.83	5.13	9.26
	MAE	5.34	6.83	5.46	4.15	6.82
200	RMSE	6.85	8.51	6.93	5.08	9.56
	MAE	5.34	6.72	5.53	4.1	7.12
500	RMSE	7.27	8.39	6.98	5.44	9.85
	MAE	5.69	6.62	5.58	4.38	7.5
800	RMSE	7.51	8.31	6.98	5.71	9.89
	MAE	5.89	6.57	5.58	4.59	7.62
1000	RMSE	7.62	8.32	6.99	5.8	9.94
	MAE	5.97	6.56	5.58	4.66	7.69
3000	RMSE	8	8.24	7.01	6.28	10.1
	MAE	6.26	6.51	5.59	5.03	7.84

4. Real life application

The dataset used for the analysis consists of 96 observations and 12 variables related to the prices of cashew nuts collected from Dakshina Kannada. The variables considered are raw length, raw breadth, raw thickness, raw width, kernel length, kernel breadth, kernel thickness, kernel width, net count, sinkers count, moisture, out turn and price of the kernel. The price of the kernel is calculated based on the quality of the kernel obtained after de-shelling the raw cashew nuts. In this study price of the kernel is taken as the response variable. Initially, a sample of 5 kg was drawn from each lot at different spots and then by hand halving method, a final sample of 1kg was drawn from these samples. Similarly, 96 such representative samples were drawn. For each sample of 1 kg, measurements on 12 variables mentioned is recorded. The dataset is randomly split into training and testing data sets in a ratio of 70:30. Each experiment is repeated five times with randomly selected test sets and train sets. The average performance over 5-fold validation is reported in Table 3. The performance of different regression models RT, KNN, MLR, SVR and hybrid RT-MLR model were recorded. In the analysis, we used most of the default arguments present in the packages.

Table 3: Performance measures for different regression models on test set

Regression model	RMSE	R^2
RT	29.24	58.57
KNN	26.70	68.67
MLR	33.23	45.24
SVR	30.52	62.65
RT-MLR	14.49	88.15

Table 3 shows that proposed hybrid RT-MLR model outperforms all other regression models with a significant margin based on RMSE and R^2. Thus, the proposed model can be used as an effective tool to fetch the most appropriate price of cashew nuts.

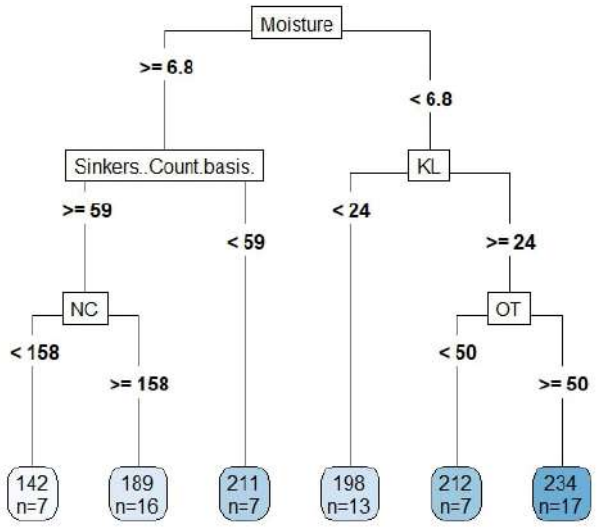


Figure 2: RT-MLR hybrid tree for cashew nuts price prediction

The above hybrid RT-MLR model suggests that the most important predictor variable for cashew nuts price is moisture content in the raw cashew nuts. Also proposed model inherently searches the interaction effects, as seen above. The interpretation of these effects is straightforward.

According to Figure 2, if Moisture $<$ **6.8** and Kernel length $\geq$ **24** $\Longrightarrow$ *High price* and if Moisture $\geq$ **6.8** and Sinkers count $<$ **59** $\Longrightarrow$ *High price*.

The proposed hybrid RT-MLR model is used to predict the cashew nut prices and to identify important casual variables and relationships. To get the optimal price for the cashew nuts, proposed model recommends checking the moisture level, kernel length and sinkers count and decide the price for the cashew nuts as shown above. Since out turn cannot be controlled at the time of purchase or selling, only controllable parameters are given based on the regression analysis performed using the hybrid model. The hybrid model along with improved accuracy, also helped the buyers and sellers to decide the most reasonable price for the cashew nut.

5. Conclusion

The main objective of this paper is to develop a hybrid model for improving prediction accuracy. This paper proposes a hybrid regression model based on the RT and MLR model. The proposed model also successfully overcomes one of the main disadvantage of the RT. The prediction performance of the proposed hybrid model is compared with many popular regression models through a simulation study. The simulation results indicate that the proposed hybrid model outperforms all other models when observations are generated from a normal distribution. The simulation results also demonstrate that the proposed hybrid model is fairly robust. The empirical study shows that hybrid model helped the cashew nuts buyers and farmers to decide the most appropriate price for the cashew nuts with improved prediction accuracy. The main advantage of this model is its easy interpretability. The proposed hybrid RT-MLR model can be used for handling both linear and non-linear datasets effectively. The proposed model approach is extended for classification problems as future work.

References

Bennett, K. P. and Blue, J. (1998). A support vector machine approach to decision trees, *In: 1998 IEEE International Joint Conference on Neural Networks Proceedings. IEEE World Congress on Computational Intelligence (Cat. No.98CH36227)*, **3**, 2396–2401.

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and Regression Trees*. CRC Press, Boca Raton.

Chakraborty, T., Chakraborty, A. K., and Mansoor, Z. (2019). A hybrid regression model for water quality prediction. *OPSEARCH*, **56**, 1167–1178.

Chakraborty, T., Chakraborty, A. K., and Chattopadhyay, S. (2019). A novel distribution-free hybrid regression model for manufacturing process efficiency improvement. *Journal of computational and Applied Mathematics*, **362**, 130–142.

Chakraborty, T., Chattopadhyay, S., and Chakraborty, A. K. (2018). A novel hybridization of classification trees and artificial neural networks for selection of students in a business school. *Opsearch*, **55**, 434–446

Chang, F. and Liu, C. C. (2012). Decision tree as an accelerator for support vector machines. *In: Ding, X. (ed.) Advances in Character Recognition. IntechOpen*, Rijeka.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, **20**, 273–297

Kumar, M. A. and Gopal, M. (2010). A hybrid SVM based decision tree. *Pattern Recognition*, **43**, 3977–3987.

Lola, M. S., Ramlee, M. N. A., Gunalan, G. S., Zainuddin, N. H., Zakariya, R., Idris, M. S., and Khalil, I. (2016). Improved the Prediction of Multiple Linear Regression Model Performance Using the Hybrid Approach: A Case Study of Chlorophyll-a at the Offshore Kuala Terengganu. *Open Journal of Statistics*, **6**, 789–804.

Mienyea, I. D., Suna, Y., and Wangb, Z. (2019). Prediction performance of improved decision tree-based algorithms: a review, *2nd International Conference on Sustainable Materials Processing and Manufacturing*.

Scholkopf, B. and Smola, A. J. (2002). *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, Massachusetts.

Publisher
Society of Statistics, Computer and Applications
Registered Office: I-1703, Chittaranjan Park, New Delhi- 110019, INDIA
Mailing Address: B-133, Ground Floor, Chittaranjan Park, New Delhi-110019, INDIA
Tele: 011 - 40517662
https://ssca.org.in/
statapp1999@gmail.com
2023

Printed by : Galaxy Studio & Graphics
Mob: +91 9818 35 2203, +91 9582 94 1203