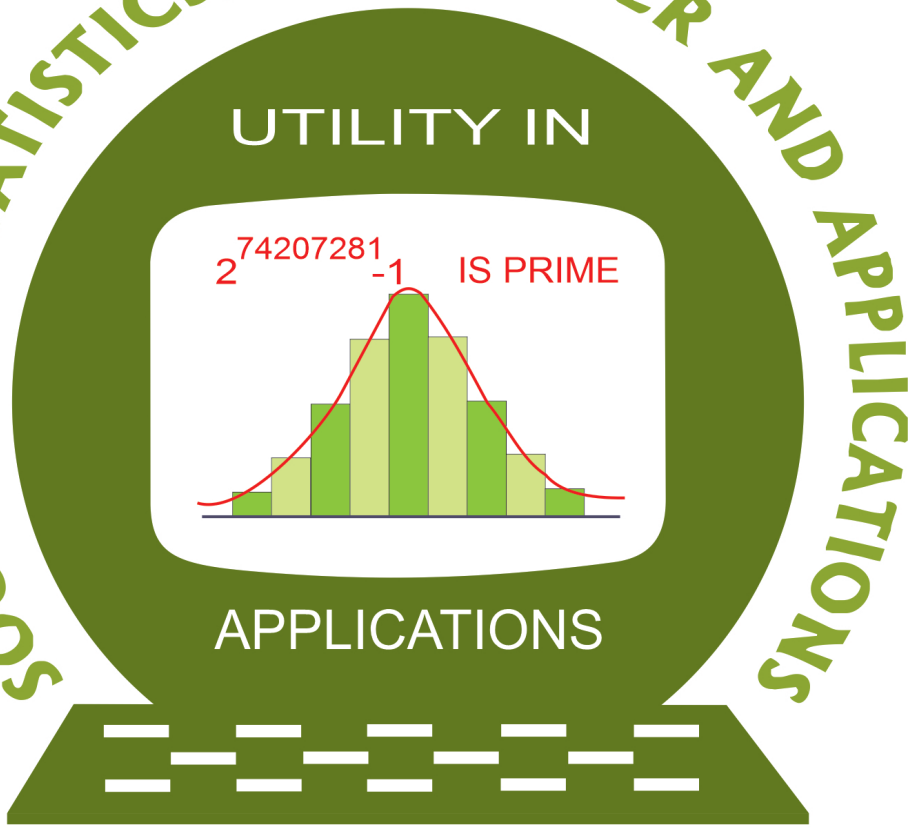
ISSN 2454-7395(online)

STATISTICS AND APPLICATIONS



FOUNDED 1998

Journal of the Society of Statistics, Computer and Applications
https://ssca.org.in/journal.html
Volume 23, No. 1, 2025 (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
K.J.S. Satyasai, Pankaj Mittal, Prithvi Yadav, R.B. Barman
R.C. Agrawal, Rahul Mukerjee, Rajpal Singh

Vice Presidents
A. Dhandapani, Manish Kumar Sharma, Manisha Pal, P. Venkatesan
Praggya Das, Ramana D. 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

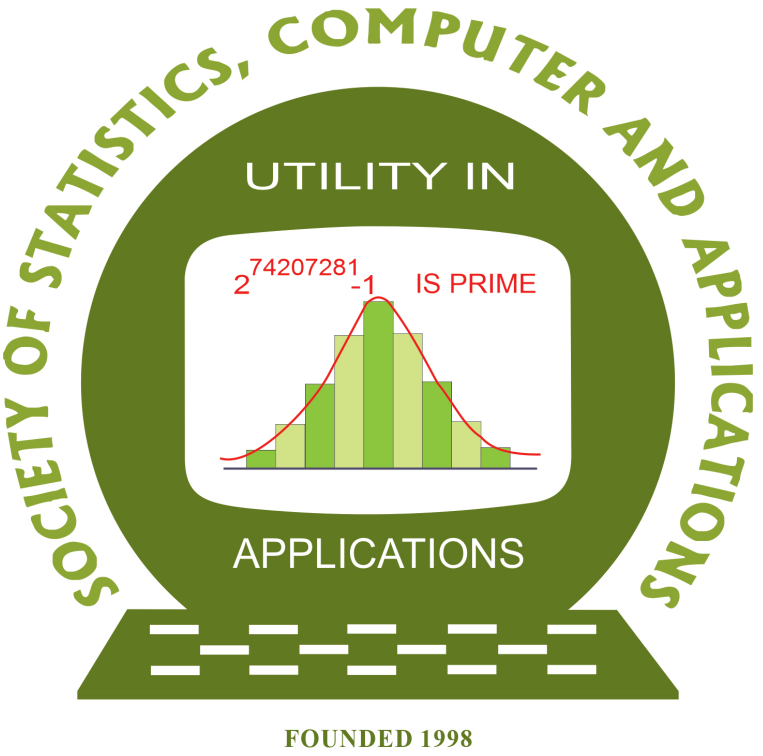
Council Members
B. Re. Victor Babu, Banti Kumar, Bishal Gurung, Imran Khan
Mukesh Kumar, Parmil Kumar, Piyush Kant Rai, Rajni Jain
Rakhi Singh, Raosaheb V. Latpate, Renu Kaul, Shalini Chandra
Sukanta Dash, V.M. Chako, Vishnu Vardhan R.

Ex-Officio Members (By Designation)
Director, ICAR-Indian Agricultural Statistics Research Institute, New Delhi
Chair Editor, Statistics and Applications
Executive Editors, 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 23, No. 1, 2025 (New Series)

Statistics and Applications

Volume 23, No. 1, 2025 (New Series)

Editorial Panel

Chair Editor

V.K. Gupta, Former ICAR National Professor at IASRI, Library Avenue, Pusa, New Delhi -110012; vkgupta_1751@yahoo.co.in

Executive Editors

Durba Bhattacharya, Head, Department of Statistics, St. Xavier's College (Autonomous), Kolkata - 700016; durba0904@gmail.com; durba@sxccal.edu

Rajender Parsad, ICAR-IASRI, Library Avenue, Pusa, New Delhi - 110012; rajender1066@yahoo.co.in; rajender.parsad@icar.gov.in

Editors

Baidya Nath Mandal, Managing Editor, ICAR-Indian Agricultural Research Institute Gauria Karma, Hazaribagh-825405, Jharkhand; mandal.stat@gmail.com

R. Vishnu Vardhan, Managing Editor, Department of Statistics, Pondicherry University, Puducherry - 605014; vrstatsguru@gmail.com

Jyoti Gangwani, Production Executive, Formerly at ICAR-IASRI, Library Avenue, New Delhi 110012; jyoti0264@yahoo.co.in

Associate Editors

Abhyuday Mandal, Professor and Undergraduate Coordinator, Department of Statistics, University of Georgia, Athens, GA 30602; amandal@stat.uga.edu

Ajay Gupta, Wireless Sensornets Laboratory, Western Michigan University, Kalamazoo, MI- 49008-5466, USA; ajay.gupta@wmich.edu

Anirban Chakraborty, School of Computational and Integrative Sciences and School of Sanskrit and Indic Studies, Jawaharlal Nehru University, New Delhi - 110067; anirban.chakraborty@gmail.com

Ashish Das, 210-C, Department of Mathematics, Indian Institute of Technology Bombay, Mumbai - 400076; ashish@math.iitb.ac.in; ashishdas.das@gmail.com

D.S. Yadav, Institute of Engineering and Technology, Department of Computer Science and Engineering, Lucknow- 226021; dsyadav@ietlucknow.ac.in

David Banks, Department of Statistical Science, Duke University, Durham, NC27708-0251 USA; david.banks@duke.edu

Deepayan Sarkar, Indian Statistical Institute, Delhi Centre, 7 SJS Sansanwal Marg, New Delhi - 110016; deepayan.sarkar@gmail.com; deepayan@isid.ac.in

Feng Shun Chai, Institute of Statistical Science, Academia Sinica, 128 Academia Road, Section 2, Nankang, Taipei -11529, Taiwan, R.O.C.; fschai@stat.sinica.edu.tw

Hanxiang Peng, Department of Mathematical Science, Purdue School of Science, Indiana University, Purdue University Indianapolis, LD224B USA; hpeng02@yahoo.com

Indranil Mukhopadhyay, Professor and Head, Human Genetics Unit, Indian Statistical Institute, Kolkata, India; indranilm100@gmail.com

J.P.S. Joorel, Director INFLIBNET, Centre Infocity, Gandhinagar -382007; jpsjoorel@gmail.com

Janet Godolphin, Department of Mathematics, University of Surrey, Guildford, GU2 7XH, UK; j.godolphin@surrey.ac.uk

Jiani Yin, Associate Director at Servier Biostatistics Project Lead Boston, MA, USA; jianiyin@gmail.com
Jyotirmoy Sarkar, Department of Mathematical Sciences, Indiana University Purdue University, Indianapolis, IN 46202-3216 USA; jsarkar@iupui.edu
K. Muralidharan, Professor, Department of Statistics, faculty of Science, Maharajah Sayajirao University of Baroda, Vadodara; lmv_murali@yahoo.com
K. Srinivasa Rao, Professor, Department of Statistics, Andhra University, Visakhapatnam, Andhra Pradesh; ksraoau@gmail.com
Katarzyna Filipiak, Institute of Mathematics, Poznañ University of Technology Poland; katarzyna.filipiak@put.poznan.pl
Lu Chen, NISS-NASS, USDA, USA, Research and Development Division, Sampling and Estimation Research Section; luchen459@gmail.com
M.N. Patel, Professor and Head, Department of Statistics, School of Sciences, Gujarat University, Ahmedabad - 380009; mnpatel.stat @gmail.com
M.R. Srinivasan, Department of Statistics, University of Madras, Chepauk, Chennai-600005; mrsrin8@gmail.com
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
Pranabendu Mishra, Computer Science Division, CMI, Chennai; pranabendu@cmi.ac.in
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
Rituparna Sen, Indian Statistical Institute Bengaluru, Karnataka 560059; ritupar.sen@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
Satyaki Mazumdar, Indian Science Education and Research Kolkata, Mohanpur, Nadia-741246, West Bengal; satyaki@iiserkol.ac.in
Saumyadipta Pyne, Health Analytics Network, and Department of Statistics and Applied Probability, University of California Santa Barbara, USA; spyne@ucsb.edu, SPYNE@pitt.edu
Shuvo Bakar, Faculty of Medicine and Health, University of Sydney, Australia; shuvo.bakar@sydney.edu.au
Snehanshu Saha, Professor, Computer Science and Information System, Head - APPCAIR 9 (All campuses), BITS Pilani K K Birla Goa Campus; snehanshus@goa.bits-pilani.ac.in
Snigdhansu Chatterjee, School of Statistics, University of Minnesota, Minneapolis, MN -55455, USA; chatt019@umn.edu
Sourish Das, Data Science Group, Chennai Mathematical Institute, Siruseri, Chennai 603103; sourish.das@gmail.com
Suman Guha, Department of Statistics, Presidency University, 86/1, College Street, Kolkata, 700073; bst0404@gmail.com

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
Utkarsh Tripathi, Solventum (3M Health Care), Pittsburgh Pennsylvania, USA; utkarshbitsp@gmail.com
V. Ramasubramanian, ICAR-NAARM, Rajendranagar, Hyderabad, Telangana – 500030; ram.vaidyanathan@gmail.com

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 23, No. 1, 2025 (New Series)
https://www.ssca.org.in/journal

CONTENTS

1.	R-optimal Designs for Logistic Regression Model in Two Variables *Mahesh Kumar Panda and Tofan Kumar Biswal*	1–12
2.	Estimation for the Length Biased Log-Logistic Model under Adaptive Progressive Type II Censoring *Ranjita Pandey, Pulkit Srivastava and Sweta Shukla*	13–32
3.	R-optimal Designs for Gamma Regression Model with Two Parameters *Mahesh Kumar Panda, Tofan Kumar Biswal and V. K. Gupta*	33–53
4.	Discrete Type I Half-Logistic Weibull Distribution and its Properties *M. Girish Babu and K. Jayakumar*	55–78
5.	Bulk Queuing Model with Reneging of Customers and their Retention *Shejal Gupta, Pradeep K. Joshi and K. N. Rajeshwari*	79–88
6.	Inference Techniques, Properties, and Applications of the T-Marshall-Olkin X Family of Distributions *Meenu Jose and Lishamol Tomy*	89–122
7.	Competing Risks Models with Multistate and Intermediate States *A. M. Rangoli, A. S. Talawar and R. P. Agadi*	123–137
8.	Extinction and Stationary Distribution of a Stochastic $SEII_aI_qHR$ Epidemic Model with Intervention *Tamalendu Das, Tridip Sardar and Sourav Rana*	139–175
9.	Sample Size Determination for Clinical Studies when Crisp Inputs are not available *Sai Sarada Vedururu, R.Vishnu Vardhan and K.V.S. Sarma*	177–187
10.	Construction of Order-of-Addition-Orthogonal Array Designs *Muhsina A., Baidya Nath Mandal, Rajender Parsad and Sukanta Dash*	189–196
11.	A Hybrid ARIMA-GARCH Type Copula Approach for Agricultural Price Forecasting *B. Manjunatha, Ranjit Kumar Paul, Ramasubramanian V., Amrit Kumar Paul, Md. Yeasin, Mrinmoy Ray, G. Avinash and Chandan Kumar Deb*	197–216

12.	Environmentally Responsible Index Tracking: Maintaining Performance while Reducing Carbon Footprint of the Portfolio *Lakshmi M. V., Soudeep Deb and Rituparna Sen*	217–223
13.	A Comprehensive Review of Data Science, Artificial Intelligence, and Big Data Analytics in Indian Official Statistics *Prasily P. and Manoharan M.*	225–247
14.	Three-State Markov Probability Distributions for the Stock Price Prediction *Tirupathi Rao Padi, Sarode Rekha, and Gulbadin Farooq Dar*	249–272
15.	An Economic Analysis of Two-Node Tandem Queue with Feedback *Ankita Roy Chowdhury and Indra*	273–281
16.	Bayesian Estimation of Scale Parameter of Inverted Kumaraswamy Distribution under Various Combinations of Different Priors and Loss Functions *Ableen Kaur, Parmil Kumar and Hemani Sharma*	283–301
17.	Determining Optimal Threshold and Some Inferential Procedures for a Skewed ROC Model in the Binary Classification Framework *Sandhya Singh, Saebugari Balaswamy and R.Vishnu Vardhan*	303–320
18.	Nonparametric Estimation of Extropy-Related Measures with Length-Biased Data *R. Dhanya Nair and E. I. Abdul Sathar*	321–334
19.	Mathematical Model for Spread of COVID-19 Virus using Fractional Order Approach *Gajanan S. Solanke and Deepak B. Pachpatte*	335–346
20.	Neutrosophic Marshall Olkin Extended Burr-XII Distribution: Theoretical Framework and Applications with Multiple Survival Time Data Sets *Shakila Bashir, Bushra Masood and Muhammad Aslam*	347–364
21.	Prediction of COVID-19 Disease Progression in India under the Effect of National Lockdown *Sourish Das*	365–383

SHORTER COMMUNICATION

22.	Understanding Fellegi Scheme for Sample Size Three *Yumnam Menon Singh and Opendra Salam*	385–388

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 23, No. 1, 2025 (New Series), pp 1–12
https://www.ssca.org.in/journal

R-optimal Designs for Logistic Regression Model in Two Variables

Mahesh Kumar Panda[1] **and Tofan Kumar Biswal**[2]
[1]*Department of Statistics, Ravenshaw University, Cuttack-753003*
[2]*Department of Statistics, Central University of Odisha, Sunabeda-763004*

Received: 24 June 2023; Revised: 08 February 2024; Accepted: 04 April 2024

Abstract

This article obtains locally R-optimal designs for a logistic regression model with two explanatory variables. The R-optimality criterion has been proposed in the literature as an alternative to the most frequently used D-optimality criterion when the experimenter wishes to minimize the volume of the confidence region for unknown parameters based on Bonferroni t-intervals. The necessary and sufficient conditions of this optimality criterion are confirmed through the equivalence theorem.

Key words: R-optimal design; Logistic regression model; D-optimality criterion; Bonferroni t-intervals; Equivalence theorem.

AMS Subject Classifications: 62K05

1. Introduction

The Generalized Linear Models (GLMs) are mostly used in those experiments where the responses are categorical type. These models are broadly applied in various types of studies when the experimenter wishes : (i) to estimate individual treatment effects in a multicenter clinical trial (*see* Lee and Nelder, 2002), (ii) to investigate the pattern of distribution of important tree species, and (iii) to identify the relationship between the risk of HIV (Human immunodeficiency virus) infection and the number of contacts with other partners and explanatory variables (*see* Jewell and Shiboski, 1992). McCullagh and Nelder (1989) have provided a detailed discussion on the analysis of data using GLMs and their application in different interdisciplinary areas.

The basic objective of finding an optimal design based on a certain criterion is to discuss statistical inference about the response of interest by selecting the control variable appropriately. The values of the control variables are chosen to minimize the variability of the estimators of the unknown parameters involved with the regression model. The pioneering work on optimal design was laid out by Kiefer (1959) and Kiefer and Wolfowitz (1959). The task of finding the optimal design for the GLM becomes quite challenging as the information matrix depends upon the unknown parameters *i.e.*, to find the best design to estimate the unknown parameters and yet one has to know the parameters to obtain the best design.

Corresponding Author: Mahesh Kumar Panda
Email: mahesh2123ster@gmail.com

Chernoff (1953) proposed an approach that targets obtaining a local optimal design for a best guess value of the parameter.

For a logistic model with two variables, Abdelbasit and Plackett (1983) established that a D-optimal k-point design is a 2-point design when k is even and a 3-point design when k is odd. Minkin (1987) modified the result of the D-optimal design by relaxing the various constraints imposed on the design space. Chaloner and Larntzin (1989) discussed Bayesian D-optimal designs for the logistic regression model. Using a geometric approach, Ford *et al.* (1992) obtained C-optimal and D-optimal designs for the discussed model. Sitter and Wu (1993) obtained D-, A-, and F-optimal designs for the logistic model, while Dette and Haines (1994) found E-optimal designs for the same model. Mathew and Sinha (2002) derived a unified approach of D-, A-, and E- optimal designs for binary data under the logistic model with two parameters. Woods *et al.* (2006), Dror and Steinberg (2006), and McGree and Eccleston (2008) reported optimal designs for two variable binary logistic models with interaction. These designs were constructed by using numerical methods. In this article, we obtain locally R-optimal designs for a logistic regression model with two explanatory variables. Dette (1997) proposed the R-optimality criterion in the literature as an alternative to the most frequently used D-optimality criterion. He recommended that an experimenter can prefer the R-optimality criterion in comparison to the D-optimality criterion when he/she wishes to minimize the volume of the confidence region for unknown parameters based on the Bonferroni t-intervals.

The rest of the article is organized as follows. Section 2 provides the preliminaries. In Section 3, we obtain R-optimal designs for the logistic model with two variables. In Section 4, we discuss the robustness of the proposed optimal design through a simulation study. Finally, the article is concluded with some discussion and conclusions in Section 5.

2. Preliminaries

Let us consider a binary response variable Y which follows a Bernoulli distribution and takes two values *i.e.* it takes value 1 for a success/positive response and 0 for a failure/negative response. If the response variable Y is related to the explanatory variables x_1 and x_2 through the two-variable binary logistic model, then the probability of success, p, can be expressed in terms of the logit

$$\mu = logit(p) = ln\frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 \tag{1}$$

where x_1 and x_2 are considered to be concentrations of the doses of two drugs with $x_1 \geq 0$, and $x_2 \geq 0$. In addition, the probability of a positive response is expected to increase with dose concentrations for both drugs, and thus β_1 and β_2 can be considered a strictly positive value [*see* Haines *et al.* (2007), and Haines *et al.* (2018)]. Due to practical considerations, the values of the parameter β_0 may be chosen as negative values in different experiments. Based on the scaled doses *i.e.* $z_1 = \beta_1 x_1$ and $z_2 = \beta_2 x_2$, the model Equations (1) can be expressed as

$$logit(p) = \beta_0 + z_1 + z_2 \qquad z_1 \geq 0 \ \textit{and} \ z_2 \geq 0. \tag{2}$$

Consider an approximate design ξ that assigns weights w_i on the distinct points $\boldsymbol{z_i} = (z_{1i}, z_{2i})$ for $i = 1, 2, \ldots, r$ is denoted by

$$\xi = \begin{Bmatrix} (z_{11}, z_{21}) & \ldots & (z_{1r}, z_{2r}) \\ w_1 & \ldots & w_r \end{Bmatrix}, \ \textit{where} \ 0 < w_i < 1 \ \textit{and} \ \sum_{i=1}^{r} w_i = 1.$$

The information matrix for the model Equation (2) based on the above design is given by

$$\boldsymbol{M}(\xi) = \sum_{i=1}^{r} w_i \boldsymbol{f}(\boldsymbol{z_i}) \boldsymbol{f'}(\boldsymbol{z_i}) \tag{3}$$

where

$$\boldsymbol{f}(\boldsymbol{z})\boldsymbol{f'}(\boldsymbol{z}) = k \begin{bmatrix} 1 & z_1 & z_2 \\ z_1 & z_1^2 & z_1 z_2 \\ z_2 & z_1 z_2 & z_2^2 \end{bmatrix}$$

with $k = \dfrac{e^{\mu}}{(1+e^{\mu})^2}$, $\boldsymbol{f}(\boldsymbol{z}) = \dfrac{e^{\mu/2}}{(1+e^{\mu})}(1, z_1, z_2)$ and $\mu = \beta_0 + z_1 + z_2$.

Selection of initial designs: To obtain the R-optimal design for the model Equation (2) we consider the support points of 3-point and 4-point D-optimal designs (*see* Haines, 2007) and define them as follows:

$$\text{3-point design} : \xi = \begin{Bmatrix} (0,0) & (\mu - \beta_0, 0) & (0, \mu - \beta_0) \\ 1 - w & \dfrac{w}{2} & \dfrac{w}{2} \end{Bmatrix} \tag{4}$$

$$\text{4-point design} : \xi_1 = \begin{Bmatrix} (-\mu - \beta_0, 0) & (0, -\mu - \beta_0) & (\mu - \beta_0, 0) & (0, \mu - \beta_0) \\ w & w & \dfrac{1}{2} - w & \dfrac{1}{2} - w \end{Bmatrix} \tag{5}$$

respectively. The support points of the design ξ_1 are having complimentary μ-values. These points are located on the boundary of the design space on lines of constant. Further, the weights allocated to these points are based on the symmetric position of the support points.

R-optimal design: A design $\xi^* \in \Omega$ with a non-singular matrix $\boldsymbol{M}(\xi^*)$ is called R-optimal for the model equation (3) if it minimizes

$$\Psi(\xi) = \prod_{i=1}^{q} (\boldsymbol{M}^{-1}(\xi))_{ii} = \prod_{i=1}^{q} \boldsymbol{e_i'} \boldsymbol{M}^{-1}(\xi) \boldsymbol{e_i} \tag{6}$$

for all $\xi \in \Omega$, where $\boldsymbol{e_i}$ denotes the ith unit vector in R^q where q is the number of parameters associated with the model Equation (2). The necessary and sufficient conditions for the R-optimality can be verified using the following equivalence theorem. For further details, one can refer to the article of Dette (1997).

Theorem 1: For model Equation (2) let

$$\varphi(\boldsymbol{z}, \xi) = \boldsymbol{f}(\boldsymbol{z}) \boldsymbol{M}^{-1}(\xi) \left(\sum_{i=1}^{q} \frac{\boldsymbol{e}_i \boldsymbol{e}_i'}{\boldsymbol{e}_i' \boldsymbol{M}^{-1}(\xi) \boldsymbol{e}_i} \right) \boldsymbol{M}^{-1}(\xi) \boldsymbol{f'}(\boldsymbol{z}). \tag{7}$$

A design $\xi^* \in \Omega$ is R-optimal if and only if

$$\sup_{z \in \Delta} \varphi(z, \xi^*) = q$$

with equality holds at the support points of ξ^*. Here Δ is the experimental region of interest.

3. R-optimal designs

In this section, we obtain locally R-optimal designs which minimize the product of the diagonal elements of the information matrix at best guesses of the unknown parameters β_0, β_1, and β_2.

3.1. Designs based on 3 points

Consider a 3-point design ξ of the form given by Equation (4) and assume that $\mu > \beta_0$ whenever $\beta_0 < 0$ and $\mu < \beta_0$ whenever $\beta_0 > 0$. Then we have the following theorem.

Theorem 2: The design ξ^* that assigns a weight of 0.2324 to the point $(\mu - \beta_0, 0)$, 0.5352 to the point (0, 0), and 0.2324 to the point $(0, \mu - \beta_0)$ in Δ is an R-optimal design where

$$\Delta = \{(z_1, z_2) : z_1 \geq 0, z_2 \geq 0, z_1 + z_2 \leq 3.7422\}.$$

Proof: The information matrix for the model Equation (2) at the three-point design ξ defined in Equation (4) is given by

$$\boldsymbol{M}(\xi) = \begin{bmatrix} \dfrac{e^{\mu}}{(1+e^{\mu})^2} & \dfrac{e^{\mu} w(\mu-\beta_0)}{2(1+e^{\mu})^2} & \dfrac{e^{\mu} w(\mu-\beta_0)}{2(1+e^{\mu})^2} \\ \dfrac{e^{\mu} w(\mu-\beta_0)}{2(1+e^{\mu})^2} & \dfrac{e^{\mu} w(\mu-\beta_0)^2}{2(1+e^{\mu})^2} & 0 \\ \dfrac{e^{\mu} w(\mu-\beta_0)}{2(1+e^{\mu})^2} & 0 & \dfrac{e^{\mu} w(\mu-\beta_0)^2}{2(1+e^{\mu})^2} \end{bmatrix}.$$

The inverse of the above information matrix is given by

$$\boldsymbol{M}^{-1}(\xi) = \begin{bmatrix} a & b & b \\ b & c & d \\ b & d & c \end{bmatrix} \tag{8}$$

with

$$a = \frac{-2(1+cosh(\mu))}{-1+w}, \quad c = \frac{2(-2+w)(1+cosh(\mu))}{(\beta_0-\mu)^2(-1+w)w},$$

$$b = \frac{-2(1+cosh(\mu))}{(\beta_0-\mu)(-1+w)}, \ \ and \ \ d = \frac{-2(1+cosh(\mu))}{(\beta_0-\mu)^2(-1+w)}.$$

Using Equation (8), we obtain the function

$$\Psi(\xi) = \frac{-8(-2+w)^2(1+cosh(\mu))^3}{(\beta_0-\mu)^4(-1+w)^3w^2}. \tag{9}$$

Next, we wish to minimize $\Psi(\xi)$ *w.r.t.* μ and w for that we obtain the partial derivatives of Equation (9) *w.r.t.* μ and w and set them equal to 0. Then we get

$$\frac{d}{d\mu}\Psi(\xi) = \frac{-8(-2+w)^2(1+cosh(\mu))^2(4+4cosh(\mu)+3(\beta_0-\mu)sinh(\mu))}{(\beta_0-\mu)^5(-1+w)^3w^2} = 0, \tag{10}$$

$$\frac{d}{dw}\Psi(\xi) = \frac{64(-2+w)(4+w(-10+3w))cosh\left(\frac{\mu}{2}\right)^6}{(3+\mu)^4(-1+w)^4w^3} = 0. \tag{11}$$

Here $cosh(\mu)$ and $sinh(\mu)$ are defined as the cosine and sine hyperbolic functions evaluated at μ. Next, Equation (10) leads to the following cases:

(i) $w = 2$,

(ii) $cosh(\mu) = -1$,

(iii) $4 + 4cosh(\mu) + 3(\beta_0 - \mu)sinh(\mu) = 0$,

and Equation (11) leads to the following cases:

(iv) $w = 2$,

(v) $cosh\left(\frac{\mu}{2}\right) = 0$,

and (vi) 4+w(-10+3w) = 0.

Out of these above-mentioned cases, the four cases *i.e.* (i), (ii), (iv), and (v) are the absurd cases. Therefore, we need to consider cases (iii) and (vi) only. Case (iii) implies

$$\begin{aligned} &4 + 4cosh(\mu) + 3(\beta_0 - \mu)sinh(\mu) = 0 \\ \Rightarrow &\beta_0 - \mu = \frac{-4}{3sinh(\mu)} - \frac{4coth(\mu)}{3} \\ \Rightarrow &\beta_0 - \mu = \frac{-4cosech(\mu)}{3} - \frac{4coth(\mu)}{3} \\ \Rightarrow &\beta_0 = \mu - \frac{4}{3}[cosech(\mu) - coth(\mu)], \end{aligned} \tag{12}$$

where the functions $cosech(\mu)$ and $coth(\mu)$ are the cosecant and cotangent hyperbolic functions evaluated at μ. Further, considering the first four terms of the Taylor series expansion of $cosech(\mu)$, and $coth(\mu)$ in Equation (12), we get the following

$$\begin{aligned} \beta_0 &= \mu - \frac{4}{3}\left[\frac{1}{\mu} - \frac{\mu}{6} + \frac{7\mu^3}{360} - \frac{31\mu^5}{15120} + \dots\right] - \frac{4}{3}\left[\frac{1}{\mu} + \frac{\mu}{3} - \frac{\mu^3}{45} + \frac{2\mu^5}{45} + \dots\right] \\ \Rightarrow \beta_0 &= \frac{703\mu^5}{11340} + \frac{\mu^3}{270} + \frac{7\mu}{9} - \frac{8}{3\mu}\,. \end{aligned} \tag{13}$$

Next considering case (vi), we get two values of w out of which one value is feasible *i.e.*, the optimal value of w denoted by

$$w^* = \frac{5-\sqrt{13}}{3} = 0.4648\,. \tag{14}$$

Here the optimal value μ should satisfy Equation (13). From the numerical solutions obtained for Equation (13), we see that there is a unique solution exists for all values $\mu > \beta_0$. Let us denote the solution by μ^*. As the solution can not be represented in an explicit form thus we provide the optimal values μ^* for some selected values of β_0 in Table 1.

The necessary and sufficient condition of the locally R-optimal design *i.e.* $\sup_{\boldsymbol{z}\in\Delta} \varphi(\boldsymbol{z},\xi^*) = q$ is confirmed by using the equivalence theorem which is as follows:

$$\begin{aligned}\varphi(\boldsymbol{z},\xi^*) =& k\Big\{a + bz_1 + bz_2 - \frac{(\beta_0-\mu)w(b+cz_1+dz_2)}{-2+w} \\ &- \frac{(\beta_0-\mu)w(b+dz_1+cz_2)}{-2+w} + z_2\Big(b + dz_1 + cz_2 \\ &- \frac{w(b+cz_1+dz_2)}{-2+w} + \frac{a+bz_1+bz_2}{\beta_0-\mu}\Big) + z_1\Big(b + cz_1 + dz_2 \\ &+ \frac{a+bz_1+bz_2}{\beta_0-\mu} - \frac{w(b+dz_1+cz_2)}{-2+w}\Big)\Big\}. \end{aligned} \tag{15}$$

Next, we provide the values of $\varphi(\boldsymbol{z},\xi^*)$ for some selected values of z_1 and z_2 in Table 2. We verify equivalence theorem for locally R-optimal design ξ^* by plotting a 3-dimensional plot of $\varphi(\boldsymbol{z},\xi^*)$ against $z_1 \geq 0$ and $z_2 \geq 0$ within the region Δ (*see* Figure 1). This proves Theorem 2. □

Table 1: Values of μ^* for selected β_0 for 3-point designs

β_0	-3	-2.5	-2	-1.5	-1	-0.5	0
μ^*	0.7422	0.8386	0.9543	1.0896	1.2392	1.3917	1.5355

Table 2: Values of $\varphi(\boldsymbol{z},\xi^*)$ for different values of z_1 and z_2

z_1	z_2	$\varphi(\boldsymbol{z},\xi^*)$
0	0	3
0	3.742231	3
3.742231	0	3
3.5	0.5	2.88382
3	1	2.34827
2.5	1.5	2.02694
1.87112	1.87112	1.5
2	1	0.774831
0.5	2	0.663202
1.25	1.45	0.494252

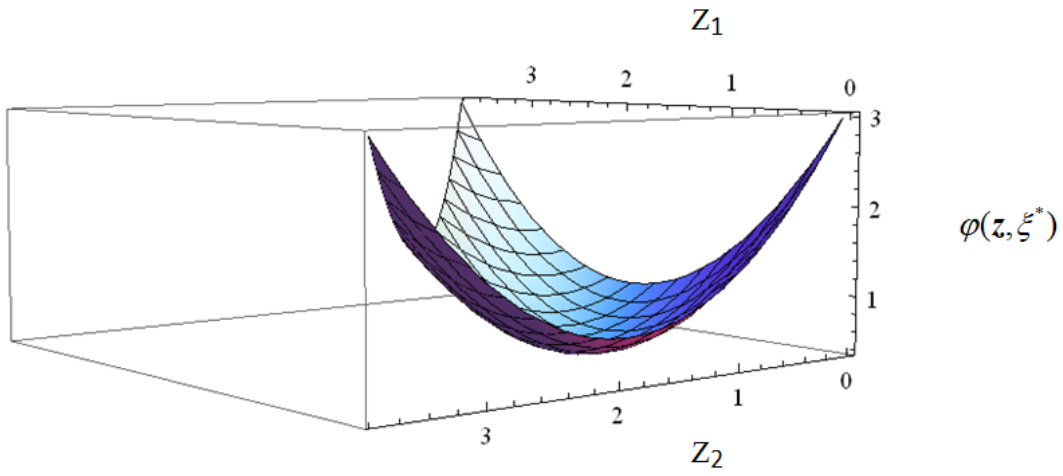


Figure 1: Plot of the $\varphi(\boldsymbol{z}, \xi^*)$ against z_1 and z_2

3.2. Designs based on 4 points

In this section, we consider a 4-point design ξ_1 of the form given by Equation (5) and assume that $0 \leq \mu \leq -\beta_0$. Then we have the following theorem.

Theorem 3: For the model Equation (2), there exists no mass-symmetric design of the form ξ_1 based on the four support points given by Equation (5).

Proof: The information matrix for the model Equation (2) at the four-point design ξ_1 is given by

$$\boldsymbol{M}(\xi_1) = \begin{bmatrix} M_{11} & M_{12} & M_{13} \\ M_{21} & M_{22} & M_{23} \\ M_{31} & M_{32} & M_{33} \end{bmatrix}$$

where

$$M_{11} = \frac{e^{\mu}}{(1+e^{\mu})^2},$$

$$M_{12} = M_{21} = M_{13} = M_{31} = \frac{e^{\mu}}{(1+e^{\mu})^2}\left\{\frac{\mu}{2} - \frac{\beta_0}{2} - 2\mu w\right\},$$

$$M_{22} = M_{33} = \frac{e^{\mu}}{(1+e^{\mu})^2}\left\{-2\mu^2 w - 2w\beta_0^2 + 4w\mu\beta_0 + \frac{\mu^2}{2} + \frac{\beta_0^2}{2} - \mu\beta_0\right\},$$

and $M_{23} = M_{32} = 0$.

The inverse of the above information matrix is

$$\boldsymbol{M}^{-1}(\xi_1) = \begin{bmatrix} M_{11}^{+} & M_{12}^{+} & M_{13}^{+} \\ M_{21}^{+} & M_{22}^{+} & M_{23}^{+} \\ M_{31}^{+} & M_{32}^{+} & M_{33}^{+} \end{bmatrix} \tag{16}$$

with

$$M_{11}^{+} = -\frac{((\beta_0-\mu)^2+8\beta_0\mu w)(1+cosh(\mu))}{4\mu^2 w(-1+2w)},$$

$$M_{12}^{+} = M_{21}^{+} = M_{13}^{+} = M_{31}^{+} = -\frac{(\beta_0+\mu(-1+4w))(1+cosh(\mu))}{4\mu^2 w(-1+2w)},$$

$$M_{22}^{+} = M_{33}^{+} = -\frac{e^{-\mu}(1+e^{\mu})^2((\beta_0-\mu)^2+8\mu(\beta_0-\mu)w-16\mu^2w^2)}{8\mu^2 w(-1+2w)((\beta_0-\mu)^2+8\beta_0\mu w)},$$

and $$M_{23}^{+} = M_{32}^{+} = -\frac{(\beta_0+\mu(-1+4w)^2)(1+cosh(\mu))}{4\mu^2 w(-1+2w)((\beta_0-\mu)^2+8\beta_0\mu w)}.$$

Using Equation (6), we obtain the function

$$\Psi(\xi_1) = \frac{e^{-3\mu}(1+e^{\mu})^6((\beta_0-\mu)^2+8\mu(\beta_0+\mu)w-16\mu^2w^2)^2}{512\mu^2w^2(-1+2w)^3((\beta_0-\mu)^2+8\beta_0\mu w)}. \tag{17}$$

Next, we wish to minimize $\Psi(\xi_1)$ *w.r.t.* μ and w for that we obtain the partial derivatives of Equation (17) *w.r.t.* μ and w and set them equal to 0. Here we also replace the functions $sinh(\mu/2)$ and $cosh(\mu/2)$ by the first three terms of their Taylor series expansion respectively. Then, we get

$$\frac{d}{d\mu}\Psi(\xi_1) = -\frac{\kappa_1(\mu,\beta_0,w)}{\kappa_2(\mu,\beta_0,w)} = 0\,, \tag{18}$$

$$\frac{d}{dw}\Psi(\xi_1) = -\frac{\lambda_1(\mu,\beta_0,w)}{\lambda_2(\mu,\beta_0,w)} = 0 \tag{19}$$

where

$$\begin{aligned}\kappa_1(\mu,\beta_0,w) =& \Big(e^{\frac{-5\mu}{2}}(1+e^{\mu})^5((-\beta_0-\mu)^2-8\mu(\beta_0+\mu)w+16\mu^2w^2)\\ &\Big(2((3\beta_0-2\mu)(\beta_0-\mu)^3+4(\beta_0-\mu)^2\mu(11\beta_0+4\mu)w\\ &+16\mu^2(9\beta_0^2+9\beta_0\mu-2\mu^2)w^2-192\beta_0\mu^3w^3)cosh(\frac{\mu}{2})\\ &+3\mu((\beta_0-\mu)^2+8\beta_0\mu w)(-(\beta_0-\mu)^2-8\mu(\beta_0+\mu)\\ &w+16\mu^2w^2)sinh(\frac{\mu}{2})\Big)\Big),\end{aligned}$$

$$\kappa_2(\mu,\beta_0,w) = 256\mu^7w^3(-1+2w)^3((\beta_0-\mu)^2+8\beta_0\mu w)^2,$$

$$\begin{aligned}\lambda_1(\mu,\beta_0,w) =& e^{-3\mu}(1+e^{\mu})^6((-\beta_0-\mu)^2+8\mu(\beta_0+\mu)w-16\mu^2w^2)\\ &(3\beta_0^4(-1+4w)-12\beta_0\mu^3(1-4w)^2(-1-2w+4w^2)\\ &+18\beta_0^2\mu^2(-1+4w)(1-4w+8w^2)+4\beta_0^3\mu(3\\ &+22w(-1+2w)+\mu^4(-3+4w(1+4(3-4w)w))),\end{aligned}$$

and $\lambda_2(\mu, \beta_0, w) = 512\mu^6(-1+2w)^4w^4((\beta_0-\mu)^2+8\beta_0\mu w)^2$.

Equation (17) leads to the following cases:

(a) $e^{\frac{-5\mu}{2}} = 0$,

(b) $1 + e^{\mu} = 0$,

(c) $-(\beta_0-\mu)^2 - 8\mu(\beta_0+\mu)w + 16\mu^2w^2 = 0$,

(d) $\Big(2((3\beta_0-2\mu)(\beta_0-\mu)^3 + 4(\beta_0-\mu)^2\mu(11\beta_0+4\mu)w + 16\mu^2(9\beta_0^2+9\beta_0\mu-2\mu^2)w^2 - 192\beta_0\mu^3w^3)cosh(\frac{\mu}{2}) + 3\mu((\beta_0-\mu)^2+8\beta_0\mu w)(-(\beta_0-\mu)^2-8\mu(\beta_0+\mu)w+16\mu^2w^2)sinh(\frac{\mu}{2})\Big) = 0$,

and Equation (18) leads to the following cases:

(e) $e^{-3\mu} = 0$,

(f) $(1+e^{\mu})^6 = 0$,

(g) $(-\beta_0-\mu)^2 + 8\mu(\beta_0+\mu)w - 16\mu^2w^2) = 0$,

(h) $(3\beta_0^4(-1+4w) - 12\beta_0\mu^3(1-4w)^2(-1-2w+4w^2) + 18\beta_0^2\mu^2(-1+4w)(-1-4w+8w^2) + 4\beta_0^3\mu(3+22w(-1+2w) + \mu^4(-3+4w(1+4(3-4w)w))) = 0$.

Out of the above-mentioned cases (a), (b), (e), and (f) are the absurd cases. Case (c) leads to two possible values of μ *i.e.*

$$\mu = \frac{-\beta_0 + 4\beta_0 w \pm 4\sqrt{-\beta_0^2 w + 2\beta_0^2 w^2}}{16w^2 - 8w - 1}\,. \tag{20}$$

However, the values given in Equation (19) will be real provided $w \geq 1/2$ which is again meaningless. Further, by solving the pair of Equations corresponding to cases (c) and (g) we get $w = 1/2$ which is not permissible. Next, we observe that the solutions of Equations corresponding to cases (d) and (h) (for different values of β_0) do not satisfy the restrictions $0 < w < 1/2$, and $0 < \mu < -\beta_0$. This indicates that there does not exist a four-point mass symmetric R-optimal design of the form ξ_1 for the model Equation (2). □

4. Robustness and simulation study

In this section, we examine the robustness of the proposed optimal design through a simulation study. First of all, we generate a sample of 50 observations of the unknown parameter β_0 from the $U(-10, 10)$ distribution and obtain the corresponding value of μ using Equation (13) by considering the assumptions about the parameter β_0 and μ as discussed in section 3.1. Next, for the pair of values of (β_0, μ) we find the supremum value of $\varphi(z, \xi^*)$

over the set Δ using Equation (2.7). The values of β_0, μ and $\sup_{z\in\Delta} \varphi(z, \xi^*)$ are shown in Table 3 (Appendix I). From Table 3, we observe that the value of $\sup_{z\in\Delta} \varphi(z, \xi^*)$ is equal to 3 and it exists at all the support points of optimal design ξ^* as defined in Equation (4). This shows that the necessary and sufficient condition of the locally R-optimal design *i.e.* the equivalence theorem is satisfied for different values of β_0. Thus, it can be concluded that the proposed optimum design is robust or insensitive toward variation in parameter values.

5. Discussion and conclusions

In the literature on the construction of optimal designs, the widely used optimality criterion is the D-optimality criterion. An experimenter decides to consider the D-optimality criterion when he/she is interested in the confidence ellipsoid of the estimators of the unknown parameters However, if the experimenter wishes to construct a rectangular confidence region then he/she should prefer an R-optimal design instead of a D-optimal design.

This present article obtains locally R-optimal designs for the logistic regression model in two variables subject to the constraint that the values of the variables are greater than or equal to zero. It is observed that the constructed designs depend upon the two unknown parameters through a scaled transformation of the explanatory variables whereas the intercept parameter β_0 provides the basic structure of the design.

Haines *et al.* (2018) have obtained D-optimal designs for the two-variable binary logistic regression model with interaction where the design points consist of an origin, two axial points, and a ray point, which lies within the design space that accommodates interaction. In this article, it is assumed that equal weights are assigned to each of the design points. An interesting research problem is to investigate locally R-optimal designs for the same model. For this purpose, the design points proposed by Haines *et al.* (2018) can be used. This shall be an interesting and challenging research problem as the weights assigned to each design point in the case of locally R-optimal designs may not be the same. We look forward to exploring this open problem in future research.

Acknowledgments

We thank the anonymous reviewers for their comments and suggestions that helped improve the manuscript.

Conflict of interest

The authors do not have any financial or non-financial conflict of interest to declare for the research work included in this article.

References

Abdelbasit, K. M. and Plackett, R. L. (1983). Experimental design for binary data. *Journal of the American Statistical Association*, **78**, 90-98.

Atkinson, A. C. and Haines, L. M. (1996). 14 designs for nonlinear and generalized linear models. *Handbook of Statistics*, **13**, 437-475.

Chaloner, K. and Larntz, K. (1989). Optimal Bayesian design applied to logistic regression experiments. *Journal of Statistical Planning and Inference*, **21**, 191-208.

Chernoff, H. (1953). Locally optimal designs for estimating parameters. *The Annals of Mathematical Statistics*, **24**, 586-602.

Dette, H. and Haines, L. M. (1994). E-optimal designs for linear and nonlinear models with two parameters. *Biometrika*, **81**, 739-754.

Dette, H. (1997). Designing experiments with respect to some 'standardized' optimality criteria. *Journal of Royal Statistical Society, Series B (Methodological)*, **59**, 97-110.

Dror, H. A. and Steinberg, D. M. (2006). Robust experimental design for multivariate generalized linear models. *Technometrics*, **48**, 520-529.

Ford, I., Torsney, B., and Wu, C. J. (1992). The use of a canonical form in the construction of locally optimal designs for non-linear problems. *Journal of the Royal Statistical Society, Series B (Methodological)*, **54**, 569-583.

Haines, L. M., Kabera, G., Ndlovu, P., and O'Brien, T. E. (2007). D-optimal designs for logistic regression in two variables. In *mODa 8-Advances in Model-Oriented Design and Analysis: Proceedings of the 8th International Workshop in Model-Oriented Design and Analysis held in Almagro, Spain, June 4–8, 2007* (pp. 91-98). Physica-Verlag HD.

Haines, L. M. and Kabera, G. M. (2018). D-optimal designs for the two-variable binary logistic regression model with interaction. *Journal of Statistical Planning and Inference*, **193**, 136-150.

He, L. and Yue, R. X. (2017). R-optimal designs for multi-factor models with heteroscedastic errors. *Metrika*, **80**, 717-732.

He, L. and Yue, R. X. (2019). R-optimality criterion for regression models with asymmetric errors. *Journal of Statistical Planning and Inference*, **199**, 318-326.

Kiefer, J. (1959). Optimum experimental designs. *Journal of the Royal Statistical Society, Series B (Methodological)*, **21**, 272-304.

Kiefer, J. and Wolfowitz, J. (1959). Optimal designs in regression problems. *Annals of Mathematical Statistics*, **30**, 271-294.

Lee, Y. and Nelder, J. A. (2002). Analysis of ulcer data using hierarchical generalized linear models. *Statistics in Medicine*, **21**, 191-202.

Liu, P., Gao, L. L., and Zhou, J. (2022). R-optimal designs for multi-response regression models with multi-factors. *Communications in Statistics-Theory and Methods*, **51**, 340-355.

Mathew, T. and Sinha, B. K. (2001). Optimal designs for binary data under logistic regression. *Journal of Statistical Planning and Inference*, **93**, 295-307.

McCullough, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman and Hall/CRC, 2nd ed.

McGree, J. M., and Eccleston, J. A. (2008). Probability-based optimal design. *Australian and New Zealand Journal of Statistics*, **50**, 13-28.

Minkin, S. (1987). Optimal designs for binary data. *Journal of the American Statistical Association*, **82**, 1098-1103.

Shiboski, S. C. and Jewell, N. P. (1992). Statistical analysis of the time dependence of HIV infectivity based on partner study data. *Journal of the American Statistical Association*, **87**, 360-372.

Silvey, S. D. (1980). *Optimal Design*. Chapman and Hall/CRC.

Sitter, R. R. and Wu, C. F. J. (1993). Optimal designs for binary response experiments: Fieller, D, and A criteria. *Scandinavian Journal of Statistics*, **20**, 329-341.

Woods, D. C., Lewis, S. M., Eccleston, J. A., and Russell, K. G. (2006). Designs for generalized linear models with several variables and model uncertainty. *Technometrics*, **48**, 284-292.

Appendix-I

Table 3: Values β_0, μ and $\sup_{z\in\Delta} \varphi(\boldsymbol{z}, \xi^*)$

S.N.	β_0	μ	$\sup_{z\in\Delta} \varphi(\boldsymbol{z}, \xi^*)$	S.N.	β_0	μ	$\sup_{z\in\Delta} \varphi(\boldsymbol{z}, \xi^*)$
1	-5.6013	0.448	3	26	6.5924	2.4696	3
2	-3.1882	0.7104	3	27	-8.9667	2.6341	3
3	-4.7036	0.5216	3	28	1.3853	1.854	3
4	-3.9974	0.5969	3	29	-5.1861	0.4795	3
5	-6.5366	0.3898	3	30	-1.4474	1.1048	3
6	-9.3569	0.2785	3	31	-1.7531	1.0188	3
7	8.5955	2.6108	3	32	-9.2929	0.2803	3
8	-9.9134	0.2635	3	33	-3.2952	0.6934	3
9	-4.1403	0.5801	3	34	-8.1109	0.319	3
10	-5.7407	0.4383	3	35	8.8737	2.6283	3
11	9.2823	2.6534	3	36	-0.41	1.4185	3
12	5.6955	2.3958	3	37	9.629	2.674	3
13	-4.7861	0.5139	3	38	9.036	2.6384	3
14	-0.4474	1.4074	3	39	7.3492	2.5263	3
15	7.1903	2.5148	3	40	-0.4009	1.4212	3
16	-2.9595	0.7493	3	41	-4.3566	0.5563	3
17	6.2333	2.441	3	42	-3.2532	0.7	3
18	-3.0336	0.7363	3	43	-8.1627	0.317	3
19	7.5542	2.5409	3	44	-0.1417	1.4961	3
20	-2.4755	0.8437	3	45	-0.0297	1.5273	3
21	1.822	1.9327	3	46	-1.8701	0.9876	3
22	6.2163	2.4396	3	47	-9.2231	0.2823	3
23	3.6757	2.1933	3	48	-7.3585	0.3494	3
24	1.4215	1.8608	3	49	-3.5446	0.6562	3
25	1.675	1.9071	3	50	-1.4147	1.1143	3

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 23, No. 1, 2025 (New Series), pp 13–32
https://www.ssca.org.in/journal

Estimation for the Length Biased Log-Logistic Model under Adaptive Progressive Type II Censoring

Ranjita Pandey[1], **Pulkit Srivastava**[1] **and Sweta Shukla**[2]
[1]*Department of Statistics, University of Delhi, Delhi, India*
[2]*Department of Mathematics, GLA University, Mathura, India*

Received: 22 November 2022; Revised: 07 December 2023; Accepted: 06 April 2024

Abstract

In this paper, point estimation of shape and scale parameters of length-biased log-logistic distribution under adaptive progressive type II censoring is addressed using Bayesian and non-Bayesian approaches. Maximum Likelihood estimators are proposed and evaluated using Newton-Raphson numerical approximation method. Asymptotic confidence interval and parametric bootstrap confidence intervals are also constructed. Parametric Bayes estimators are proposed under three different loss functions using Markov Chain Monte Carlo iterative method. Credible intervals and Highest Posterior Density region are also constructed. Simulation study for different sample sizes and different censoring schemes is carried out to establish utility of the proposed decision-theoretic strategies. A real dataset has also been analyzed to reinforce the simulated results.

Key words: Length-biased log-logistic distribution; Adaptive progressive type II censoring; Bootstrap confidence intervals; Highest Posterior Density region; Markov Chain Monte Carlo.

AMS Subject Classifications: 62F15, 65C05

1. Introduction

The main purpose of any censoring plan is to reduce time duration while optimizing the total experimental cost. A balanced censoring strategy might take into account the length of the experiment, the number of units involved, and the effectiveness of statistical inference drawn from the study's outcomes. The basic censoring schemes are time and failure censoring. Progressive censoring schemes have flexibility of removing additional live (functioning and good) items at the observed actual failure times. Ng *et. al* (2009) proposed a new and more flexible censoring scheme known as adaptive progressive type II censoring (APT-IIC) scheme which is combination of the type I and progressive type II censoring plans. APT-IIC ensures a desired number of failed observations in a tested sample within a prescribed duration of the experiment.

Consider n test units in a life-test. Let m be the desired counts of failed units in the observed sample. Let $R = (R_1, R_2, \ldots, R_m)$ be the pre-determined intermittent withdrawals

Corresponding Author: Pulkit Srivastava
Email: pulkit.stats@gmail.com

under progressive censoring scheme such that the experiment span is pre-fixed at time T. Let k be the total number of observed failure times before the pre-determined time T *i.e.* $X_{k:m:n} \leq T \leq X_{k+1:m:n}$; $k = 0, 1, \ldots, m$ where $X_{0:m:n} = 0$ and $X_{m+1:m:n} = \infty$. If the total experiment time exceeds the ideal test time T, then $R_{k+1} = R_{k+2} = \cdots = R_{m-1} = 0$ and $R_m = n - m - \sum_{i=1}^{k} R_i$. In this situation, no surviving units get chance to be removed except at the time of m^{th} failure. This condition helps to accelerate the experiment so that it ends as soon as possible.

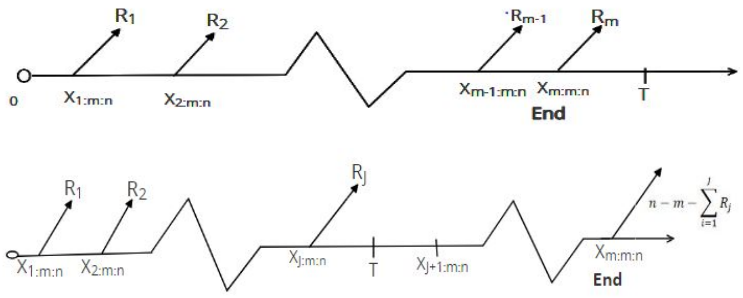


Figure 1: A visual of the Adaptive Progressive Type II censoring scheme

Inferential studies for different life-time models under APT-IIC scheme are undertaken by various authors. Parameter estimation of exponential distribution under APT-IIC has been considered by Ng *et. al* (2009). Burr type XII distribution was considered by Amein (2016) for estimation of unknown parameters under APT-IIC. Sobhi and Soliman (2016) considered classical and Bayes estimation of the exponentiated Weibull model. Similarly, parameter estimation of exponential, generalized exponential, exponentiated exponential, generalized inverted exponential distributions under APT-IIC have been considered by Ng *et. al* (2009), El-Din *et. al* (2017), Ateya and Mohammed (2017) and Soliman *et. al* (2020) respectively. Maximum product spacing and the maximum likelihood estimation of parameters of generalized Rayleigh distribution and Weibull distribution was discussed by Almetwally *et. al* (2019) and Almetwally *et. al* (2020) respectively. Some other distributions under APT-IIC are: Generalized Pareto distribution by Mahmoud *et. al* (2013), Kumaraswamy distribution by Almalki *et. al* (2022), new Weibull-Pareto distribution by EL-Sagheer *et. al* (2018), Kumaraswamy-exponential distribution by Mohan and Chako (2021), Truncated normal distribution by Chen and Gui (2020), generalized Gompertz distribution by Amein *et. al* (2020), extreme value distribution by Ye *et. al* (2014), exponentiated power Lindley by Ahmad *et. al* (2021), exponentiated half-logistic distribution by Xiong and Gui (2021), exponentiated Pareto distribution by Wang and Gui (2021), asymmetric power hazard distribution by El-Morshedy *et. al* (2022).

The present paper focuses on a length biased model which is defined in section 2. Maximum likelihood estimation (MLE) along with Asymptotic Confidence Interval (ACI) is derived in section 3. Section 4 describes Bayes estimation under three loss functions namely squared error loss function (SELF), general entropy loss function (GELF) and linear exponential loss function (LINEX). In addition, the corresponding Bayesian credible intervals (BCI) and highest probability density intervals (HPD) are also calculated. Markov Chain Monte Carlo (MCMC) approximations are detailed in section 5. A real data set illustrates the developed theory in section 6. Concluding remarks are given in section 7.

2. The model

Weighted distributions (WD) were first proposed by Fisher (1934). WD emerge when information from any stochastic process are produced using a predetermined weight function. Compared to the original distributions, WD are more adaptable and as a result, they are helpful in several fields including ecology, biometry, environmental sciences, survivability, and reliability analysis. When the weight function, say $w(x)$, depends on the length of the units of interest, *i.e.* $w(x) = x$, the resulting distribution is termed as *length biased distribution* (LBD). Although LBD does not add any additional parameters to the model, it does provide it more flexibility. There are LB versions of a number of distributions accessible in statistical literature. Patil and Rao (1978) introduced LB versions of many basic distributions such as log-normal, gamma, Pareto, beta. Das and Roy (2011) discussed LB weighted Weibull distribution. Some other works on LB versions of different distributions are: LB weighted generalized Rayleigh distribution (Das and Roy, 2011), LB beta distribution (Mir et.al, 2013), LB Weibull distribution (Pandya *et. al*, 2013), LB exponentiated inverted Weibull (Seenoi *et. al*, 2014), LB weighted Lomax distribution (Ahmad *et. al*, 2016), LB Inverse Rayleigh distribution (Pandey and Kumari, 2016), LB weighted Erlang distribution (Reyad *et. al*, 2017), LB Sushila distribution (Rather and Subramanian, 2018), LB weighted Lomax distributions (Karimi and Nasiri, 2018), LB Erlang–truncated exponential distribution (Rather and Subramanian, 2019) and many more.

Recently Pandey *et. al* (2021) introduced LB Log Logistic distribution ($LBLL(\alpha, \beta)$) as a lifetime model. The pdf of ($LBLL(\alpha, \beta)$) is given as

$$f(x;\alpha,\beta) = \frac{\left(\frac{\beta}{\alpha}\right)\left(\frac{x}{\alpha}\right)^{\beta}}{\left\{1+\left(\frac{x}{\alpha}\right)^{\beta}\right\}^{2}} \frac{\sin\left(\frac{\pi}{\beta}\right)}{\left(\frac{\pi}{\beta}\right)} \quad \text{for } x,\ \alpha,\ \beta > 0 \tag{1}$$

The corresponding cdf can be obtained as (see Pandey *et. al* (2021))

$$F(x) = \int_{x}^{\infty} f(t;\alpha,\beta)dt$$

$$F(x) = \frac{\sin\left(\frac{\pi}{\beta}\right)}{\left(\frac{\pi}{\beta}\right)} \frac{1}{\beta}\left(\frac{x}{\alpha}\right)^{1-\beta} \log\left(1+\left(\frac{x}{\alpha}\right)^{\beta}\right) - \frac{\left(\frac{x}{\alpha}\right)}{1+\left(\frac{x}{\alpha}\right)^{\beta}} - \left(\frac{1-\beta}{\beta}\right)\left[\left(\frac{x}{\alpha}\right) + \sum_{u=1}^{\infty} \frac{(-1)^{u}\left(\frac{x}{\alpha}\right)^{1+u\beta}}{u\,(1+u\beta)}\right] \quad \text{for } x,\ \alpha,\ \beta > 0 \tag{2}$$

3. Classical point and interval estimation

Under APT-IIC, n, m, R, T be fixed before the experiment begins. Lifetime distribution is assumed to follow pdf $f(x;\Theta)$and corresponding cdf $F(x;\Theta)$, where Θ represents a vector of parameters. The likelihood function under APT-IIC scheme (Ng *et al.*, 2009) is given as

$$L(\Theta;t) = C_J \left(\prod_{i=1}^{m} f\left(t_i;\Theta\right)\right)\left(\prod_{i=1}^{k}\left(1-F\left(t_i;\Theta\right)\right)^{R_i}\right)\left(1-F\left(t_m;\Theta\right)\right)^{n-m-\sum_{i=1}^{k}R_i}$$
$$0 < t_1 < t_2 < \cdots < t_m < \infty \quad (3)$$

where

$$C_J = \prod_{i=1}^{m}\left(n-i+1-\sum_{j=1}^{\max\{i-1,k\}} R_j\right)$$

The likelihood function for $LBLL(\alpha,\beta)$ whose pdf and cdf are given by (1) and (2) respectively, under APT-IIC is given as

$$L\left(\beta,\alpha;t\right) = C_J\left(\prod_{i=1}^{m}\left\{E_1\right\}\right)\left(\prod_{i=1}^{J}\left\{E_2\right\}^{R_i}\right)\left(\left\{E_3\right\}^{n-m-\sum_{i=1}^{k}R_i}\right) \quad (4)$$

where

$$E_1 = \frac{\left(\frac{\beta}{\alpha}\right)\left(\frac{t_i}{\alpha}\right)^{\beta}}{\left\{1+\left(\frac{t_i}{\alpha}\right)^{\beta}\right\}^2}\frac{\sin\left(\frac{\pi}{\beta}\right)}{\left(\frac{\pi}{\beta}\right)}$$

$$E_2 = 1 - \frac{\sin\left(\frac{\pi}{\beta}\right)}{\left(\frac{\pi}{\beta}\right)}\frac{1}{\beta}\left(\frac{t_i}{\alpha}\right)^{1-\beta}\log\left(1+\left(\frac{t_i}{\alpha}\right)^{\beta}\right) + \frac{\left(\frac{t_i}{\alpha}\right)}{1+\left(\frac{t_i}{\alpha}\right)^{\beta}} + \left(\frac{1-\beta}{\beta}\right)\left[\left(\frac{t_i}{\alpha}\right)+\sum_{r=1}^{\infty}\frac{(-1)^r\left(\frac{t_i}{\alpha}\right)^{1+r\beta}}{r\left(1+r\beta\right)}\right]$$

$$E_3 = 1 - \frac{\sin\left(\frac{\pi}{\beta}\right)}{\left(\frac{\pi}{\beta}\right)}\frac{1}{\beta}\left(\frac{t_m}{\alpha}\right)^{1-\beta}\log\left(1+\left(\frac{t_m}{\alpha}\right)^{\beta}\right) + \frac{\left(\frac{t_m}{\alpha}\right)}{1+\left(\frac{t_m}{\alpha}\right)^{\beta}} + \left(\frac{1-\beta}{\beta}\right)\left[\left(\frac{t_m}{\alpha}\right)+\sum_{r=1}^{\infty}\frac{(-1)^r\left(\frac{t_m}{\alpha}\right)^{1+r\beta}}{r\left(1+r\beta\right)}\right]$$

The corresponding log likelihood function is written as

$$\ln L = const + \sum_{i=1}^{m}\ln\left\{E_1\right\} + \sum_{i=i}^{J} R_i \ln\left\{E_2\right\} + \left(n-m-\sum_{i=1}^{k}R_i\right)\ln\left\{E_3\right\} \quad (5)$$

Partially differentiating (5) with respect to (w.r.t.) α yields

$$\frac{\partial \ln L}{\partial \alpha} = \frac{\partial}{\partial \alpha}\left\{\sum_{i=1}^{m}\ln\left\{E_1\right\}\right\} + \frac{\partial}{\partial \alpha}\left\{\sum_{i=i}^{J} R_i \ln\left\{E_2\right\}\right\} + \left(n-m-\sum_{i=1}^{k}R_i\right)\frac{\partial}{\partial \alpha}\left\{\ln\left\{E_3\right\}\right\} \quad (6)$$

$\hat{\alpha}$, mle of α is the value for which $\frac{\partial \ln L}{\partial \alpha} = 0$ and $\left.\frac{\partial^2 \ln L}{\partial \alpha^2}\right|_{\alpha=\hat{\alpha}} < 0$.

Similarly, partial differentiation of (5) w.r.t. β yields

$$\frac{\partial \ln L}{\partial \beta} = \frac{\partial}{\partial \beta}\left\{\sum_{i=1}^{m} \ln\{E_1\}\right\} + \frac{\partial}{\partial \beta}\left\{\sum_{i=i}^{J} R_i \ln\{E_2\}\right\} + \left(n - m - \sum_{i=1}^{k} R_i\right)\frac{\partial}{\partial \beta}\{\ln\{E_3\}\} \quad (7)$$

$\hat{\beta}$, mle of β, is the value for which $\frac{\partial \ln L}{\partial \beta} = 0$ and $\frac{\partial^2 \ln L}{\partial \beta^2}\Big|_{\beta=\hat{\beta}} < 0$.

The solution of the system of nonlinear equations (6)-(7) of the first partial derivatives of log-likelihood function w.r.t. parameters cannot be obtained in closed form . Therefore, to find approximate MLE*s*, a numerical method like Newton-Raphson (N-R) method is used.

Let $\widehat{\lambda} = (\widehat{\alpha}, \widehat{\beta})$ be denoting the *mle* of $\lambda = (\alpha, \beta)$ and $I(\lambda)$ is Fisher's Information matrix, *i.e.*

$$I(\lambda) = -\frac{1}{n}\begin{bmatrix} E\left(\frac{\partial^2 \log L}{\partial \alpha^2}\right) & E\left(\frac{\partial^2 \log L}{\partial \alpha \partial \beta}\right) \\ E\left(\frac{\partial^2 \log L}{\partial \alpha \partial \beta}\right) & E\left(\frac{\partial^2 \log L}{\partial \beta^2}\right) \end{bmatrix} \quad (8)$$

Matrix elements for $I(\lambda)$, given in (8) are defined under APT-IIC as under,

$$\frac{\partial^2 \ln L}{\partial \alpha^2} = \frac{\partial^2}{\partial \alpha^2}\left\{\sum_{i=1}^{m} \ln\{E_1\}\right\} + \frac{\partial^2}{\partial \alpha^2}\left\{\sum_{i=i}^{J} R_i \ln\{E_2\}\right\} + \left(n - m - \sum_{i=1}^{k} R_i\right)\frac{\partial^2}{\partial \alpha^2}\{\ln\{E_3\}\} \quad (9)$$

$$\frac{\partial^2 \ln L}{\partial \beta^2} = \frac{\partial^2}{\partial \beta^2}\left\{\sum_{i=1}^{m} \ln\{E_1\}\right\} + \frac{\partial^2}{\partial \beta^2}\left\{\sum_{i=i}^{J} R_i \ln\{E_2\}\right\} + \left(n - m - \sum_{i=1}^{k} R_i\right)\frac{\partial^2}{\partial \beta^2}\{\ln\{E_3\}\} \quad (10)$$

$$\frac{\partial^2 \ln L}{\partial \alpha \partial \beta} = \frac{\partial^2}{\partial \alpha \partial \beta}\left\{\sum_{i=1}^{m} \ln\{E_1\}\right\} + \frac{\partial^2}{\partial \alpha \partial \beta}\left\{\sum_{i=i}^{J} R_i \ln\{E_2\}\right\} + \left(n - m - \sum_{i=1}^{k} R_i\right)\frac{\partial^2}{\partial \alpha \partial \beta}\{\ln\{E_3\}\} \quad (11)$$

As it is evident from (9)-(11), the expectation of Hessian matrix is complicated due to presence of mathematically intractable terms. Since the parameter vector λ is unknown, hence using uniqueness property of *mle*, we estimate $I^{-1}(\lambda)$ by $I^{-1}(\widehat{\lambda})$. This provides ACI for the unknown parameters α and β given as

$$\left(\hat{\alpha} - z_{\frac{\xi}{2}}\sqrt{var\,(\hat{\alpha})}, \hat{\alpha} + z_{\frac{\xi}{2}}\sqrt{var\,(\hat{\alpha})}\right)$$
$$\left(\hat{\beta} - z_{\frac{\xi}{2}}\sqrt{var\left(\hat{\beta}\right)}, \hat{\beta} + z_{\frac{\xi}{2}}\sqrt{var\left(\hat{\beta}\right)}\right)$$

where $var\,(\hat{\alpha})$ and $var\left(\hat{\beta}\right)$ are the estimated variances of $\hat{\alpha}$ and $\hat{\beta}$ given by the main diagonal elements of $I^{-1}(\widehat{\lambda})$ and $z_{\frac{\xi}{2}}$ represents the right tail probability for standard normal distribution.

4. Bayesian estimation

Availability of prior information about concerned parameters, enables alternate Bayesian inferential approach. In this paper, Bayes estimators (BE*s*) of the unknown parameters (α, β) are proposed under SELF, GELF and LINEX loss function. In addition, the corresponding

BCI and HPD interval are also calculated. Prior distributions for the unknown independent parameters α and β are taken to be non-informative prior and gamma prior respectively.

$$p(\alpha) = \frac{1}{\alpha}; \quad \alpha > 0$$
$$p(\beta) = \frac{c^d}{\Gamma d}\beta^{d-1}\exp(-c\beta); \quad \beta, c, d > 0$$

where c and d are hyper parameters. Assuming the independence of the scale and shape parameters, the joint prior distribution of α and β is written as

$$p(\alpha, \beta) = \frac{c^d}{\alpha\Gamma d}\beta^{d-1}\exp(-c\beta); \quad \alpha, \beta, c, d > 0 \tag{12}$$

Joint posterior distribution of α and β is

$$p(\alpha, \beta|x) \propto \left(\prod_{i=1}^{m}\{E_1\}\right)\left(\prod_{i=1}^{J}\{E_2\}^{R_i}\right)\left(\{E_3\}^{n-m-\sum_{i=1}^{k}R_i}\right)\{E_5\} \tag{13}$$

where

$$E_5 = \frac{\beta^{d-1}}{\alpha}\exp(-c\beta)$$

4.1. Marginal posterior distributions

Marginal posterior distribution of unknown parameter α

$$p(\alpha|x, \beta) \propto \int_0^\infty \left(\prod_{i=1}^{m}\{E_1\}\right)\left(\prod_{i=1}^{J}\{E_2\}^{R_i}\right)\left(\{E_3\}^{n-m-\sum_{i=1}^{k}R_i}\right)\{E_5\}\, d\beta \tag{14}$$

Marginal posterior distribution of unknown parameter β

$$p(\beta|x, \alpha) \propto \int_0^\infty \left(\prod_{i=1}^{m}\{E_1\}\right)\left(\prod_{i=1}^{J}\{E_2\}^{R_i}\right)\left(\{E_3\}^{n-m-\sum_{i=1}^{k}R_i}\right)\{E_5\}\, d\alpha \tag{15}$$

4.2. Parametric Bayes estimators under different loss functions

The BE of an unknown parameter depends on the form of the loss function. We obtain the expressions for BEs of unknown parameters under different loss functions. SELF, a symmetric loss function, weighs underestimation (UE) and overestimation (OE) equally. GELF (Calabria and Pulcini, 1996) and LINEX (Varian, 1975) are asymmetric in respect of UE and OE being assigned different degrees of seriousness.

1. SELF
   BE of the unknown parameter α, the posterior mean, is given as

$$\tilde{\alpha}_{BS} \propto \alpha \iint \left(\prod_{i=1}^{m}\{E_1\}\right)\left(\prod_{i=1}^{J}\{E_2\}^{R_i}\right)\left(\{E_3\}^{n-m-\sum_{i=1}^{k}R_i}\right)\{E_5\}\, d\beta d\alpha \tag{16}$$

BE of unknown parameter β is given as

$$\tilde{\beta}_{BS} \propto \beta \iint \left(\prod_{i=1}^{m} \{E_1\}\right) \left(\prod_{i=1}^{J} \{E_2\}^{R_i}\right) \left(\{E_3\}^{n-m-\sum_{i=1}^{k} R_i}\right) \{E_5\}\, d\alpha d\beta \tag{17}$$

2. GELF
BE of unknown parameter α is given as

$$(\tilde{\alpha}_{BG})^{-q} \propto \iint \alpha^{-q} \left(\prod_{i=1}^{m} \{E_1\}\right) \left(\prod_{i=1}^{J} \{E_2\}^{R_i}\right) \left(\{E_3\}^{n-m-\sum_{i=1}^{k} R_i}\right) \{E_5\}\, d\beta d\alpha \tag{18}$$

BE of unknown parameter β is given as

$$(\tilde{\beta}_{BG})^{-q} \propto \iint \beta^{-q} \left(\prod_{i=1}^{m} \{E_1\}\right) \left(\prod_{i=1}^{J} \{E_2\}^{R_i}\right) \left(\{E_3\}^{n-m-\sum_{i=1}^{k} R_i}\right) \{E_5\}\, d\alpha d\beta \tag{19}$$

3. LINEX
BE of unknown parameter α is given as

$$\tilde{\alpha}_{BLL} \propto \frac{1}{q} \ln \iint e^{-q\alpha} \left(\prod_{i=1}^{m} \{E_1\}\right) \left(\prod_{i=1}^{J} \{E_2\}^{R_i}\right) \left(\{E_3\}^{n-m-\sum_{i=1}^{k} R_i}\right) \{E_5\}\, d\beta d\alpha \tag{20}$$

BE of unknown parameter β is given as

$$\tilde{\beta}_{BLL} \propto \frac{1}{q} \ln \iint e^{-q\beta} \left(\prod_{i=1}^{m} \{E_1\}\right) \left(\prod_{i=1}^{J} \{E_2\}^{R_i}\right) \left(\{E_3\}^{n-m-\sum_{i=1}^{k} R_i}\right) \{E_5\}\, d\alpha d\beta \tag{21}$$

5. Markov Chain Monte Carlo approximation

Markov Chain Monte Carlo (MCMC) technique approximates the complex expressions of posterior distribution and BE*s* which are not available in closed form. We use MCMC iteration such that the Gibbs sampler nests Metropolis-Hastings (M-H) algorithms (Metropolis *et. al*, 1953; Hastings, 1970). Convergence of Markov chain simulation is achieved by choosing a starting value which is nearer to the true value. Initial M simulated variates are omitted to shake off the transient influence of arbitrary initial values. The desired posterior sample is thus the residual set corresponding to position $i, i = M+1, \ldots, N$, for sufficiently large N.
BE*s* of the unknown parameters under SELF are given by

$$\begin{aligned} \tilde{\alpha}_{BSMC} &= \frac{1}{N-M} \sum_{i=M+1}^{N} \alpha_i \\ \tilde{\beta}_{BSMC} &= \frac{1}{N-M} \sum_{i=M+1}^{N} \beta_i \end{aligned} \tag{22}$$

Also, the approximate BEs of the unknown parameters under GELF are given by

$$\tilde{\alpha}_{BGMC} = \left(\frac{1}{N-M} \sum_{i=M+1}^{N} {\alpha_i}^{-q} \right)^{-\frac{1}{q}}$$
$$\tilde{\beta}_{BGMC} = \left(\frac{1}{N-M} \sum_{i=M+1}^{N} {\beta_i}^{-q} \right)^{-\frac{1}{q}} \tag{23}$$

where $q > 0$ represents OE $\left(\tilde{\alpha}_{BG_1MC}, \tilde{\beta}_{BG_1MC}\right)$ and $q < 0$ represents UE $\left(\tilde{\alpha}_{BG_2MC}, \tilde{\beta}_{BG_2MC}\right)$. The approximate BE$s$ of the unknown parameters under LINEX are given by

$$\tilde{\alpha}_{BLMC} = -\frac{1}{q} \log \left(\frac{1}{N-M} \sum_{i=M+1}^{N} e^{-q\alpha_i} \right)$$
$$\tilde{\beta}_{BLMC} = -\frac{1}{q} \log \left(\frac{1}{N-M} \sum_{i=M+1}^{N} e^{-q\beta_i} \right) \tag{24}$$

where $q > 0$ represents OE $\left(\tilde{\alpha}_{BL_1MC}, \tilde{\beta}_{BL_1MC}\right)$ and $q < 0$ represents UE $\left(\tilde{\alpha}_{BL_2MC}, \tilde{\beta}_{BL_2MC}\right)$.

6. Simulation study

In this section, foremost data from LBLLD is generated via simulation. Next, APT-IIC samples from the obtained LBLLD data is extracted by following procedure of Balakrishnan and Sandhu (1995) and Ng *et al.* (2009). The algorithm described below resamples according to APT-IIC from continuous lifetime distribution.

1. Set the values of n, m, Θ, T and $R = (R_1, R_2, \ldots, R_m)$, as desired by the sample situation.

2. Simulate m random variables from $U(0,1)$ as $U_1, U_2, \ldots, U_m$.

3. Set $W_i = U_i^{1/(i+R_m+R_{m-1}+\cdots+R_{m-i+1})}$ for $i = 1, 2, \ldots, m$.

4. Set $V_i = 1 - W_m W_{m-1} \cdots W_{m-i+1}$ for $i = 1, 2, \ldots, m$. Then $V_1, V_2, \ldots, V_m$ is the m progressive type II censored sample from $U(0,1)$.

5. Set $X_i = F^{-1}(V_i; \Theta)$ for $i = 1, 2, \ldots, m$, where $F^{-1}(.; \Theta)$ is the quantile function of the lifetime distribution. Thus $X_1, X_2, \ldots, X_m$ represent the required m progressive type II censored sample from the specified distribution $F(.)$.

6. Next, identify the value of k, where $X_{k:m:n} < T < X_{k+1:m:n}$ and discard the sample $X_{k+2:m:n}, \cdots, X_{m:m:n}$.

7. Simulate the first $m-k-1$ order statistics from a truncated distribution considered as $\frac{f(x)}{1-F(x_{k+1:m:n})}$ with sample size $\left(n - \sum_{i=1}^{k} R_i - J - 1\right)$ as $X_{k+2:m:n}, X_{k+3:m:n} \cdots, X_{m:m:n}$.

Two random samples of sizes $n = 30,\ 50$ have been generated from $LBLL(\alpha, \beta)$ by setting $\alpha = 1.5,\ \beta = 3.2$. Three different preset values for $T = 4.5, 5, 5.5$ are taken. MLEs are computed from these samples through numerical approximation N-R method in R software. OpenBUGS is utilised for generating posterior samples using MCMC by fixing the hyper parameters at $b = 2, c = 4$. 10000 samples with 2000 samples for burn-in period are generated. We have taken $q = 2$ for OE and $q = -2$ for UE. Bayes estimates under MCMC have been calculated using (22)-(24).

Table 1 represents the APT-IIC schemes which we have used in simulation. Estimated values of scale parameter α and shape parameter β with the associated mean square error (MSE) of MLEs and BEs under three loss functions for different combinations of (n, T) are presented in Tables (2)-(4) respectively. Following inferences based on these tables:

1. For unknown scale parameter α, Bayes estimates give values which are closer to true values for all the three values of T. For some censoring schemes, MLEs also give better estimates in terms of minimum MSEs. Among Bayes estimates, LINEX OE gives values with higher precision for most of the censoring scheme.

2. Same pattern can be seen for the unknown shape parameter β also. Among Bayes estimates, LINEX UE gives better values as they have minimum MSEs than others.

Tables (5)-(6) represent the LL, UL and AL of ACI and BCI, HPD1 and HPD2 of the parameters under study for the three selected values of T respectively. The following relationship is obtained for both unknown parameters

$$HPD1_{AL} < HPD2_{AL} < BCI_{AL} < ACI_{AL}$$

This is true for all the three values of T. Here, HPD1 refers 89%HPD and HPD2 refers 95%HPD intervals. AL of all intervals are decreased as we increase the value of m for different censoring schemes.

Table 1: Progressive type II censoring schemes used in simulation

n	m	**CS**	**R**
30	**10**	**CS[1]**	2*10
		CS[2]	0*3, 5*4, 0*3
	20	**CS[3]**	0*5, 1*10, 0*5
		CS[4]	0*8, 2*5, 0*7
50	**25**	**CS[5]**	1*25
		CS[6]	0*10, 5*5, 0*10
	35	**CS[7]**	0*20, 1*15
		CS[8]	1*15, 0*20

Table 2: Different point estimates of unknown parameters for $T = 4.5$

n	m	CS		MLE		Bayes Estimates									
						SELF		GELF				LINEX			
								α		β		α		β	
				α	β	α	β	OE	UE	OE	UE	OE	UE	OE	UE
30	10	CS[1]	Est.	1.4192	3.4185	1.4434	3.352	1.4348	1.4463	3.3245	3.3612	1.4352	1.4515	3.292	3.4153
			MSE	0.0354	0.2904	0.0033	0.0235	0.0043	0.0029	0.0159	0.0264	0.0043	0.0024	0.0088	0.047
		CS[2]	Est.	1.1054	3.0748	1.1285	3.1237	1.1212	1.1311	3.0949	3.1334	1.123	1.1344	3.0655	3.1866
			MSE	0.1933	0.266	0.138	0.0063	0.1435	0.1362	0.0115	0.0049	0.1422	0.1337	0.0185	0.0008
	20	CS[3]	Est.	1.3804	3.5995	1.4535	4.3541	1.4522	1.454	4.3415	4.3583	1.4522	1.4548	4.3179	4.391
			MSE	0.0414	0.4477	0.0022	1.3323	0.0023	0.0021	1.3034	1.342	0.0023	0.002	1.2501	1.4188
		CS[4]	Est.	1.3321	3.5215	1.399	3.8636	1.3976	1.3994	3.8548	3.8666	1.3977	1.4002	3.8412	3.8864
			MSE	0.0551	0.3772	0.0102	0.4406	0.0105	0.0101	0.429	0.4445	0.0105	0.01	0.4113	0.4714
50	25	CS[5]	Est.	1.5107	3.7999	1.6384	3.8142	1.6371	1.6388	3.8077	3.8164	1.6369	1.6399	3.7977	3.831
			MSE	0.0178	0.5217	0.0192	0.3774	0.0188	0.0193	0.3694	0.3801	0.0188	0.0196	0.3574	0.3984
		CS[6]	Est.	1.2228	3.3717	1.1115	2.8917	1.1097	1.1121	2.8876	2.893	1.1102	1.1129	2.8839	2.8995
			MSE	0.0938	0.3066	0.1509	0.0951	0.1524	0.1504	0.0976	0.0943	0.152	0.1499	0.1	0.0903
	35	CS[7]	Est.	1.3492	3.9822	1.2338	3.4561	1.2332	1.234	3.4536	3.457	1.2333	1.2342	3.4504	3.4619
			MSE	0.0397	0.8116	0.0709	0.0656	0.0712	0.0708	0.0644	0.0661	0.0711	0.0706	0.0627	0.0686
		CS[8]	Est.	1.7086	3.9886	1.731	3.7049	1.7302	1.7313	3.7017	3.7059	1.73	1.732	3.6971	3.7127
			MSE	0.0575	0.7673	0.0534	0.255	0.053	0.0535	0.2518	0.256	0.0529	0.0538	0.2472	0.2629

Table 3: Different point estimates of unknown parameters for $T = 5$

n	m	CS		MLE		Bayes Estimates									
						SELF		GELF				LINEX			
								α		β		α		β	
				α	β	α	β	OE	UE	OE	UE	OE	UE	OE	UE
30	10	CS[1]	Est.	1.4348	3.3993	1.2226	3.1913	1.2148	1.2252	3.1674	3.1993	1.2163	1.229	3.1417	3.2437
			MSE	0.0383	0.2994	0.077	0.0006	0.0814	0.0755	0.0016	0.0006	0.0805	0.0735	0.0039	0.0026
		CS[2]	Est.	1.0874	2.9197	1.2066	3.1374	1.1966	1.2099	3.105	3.1482	1.1986	1.2147	3.0721	3.209
			MSE	0.2049	0.279	0.0861	0.0043	0.0921	0.0842	0.0094	0.003	0.0909	0.0814	0.0167	0.0005
	20	CS[3]	Est.	1.3757	3.4025	1.3706	3.2105	1.368	1.3715	3.2034	3.2128	1.3683	1.373	3.1955	3.2257
			MSE	0.0443	0.276	0.0168	0.0002	0.0174	0.0165	0.0001	0.0003	0.0174	0.0161	0.0001	0.0008
		CS[4]	Est.	1.3289	3.3174	1.3549	2.9123	1.3516	1.356	2.9064	2.9143	1.352	1.3578	2.9009	2.9238
			MSE	0.0588	0.2396	0.0211	0.0828	0.022	0.0208	0.0863	0.0817	0.0219	0.0202	0.0895	0.0763
50	25	CS[5]	Est.	1.5069	3.6342	1.6052	3.916	1.604	1.6056	3.9096	3.9181	1.6039	1.6064	3.8995	3.9326
			MSE	0.0185	0.3164	0.0111	0.5127	0.0108	0.0112	0.5037	0.5158	0.0108	0.0113	0.4894	0.5368
		CS[6]	Est.	1.2098	3.1185	1.146	3.2756	1.1445	1.1465	3.2698	3.2775	1.1448	1.1472	3.2629	3.2883
			MSE	0.1025	0.159	0.1253	0.0058	0.1264	0.1249	0.005	0.0061	0.1262	0.1245	0.0041	0.0079
	35	CS[7]	Est.	1.3446	3.843	1.3981	3.1858	1.3972	1.3984	3.1836	3.1866	1.3973	1.3989	3.1811	3.1905
			MSE	0.0402	0.5686	0.0104	0.0002	0.0106	0.0103	0.0003	0.0002	0.0106	0.0102	0.0004	0.0001
		CS[8]	Est.	1.7291	3.7901	1.7277	4.1016	1.7271	1.7279	4.0981	4.1028	1.727	1.7284	4.092	4.1113
			MSE	0.0684	0.4845	0.0518	0.813	0.0516	0.0519	0.8066	0.8151	0.0515	0.0521	0.7957	0.8305

Table 4: Different point estimates of unknown parameters for $T = 5.5$

n	m	CS		MLE		Bayes Estimates									
						SELF		GELF				LINEX			
								α		β		α		β	
				α	β	α	β	OE	UE	OE	UE	OE	UE	OE	UE
30	10	CS[1]	Est.	1.4335	3.3875	1.376	2.8097	1.3612	1.3808	2.7892	2.8165	1.3627	1.3893	2.7722	2.849
			MSE	0.0371	0.2799	0.0155	0.1526	0.0194	0.0143	0.169	0.1473	0.019	0.0124	0.1833	0.1235
		CS[2]	Est.	1.0831	2.8149	1.0673	2.661	1.0635	1.0686	2.6443	2.6666	1.0646	1.0703	2.632	2.6918
			MSE	0.211	0.3229	0.1872	0.2908	0.1905	0.1861	0.309	0.2848	0.1896	0.1847	0.3228	0.2586
	20	CS[3]	Est.	1.352	3.2312	1.4541	3.854	1.4519	1.4548	3.8415	3.8582	1.452	1.4562	3.8222	3.8866
			MSE	0.0519	0.1687	0.0021	0.428	0.0023	0.0021	0.4118	0.4335	0.0023	0.0019	0.3874	0.4717
		CS[4]	Est.	1.3035	3.1502	1.294	3.197	1.2917	1.2947	3.1901	3.1994	1.292	1.2959	3.1823	3.212
			MSE	0.0668	0.1693	0.0425	0.0001	0.0434	0.0421	0.0002	0.0001	0.0433	0.0416	0.0004	0.0002
50	25	CS[5]	Est.	1.4992	3.5115	1.6836	3.5613	1.6821	1.684	3.556	3.5631	1.6819	1.6852	3.5488	3.5738
			MSE	0.0169	0.2139	0.0337	0.1306	0.0332	0.0339	0.1268	0.1319	0.0331	0.0343	0.1217	0.1398
		CS[6]	Est.	1.1935	2.9376	1.3022	3.168	1.3008	1.3027	3.1637	3.1694	1.301	1.3035	3.1591	3.1769
			MSE	0.1116	0.1741	0.0391	0.0011	0.0397	0.0389	0.0014	0.001	0.0396	0.0386	0.0017	0.0006
	35	CS[7]	Est.	1.3395	3.7328	1.3375	3.5832	1.3369	1.3378	3.5806	3.584	1.337	1.3381	3.577	3.5894
			MSE	0.0442	0.4306	0.0264	0.1469	0.0266	0.0263	0.1449	0.1475	0.0266	0.0262	0.1421	0.1517
		CS[8]	Est.	1.7275	3.6091	1.7821	3.7001	1.7812	1.7824	3.6971	3.7011	1.7811	1.7831	3.6926	3.7076
			MSE	0.0681	0.2788	0.0796	0.2502	0.0791	0.0797	0.2471	0.2512	0.079	0.0801	0.2427	0.2578

Table 5: Different interval estimates for scale parameters

n	m	CS		T=4.5				T=5				T=5.5			
				ACI	BCI	HPD1	HPD2	ACI	BCI	HPD1	HPD2	ACI	BCI	HPD1	HPD2
30	10	CS[1]	LL	0.855	1.262	1.299	1.268	0.86	1.066	1.088	1.062	0.857	1.150	1.193	1.151
			UL	1.982	1.621	1.589	1.625	2.009	1.382	1.347	1.378	2.009	1.602	1.560	1.603
			AL	1.126	0.359	0.290	0.357	1.149	0.316	0.259	0.316	1.152	0.452	0.367	0.452
		CS[2]	LL	0.630	1.010	1.002	1.000	0.590	1.039	1.060	1.030	0.567	1.003	1.000	1.000
			UL	1.580	1.295	1.230	1.267	1.584	1.388	1.347	1.376	1.598	1.201	1.139	1.174
			AL	0.949	0.285	0.228	0.267	0.994	0.349	0.287	0.346	1.030	0.198	0.139	0.174
	20	CS[3]	LL	1.025	1.382	1.391	1.382	0.999	1.276	1.292	1.276	0.959	1.362	1.376	1.360
			UL	1.735	1.525	1.508	1.524	1.752	1.466	1.447	1.465	1.744	1.544	1.523	1.542
			AL	0.709	0.143	0.117	0.142	0.752	0.190	0.155	0.189	0.784	0.182	0.147	0.182
		CS[4]	LL	0.986	1.327	1.338	1.326	0.960	1.250	1.271	1.246	0.920	1.206	1.223	1.205
			UL	1.677	1.468	1.453	1.467	1.697	1.462	1.442	1.458	1.686	1.382	1.364	1.380
			AL	0.690	0.141	0.115	0.141	0.736	0.212	0.171	0.212	0.766	0.176	0.141	0.175
50	25	CS[5]	LL	1.185	1.563	1.576	1.567	1.166	1.535	1.548	1.534	1.146	1.603	1.620	1.602
			UL	1.835	1.713	1.697	1.716	1.847	1.674	1.662	1.673	1.852	1.762	1.747	1.760
			AL	0.649	0.150	0.121	0.149	0.681	0.139	0.114	0.139	0.706	0.159	0.127	0.158
		CS[6]	LL	0.923	1.039	1.054	1.037	0.887	1.078	1.092	1.075	0.852	1.233	1.248	1.231
			UL	1.522	1.185	1.172	1.182	1.532	1.215	1.203	1.212	1.534	1.371	1.359	1.368
			AL	0.598	0.146	0.118	0.145	0.644	0.137	0.111	0.137	0.682	0.138	0.111	0.137
	35	CS[7]	LL	1.122	1.192	1.199	1.193	1.11	1.344	1.353	1.342	1.098	1.292	1.300	1.289
			UL	1.576	1.277	1.267	1.277	1.578	1.454	1.442	1.452	1.58	1.385	1.376	1.382
			AL	0.453	0.085	0.068	0.084	0.468	0.110	0.089	0.110	0.482	0.093	0.076	0.093
		CS[8]	LL	1.413	1.670	1.681	1.673	1.410	1.677	1.685	1.676	1.389	1.719	1.729	1.719
			UL	2.003	1.791	1.779	1.793	2.047	1.779	1.767	1.777	2.065	1.844	1.830	1.843
			AL	0.590	0.121	0.098	0.120	0.637	0.102	0.082	0.101	0.675	0.125	0.101	0.124

Table 6: Different interval estimates for shape parameters

n	m	CS		T=4.5				T=5				T=5.5			
				ACI	BCI	HPD1	HPD2	ACI	BCI	HPD1	HPD2	ACI	BCI	HPD1	HPD2
30	10	CS[1]	LL	1.846	2.880	2.954	2.874	1.837	2.764	2.823	2.744	1.831	2.444	2.486	2.431
			UL	4.990	3.855	3.742	3.844	4.960	3.658	3.553	3.633	4.943	3.207	3.109	3.190
			AL	3.144	0.975	0.788	0.970	3.123	0.894	0.730	0.889	3.112	0.763	0.623	0.759
		CS[2]	LL	1.738	2.663	2.709	2.649	1.674	2.659	2.716	2.646	1.633	2.343	2.376	2.333
			UL	4.411	3.631	3.497	3.612	4.165	3.685	3.550	3.668	3.996	3.019	2.925	3.004
			AL	2.673	0.968	0.788	0.963	2.491	1.026	0.834	1.022	2.362	0.676	0.549	0.671
	20	CS[3]	LL	2.426	3.972	4.050	3.970	2.312	2.972	3.007	2.970	2.212	3.501	3.564	3.494
			UL	4.772	4.740	4.668	4.737	4.492	3.456	3.401	3.453	4.250	4.217	4.141	4.209
			AL	2.345	0.768	0.618	0.767	2.179	0.484	0.394	0.483	2.037	0.716	0.577	0.715
		CS[4]	LL	2.382	3.567	3.615	3.567	2.265	2.702	2.732	2.696	2.169	2.957	2.999	2.950
			UL	4.660	4.162	4.096	4.162	4.369	3.126	3.076	3.118	4.131	3.440	3.392	3.430
			AL	2.277	0.595	0.481	0.595	2.104	0.424	0.344	0.422	1.962	0.483	0.393	0.480
50	25	CS[5]	LL	2.665	3.561	3.606	3.562	2.561	3.660	3.702	3.652	2.484	3.344	3.375	3.339
			UL	4.934	4.074	4.020	4.074	4.706	4.171	4.116	4.161	4.538	3.787	3.734	3.780
			AL	2.269	0.513	0.414	0.512	2.145	0.511	0.414	0.509	2.054	0.443	0.359	0.441
		CS[6]	LL	2.412	2.720	2.745	2.710	2.254	3.052	3.096	3.053	2.142	2.984	3.008	2.981
			UL	4.331	3.069	3.028	3.058	3.982	3.506	3.459	3.506	3.732	3.355	3.310	3.351
			AL	1.918	0.349	0.283	0.348	1.727	0.454	0.363	0.453	1.589	0.371	0.302	0.370
	35	CS[7]	LL	2.971	3.305	3.330	3.301	2.875	3.053	3.069	3.052	2.799	3.426	3.454	3.430
			UL	4.993	3.605	3.573	3.601	4.810	3.321	3.288	3.319	4.666	3.740	3.707	3.743
			AL	2.022	0.300	0.243	0.300	1.935	0.268	0.219	0.267	1.867	0.314	0.253	0.313
		CS[8]	LL	2.971	3.534	3.557	3.531	2.834	3.907	3.943	3.904	2.711	3.531	3.556	3.533
			UL	5.006	3.878	3.839	3.875	4.745	4.299	4.258	4.294	4.507	3.874	3.835	3.874
			AL	2.035	0.344	0.282	0.344	1.910	0.392	0.315	0.390	1.796	0.343	0.279	0.341

7. Real data illustration

A real dataset is taken from Teza (2015, ch 4). Data describes the mechanical properties such as initial rate of absorption, water absorption, dry density and compressive strength of 50 units of clay bricks and fly ash bricks. This data set has also been analysed by Nagamani *et. al* (2021) for estimating common scale parameter of two logistic populations.

In the present paper, we have taken the data on compressive strength of fly ash bricks to illustrate the proposed method. The uncensored data are composed of 50 observations (3.62, 4.74, 9.88, 5.93, 6.09, 6.94, 6.32, 5.30, 5.14, 4.55, 4.03, 7.36, 3.57, 3.98, 4.03, 4.74, 7.32, 3.23, 5.38, 7.18, 6.07, 3.62, 6.64, 5.58, 5.23, 3.95, 5.86, 5.58, 6.97, 5.05, 4.35, 4.55, 4.79, 4.03, 4.74, 7.58, 3.62, 6.01, 3.99, 6.04, 4.74, 7.21, 3.61, 5.69, 7.21, 6.40, 3.55, 8.70, 4.35, 7). Table (7) indicates that LBLLD is suitable for the given data set based on negative log likelihood and three information criteria.

Table 7: Fitting of data to three different distributions

Sr no.	Reliability model	-LogL	AIC	BIC	AICC
1.	Logistic β=scale α=location	91.350	186.701	190.525	186.956
2.	Log logistic β=shape α=scale	89.547	183.094	186.918	183.349
3.	LBLL β=shape α=scale	89.493	182.986	186.81	183.241

Further two APT-II censored samples for $T = 3.99, 7.18$ are extracted with $n = 50, m = 30, R = (0^{*10}, 2^{*10}, 0^{*10})$. The censored samples thus obtained are (3.23, 3.55, 3.57, 3.61, 3.62, 3.62, 3.62, 3.95, 3.98, 3.99, 4.03, 4.03, 4.03, 4.35, 4.35, 4.55, 4.55, 4.74, 4.74, 4.74, 4.74, 4.79, 5.05, 5.14, 5.23, 5.30, 5.38, 5.58, 5.58, 5.69) and (3.23, 3.55, 3.57, 3.61, 3.62, 3.62, 3.62, 3.95, 3.98, 3.99, 4.03, 4.35, 4.55, 4.74, 5.05, 5.05, 5.14, 5.23, 5.30, 5.30, 5.38, 5.58, 5.58, 5.58, 5.69, 5.93, 6.07, 6.40, 6.97, 7.18). MLEs and Bayes estimates of the unknown parameters are given in Table (8) for the selected values of T. LL, UL, and AL of different confidence and credible intervals for the unknown scale and shape parameters are given in Table (9) for the selected values of T. Among Bayesian intervals, HPD1 interval has shortest length.

ACI is shortest classical interval followed by Boot-p and Boot-t. The following relationship can be seen for scale parameter

$$HPD1_{AL} < HPD2_{AL} < BCI_{AL} < ACI_{AL} < Boot - p_{AL} < Boot - t_{AL}$$

AL of intervals are increased amomg classical intervals while it is decreased among Bayesian intervals as we increase the value of T. Similarly, for shape parameter, AL is decreased for

Table 8: MLEs and Bayes estimates under APT-IIC real data for both values of T

(n,m)	T		MLE	Bayes Estimates				
				SELF	GELF		LINEX	
(50,30)	3.99	α	4.2667	4.2484	4.2476	4.2486	4.2461	4.2506
		β	10.4737	9.5466	9.5137	9.5574	9.3386	9.7536
	7.18	α	4.502	4.4906	4.4899	4.4908	4.4886	4.4927
		β	7.6081	7.3946	7.3837	7.3982	7.341	7.448

all intervals with increment in the value of T. For β, we get

$$HPD1_{AL} < HPD2_{AL} < BCI_{AL} < Boot - t_{AL} < Boot - p_{AL} < ACI_{AL}$$

This can be seen in intervals plots (Figure 2-3). MCMC trace plots are presented in Figure 4.

Table 9: Different intervals of unknown parameters under APT-IIC real data

T			ACI	BOOT-T	BOOT-P	BCI	HPD1	HPD2
α	3.99	**LL**	3.999	3.945	4.007	4.158	4.171	4.163
		UL	4.533	4.559	4.565	4.341	4.321	4.345
		AL	0.534	0.614	0.558	0.183	0.15	0.182
	7.18	**LL**	4.103	3.972	4.096	4.405	4.42	4.408
		UL	4.901	4.978	5.008	4.58	4.561	4.581
		AL	0.798	1.006	0.912	0.175	0.141	0.173
β	3.99	**LL**	7.489	8.25	9.04	8.65	8.806	8.64
		UL	13.458	12.17	13.37	10.44	10.24	10.41
		AL	5.968	3.92	4.33	1.79	1.434	1.77
	7.18	**LL**	5.471	5.918	6.637	6.942	7.041	6.926
		UL	9.744	8.742	9.984	7.849	7.791	7.826
		AL	4.273	2.824	3.347	0.907	0.75	0.9

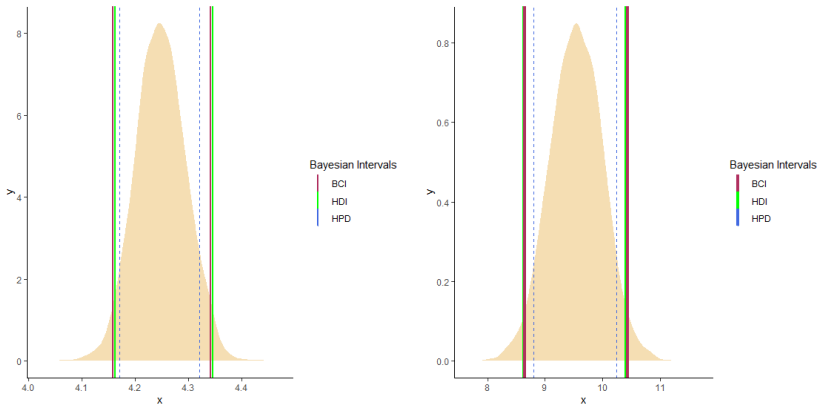


Figure 2: Interval plot for *T=3.99*

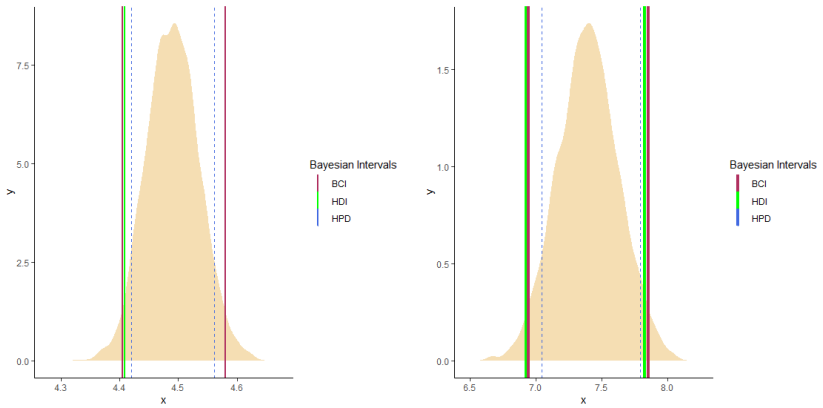


Figure 3: Interval plot for *T=7.18*

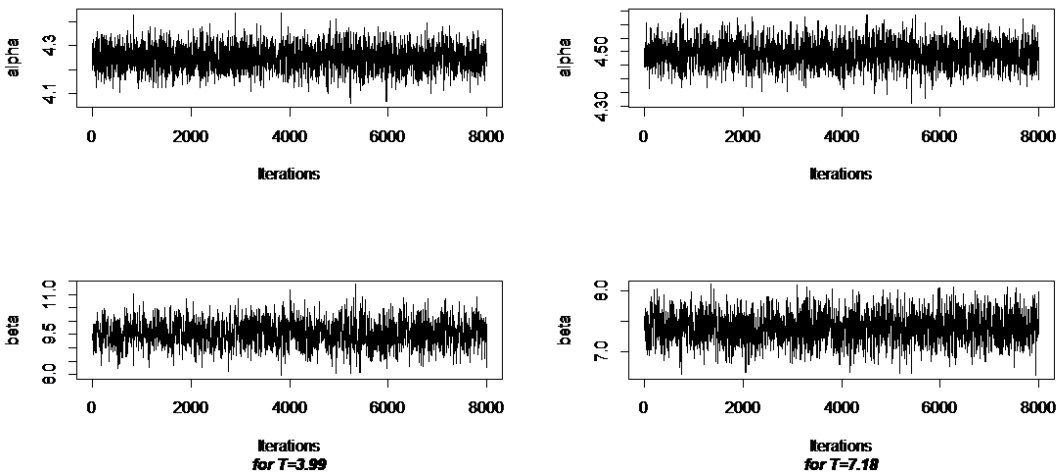


Figure 4: MCMC trace plot for *T=3.99* and *T=7.18*

8. Conclusion

In this paper, we have considered the point and interval estimations of the parameters of the LBLLD based on an APT-IIC scheme for Bayes and non-Bayes settings. This censoring scheme allows us to choose the next censoring number taking into account both the previous censoring numbers and previous failure times. The MLEs, the bootstrap confidence intervals and the ACIs based on the observed Fisher information matrix have been discussed. We assume the Jefferys and gamma priors for the unknown scale and shape parameters respectively and provide the Bayes estimators under the assumptions of SELF, GELF and LINEX loss functions. It is also found that when both parameters are unknown, the expressions for Bayes estimates cannot be obtained in explicit form. The Gibbs sampling technique is employed to generate MCMC samples. Credible intervals and HPD intervals have also been constructed. A real life example is discussed to verify the proposed methodology. The performance of different methods is compared via a Monte Carlo simulation.

Acknowledgments

First author gratefully acknowledges IoE grant from University of Delhi.

Conflict of interest

The authors do not have any financial or non-financial conflict of interest to declare for the research work included in this article.

References

Ahmad, A., Ahmad, S. P., and Ahmed, A. (2016). Length biased weighted Lomax distribution: statistical properties and application. *Pakistan Journal of Statistics and Operation Research*, **12**, 245-255.

Almalki, S. J., Farghal, A. W. A., Rastogi, M. K., and Abd-Elmougod, G. A. (2022). Partially constant-stress accelerated life tests model for parameters estimation of Kumaraswamy distribution under adaptive Type-II progressive censoring. *Alexandria Engineering Journal*, **61**, 5133-5143.

Almetwally, E. M., Almongy, H. M., and ElSherpieny, E. A. (2019). Adaptive type-II progressive censoring schemes based on maximum product spacing with application of generalized Rayleigh distribution. *Journal of Data Science*, **17**, 802-831.

Almetwally, E. M., Almongy, H. M., Rastogi, M. K., and Ibrahim, M. (2020). Maximum product spacing estimation of Weibull distribution under adaptive type-II progressive censoring schemes. *Annals of Data Science*, **7**, 257-279.

Amein, M. M. (2017). Estimation for unknown parameters of the extended Burr type-XII distribution based on an adaptive type-II progressive censoring scheme. Global *Journal of Pure and Applied Mathematics*, **13**, 7709-7724.

Amein, M. M., El-Saady, M., and Khalil, N. (2020). Estimation for generalized Gompertz distribution based on adaptive type-II progressive censored scheme. *Journal of Applied Probability*, **15**, 45-59.

Ateya, S. F. and Mohammed, H. S. (2017). Statistical inferences based on an adaptive progressive type-II censoring from exponentiated exponential distribution. *Journal of the Egyptian Mathematical Society*, **25**, 393-399.

Balakrishnan, N. and Sandhu, R. (1995). A simple simulation algorithm for generating progressive type-II censored samples. *American Statistical Association*, **49**, 229–230.

Calabria and Pulcini. (1996). Point estimation under asymmetric loss functions for left-truncated exponential samples. *Communications in Statistics-Theory and Methods*, **25**, 585-600.

Chen, S. and Gui, W. (2020). Estimation of unknown parameters of truncated Normal distribution under adaptive progressive type II censoring scheme. *Mathematics*, **9**, 49.

Das, K. K. and Roy, T. D. (2011). Applicability of length biased weighted generalized Rayleigh distribution. *Advances in Applied Science Research*, **2**, 320-327.

Das, K. K. and Roy, T. D. (2011). On some length-biased weighted Weibull distribution. *Advances in Applied Science Research*, **2**, 465-475.

El-Morshedy, M., El-Sagheer, R. M., Eliwa, M. S., and Alqahtani, K. M. (2022). Asymmetric power hazard distribution for COVID-19 mortality rate under adaptive type-II progressive censoring: theory and inferences. *Computational Intelligence and Neuroscience*, 2022.

El-Sagheer, R. M., Mahmoud, M. A., and Abdallah, S. H. (2018). Statistical inferences for new Weibull-Pareto distribution under an adaptive type-II progressive censored data. *Journal of Statistics and Management Systems*, **21**, 1021-1057.

Haj Ahmad, H., Salah, M. M., Eliwa, M. S., Ali Alhussain, Z., Almetwally, E. M., and Ahmed, E. A. (2021). Bayesian and non-Bayesian inference under adaptive type-II progressive censored sample with exponentiated power Lindley distribution. *Journal of Applied Statistics*, 1-21.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97-109.

Karimi, H. and Nasiri, P. (2018). Estimation parameter of $R = P(Y < X)$ for length-biased weighted Lomax distributions in the presence of outliers. *Mathematical and Computational Applications*, **23**, 9.

Mahmoud, M. A., Soliman, A. A., Abd Ellah, A. H., and El-Sagheer, R. M. (2013). Estimation of generalized Pareto under an adaptive type-II progressive censoring. *Intelligent Information Management*, **5**, 73-83.

Metropolis, N., A. W. Rosenbluth, M.N. Rosenbluth, A. H. Teller., and E. Teller. (1953). Equation of state calculations by fast computing machines. *Journal of Chemical Physics*, **21**, 1087-1091.

Mir, K. A., Ahmed, A., and Reshi, J. A. (2013). Structural properties of length biased beta distribution of first kind. *American Journal of Engineering Research*, **2**, 1-6.

Mohan, R. and Chacko, M. (2021). Estimation of parameters of Kumaraswamy-exponential distribution based on adaptive type-II progressive censored schemes. *Journal of Statistical Computation and Simulation*, **91**, 81-107.

Mohie El-Din, M. M. M., Amein, M. M., Shafay, A. R., and Mohamed, S. (2017). Estimation of generalized exponential distribution based on an adaptive progressively type-II censored sample. *Journal of Statistical Computation and Simulation*, **87**, 1292-1304.

Nagamani, N., Tripathy, M. R., and Kumar, S. (2020). Estimating common scale parameter of two Logistic populations: a Bayesian study. *American Journal of Mathematical and Management Sciences*, **40**, 44-67.

Ng, H. K. T., Kundu, D., and Chan, P. S. (2009). Statistical analysis of exponential lifetimes under an adaptive Type-II progressive censoring scheme. *Naval Research Logistics (NRL),* **56**, 687-698.

Pandey, R. and Kumari, N. (2016). A new lifetime distribution for modeling monotonic decreasing survival patterns. *Journal of Reliability and Statistical Studies*, **9**, 53-70.

Pandey, R., Srivastava, P., and Kumari, N. (2021). On some inferential aspects of length biased log-logistic model. *International Journal of System Assurance Engineering and Management*, **12**, 154-163.

Pandya, M., Pandya, S., and Andhira, P. (2013). Bayes estimation of Weibull length biased distribution. *Asian Journal of Current Engineering and Maths*, **2**, 44-49.

Patil, G. P. and Rao, C. R. (1978). Weighted distributions and size-biased sampling with applications to wildlife populations and human families. *Biometrics*, **34**, 179-189.

Rather, A. A. and Subramanian, C. (2018). Length biased Sushila distribution. *Universal Review*, **7**, 1010-1023.

Rather, A. A. and Subramanian, C. (2019). The length-biased erlang–truncated exponential distribution with life time data. *Journal of Information and Computational Science*, **9**, 340-355.

Reyad, H. M., Othman, S. A., and Moussa, A. A. (2017). The length-biased weighted Erlang distribution. *Asian Research Journal of Mathematics*, **3**, 1-15.

Seenoi, P., Supapakorn, T., and Bodhisuwan, W. (2014). The length-biased exponentiated inverted Weibull distribution. *International Journal of Pure and Applied Mathematics*, **92**, 191-206.

Sobhi, M. M. A. and Soliman, A. A. (2016). Estimation for the exponentiated Weibull model with adaptive Type-II progressive censored schemes. *Applied Mathematical Modelling*, **40**, 1180-1192.

Soliman, A. A., Ahmed, E. A., Farghal, A. W. A., and AL-Shibany, A. A. (2020). Estimation of generalized inverted exponential distribution based on adaptive type-II progressive censoring data. *Journal of Statistics Applications and Probability*, **9**, 215-230.

Teza, P. R. R. (2015). *Studies on Mechanical Properties of Brick Masonry* [Tech. Research Thesis]. Department of Civil Engineering, National Institute of Technology Rourkela.

Varian, H.R. (1975) A Bayesian approach to real estate assessment. In Savage, L.J., Feinberg, S.E. and Zellner, A., Eds., *Studies in Bayesian Econometrics and Statistics: In Honor of L. J. Savage*, North-Holland Pub. Co., 195-208.

Wang, K. and Gui, W. (2021). Statistical inference of exponentiated Pareto distribution under adaptive type-II progressive censored schemes. *Communications in Statistics-Simulation and Computation*, 1-32.

Xiong, Z. and Gui, W. (2021). Classical and Bayesian Inference of an exponentiated half-Logistic distribution under adaptive type II progressive censoring. *Entropy*, **23**, 1558.

Ye, Z. S., Chan, P. S., Xie, M., and Ng, H. K. T. (2014). Statistical inference for the extreme value distribution under adaptive Type-II progressive censoring schemes. *Journal of Statistical Computation and Simulation*, **84**, 1099-1114.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 23, No. 1, 2025 (New Series), pp 33–53
https://www.ssca.org.in/journal

R-optimal Designs for Gamma Regression Model with Two Parameters

Mahesh Kumar Panda[1], **Tofan Kumar Biswal**[2] **and V. K. Gupta**[3]
[1]*Department of Statistics, Ravenshaw University, Cuttack-753003*
[2]*Department of Statistics, Central University of Odisha, Sunabeda-763004*
[3]*Formerly at ICAR-IASRI, New Delhi 110012*

Received: 2 September 2023; Revised: 23 February 2024; Accepted: 12 April 2024

Abstract

This article finds locally R-optimal designs for the gamma regression model having two parameters using the inverse link function. The R-optimality criterion has been proposed in the literature as an alternative criterion to the well-known D-optimality criterion when the target is to minimize the volume of the confidence region for unknown parameters based on the Bonferroni t-intervals. The optimality of the proposed designs is confirmed using the corresponding equivalence theorem.

Key words: Locally R-optimal design; Gamma regression model; Inverse link function; Bonferroni t-intervals; Equivalence theorem.

AMS Subject Classifications: 62K05

1. Introduction

The Generalized Linear Model (GLM), introduced by Nelder and Wedderburn (1972) is a generalized version of the ordinary linear regression model. The GLM has extensive applications in various disciplines of science such as clinical trials, engineering, reliability, survival analysis, image analysis, bioinformatics, economics, insurance, agriculture, and industry. For more details on the applications of GLM, one can refer to the articles of Bailey *et al.* (1960), Myers and Montgomery (1997), de Jong and Heller (2008), Fox (2015), and Goldburd (2016).

The Gamma regression model is a particular form of GLM. This model is useful when the responses are continuous, non-negative, and right-skewed type. There are many instances in the literature where the gamma model with an appropriate link function has been used to analyze the real data. The data analysis of car insurance claims (*pg.* 296, McCullagh and Nelder, 1989) and clotting times of blood (*pg.* 300, McCullagh and Nelder, 1989) was carried out by fitting a first-order Gamma model with the natural link function. Anderson *et al.* (2010) used a first-order gamma model with a natural link function to analyze the reaction time taken by the elders to recognize words on a computer monitor.

Corresponding Author: Mahesh Kumar Panda
Email: mahesh2123ster@gmail.com

In experimental design, the target for constructing an optimal design is to make the predicted response closer to the mean response over a certain region of interest based on a specific criterion of interest. For the seminal work on optimal designs, one can refer to the work of Kiefer and Wolfowitz (1959), and Kiefer (1959). In the case of GLM, finding the optimal designs becomes a very difficult task because the optimal design depends on the unknown values of the model parameters. In this context, Chernoff (1953) proposed an alternative way of finding optimal design by starting with an initial guess value of parameter values that can lead to locally optimal designs.

Ford *et al.* (1992) obtained a locally D-optimal design for the Gamma regression model that involves a single factor. Subsequently, Burridge and Sebastiani(1992) found the locally D-optimal design for the Gamma model with two factors but without an intercept. Burridge and Sebastiani (1994) obtained the same D-optimal design for the Gamma regression model which involves multiple factors. Aminenjad and Jafari (2017) found Bayesian A- and D-optimal designs for the Gamma model with inverse link function by considering various prior distributions such as Normal, Half-normal, Gamma, and Uniform distributions. Gaffke *et al.* (2019) provided analytical solutions to derive locally D- and A-optimal designs for the Gamma models that involve intercept terms. They also established that the derived designs are essentially a complete class of designs. Idais and Schwabe (2021) found locally D- and A-optimal designs for the Gamma models having linear predictors without intercept. Idais (2021) obtained D-, A-, and Kiefer's Φ_k -criteria optimality for vertex-type designs.

In experimental design, the D-optimality criterion is the most widely used optimal design criterion. The geometrical interpretation of the D-optimality criterion is to minimize the volume of the confidence ellipsoid region of the unknown parameters (*see* Silvey, 1980). However, computation of the D-optimal design for a regression model becomes simple if the number of parameters associated with the given model is small, let's say 2 or 3. In this perspective, an alternative design known as the R-optimal design was introduced by Dette (1997). This design aims at minimizing the volume of the Bonferroni t-intervals. Recently, many authors have obtained R-optimal designs for different types of regression models *e.g.*, second-order response surface models (Liu *et al.*, 2016), multi-factor models with heteroscedastic errors (He and Yue, 2017), multi-response regression models with multiple factors (Liu *et al.*, 2022), and models with mixture experiments (Panda, 2021; Panda and Sahoo, 2024). To the best of our knowledge, the construction of R-optimal designs for GLM has not been discussed yet in the literature except for the work of Panda and Biswal (2024). In this context, the present article aims to construct locally R-optimal designs for the Gamma Model with two parameters including the intercept parameter.

The rest of the article is organized as follows. Section 2 provides the model specification as well as brief details on locally R-optimal designs. In Section 3, we obtain R-optimal designs for the Gamma model with two parameters. Finally, the article is concluded with some discussions and conclusions in Section 4.

2. Model specification and locally R-optimal designs

Let the response variables $Y_1, Y_2, \ldots, Y_n$ are assumed to be independent gamma-distributed random variables *i.e.,* the probability density function (*p.d.f.*) of each Y_i

$$p(y_i;\nu) = \frac{1}{\Gamma(\nu)} y_i^{\nu-1} e^{-y_i},\ y_i,\ \nu > 0,\ i = 1, 2, \ldots, n\,. \tag{1}$$

Here ν is the shape parameter associated with the *p.d.f* as specified in equation (1). It is assumed to be known and the same for all y_i. However the expected value *i.e.* μ_i depends on the values of $\boldsymbol{x_i}$ the covariate of $\boldsymbol{x}$. The canonical link for the Gamma distribution given by Equation (1) is the inverse link function defined as

$$\eta_i = \frac{\nu}{\mu_i},\ \ where\ \ \eta_i = \boldsymbol{g'(x_i)\beta},\ i = 1, 2, \ldots, n \tag{2}$$

is the linear predictor. In Equation (2), $\boldsymbol{g} = [g_1, g_2, \ldots, g_p]'$ is a p-dimensional vector valued function defined on a domain set $\Xi \subset \mathbb{R}^t, t \geq 1$. Here the component functions $g_1, g_2, \ldots, g_p$ are assumed to be linearly independent, and $\boldsymbol{\beta} \in \mathbb{R}^p$ are assumed to be a p-dimensional vector consisting of unknown parameters associated with the model Equation (2).

In this case, the variance function of the gamma distribution is $Var(Y) = \nu^{-1}\mu^2$ therefore the intensity function at a particular point $\boldsymbol{x} \in \Xi$ (*see* Atkison and Woods, 2015) can be defined as

$$u_0(\boldsymbol{x}, \boldsymbol{\beta}) = \left(Var(Y) \left(\frac{d\eta}{d\mu} \right)^2 \right)^{-1} = \nu(\boldsymbol{g'(x)\beta})^{-2}. \tag{3}$$

As the gamma-distributed responses are continuous and non-negative and thus for a given experimental region Ξ we assume throughout that the parameter vector $\boldsymbol{\beta}$ satisfies

$$\boldsymbol{g'(x)\beta} > 0\ \ for\ all\ \ \boldsymbol{x} \in \Xi\,. \tag{4}$$

For the model Equation (2), the Fisher information matrix at $\boldsymbol{x}$ and $\boldsymbol{\beta}$ can be defined as

$$\boldsymbol{M(x, \beta)} = u(\boldsymbol{x, \beta})\boldsymbol{g(x)g'(x)} \quad where \quad u(\boldsymbol{x}, \beta) = (\boldsymbol{g'(x)\beta})^{-2}. \tag{5}$$

For more details about the assumption made in Equation (4) and the information matrix defined in Equation (5), one can refer to the articles of Gaffke *et al.* (2019) and Idais *et al.* (2021).

For a given parameter value, let us define $\boldsymbol{g}_\beta$ as the local regression function then

$$\boldsymbol{g}_\beta(\boldsymbol{x}) = (\boldsymbol{g'(x)\beta})^{-1}\boldsymbol{g(x)}\ \ forall\ \ \boldsymbol{x} \in \Xi\,. \tag{6}$$

Using Equation (6), the Fisher information matrix in model Equation (5) can be rewritten as

$$\boldsymbol{M(x, \beta)} = \boldsymbol{g(x)g'(x)}. \tag{7}$$

To obtain the R-optimal design for the model Equation (2), we consider the approximate design $\xi \in \Omega$ (Ω the set of all approximate designs) of the form

$$\xi = \begin{Bmatrix} \boldsymbol{x}_1 & \dots & \boldsymbol{x}_s \\ w_1 & \dots & w_s \end{Bmatrix}, \;\; w_i(>0) \;\; and \;\; \sum_{i=1}^{s} w_i = 1 \tag{8}$$

where $\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_s \in \Xi$ are the 's' distinct points and w_i is the weight associated with the point $\boldsymbol{x_i}$ for $i = 1, 2, \dots, s$. For the model Equation (2), the Fisher information matrix of a design ξ at parameter vector $\boldsymbol{\beta}$ is defined as

$$\boldsymbol{M}(\xi, \boldsymbol{\beta}) = \sum_{i=1}^{s} w_i \boldsymbol{M}(\boldsymbol{x}_i, \boldsymbol{\beta}). \tag{9}$$

R-optimal design: A design $\xi \in \Omega$ with a non-singular information matrix $\boldsymbol{M}(\xi)$ is called R-optimal for the model Equation (2) if it minimizes

$$\phi(\xi) = \prod_{i=1}^{p} (\boldsymbol{M}^{-1}(\xi))_{ii} = \prod_{i=1}^{p} \boldsymbol{e_i'} \boldsymbol{M}^{-1}(\xi) \boldsymbol{e_i} \tag{10}$$

for all $\xi \in \Omega$. Here $\boldsymbol{e_i}$ denotes the i^{th} unit vector in $\mathbb{R}^p$, where p is the number of unknown parameters associated with the model Equation (2). The necessary and sufficient conditions for the R-optimality will be examined using the following equivalence theorem. For further details, one can refer to the article of Dette (1997).

Theorem 1: For model Equation (2), let

$$\varphi(\boldsymbol{x}, \xi) = \boldsymbol{g'(x)} \boldsymbol{M}^{-1}(\xi) \left(\sum_{i=1}^{p} \frac{\boldsymbol{e_i} \boldsymbol{e_i'}}{\boldsymbol{e_i} \boldsymbol{M}^{-1}(\xi^*) \boldsymbol{e_i'}} \right) \boldsymbol{M}^{-1}(\xi) \boldsymbol{g(x)}. \tag{11}$$

A design $\xi^* \in \Omega$ is R-optimal if and only if

$$\sup_{\boldsymbol{x} \in \Xi} \; \varphi(\boldsymbol{x}, \xi^*) = p$$

with equality attained at the support points of ξ^*.

3. R-optimal designs

In this section, we obtain locally R-optimal designs for the model Equation (2) that involves two unknown parameters including the intercept parameter. Thus the assumption in Equation (4) becomes

$$\boldsymbol{g'}(x)\boldsymbol{\beta} = \beta_0 + \beta_1 x > 0$$

for all $x \in R$. Here, we restrict our search to two-, three-, and four-support points design by considering discrete values of β_0 and β_1 in the arbitrarily chosen intervals $[0, 10]$ and $[0, 100]$ respectively.

3.1. Design based on two support points

Let us consider a 2-point design ξ of the form

$$\xi = \begin{Bmatrix} a & b \\ w & 1-w \end{Bmatrix}, \quad \text{where} \;\; 0 < w < 1. \tag{12}$$

Theorem 2: The design ξ^* that assigns a weight of w^* to the point a^* and $1 - w^*$ to the point b^* in $\mathbb{R}$ [the numerical values of a^*, b^* and w^* are given in Table 1 (Appendix-I)] is an R-optimal design.

Proof: Using Equation (9), the information matrix for the model Equation (2) at the two-point design ξ will be

$$\boldsymbol{M}(\xi) = \begin{bmatrix} \dfrac{1-w}{(\beta_0+b\beta_1)^2} + \dfrac{w}{(\beta_0+a\beta_1)^2} & \dfrac{b(1-w)}{(\beta_0+b\beta_1)^2} + \dfrac{aw}{(\beta_0+a\beta_1)^2} \\ \dfrac{b(1-w)}{(\beta_0+b\beta_1)^2} + \dfrac{aw}{(\beta_0+a\beta_1)^2} & \dfrac{b^2(1-w)}{(\beta_0+b\beta_1)^2} + \dfrac{a^2w}{(\beta_0+a\beta_1)^2} \end{bmatrix}.$$

The inverse of the information matrix $\boldsymbol{M}(\xi)$ is given by

$$\boldsymbol{M}^{-1}(\xi) = \begin{bmatrix} m_{11} & m_{12} \\ m_{21} & m_{22} \end{bmatrix} \tag{13}$$

where

$$m_{11} = \frac{-b^2(\beta_0+a\beta_1)^2 + (b-a)\beta_0((a+b)\beta_0 + 2ab\beta_1)w}{(a-b)^2(-1+w)w},$$

$$m_{12} = m_{21} = \frac{b(\beta_0+a\beta_1)^2 + (b-a)(-\beta_0^2 + ab\beta_1^2)w}{(a-b)^2(-1+w)w},$$

and $$m_{22} = \frac{-(\beta_0+a\beta_1)^2 + (a-b)\beta_1(2\beta_0 + (a+b)\beta_1)w}{(a-b)^2(-1+w)w}.$$

Using Equation (10), we obtain

$$\phi(\xi) = \frac{\left[\left\{-b^2(\beta_0+a\beta_1)^2 + (b-a)\beta_0((a+b)\beta_0 + 2ab\beta_1)w\right\} \times \left\{-(\beta_0+a\beta_1)^2 + (a-b)\beta_1(2\beta_0 + (a+b)\beta_1)w\right\}\right]}{(a-b)^4(-1+w)^2w^2}. \tag{14}$$

Now, the problem is to minimize the function $\phi(\xi)$ *w.r.t* a, b and w for given values of β_0 and β_1. This is done using the "fminsearch" function of Matlab software and getting the optimal values a^*, b^* and w^* by discrete values of β_0 and β_1 in the arbitrarily chosen intervals $[0, 10]$ and $[0, 100]$ respectively. The numerical values a^*, b^* and w^* are given in Table 1 (Appendix-I).

Next, by using Equation (13) we derive the quadratic form as specified in Equation (11) which is as follows:

$$\varphi(\boldsymbol{x},\xi) = \frac{1}{(\beta_0+\beta_1 x)^2}\Big\{m_{11}+m_{12}x+\frac{(b(\beta_0+a\beta_1)^2+(b-a)(-\beta_0^2+ab\beta_1^2)w)(m_{12}+m_{22}x)}{-(\beta_0+a\beta_1)^2+(a-b)\beta_1(2\beta_0+(a+b)\beta_1)w} + x\left(m_{12}+m_{22}x+\frac{(b(\beta_0+a\beta_1)^2+(b-a)(-\beta_0^2+ab\beta_1^2)w)(m_{11}+m_{12}x)}{-b^2(\beta_0+a\beta_1)^2+(b-a)\beta_0((a+b)\beta_0+2ab\beta_1)w}\right)\Big\}. \tag{15}$$

Replacing the numerical values of a^*, b^* and w^* in Equation (15) and using the "fminsearch" function of Matlab software we find $\sup_{\boldsymbol{x}\in\boldsymbol{R}} \varphi(\boldsymbol{x},\xi^*) = 2$. Thus the necessary and sufficient condition of the equivalence theorem is established. This proves Theorem 2. □

3.2. Design based on three support points

Let us consider a 3-point design ξ of the form

$$\xi = \begin{Bmatrix} a & b & c \\ w/2 & 1-w & w/2 \end{Bmatrix}, \quad where \;\; 0<w<1. \tag{16}$$

Theorem 3: The design ξ^* that assigns a weight of $w^*/2$ to the point a^*, $1-w^*$ to the point b^*, and $w^*/2$ to the point c^* in $\mathbb{R}$ [the numerical values of a^*, b^*, c^* and w^* are given in Table 2 (Appendix-I)] is an R-optimal design.

Proof: Using Equation (9), the information matrix for the model Equation (2) at the three-point design ξ will be

$$\boldsymbol{M}(\xi) = \begin{bmatrix} \frac{1-w}{(\beta_0+b\beta_1)^2}+\frac{w}{2(\beta_0+a\beta_1)^2}+\frac{w}{2(\beta_0+c\beta_1)^2} & \frac{b(1-w)}{(\beta_0+b\beta_1)^2}+\frac{aw}{2(\beta_0+a\beta_1)^2}+\frac{cw}{2(\beta_0+c\beta_1)^2} \\ \frac{b(1-w)}{(\beta_0+b\beta_1)^2}+\frac{aw}{2(\beta_0+a\beta_1)^2}+\frac{cw}{2(\beta_0+c\beta_1)^2} & \frac{b^2(1-w)}{(\beta_0+b\beta_1)^2}+\frac{a^2w}{2(\beta_0+a\beta_1)^2}+\frac{c^2w}{2(\beta_0+c\beta_1)^2} \end{bmatrix}.$$

The inverse of the information matrix $\boldsymbol{M}(\xi)$ is given by

$$\boldsymbol{M}^{-1}(\xi) = \begin{bmatrix} m_{11}^* & m_{12}^* \\ m_{21}^* & m_{22}^* \end{bmatrix} \tag{17}$$

where

$$m_{11}^* = \frac{\alpha_1}{\alpha_2+(\alpha_3\times\alpha_4)},$$

$$m_{12}^* = m_{21}^* = \frac{\alpha_5}{\alpha_2+(\alpha_3\times\alpha_4)},$$

and $\; m_{22}^* = \dfrac{2\alpha_6}{\alpha_2+(\alpha_3\times\alpha_4)},$

with

$$\alpha_1 = 2\left(-\frac{2b^2(1-w)}{(\beta_0+b\beta_1)^2}+\frac{a^2w}{(\beta_0+a\beta_1)^2}+\frac{c^2w}{(\beta_0+c\beta_1)^2}\right),$$

$$\alpha_2 = -\left(-\frac{2b(w-1)}{(\beta_0+b\beta_1)^2}+\frac{aw}{(\beta_0+a\beta_1)^2}+\frac{cw}{(\beta_0+c\beta_1)^2}\right)^2,$$

$$\alpha_3 = \left(-\frac{2(w-1)}{(\beta_0+b\beta_1)^2}+\frac{w}{(\beta_0+a\beta_1)^2}+\frac{w}{(\beta_0+c\beta_1)^2}\right),$$

$$\alpha_4 = \left(-\frac{2b^2(w-1)}{(\beta_0+b\beta_1)^2}+\frac{a^2w}{(\beta_0+a\beta_1)^2}+\frac{c^2w}{(\beta_0+c\beta_1)^2}\right),$$

$$\alpha_5 = 4\left(\frac{b(w-1)}{(\beta_0+b\beta_1)^2}+\frac{1}{2}\left(-\frac{a}{(\beta_0+a\beta_1)^2}-\frac{c}{(\beta_0+c\beta_1)^2}\right)w\right),$$

and $$\alpha_6 = \left(-\frac{2(1-w)}{(\beta_0+b\beta_1)^2}+\frac{w}{(\beta_0+a\beta_1)^2}+\frac{w}{(\beta_0+c\beta_1)^2}\right).$$

Using Equation (10), we obtain the function

$$\phi(\xi) = \frac{4(\alpha_3\times\alpha_4)}{\{\alpha_2-(\alpha_3\times\alpha_4)\}^2}\,. \tag{18}$$

Next, we need to minimize the function $\phi(\xi)$ *w.r.t* a, b, c and w for given values of β_0 and β_1. This is achieved by using the "fminsearch" function of Matlab software and getting the optimal values a^*, b^*, c^* and w^* by discrete values of β_0 and β_1 in the arbitrarily chosen intervals $[0,10]$ and $[0,100]$ respectively. The numerical values a^*, b^*, c^* and w^* are given in Table 2 (Appendix-I).

Next, by using Equation (17) we derive the quadratic form as specified in Equation (11) which is as follows:

$$\begin{aligned}\varphi(\boldsymbol{x},\xi) = \frac{1}{(\beta_0+b\beta_1)^2}\Bigg\{&m_{11}^*+m_{21}^*x+\left(\frac{2(\alpha_7)(m_{21}^*+m_{22}^*x)}{(\alpha_6)}\right)\\&+x\left(m_{21}^*+m_{22}^*x+\left(\frac{2(\alpha_7)(m_{11}^*+m_{21}^*x)}{(\alpha_4)}\right)\right)\Bigg\}\end{aligned} \tag{19}$$

with

$$\alpha_7 = \frac{b(w-1)}{(\beta_0+b\beta_1)^2}+\frac{1}{2}\left(-\frac{a}{(\beta_0+a\beta_1)^2}-\frac{c}{(\beta_0+c\beta_1)^2}\right)w.$$

Replacing the numerical values of a^*, b^*, c^* and w^* in Equation (19) and using the "fminsearch" function in Matlab software we find $\sup_{\boldsymbol{x}\in\boldsymbol{R}}\varphi(\boldsymbol{x},\xi^*) = 2$. Thus the necessary and sufficient condition of the equivalence theorem is established. This proves Theorem 3. □

3.3. Design based on four support points

Let us consider a 4-point design ξ of the form

$$\xi = \begin{Bmatrix} a & b & c & d \\ w & \left(\frac{1}{2} - w\right) & \left(\frac{1}{2} - w\right) & w \end{Bmatrix}, \quad where \;\; 0 < w < 1. \tag{20}$$

Theorem 4: The design ξ^* that assigns a weight of w^* to the point a^*, $(1/2) - w^*$ to the point b^*, $(1/2) - w^*$ to the point c^* and w^* to the point d^* in $\mathbb{R}$ [the numerical values of a^*, b^*, c^*, d^* and w^* are given in Table 3 (Appendix-I)] is an R-optimal design.

Proof: Using Equation (9), the information matrix for the model Equation (2) at the four-point design ξ will be

$$\boldsymbol{M}(\xi) = \begin{bmatrix} \lambda_1 & \lambda_2 \\ \lambda_2 & \lambda_3 \end{bmatrix}$$

where

$$\lambda_1 = \frac{\frac{1}{2} - w}{(\beta_0 + b\beta_1)^2} + \frac{\frac{1}{2} - w}{(\beta_0 + c\beta_1)^2} + \frac{w}{(\beta_0 + a\beta_1)^2} + \frac{w}{(\beta_0 + d\beta_1)^2},$$

$$\lambda_2 = \frac{b\left(\frac{1}{2} - w\right)}{(\beta_0 + b\beta_1)^2} + \frac{c\left(\frac{1}{2} - w\right)}{(\beta_0 + c\beta_1)^2} + \frac{aw}{(\beta_0 + a\beta_1)^2} + \frac{dw}{(\beta_0 + d\beta_1)^2},$$

and $$\lambda_3 = \frac{b^2\left(\frac{1}{2} - w\right)}{(\beta_0 + b\beta_1)^2} + \frac{c^2\left(\frac{1}{2} - w\right)}{(\beta_0 + c\beta_1)^2} + \frac{a^2 w}{(\beta_0 + a\beta_1)^2} + \frac{d^2 w}{(\beta_0 + d\beta_1)^2}.$$

The inverse of the information matrix $\boldsymbol{M}(\xi)$ is given by

$$\boldsymbol{M}^{-1}(\xi) = \begin{bmatrix} m_{11}^+ & m_{12}^+ \\ m_{21}^+ & m_{22}^+ \end{bmatrix} \tag{21}$$

with

$$m_{11}^+ = \frac{\lambda_3}{-(\lambda_2)^2 + (\lambda_1 \times \lambda_3)},$$

$$m_{12}^+ = m_{21}^+ = \frac{\lambda_4}{2\{-(\lambda_2)^2 + (\lambda_1 \times \lambda_3)\}},$$

and $$m_{22}^+ = \frac{\lambda_1}{-(\lambda_2)^2 + (\lambda_1 \times \lambda_3)} \; .$$

Using Equation (10), we obtain the function

$$\phi(\xi) = \frac{\lambda_1 \times \lambda_3}{\{(\lambda_2)^2 - (\lambda_1 \times \lambda_3)\}^2} \; . \tag{22}$$

Now, the problem reduces to minimizing the function $\phi(\xi)$ *w.r.t* a, b, c, d and w for given values of β_0 and β_1. This is achieved by using the "fminsearch" function of Matlab software and getting the optimal values a^*, b^*, c^*, d^* and w^* by discrete values of β_0 and β_1 in the arbitrarily chosen intervals $[0, 10]$ and $[0, 100]$ respectively. The numerical values a^*, b^*, c^*, d^* and w^* are given in Table 3 (Appendix-I).

Next, by using Equation (21) we derive the quadratic form as specified in Equation (11) which is as follows:

$$\varphi(\boldsymbol{x}, \xi) = \frac{1}{(\beta_0 + \beta_1 x)^2}\Bigg\{m_{11}^{+} + m_{12}^{+}x + \frac{\lambda_4(m_{12}^{+} + m_{22}^{+}x)}{2\lambda_1} + x\Big(m_{12}^{+} + m_{22}^{+}x + \frac{\lambda_4(m_{11}^{+} + m_{12}^{+}x)}{2\lambda_3}\Big)\Bigg\} \tag{23}$$

with $\lambda_4 = \dfrac{-2aw}{(\beta_0 + a\beta_1)^2} + \dfrac{-2dw}{(\beta_0 + d\beta_1)^2} + \dfrac{b(2w-1)}{(\beta_0 + b\beta_1)^2} + \dfrac{c(2w-1)}{(\beta_0 + c\beta_1)^2}$.

Replacing the numerical values of a^*, b^*, c^*, d^* and w^* in Equation (23) and using the "fminsearch" function of Matlab software we find $\sup_{\boldsymbol{x}\in\boldsymbol{R}} \varphi(\boldsymbol{x}, \xi^*) = 2$. Thus the necessary and sufficient condition of the equivalence theorem is established. This proves Theorem 4. □

4. Discussion and conclusion

This article finds locally R-optimal designs for two parameters Gamma regression model when the model is associated with inverse link function based on two-, three-, and four-support point designs. The support points of the optimal designs and the weights assigned to these points are calculated numerically using the "fminsearch" function of Matlab software whereas the necessary and sufficient condition of R-optimality *i.e.* the equivalence theorem is also established at the support points of the R-optimal design using "fminsearch" function of Matlab software. A catalog of support points and the weight assigned to each of the support points corresponding to R-optimal designs are listed in Table 1, Table 2, and Table 3 (Appendix I). These Tables provide the solutions for only those values of β_0 and β_1 when the equivalence theorem is satisfied.

The present work considers three types of designs : (i) two-point designs (ii) three-point designs where equal weights are assigned to one-pair of support points (iii) four-point designs where equal weights are assigned to two-pair of support points. In all these cases, we observe that the equivalence theorem does not hold for many discrete values of the unknown parameters which indicates that the proposed designs are sensitive towards the R-optimality criterion with the varying parameter choices. However, when we relax the assumption of equal weights the optimal search criterion does not converge to any solution as the problem becomes complicated with an increase in the number of unknown entities (support points and weights). Therefore, more research work is required especially to propose an alternative optimal search criterion that can converge to a finite solution that satisfies the weight restriction as well. Nevertheless, the present work provides the necessary motivation to find the solution of local R-optimal designs for GLM when the parameters take continuous values.

For the two-support points design, we find that the support points lie in the third

quadrant of the two-dimensional space. The values of the first coordinate and second coordinate of the support points are approximately equal.

One can extend this idea to obtain R-optimal designs for the Gamma model with more than two parameters.

Conflict of interest

The authors do not have any financial or non-financial conflict of interest to declare for the research work included in this article.

References

Aminnejad, M. and Jafari, H. (2017). Bayesian A-and D-optimal designs for gamma regression model with inverse link function. *Communications in Statistics-Simulation and Computation*, **46**, 8166-8189.

Anderson, C. J., Verkuilen, J., and Johnson, T. (2010). *Applied Generalized Linear Mixed Models: Continuous and Discrete Data*. Springer, New York.

Atkinson, A. C. and Woods, D. C. (2015). Designs for Generalized Linear Models. *Handbook of Design and Analysis of Experiments*, **7**, 471-514.

Bailey, R. A. and Simon, L. J. (1960). Two studies in automobile insurance ratemaking. *ASTIN Bulletin: The Journal of the IAA*, **1**, 192-217.

Burridge, J. and Sebastiani, P. (1992). Optimal designs for generalized linear models. *Journal of the Italian Statistical Society*, **1**, 183-202.

Burridge, J. and Sebastiani, P. (1994). D-optimal designs for generalized linear models with variance proportional to the square of the mean. *Biometrika*, **81**, 295-304.

Chernoff, H. (1953). Locally optimal designs for estimating parameters. *The Annals of Mathematical Statistics*, 586-602.

De Jong, P. and Heller, G. Z. (2008). *Generalized Linear Models for Insurance Data*. Cambridge University Press.

Dette, H. (1997). Designing experiments with respect to some 'standardized' optimality criteria. *Journal of Royal Statistical Society, Series B (Methodological)*, **59**, 97-110.

Ford, I., Torsney, B., and Wu, C. J. (1992). The use of a canonical form in the construction of locally optimal designs for non-linear problems. *Journal of the Royal Statistical Society, Series B (Methodological)*, **54**, 569-583.

Fox, J. (2015). *Applied Regression Analysis and Generalized Linear Models*. Sage Publications.

Gaffke, N., Idais, O., and Schwabe, R. (2019). Locally optimal designs for gamma models. *Journal of Statistical Planning and Inference*, **203**, 199-214.

Goldburd, M., Khare, A., Tevet, D., and Guller, D. (2016). Generalized linear models for insurance rating. *Casualty Actuarial Society, CAS Monographs Series*, **5**.

He, L. and Yue, R. X. (2017). R-optimal designs for multi-factor models with heteroscedastic errors. *Metrika*, **80**, 717-732.

He, L. and Yue, R. X. (2019). R-optimality criterion for regression models with asymmetric errors. *Journal of Statistical Planning and Inference*, **199**, 318-326.

Idais, O. (2021). On local optimality of vertex type designs in generalized linear models. *Statistical Papers*, **62**, 1871-1898.

Idais, O. and Schwabe, R. (2021). Analytic solutions for locally optimal designs for gamma models having linear predictors without intercept. *Metrika*, **84**, 1-26.

Kiefer, J. (1959). Optimum experimental designs. *Journal of the Royal Statistical Society: Series B (Methodological)*, **21**, 272-304.

Kiefer, J. and Wolfowitz, J. (1959). Optimal designs in regression problems. *The Annals of Mathematical Statistics*, **30**, 271-294.

Liu, X., Yue, R. X., Xu, J., and Chatterjee, K. (2016). Algorithmic construction of R-optimal designs for second-order response surface models. *Journal of Statistical Planning and Inference*, **178**, 61-69.

Liu, P., Gao, L. L., and Zhou, J. (2022). R-optimal designs for multi-response regression models with multi-factors. *Communications in Statistics-Theory and Methods*, **51**, 340-355.

McCullough, P. and Nelder, J. A. (1989). *Generalized Linear Models*. Chapman and Hall/CRC, 2nd ed.

Myers, R. H. and Montgomery, D. C. (1997). A tutorial on generalized linear models. *Journal of Quality Technology*, **29**, 274-291.

Nelder, J. A. and Wedderburn, R. W. (1972). Generalized linear models. *Journal of the Royal Statistical Society Series A: Statistics in Society*, **135**, 370-384.

Panda, M. K. (2021). R-Optimal Designs for Canonical Polynomial Models with Mixture Experiments. *Calcutta Statistical Association Bulletin*, **73**, 146-161.

Panda, M. K. and Sahoo, R. P. (2024). R-optimal designs for linear log contrast model with mixture experiments. *Communications in Statistics-Theory and Methods*, **53**, 2355-2368.

Panda, M. K. and Biswal, T. K. (2024). R-optimal designs for logistic regression model with two variables. *Statistics and Applications*, *Accepted*.

Silvey, S. D. (1980). *Optimal Design*. Chapman and Hall/CRC.

Appendix-I

Table 1: R-optimal design for Gamma regression model with two parameters (Two support points)

$\boldsymbol{\beta}$	$\beta_0 = 1,\ \beta_1 = 2$	$\beta_0 = 1,\ \beta_1 = 3$	$\beta_0 = 1,\ \beta_1 = 4$	$\beta_0 = 1,\ \beta_1 = 5$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.5000 & -0.4999 \\ 0.1057 & 0.8943 \end{pmatrix}$	$\begin{pmatrix} -0.3333 & -0.3333 \\ 0.3042 & 0.6958 \end{pmatrix}$	$\begin{pmatrix} -0.2500 & -0.2499 \\ 0.5512 & 0.4488 \end{pmatrix}$	$\begin{pmatrix} -0.2000 & -0.1999 \\ 0.6098 & 0.3092 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 1,\ \beta_1 = 6$	$\beta_0 = 1,\ \beta_1 = 7$	$\beta_0 = 1,\ \beta_1 = 8$	$\beta_0 = 1,\ \beta_1 = 9$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.1666 & -0.1666 \\ 0.6084 & 0.3916 \end{pmatrix}$	$\begin{pmatrix} -0.1428 & -0.1428 \\ 0.6067 & 0.3933 \end{pmatrix}$	$\begin{pmatrix} -0.1250 & -0.1250 \\ 0.5672 & 0.4328 \end{pmatrix}$	$\begin{pmatrix} -0.1111 & -0.1111 \\ 0.5824 & 0.4176 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 1,\ \beta_1 = 10$	$\beta_0 = 1,\ \beta_1 = 11$	$\beta_0 = 1,\ \beta_1 = 12$	$\beta_0 = 1,\ \beta_1 = 13$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.1000 & -0.0999 \\ 0.6891 & 0.3109 \end{pmatrix}$	$\begin{pmatrix} -0.0909 & -0.0909 \\ 0.6094 & 0.3906 \end{pmatrix}$	$\begin{pmatrix} -0.0833 & -0.0833 \\ 0.6905 & 0.3095 \end{pmatrix}$	$\begin{pmatrix} -0.0769 & -0.0769 \\ 0.6137 & 0.3863 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 1,\ \beta_1 = 14$	$\beta_0 = 1,\ \beta_1 = 15$	$\beta_0 = 2,\ \beta_1 = 4$	$\beta_0 = 2,\ \beta_1 = 5$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.0714 & -0.0714 \\ 0.6344 & 0.3656 \end{pmatrix}$	$\begin{pmatrix} -0.0666 & -0.0666 \\ 0.6135 & 0.3865 \end{pmatrix}$	$\begin{pmatrix} -0.5000 & -0.4998 \\ 0.1058 & 0.8942 \end{pmatrix}$	$\begin{pmatrix} -0.4000 & -0.3999 \\ 0.2226 & 0.7774 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 2,\ \beta_1 = 6$	$\beta_0 = 2,\ \beta_1 = 7$	$\beta_0 = 2,\ \beta_1 = 8$	$\beta_0 = 2,\ \beta_1 = 9$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.3333 & -0.3333 \\ 0.3043 & 0.6957 \end{pmatrix}$	$\begin{pmatrix} -0.2857 & -0.2856 \\ 0.1965 & 0.8035 \end{pmatrix}$	$\begin{pmatrix} -0.2500 & -0.2499 \\ 0.5512 & 0.4488 \end{pmatrix}$	$\begin{pmatrix} -0.2222 & -0.2222 \\ 0.5871 & 0.4129 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 2,\ \beta_1 = 10$	$\beta_0 = 2,\ \beta_1 = 11$	$\beta_0 = 2,\ \beta_1 = 12$	$\beta_0 = 2,\ \beta_1 = 13$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.2000 & -0.1999 \\ 0.6098 & 0.3902 \end{pmatrix}$	$\begin{pmatrix} -0.1818 & -0.1818 \\ 0.6003 & 0.3997 \end{pmatrix}$	$\begin{pmatrix} -0.1666 & -0.1666 \\ 0.6148 & 0.3852 \end{pmatrix}$	$\begin{pmatrix} -0.1538 & -0.1538 \\ 0.6397 & 0.3603 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 2,\ \beta_1 = 14$	$\beta_0 = 2,\ \beta_1 = 15$	$\beta_0 = 3,\ \beta_1 = 6$	$\beta_0 = 3,\ \beta_1 = 7$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.1428 & -0.1428 \\ 0.6067 & 0.3933 \end{pmatrix}$	$\begin{pmatrix} -0.1333 & -0.1333 \\ 0.6100 & 0.3900 \end{pmatrix}$	$\begin{pmatrix} -0.5000 & -0.4997 \\ 0.1058 & 0.8942 \end{pmatrix}$	$\begin{pmatrix} -0.4285 & -0.4285 \\ 0.1702 & 0.8298 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 3,\ \beta_1 = 8$	$\beta_0 = 3,\ \beta_1 = 9$	$\beta_0 = 3,\ \beta_1 = 10$	$\beta_0 = 3,\ \beta_1 = 11$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.3750 & -0.3749 \\ 0.2920 & 0.7080 \end{pmatrix}$	$\begin{pmatrix} -0.3333 & -0.3333 \\ 0.3043 & 0.6957 \end{pmatrix}$	$\begin{pmatrix} -0.3000 & -0.2999 \\ 0.3204 & 0.6796 \end{pmatrix}$	$\begin{pmatrix} -0.2727 & -0.2727 \\ 0.6455 & 0.3545 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 3,\ \beta_1 = 12$	$\beta_0 = 3,\ \beta_1 = 13$	$\beta_0 = 3,\ \beta_1 = 14$	$\beta_0 = 3,\ \beta_1 = 15$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.2500 & -0.2499 \\ 0.5512 & 0.4488 \end{pmatrix}$	$\begin{pmatrix} -0.2307 & -0.2307 \\ 0.5830 & 0.4170 \end{pmatrix}$	$\begin{pmatrix} -0.2142 & -0.2142 \\ 0.6218 & 0.3782 \end{pmatrix}$	$\begin{pmatrix} -0.2000 & -0.1999 \\ 0.6098 & 0.3902 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 4,\ \beta_1 = 7$	$\beta_0 = 4,\ \beta_1 = 8$	$\beta_0 = 4,\ \beta_1 = 10$	$\beta_0 = 4,\ \beta_1 = 11$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.5714 & -0.5715 \\ 0.0837 & 0.9163 \end{pmatrix}$	$\begin{pmatrix} -0.4999 & -0.5000 \\ 0.1061 & 0.8939 \end{pmatrix}$	$\begin{pmatrix} -0.4000 & -0.3999 \\ 0.2226 & 0.7774 \end{pmatrix}$	$\begin{pmatrix} -0.3636 & -0.3636 \\ 0.2373 & 0.7627 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 4,\ \beta_1 = 12$	$\beta_0 = 4,\ \beta_1 = 13$	$\beta_0 = 4,\ \beta_1 = 14$	$\beta_0 = 4,\ \beta_1 = 15$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.3333 & -0.3333 \\ 0.3043 & 0.6957 \end{pmatrix}$	$\begin{pmatrix} -0.3076 & -0.3076 \\ 0.3441 & 0.6559 \end{pmatrix}$	$\begin{pmatrix} -0.2857 & -0.2856 \\ 0.1965 & 0.8035 \end{pmatrix}$	$\begin{pmatrix} -0.2666 & -0.2666 \\ 0.5384 & 0.4616 \end{pmatrix}$

Table 1: Continued

$\boldsymbol{\beta}$	$\beta_0 = 5,\ \beta_1 = 10$	$\beta_0 = 5,\ \beta_1 = 11$	$\beta_0 = 5,\ \beta_1 = 12$	$\beta_0 = 5,\ \beta_1 = 13$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.5000 & -0.4997 \\ 0.1058 & 0.8942 \end{pmatrix}$	$\begin{pmatrix} -0.4545 & -0.4544 \\ 0.0894 & 0.9106 \end{pmatrix}$	$\begin{pmatrix} -0.4166 & -0.4167 \\ 0.1951 & 0.8049 \end{pmatrix}$	$\begin{pmatrix} -0.3846 & -0.3846 \\ 0.3048 & 0.6952 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 5,\ \beta_1 = 14$	$\beta_0 = 5,\ \beta_1 = 15$	$\beta_0 = 5,\ \beta_1 = 16$	$\beta_0 = 5,\ \beta_1 = 17$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.3571 & -0.3571 \\ 0.3731 & 0.6269 \end{pmatrix}$	$\begin{pmatrix} -0.3333 & -0.3333 \\ 0.3043 & 0.6957 \end{pmatrix}$	$\begin{pmatrix} -0.3125 & -0.3124 \\ 0.3168 & 0.6832 \end{pmatrix}$	$\begin{pmatrix} -0.2941 & -0.2941 \\ 0.3197 & 0.6803 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 5,\ \beta_1 = 18$	$\beta_0 = 5,\ \beta_1 = 19$	$\beta_0 = 5,\ \beta_1 = 20$	$\beta_0 = 6,\ \beta_1 = 11$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.2777 & -0.2777 \\ 0.5890 & 0.4110 \end{pmatrix}$	$\begin{pmatrix} -0.2631 & -0.2631 \\ 0.6598 & 0.3402 \end{pmatrix}$	$\begin{pmatrix} -0.2500 & -0.2499 \\ 0.5512 & 0.4488 \end{pmatrix}$	$\begin{pmatrix} -0.5454 & -0.5454 \\ 0.0680 & 0.9320 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 6,\ \beta_1 = 12$	$\beta_0 = 6,\ \beta_1 = 13$	$\beta_0 = 6,\ \beta_1 = 14$	$\beta_0 = 6,\ \beta_1 = 15$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.5000 & -0.4997 \\ 0.1058 & 0.8942 \end{pmatrix}$	$\begin{pmatrix} -0.4615 & -0.4615 \\ 0.1378 & 0.8622 \end{pmatrix}$	$\begin{pmatrix} -0.4285 & -0.4285 \\ 0.1702 & 0.8298 \end{pmatrix}$	$\begin{pmatrix} -0.3999 & -0.4000 \\ 0.2226 & 0.7774 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 6,\ \beta_1 = 16$	$\beta_0 = 6,\ \beta_1 = 17$	$\beta_0 = 6,\ \beta_1 = 18$	$\beta_0 = 6,\ \beta_1 = 19$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.3750 & -0.3749 \\ 0.2920 & 0.7080 \end{pmatrix}$	$\begin{pmatrix} -0.3529 & -0.3528 \\ 0.2523 & 0.7477 \end{pmatrix}$	$\begin{pmatrix} -0.3333 & -0.3333 \\ 0.3043 & 0.6957 \end{pmatrix}$	$\begin{pmatrix} -0.3158 & -0.3157 \\ 0.3279 & 0.6721 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 6,\ \beta_1 = 20$	$\beta_0 = 7,\ \beta_1 = 12$	$\beta_0 = 7,\ \beta_1 = 13$	$\beta_0 = 7,\ \beta_1 = 14$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.3000 & -0.2999 \\ 0.3204 & 0.6796 \end{pmatrix}$	$\begin{pmatrix} -0.5833 & -0.5820 \\ 0.0341 & 0.9659 \end{pmatrix}$	$\begin{pmatrix} -0.5384 & -0.5384 \\ 0.0428 & 0.9572 \end{pmatrix}$	$\begin{pmatrix} -0.5000 & -0.4997 \\ 0.1058 & 0.8942 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 7,\ \beta_1 = 16$	$\beta_0 = 7,\ \beta_1 = 17$	$\beta_0 = 7,\ \beta_1 = 18$	$\beta_0 = 7,\ \beta_1 = 19$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.4375 & -0.4374 \\ 0.2406 & 0.7594 \end{pmatrix}$	$\begin{pmatrix} -0.4117 & -0.4117 \\ 0.2248 & 0.7752 \end{pmatrix}$	$\begin{pmatrix} -0.3888 & -0.3888 \\ 0.2329 & 0.7671 \end{pmatrix}$	$\begin{pmatrix} -0.3684 & -0.3684 \\ 0.2640 & 0.7360 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 7,\ \beta_1 = 20$	$\beta_0 = 8,\ \beta_1 = 14$	$\beta_0 = 8,\ \beta_1 = 15$	$\beta_0 = 8,\ \beta_1 = 16$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.3499 & -0.3500 \\ 0.3458 & 0.6542 \end{pmatrix}$	$\begin{pmatrix} -0.5714 & -0.5715 \\ 0.0837 & 0.9163 \end{pmatrix}$	$\begin{pmatrix} -0.5333 & -0.5332 \\ 0.0423 & 0.9577 \end{pmatrix}$	$\begin{pmatrix} -0.4999 & -0.5000 \\ 0.1061 & 0.8939 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 8,\ \beta_1 = 19$	$\beta_0 = 8,\ \beta_1 = 20$	$\beta_0 = 9,\ \beta_1 = 16$	$\beta_0 = 9,\ \beta_1 = 17$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.4210 & -0.4211 \\ 0.1934 & 0.8066 \end{pmatrix}$	$\begin{pmatrix} -0.4000 & -0.3999 \\ 0.2226 & 0.7774 \end{pmatrix}$	$\begin{pmatrix} -0.5625 & -0.5624 \\ 0.1161 & 0.8839 \end{pmatrix}$	$\begin{pmatrix} -0.5294 & -0.5294 \\ 0.0694 & 0.9306 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 9,\ \beta_1 = 18$	$\beta_0 = 9,\ \beta_1 = 19$	$\beta_0 = 9,\ \beta_1 = 20$	$\beta_0 = 10,\ \beta_1 = 19$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.5000 & -0.4997 \\ 0.1058 & 0.8942 \end{pmatrix}$	$\begin{pmatrix} -0.4737 & -0.4733 \\ 0.1565 & 0.8435 \end{pmatrix}$	$\begin{pmatrix} -0.4497 & -0.4525 \\ 0.8914 & 0.1086 \end{pmatrix}$	$\begin{pmatrix} -0.5263 & -0.5261 \\ 0.0598 & 0.9402 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 10,\ \beta_1 = 20$	-	-	-
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.5000 & -0.4997 \\ 0.1058 & 0.8942 \end{pmatrix}$	-	-	-

Table 2: R-optimal design for Gamma regression model with two parameters (Three support points)

$\boldsymbol{\beta}$	$\beta_0 = 1, \beta_1 = 1$	$\beta_0 = 1, \beta_1 = 3$	$\beta_0 = 1, \beta_1 = 4$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.9998 & -0.9791 & -1.0001 \\ 0.4588 & 0.0824 & 0.4588 \end{pmatrix}$	$\begin{pmatrix} -0.2966 & -0.3332 & -0.3333 \\ 0.3704 & 0.2592 & 0.3704 \end{pmatrix}$	$\begin{pmatrix} 2.3673 & -0.2498 & -0.2500 \\ 0.1827 & 0.3646 & 0.1827 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 1, \beta_1 = 5$	$\beta_0 = 1, \beta_1 = 6$	$\beta_0 = 1, \beta_1 = 7$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.1997 & -0.2000 & -0.5475 \\ 0.4745 & 0.0510 & 0.4745 \end{pmatrix}$	$\begin{pmatrix} -0.1664 & -0.1667 & -3.0330 \\ 0.4474 & 0.1052 & 0.4474 \end{pmatrix}$	$\begin{pmatrix} -0.1427 & -0.1428 & -1.9328 \\ 0.4364 & 0.1272 & 0.4364 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 1, \beta_1 = 14$	$\beta_0 = 1, \beta_1 = 15$	$\beta_0 = 2, \beta_1 = 1$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.9991 & -0.9072 & -1.0008 \\ 0.3528 & 0.2944 & 0.3528 \end{pmatrix}$	$\begin{pmatrix} -0.0666 & -0.0666 & -0.4346 \\ 0.3194 & 0.3403 & 0.3194 \end{pmatrix}$	$\begin{pmatrix} 4.1382 & -1.9996 & -2.0001 \\ 0.2301 & 0.5398 & 0.2301 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 2, \beta_1 = 2$	$\beta_0 = 2, \beta_1 = 3$	$\beta_0 = 2, \beta_1 = 6$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.9998 & -0.9791 & -1.0001 \\ 0.4588 & 0.0824 & 0.4588 \end{pmatrix}$	$\begin{pmatrix} -0.3689 & -0.6666 & -0.6667 \\ 0.2737 & 0.4526 & 0.2737 \end{pmatrix}$	$\begin{pmatrix} -0.3027 & -0.3333 & -0.3333 \\ 0.3641 & 0.2718 & 0.3641 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 2, \beta_1 = 7$	$\beta_0 = 2, \beta_1 = 8$	$\beta_0 = 2, \beta_1 = 9$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} 1.2836 & -0.2857 & -0.2857 \\ 0.2417 & 0.5166 & 0.2417 \end{pmatrix}$	$\begin{pmatrix} 2.3673 & -0.2498 & -0.2500 \\ 0.2827 & 0.4346 & 0.2827 \end{pmatrix}$	$\begin{pmatrix} 2.4172 & -0.2221 & -0.2222 \\ 0.1870 & 0.6260 & 0.1870 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 2, \beta_1 = 10$	$\beta_0 = 2, \beta_1 = 12$	$\beta_0 = 2, \beta_1 = 14$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.1999 & -0.2000 & -2.2390 \\ 0.4746 & 0.0508 & 0.4746 \end{pmatrix}$	$\begin{pmatrix} -0.1666 & -0.1666 & -2.1437 \\ 0.4169 & 0.1662 & 0.4169 \end{pmatrix}$	$\begin{pmatrix} -0.1428 & -0.1428 & -1.9329 \\ 0.4364 & 0.1272 & 0.4364 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 3, \beta_1 = 1$	$\beta_0 = 3, \beta_1 = 2$	$\beta_0 = 3, \beta_1 = 3$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.9973 & -2.9995 & -3.0002 \\ 0.2611 & 0.4778 & 0.2611 \end{pmatrix}$	$\begin{pmatrix} -1.4997 & -0.3850 & -1.5002 \\ 0.1797 & 0.6406 & 0.1797 \end{pmatrix}$	$\begin{pmatrix} -0.9991 & -0.9073 & -1.0008 \\ 0.3528 & 0.2944 & 0.3528 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 3, \beta_1 = 4$	$\beta_0 = 3, \beta_1 = 5$	$\beta_0 = 3, \beta_1 = 8$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.7501 & -0.4096 & -0.7498 \\ 0.4127 & 0.1746 & 0.4127 \end{pmatrix}$	$\begin{pmatrix} -0.6063 & -0.5995 & -0.6004 \\ 0.3365 & 0.3270 & 0.3365 \end{pmatrix}$	$\begin{pmatrix} 3.8301 & -0.3707 & -0.3753 \\ 0.0732 & 0.8536 & 0.0732 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 3, \beta_1 = 9$	$\beta_0 = 3, \beta_1 = 11$	$\beta_0 = 3, \beta_1 = 12$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.2966 & -0.3332 & -0.3334 \\ 0.3704 & 0.2592 & 0.3704 \end{pmatrix}$	$\begin{pmatrix} 1.4888 & -0.2727 & -0.2727 \\ 0.2294 & 0.5412 & 0.2294 \end{pmatrix}$	$\begin{pmatrix} 2.3673 & -0.2499 & -0.2500 \\ 0.1827 & 0.6346 & 0.1827 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 3, \beta_1 = 13$	$\beta_0 = 3, \beta_1 = 14$	$\beta_0 = 3, \beta_1 = 15$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} 2.0636 & -0.2306 & -0.2308 \\ 0.2044 & 0.5912 & 0.2044 \end{pmatrix}$	$\begin{pmatrix} 2.4503 & -0.2142 & -0.2143 \\ 0.1891 & 0.6218 & 0.1891 \end{pmatrix}$	$\begin{pmatrix} -0.1999 & -0.2000 & -2.239 \\ 0.4746 & 0.0508 & 0.4746 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 4, \beta_1 = 1$	$\beta_0 = 4, \beta_1 = 2$	$\beta_0 = 4, \beta_1 = 3$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -3.1435 & -4.0012 & -3.9995 \\ 0.2141 & 0.5718 & 0.2141 \end{pmatrix}$	$\begin{pmatrix} 13.4070 & -1.9995 & -2.0001 \\ 0.2284 & 0.5432 & 0.2284 \end{pmatrix}$	$\begin{pmatrix} -1.3334 & -0.2232 & -1.3332 \\ 0.2226 & 0.5480 & 0.2226 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 4, \beta_1 = 4$	$\beta_0 = 4, \beta_1 = 5$	$\beta_0 = 4, \beta_1 = 6$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.9986 & -0.9514 & -1.0013 \\ 0.3523 & 0.2954 & 0.3523 \end{pmatrix}$	$\begin{pmatrix} -0.7998 & -0.1396 & -0.8001 \\ 0.3622 & 0.2756 & 0.3622 \end{pmatrix}$	$\begin{pmatrix} -0.3687 & -0.6669 & -0.6664 \\ 0.2770 & 0.4460 & 0.2770 \end{pmatrix}$

Table 2: Continued

$\boldsymbol{\beta}$	$\beta_0 = 4,\ \beta_1 = 10$	$\beta_0 = 4,\ \beta_1 = 11$	$\beta_0 = 4,\ \beta_1 = 12$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} 1.4528 & -0.3996 & -0.4002 \\ 0.2719 & 0.4562 & 0.2719 \end{pmatrix}$	$\begin{pmatrix} 1.1967 & -0.3632 & -0.3639 \\ 0.3046 & 0.3908 & 0.3046 \end{pmatrix}$	$\begin{pmatrix} -0.2966 & -0.3328 & -0.3339 \\ 0.3704 & 0.2592 & 0.3704 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 4,\ \beta_1 = 13$	$\beta_0 = 4,\ \beta_1 = 14$	$\beta_0 = 4,\ \beta_1 = 15$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} 1.2290 & -0.3076 & -0.3077 \\ 0.2352 & 0.5296 & 0.2352 \end{pmatrix}$	$\begin{pmatrix} 1.2836 & -0.2857 & -0.2857 \\ 0.2917 & 0.5166 & 0.2917 \end{pmatrix}$	$\begin{pmatrix} 1.1672 & -0.2662 & -0.2668 \\ 0.2263 & 0.5474 & 0.2263 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 5,\ \beta_1 = 1$	$\beta_0 = 5,\ \beta_1 = 2$	$\beta_0 = 5,\ \beta_1 = 3$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -4.9990 & 8.9742 & -5.0010 \\ 0.4527 & 0.0946 & 0.4527 \end{pmatrix}$	$\begin{pmatrix} 4.3431 & -2.5042 & -2.4998 \\ 0.0265 & 0.9470 & 0.0265 \end{pmatrix}$	$\begin{pmatrix} -1.6668 & 0.4464 & -1.6664 \\ 0.2989 & 0.4022 & 0.2989 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 5,\ \beta_1 = 4$	$\beta_0 = 5,\ \beta_1 = 5$	$\beta_0 = 5,\ \beta_1 = 6$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -1.2500 & -1.2498 & -2.2318 \\ 0.1785 & 0.6430 & 0.1785 \end{pmatrix}$	$\begin{pmatrix} -0.9985 & -0.9465 & -1.0001 \\ 0.3517 & 0.2966 & 0.3517 \end{pmatrix}$	$\begin{pmatrix} -0.7045 & -0.8334 & -0.8332 \\ 0.3373 & 0.3254 & 0.3373 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 5,\ \beta_1 = 7$	$\beta_0 = 5,\ \beta_1 = 8$	$\beta_0 = 5,\ \beta_1 = 13$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.7141 & 0.3424 & -0.7141 \\ 0.4088 & 0.1824 & 0.4088 \end{pmatrix}$	$\begin{pmatrix} 0.3565 & -0.6072 & -0.6256 \\ 0.0349 & 0.9301 & 0.0349 \end{pmatrix}$	$\begin{pmatrix} 1.3649 & -0.3770 & -0.3848 \\ 0.0288 & 0.9423 & 0.02888 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 5,\ \beta_1 = 14$	$\beta_0 = 5,\ \beta_1 = 15$	$\beta_0 = 6,\ \beta_1 = 1$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} 1.3238 & -0.3570 & -0.3571 \\ 0.2666 & 0.4668 & 0.2666 \end{pmatrix}$	$\begin{pmatrix} -0.3030 & -0.3332 & -0.3334 \\ 0.3660 & 0.3680 & 0.3660 \end{pmatrix}$	$\begin{pmatrix} -5.9967 & 8.2075 & -6.0032 \\ 0.4825 & 0.0350 & 0.4825 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 6,\ \beta_1 = 2$	$\beta_0 = 6,\ \beta_1 = 3$	$\beta_0 = 6,\ \beta_1 = 4$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.9973 & -2.9996 & -3.0001 \\ 0.2611 & 0.4778 & 0.2611 \end{pmatrix}$	$\begin{pmatrix} 4.1383 & -1.9994 & -2.0002 \\ 0.2289 & 0.5422 & 0.2289 \end{pmatrix}$	$\begin{pmatrix} -1.4997 & -0.3850 & -1.5002 \\ 0.1796 & 0.6408 & 0.1796 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 6,\ \beta_1 = 6$	$\beta_0 = 6,\ \beta_1 = 7$	$\beta_0 = 6,\ \beta_1 = 8$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.9994 & -0.9073 & -1.0005 \\ 0.3528 & 0.2944 & 0.3528 \end{pmatrix}$	$\begin{pmatrix} -0.8572 & -0.8570 & -1.7236 \\ 0.3553 & 0.2894 & 0.3553 \end{pmatrix}$	$\begin{pmatrix} -0.7501 & 0.4096 & -0.7498 \\ 0.4127 & 0.1746 & 0.4127 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 6,\ \beta_1 = 9$	$\beta_0 = 6,\ \beta_1 = 10$	$\beta_0 = 6,\ \beta_1 = 13$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.3687 & -0.6664 & -0.6668 \\ 0.2770 & 0.4460 & 0.2770 \end{pmatrix}$	$\begin{pmatrix} 0.6068 & -0.5995 & -0.6004 \\ 0.3365 & 0.3270 & 0.3365 \end{pmatrix}$	$\begin{pmatrix} 1.9172 & -0.4613 & -0.4615 \\ 0.1120 & 0.7760 & 0.1120 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 6,\ \beta_1 = 15$	$\beta_0 = 7,\ \beta_1 = 1$	$\beta_0 = 7,\ \beta_1 = 2$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} 1.4524 & -0.3994 & -0.4003 \\ 0.2719 & 0.4562 & 0.2719 \end{pmatrix}$	$\begin{pmatrix} -6.9988 & 0.7763 & -7.0011 \\ 0.4372 & 0.1256 & 0.4372 \end{pmatrix}$	$\begin{pmatrix} 0.9603 & -3.5000 & -3.5000 \\ 0.3192 & 0.3616 & 0.3192 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 7,\ \beta_1 = 3$	$\beta_0 = 7,\ \beta_1 = 4$	$\beta_0 = 7,\ \beta_1 = 5$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} 4.2596 & -2.3290 & -2.3333 \\ 0.0538 & 0.8924 & 0.0538 \end{pmatrix}$	$\begin{pmatrix} -0.2754 & -1.6012 & -1.7518 \\ 0.0119 & 0.9762 & 0.0119 \end{pmatrix}$	$\begin{pmatrix} -1.3999 & 1.2023 & -1.4000 \\ 0.4997 & 0.0006 & 0.4997 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 7,\ \beta_1 = 6$	$\beta_0 = 7,\ \beta_1 = 7$	$\beta_0 = 7,\ \beta_1 = 8$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.2672 & -1.1666 & -1.1666 \\ 0.1937 & 0.6126 & 0.1937 \end{pmatrix}$	$\begin{pmatrix} -0.9993 & -0.9072 & -1.0006 \\ 0.3528 & 0.2944 & 0.3528 \end{pmatrix}$	$\begin{pmatrix} -0.8048 & -0.8752 & -1.3642 \\ 0.2562 & 0.4848 & 0.2562 \end{pmatrix}$

Table 2: Continued

$\boldsymbol{\beta}$	$\beta_0 = 7,\ \beta_1 = 9$	$\beta_0 = 7,\ \beta_1 = 10$	$\beta_0 = 7,\ \beta_1 = 11$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.7778 & -0.1762 & -0.7776 \\ 0.3684 & 0.2632 & 0.3684 \end{pmatrix}$	$\begin{pmatrix} -0.6997 & 0.1970 & -0.7002 \\ 0.3791 & 0.2418 & 0.3791 \end{pmatrix}$	$\begin{pmatrix} 1.0681 & -0.6361 & -0.6364 \\ 0.1999 & 0.6002 & 0.1999 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 7,\ \beta_1 = 13$	$\beta_0 = 7,\ \beta_1 = 15$	$\beta_0 = 8,\ \beta_1 = 1$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} 1.8182 & -0.2227 & -0.5385 \\ 0.0001 & 0.9998 & 0.0001 \end{pmatrix}$	$\begin{pmatrix} -0.4666 & -0.3031 & -0.4666 \\ 0.3535 & 0.2930 & 0.3535 \end{pmatrix}$	$\begin{pmatrix} -7.9959 & 4.9336 & -8.0040 \\ 0.2500 & 0.5000 & 0.2500 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 8,\ \beta_1 = 2$	$\beta_0 = 8,\ \beta_1 = 3$	$\beta_0 = 8,\ \beta_1 = 4$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -3.1435 & -3.9975 & -4.0009 \\ 0.2142 & 0.5716 & 0.2142 \end{pmatrix}$	$\begin{pmatrix} -2.6585 & -1.2272 & -2.6747 \\ 0.1341 & 0.7318 & 0.1341 \end{pmatrix}$	$\begin{pmatrix} 4.1382 & -2.0003 & -1.9998 \\ 0.2289 & 0.5422 & 0.2289 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 8,\ \beta_1 = 5$	$\beta_0 = 8,\ \beta_1 = 6$	$\beta_0 = 8,\ \beta_1 = 8$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -1.5998 & -1.0041 & -1.6001 \\ 0.2402 & 0.5196 & 0.2402 \end{pmatrix}$	$\begin{pmatrix} -1.3334 & -0.2232 & -1.3332 \\ 0.2260 & 0.5480 & 0.2260 \end{pmatrix}$	$\begin{pmatrix} -0.9983 & -0.9380 & -1.0016 \\ 0.3460 & 0.3080 & 0.3460 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 8,\ \beta_1 = 9$	$\beta_0 = 8,\ \beta_1 = 10$	$\beta_0 = 8,\ \beta_1 = 11$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.8892 & -0.8883 & -1.7513 \\ 0.2902 & 0.4196 & 0.2902 \end{pmatrix}$	$\begin{pmatrix} -0.7998 & 0.1396 & -0.8001 \\ 0.3621 & 0.2758 & 0.3621 \end{pmatrix}$	$\begin{pmatrix} -0.7271 & 0.1861 & -0.7274 \\ 0.3672 & 0.2256 & 0.3672 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 8,\ \beta_1 = 12$	$\beta_0 = 8,\ \beta_1 = 13$	$\beta_0 = 9,\ \beta_1 = 1$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.3687 & -0.6669 & -0.6664 \\ 0.2770 & 0.4460 & 0.2770 \end{pmatrix}$	$\begin{pmatrix} 1.3602 & -0.6118 & -0.6158 \\ 0.1059 & 0.7882 & 0.1059 \end{pmatrix}$	$\begin{pmatrix} -9.0008 & 3.9228 & -8.9991 \\ 0.4933 & 0.0135 & 0.4933 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 9,\ \beta_1 = 2$	$\beta_0 = 9,\ \beta_1 = 3$	$\beta_0 = 9,\ \beta_1 = 4$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.6459 & -4.4982 & -4.5003 \\ 0.1327 & 0.7346 & 0.1327 \end{pmatrix}$	$\begin{pmatrix} -0.9973 & -2.9958 & -3.0022 \\ 0.2617 & 0.4766 & 0.2617 \end{pmatrix}$	$\begin{pmatrix} 4.6884 & -2.2488 & -2.2501 \\ 0.1033 & 0.7934 & 0.1033 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 9,\ \beta_1 = 5$	$\beta_0 = 9,\ \beta_1 = 6$	$\beta_0 = 9,\ \beta_1 = 7$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.5318 & -1.7996 & -1.8001 \\ 0.2556 & 0.4888 & 0.2556 \end{pmatrix}$	$\begin{pmatrix} -1.4998 & -1.4731 & -1.5001 \\ 0.1908 & 0.6184 & 0.1908 \end{pmatrix}$	$\begin{pmatrix} -0.8256 & -1.2855 & -1.2858 \\ 0.2492 & 0.5016 & 0.2492 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 9,\ \beta_1 = 8$	$\beta_0 = 9,\ \beta_1 = 9$	$\beta_0 = 9,\ \beta_1 = 10$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.9108 & -1.1170 & -1.1263 \\ 0.1460 & 0.7080 & 0.1460 \end{pmatrix}$	$\begin{pmatrix} -0.9988 & -0.9072 & -1.0011 \\ 0.3528 & 0.2944 & 0.3528 \end{pmatrix}$	$\begin{pmatrix} -0.9000 & -0.8999 & -1.1354 \\ 0.2253 & 0.5494 & 0.2253 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 9,\ \beta_1 = 11$	$\beta_0 = 9,\ \beta_1 = 12$	$\beta_0 = 9,\ \beta_1 = 13$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.8180 & 1.3889 & -0.8183 \\ 0.3722 & 0.2552 & 0.3722 \end{pmatrix}$	$\begin{pmatrix} -0.7501 & 0.4096 & -0.7498 \\ 0.4127 & 0.1746 & 0.4127 \end{pmatrix}$	$\begin{pmatrix} -0.6116 & -0.6921 & -0.6924 \\ 0.3537 & 0.2926 & 0.3537 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 9,\ \beta_1 = 15$	$\beta_0 = 10,\ \beta_1 = 1$	$\beta_0 = 10,\ \beta_1 = 2$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} 0.6063 & -0.5999 & -0.6000 \\ 0.3364 & 0.3272 & 0.3364 \end{pmatrix}$	$\begin{pmatrix} -9.9992 & -0.8220 & -10.0002 \\ 0.2986 & 0.4028 & 0.2986 \end{pmatrix}$	$\begin{pmatrix} -4.9990 & 8.9742 & -5.0009 \\ 0.4528 & 0.0944 & 0.4528 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 10,\ \beta_1 = 3$	$\beta_0 = 10,\ \beta_1 = 4$	$\beta_0 = 10,\ \beta_1 = 5$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -3.3219 & -4.5466 & -3.3447 \\ 0.4872 & 0.0256 & 0.4872 \end{pmatrix}$	$\begin{pmatrix} 3.8099 & -2.4494 & -2.5000 \\ 0.0001 & 0.9998 & 0.0001 \end{pmatrix}$	$\begin{pmatrix} 4.1382 & -1.9993 & -2.0002 \\ 0.2296 & 0.5408 & 0.2296 \end{pmatrix}$

Table 2: Continued

$\boldsymbol{\beta}$	$\beta_0 = 10,\ \beta_1 = 6$	$\beta_0 = 10,\ \beta_1 = 7$	$\beta_0 = 10,\ \beta_1 = 8$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -1.6668 & -0.4464 & -1.6665 \\ 0.2989 & 0.4022 & 0.2989 \end{pmatrix}$	$\begin{pmatrix} -1.1727 & -1.4280 & -1.4296 \\ 0.3997 & 0.2006 & 0.3997 \end{pmatrix}$	$\begin{pmatrix} -1.2500 & -1.2498 & -2.2318 \\ 0.1785 & 0.6430 & 0.1785 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 10,\ \beta_1 = 9$	$\beta_0 = 10,\ \beta_1 = 10$	$\beta_0 = 10,\ \beta_1 = 11$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.6269 & -1.1046 & -1.1122 \\ 0.1294 & 0.7412 & 0.1294 \end{pmatrix}$	$\begin{pmatrix} -0.9985 & -0.9561 & -1.0013 \\ 0.3678 & 0.2644 & 0.3678 \end{pmatrix}$	$\begin{pmatrix} -0.4497 & -0.9093 & -0.9080 \\ 0.2630 & 0.4740 & 0.2630 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 10,\ \beta_1 = 12$	$\beta_0 = 10,\ \beta_1 = 13$	$\beta_0 = 10,\ \beta_1 = 14$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.7045 & -0.8334 & -0.8332 \\ 0.3373 & 0.3254 & 0.3373 \end{pmatrix}$	$\begin{pmatrix} -0.7690 & 0.2041 & -0.7693 \\ 0.3732 & 0.2536 & 0.3732 \end{pmatrix}$	$\begin{pmatrix} -0.7141 & 0.3424 & -0.7144 \\ 0.4088 & 0.1824 & 0.4088 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 10,\ \beta_1 = 15$	$\beta_0 = 10,\ \beta_1 = 16$	$\beta_0 = 10,\ \beta_1 = 17$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.3687 & -0.6665 & -0.6667 \\ 0.2759 & 0.4482 & 0.2759 \end{pmatrix}$	$\begin{pmatrix} -0.9998 & -0.9791 & -1.0001 \\ 0.4588 & 0.0824 & 0.4588 \end{pmatrix}$	$\begin{pmatrix} -0.9998 & -0.9791 & -1.0001 \\ 0.4588 & 0.0824 & 0.4588 \end{pmatrix}$

Table 3: R-optimal design for Gamma regression model with two parameters (Four support points)

$\boldsymbol{\beta}$	$\beta_0 = 1,\ \beta_1 = 8$	$\beta_0 = 1,\ \beta_1 = 10$
$\boldsymbol{x}$	$\begin{pmatrix} -0.1251 & 1.4191 & -0.1249 & 0.4328 \end{pmatrix}$	$\begin{pmatrix} -0.1000 & 0.6989 & -0.0999 & 0.4254 \end{pmatrix}$
$\boldsymbol{w}$	$\begin{pmatrix} 0.1328 & 0.3672 & 0.3672 & 0.1328 \end{pmatrix}$	$\begin{pmatrix} 0.3494 & 0.1506 & 0.1506 & 0.3494 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 1,\ \beta_1 = 11$	$\beta_0 = 2,\ \beta_1 = 3$
$\boldsymbol{x}$	$\begin{pmatrix} 0.1024 & 1.2526 & -0.0909 & -0.0909 \end{pmatrix}$	$\begin{pmatrix} -0.6668 & 1.5448 & -0.6663 & 1.0896 \end{pmatrix}$
$\boldsymbol{w}$	$\begin{pmatrix} 0.2028 & 0.2972 & 0.2972 & 0.2028 \end{pmatrix}$	$\begin{pmatrix} 0.1767 & 0.3233 & 0.3233 & 0.1767 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 2,\ \beta_1 = 9$	$\beta_0 = 2,\ \beta_1 = 15$
$\boldsymbol{x}$	$\begin{pmatrix} 0.3407 & 1.3105 & -0.2221 & -0.2222 \end{pmatrix}$	$\begin{pmatrix} -0.1333 & 0.9882 & -0.1333 & 0.0713 \end{pmatrix}$
$\boldsymbol{w}$	$\begin{pmatrix} 0.2319 & 0.2681 & 0.2681 & 0.2319 \end{pmatrix}$	$\begin{pmatrix} 0.1449 & 0.3551 & 0.3551 & 0.1449 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 2,\ \beta_1 = 16$	$\beta_0 = 2,\ \beta_1 = 17$
$\boldsymbol{x}$	$\begin{pmatrix} -0.1250 & 1.4191 & -0.1249 & 0.4328 \end{pmatrix}$	$\begin{pmatrix} -0.1176 & 1.0332 & -0.1176 & 0.3824 \end{pmatrix}$
$\boldsymbol{w}$	$\begin{pmatrix} 0.1327 & 0.3673 & 0.3673 & 0.1327 \end{pmatrix}$	$\begin{pmatrix} 0.2506 & 0.2494 & 0.2494 & 0.2506 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 2,\ \beta_1 = 20$	$\beta_0 = 3,\ \beta_1 = 18$
$\boldsymbol{x}$	$\begin{pmatrix} -0.1000 & 0.6989 & -0.0999 & 0.4254 \end{pmatrix}$	$\begin{pmatrix} -0.1450 & 0.7336 & -0.1576 & -0.1669 \end{pmatrix}$
$\boldsymbol{w}$	$\begin{pmatrix} 0.3494 & 0.1506 & 0.1506 & 0.3494 \end{pmatrix}$	$\begin{pmatrix} 0.0159 & 0.4841 & 0.4841 & 0.0159 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 3,\ \beta_1 = 26$	$\beta_0 = 3,\ \beta_1 = 30$
$\boldsymbol{x}$	$\begin{pmatrix} -0.1154 & 0.8000 & -0.1153 & 0.3216 \end{pmatrix}$	$\begin{pmatrix} -0.0999 & 0.6989 & -0.1000 & 0.4254 \end{pmatrix}$
$\boldsymbol{w}$	$\begin{pmatrix} 0.2307 & 0.2693 & 0.2693 & 0.2307 \end{pmatrix}$	$\begin{pmatrix} 0.3495 & 0.1505 & 0.1505 & 0.3495 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 3,\ \beta_1 = 32$	$\beta_0 = 4,\ \beta_1 = 6$
$\boldsymbol{x}$	$\begin{pmatrix} -0.0937 & 1.0108 & -0.0937 & 0.6547 \end{pmatrix}$	$\begin{pmatrix} -0.6668 & 1.5449 & -0.6663 & 1.0897 \end{pmatrix}$
$\boldsymbol{w}$	$\begin{pmatrix} 0.4206 & 0.0794 & 0.0794 & 0.4206 \end{pmatrix}$	$\begin{pmatrix} 0.1767 & 0.3233 & 0.3233 & 0.1767 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 4,\ \beta_1 = 15$	$\beta_0 = 4,\ \beta_1 = 20$
$\boldsymbol{x}$	$\begin{pmatrix} -0.2664 & 0.8092 & -0.2668 & 0.4288 \end{pmatrix}$	$\begin{pmatrix} 0.0882 & 0.5674 & -0.1999 & -0.2001 \end{pmatrix}$
$\boldsymbol{w}$	$\begin{pmatrix} 0.2419 & 0.2581 & 0.2581 & 0.2419 \end{pmatrix}$	$\begin{pmatrix} 0.0008 & 0.4992 & 0.4992 & 0.0008 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 4,\ \beta_1 = 22$	$\beta_0 = 4,\ \beta_1 = 30$
$\boldsymbol{x}$	$\begin{pmatrix} 0.1844 & 0.6942 & -0.1817 & -0.1818 \end{pmatrix}$	$\begin{pmatrix} -0.1333 & 0.9608 & -0.1333 & 0.4805 \end{pmatrix}$
$\boldsymbol{w}$	$\begin{pmatrix} 0.1322 & 0.3678 & 0.3678 & 0.1322 \end{pmatrix}$	$\begin{pmatrix} 0.1575 & 0.3425 & 0.3425 & 0.1575 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 4,\ \beta_1 = 31$	$\beta_0 = 4,\ \beta_1 = 32$
$\boldsymbol{x}$	$\begin{pmatrix} -0.1283 & 1.4199 & -0.1317 & 0.5303 \end{pmatrix}$	$\begin{pmatrix} -0.1250 & 1.4191 & -0.1249 & 0.4328 \end{pmatrix}$
$\boldsymbol{w}$	$\begin{pmatrix} 0.1067 & 0.3933 & 0.3933 & 0.1067 \end{pmatrix}$	$\begin{pmatrix} 0.1327 & 0.3673 & 0.3673 & 0.1327 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 4,\ \beta_1 = 33$	$\beta_0 = 4,\ \beta_1 = 34$
$\boldsymbol{x}$	$\begin{pmatrix} -0.1212 & 1.5383 & -0.1211 & 0.4779 \end{pmatrix}$	$\begin{pmatrix} -0.1176 & 1.0332 & -0.1176 & 0.3824 \end{pmatrix}$
$\boldsymbol{w}$	$\begin{pmatrix} 0.1514 & 0.3486 & 0.3486 & 0.1514 \end{pmatrix}$	$\begin{pmatrix} 0.2506 & 0.2494 & 0.2494 & 0.2506 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 4,\ \beta_1 = 35$	$\beta_0 = 5,\ \beta_1 = 19$
$\boldsymbol{x}$	$\begin{pmatrix} -0.1143 & 0.8657 & -0.1142 & 0.4394 \end{pmatrix}$	$\begin{pmatrix} -0.2630 & 1.0310 & -0.2632 & 0.4691 \end{pmatrix}$
$\boldsymbol{w}$	$\begin{pmatrix} 0.3047 & 0.1953 & 0.1953 & 0.3047 \end{pmatrix}$	$\begin{pmatrix} 0.2334 & 0.2666 & 0.2666 & 0.2334 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 5,\ \beta_1 = 26$	$\beta_0 = 5,\ \beta_1 = 27$
$\boldsymbol{x}$	$\begin{pmatrix} 0.2964 & 0.6382 & -0.1922 & -0.1923 \end{pmatrix}$	$\begin{pmatrix} 0.1441 & 0.6884 & -0.1848 & -0.1851 \end{pmatrix}$
$\boldsymbol{w}$	$\begin{pmatrix} 0.1173 & 0.3827 & 0.3827 & 0.1173 \end{pmatrix}$	$\begin{pmatrix} 0.0072 & 0.4928 & 0.4928 & 0.0072 \end{pmatrix}$

Table 3: Continued

$\boldsymbol{\beta}$	$\beta_0 = 5,\ \beta_1 = 28$	$\beta_0 = 5,\ \beta_1 = 29$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} 0.0525 & 0.5365 & -0.1770 & -0.1785 \\ 0.0011 & 0.4988 & 0.4988 & 0.0011 \end{pmatrix}$	$\begin{pmatrix} 0.2664 & 0.5590 & -0.1723 & 0.1724 \\ 0.0871 & 0.4129 & 0.4129 & 0.0871 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 5,\ \beta_1 = 30$	$\beta_0 = 5,\ \beta_1 = 31$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} 0.1447 & 0.7737 & -0.1664 & -0.1666 \\ 0.0450 & 0.4550 & 0.4550 & 0.0450 \end{pmatrix}$	$\begin{pmatrix} -0.1612 & 0.7699 & -0.1613 & 0.1804 \\ 0.0968 & 0.4032 & 0.4032 & 0.0968 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 5,\ \beta_1 = 34$	$\beta_0 = 5,\ \beta_1 = 38$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.1471 & 0.7057 & -0.1469 & 0.1660 \\ 0.1272 & 0.3728 & 0.3728 & 0.1272 \end{pmatrix}$	$\begin{pmatrix} -0.1315 & 0.5356 & -0.1126 & -0.0003 \\ 0.0002 & 0.4997 & 0.4997 & 0.0002 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 5,\ \beta_1 = 39$	$\beta_0 = 5,\ \beta_1 = 40$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.1282 & 1.6957 & -0.1282 & -0.4058 \\ 0.0359 & 0.4641 & 0.4641 & 0.0359 \end{pmatrix}$	$\begin{pmatrix} -0.1250 & 1.4191 & -0.1249 & -0.4328 \\ 0.1327 & 0.3670 & 0.3670 & 0.1327 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 5,\ \beta_1 = 41$	$\beta_0 = 5,\ \beta_1 = 42$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.1219 & 1.5634 & -0.1218 & 0.4439 \\ 0.1339 & 0.3661 & 0.3661 & 0.1339 \end{pmatrix}$	$\begin{pmatrix} -0.1190 & 1.5748 & -0.1190 & 0.4092 \\ 0.0836 & 0.4164 & 0.4164 & 0.0836 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 5,\ \beta_1 = 43$	$\beta_0 = 5,\ \beta_1 = 44$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.1162 & 1.0502 & -0.1162 & 0.4671 \\ 0.2928 & 0.2072 & 0.2072 & 0.2928 \end{pmatrix}$	$\begin{pmatrix} -0.1136 & 0.9680 & -0.1136 & 0.3829 \\ 0.2390 & 0.2610 & 0.2610 & 0.2390 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 6,\ \beta_1 = 9$	$\beta_0 = 6,\ \beta_1 = 23$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.6668 & 1.5449 & -0.6663 & 1.0897 \\ 0.1767 & 0.3233 & 0.3233 & 0.1767 \end{pmatrix}$	$\begin{pmatrix} -0.2608 & 0.9196 & -0.2608 & 0.3851 \\ 0.2032 & 0.2968 & 0.2968 & 0.2032 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 6,\ \beta_1 = 27$	$\beta_0 = 6,\ \beta_1 = 30$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} 0.2864 & 1.4903 & -0.2222 & -0.2221 \\ 0.3334 & 0.1666 & 0.1666 & 0.3334 \end{pmatrix}$	$\begin{pmatrix} -0.0270 & 0.3240 & -0.1986 & -0.2000 \\ 0.0007 & 0.4992 & 0.4992 & 0.0007 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 6,\ \beta_1 = 37$	$\beta_0 = 6,\ \beta_1 = 45$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.1622 & 0.6520 & -0.1619 & -0.2367 \\ 0.1578 & 0.3422 & 0.3422 & 0.1578 \end{pmatrix}$	$\begin{pmatrix} -0.1333 & 1.0400 & -0.1332 & 0.0001 \\ 0.1556 & 0.3444 & 0.3444 & 0.1556 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 6,\ \beta_1 = 48$	$\beta_0 = 6,\ \beta_1 = 51$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.1249 & 1.4193 & -0.1250 & 0.4328 \\ 0.1327 & 0.3673 & 0.3673 & 0.1327 \end{pmatrix}$	$\begin{pmatrix} -0.1176 & 1.0332 & -0.1176 & 0.3824 \\ 0.2506 & 0.2494 & 0.2494 & 0.2506 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 6,\ \beta_1 = 60$	$\beta_0 = 7,\ \beta_1 = 39$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.0999 & 0.6989 & -0.1000 & 0.4254 \\ 0.3495 & 0.1505 & 0.1505 & 0.3495 \end{pmatrix}$	$\begin{pmatrix} -0.0421 & 0.3061 & -0.1764 & -0.1794 \\ 0.0002 & 0.4997 & 0.4997 & 0.0002 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 7,\ \beta_1 = 40$	$\beta_0 = 7,\ \beta_1 = 50$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} 0.3548 & 0.6450 & -0.1749 & -0.1750 \\ 0.1389 & 0.3611 & 0.3611 & 0.1389 \end{pmatrix}$	$\begin{pmatrix} -0.1400 & 0.9105 & -0.1399 & 0.1282 \\ 0.0461 & 0.4539 & 0.4539 & 0.0461 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 7,\ \beta_1 = 51$	$\beta_0 = 7,\ \beta_1 = 52$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.1376 & 0.9959 & -0.1267 & 0.1245 \\ 0.0174 & 0.4826 & 0.4826 & 0.0174 \end{pmatrix}$	$\begin{pmatrix} -0.1346 & 1.2422 & -0.1345 & 0.2680 \\ 0.0185 & 0.4815 & 0.4815 & 0.0185 \end{pmatrix}$

Table 3: Continued

$\boldsymbol{\beta}$	$\beta_0 = 7,\ \beta_1 = 56$	$\beta_0 = 7,\ \beta_1 = 70$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.1250 & 1.4203 & -0.1249 & 0.4340 \\ 0.1310 & 0.3690 & 0.3690 & 0.1310 \end{pmatrix}$	$\begin{pmatrix} -0.1000 & 0.6989 & -0.0999 & 0.4254 \\ 0.3493 & 0.1506 & 0.1506 & 0.3493 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 8,\ \beta_1 = 12$	$\beta_0 = 8,\ \beta_1 = 30$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.6668 & 1.5449 & -0.6663 & 1.0897 \\ 0.1767 & 0.3233 & 0.3233 & 0.1767 \end{pmatrix}$	$\begin{pmatrix} -0.2664 & 0.8092 & -0.2668 & 0.4228 \\ 0.2419 & 0.2581 & 0.2581 & 0.2419 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 8,\ \beta_1 = 36$	$\beta_0 = 8,\ \beta_1 = 40$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} 0.3996 & 1.4467 & -0.2221 & -0.2221 \\ 0.3056 & 0.1944 & 0.1944 & 0.3056 \end{pmatrix}$	$\begin{pmatrix} 0.0882 & 0.5674 & -0.1993 & -0.2000 \\ 0.0007 & 0.4992 & 0.4992 & 0.0007 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 8,\ \beta_1 = 44$	$\beta_0 = 8,\ \beta_1 = 60$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.0286 & 0.2663 & -0.1811 & 0.1818 \\ 0.0003 & 0.4996 & 0.4996 & 0.0003 \end{pmatrix}$	$\begin{pmatrix} -0.1333 & 1.0420 & -0.1332 & 0.1441 \\ 0.1554 & 0.3446 & 0.3446 & 0.1554 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 8,\ \beta_1 = 63$	$\beta_0 = 8,\ \beta_1 = 64$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.1269 & 1.1873 & -0.1275 & 0.3266 \\ 0.0005 & 0.4994 & 0.4994 & 0.0005 \end{pmatrix}$	$\begin{pmatrix} -0.1250 & 1.4191 & -0.1249 & 0.4328 \\ 0.1327 & 0.3673 & 0.3673 & 0.1327 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 8,\ \beta_1 = 65$	$\beta_0 = 8,\ \beta_1 = 66$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.1232 & 1.4500 & -0.1221 & 0.3714 \\ 0.0731 & 0.4269 & 0.4269 & 0.0731 \end{pmatrix}$	$\begin{pmatrix} -0.1212 & 1.5383 & -0.1211 & 0.4779 \\ 0.1514 & 0.3486 & 0.3486 & 0.1514 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 8,\ \beta_1 = 68$	$\beta_0 = 8,\ \beta_1 = 70$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.1176 & 1.0332 & -0.1176 & 0.3824 \\ 0.2506 & 0.2494 & 0.2494 & 0.2506 \end{pmatrix}$	$\begin{pmatrix} -0.1143 & 0.8657 & -0.1142 & 0.4394 \\ 0.3047 & 0.1953 & 0.1953 & 0.3047 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 8,\ \beta_1 = 80$	$\beta_0 = 8,\ \beta_1 = 88$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.1000 & 0.6989 & -0.0999 & 0.4254 \\ 0.3494 & 0.1506 & 0.1506 & 0.3494 \end{pmatrix}$	$\begin{pmatrix} -0.1043 & 0.2520 & -0.0908 & -0.0909 \\ 0.1951 & 0.3049 & 0.3049 & 0.1951 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 9,\ \beta_1 = 54$	$\beta_0 = 9,\ \beta_1 = 72$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} 0.1447 & 0.7737 & -0.1663 & -0.1666 \\ 0.0450 & 0.4550 & 0.4550 & 0.0450 \end{pmatrix}$	$\begin{pmatrix} -0.1250 & 1.4191 & -0.1249 & 0.4328 \\ 0.1327 & 0.3673 & 0.3673 & 0.1327 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 9,\ \beta_1 = 78$	$\beta_0 = 9,\ \beta_1 = 90$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.1153 & 0.8000 & -0.1153 & 0.3216 \\ 0.2359 & 0.2641 & 0.2641 & 0.2359 \end{pmatrix}$	$\begin{pmatrix} -0.1000 & 0.6989 & -0.0999 & 0.4254 \\ 0.3493 & 0.1507 & 0.1507 & 0.3493 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 9,\ \beta_1 = 96$	$\beta_0 = 10,\ \beta_1 = 15$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.0939 & 1.0108 & -0.0937 & 0.6547 \\ 0.4199 & 0.0801 & 0.0801 & 0.4199 \end{pmatrix}$	$\begin{pmatrix} -0.6666 & 1.3241 & -0.6666 & 0.9085 \\ 0.1786 & 0.3214 & 0.3214 & 0.1786 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 10,\ \beta_1 = 38$	$\beta_0 = 10,\ \beta_1 = 52$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.2630 & 1.0310 & -0.2632 & 0.4691 \\ 0.2334 & 0.2666 & 0.2666 & 0.2334 \end{pmatrix}$	$\begin{pmatrix} 0.1965 & 0.6549 & -0.1922 & -0.1923 \\ 0.1199 & 0.3801 & 0.3801 & 0.1199 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 10,\ \beta_1 = 58$	$\beta_0 = 10,\ \beta_1 = 68$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} 0.2664 & 0.5590 & -0.1724 & -0.1724 \\ 0.0871 & 0.4129 & 0.4129 & 0.0871 \end{pmatrix}$	$\begin{pmatrix} -0.1471 & 0.7057 & -0.1468 & -0.1660 \\ 0.3494 & 0.1506 & 0.1506 & 0.3494 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 10,\ \beta_1 = 75$	$\beta_0 = 10,\ \beta_1 = 78$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.1333 & 0.9059 & -0.1332 & -0.0098 \\ 0.1466 & 0.3534 & 0.3534 & 0.1466 \end{pmatrix}$	$\begin{pmatrix} -0.1282 & 1.6985 & -0.1280 & 0.4065 \\ 0.0290 & 0.4710 & 0.4710 & 0.0290 \end{pmatrix}$

Table 3: Continued

$\boldsymbol{\beta}$	$\beta_0 = 10,\ \beta_1 = 81$	$\beta_0 = 10,\ \beta_1 = 82$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.1234 & 1.7453 & -0.1233 & -0.5493 \\ 0.1206 & 0.3794 & 0.3794 & 0.1206 \end{pmatrix}$	$\begin{pmatrix} -0.1219 & 1.5634 & -0.1219 & 0.4439 \\ 0.1339 & 0.3661 & 0.3661 & 0.1339 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 10,\ \beta_1 = 85$	$\beta_0 = 10,\ \beta_1 = 86$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.1176 & 1.0332 & -0.1176 & 0.3824 \\ 0.2506 & 0.2494 & 0.2494 & 0.2506 \end{pmatrix}$	$\begin{pmatrix} -0.1162 & 0.8101 & -0.1162 & 0.3439 \\ 0.2934 & 0.2066 & 0.2066 & 0.2934 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 10,\ \beta_1 = 88$	$\beta_0 = 10,\ \beta_1 = 92$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.1136 & 0.9680 & -0.1136 & 0.3829 \\ 0.2346 & 0.2654 & 0.2654 & 0.2346 \end{pmatrix}$	$\begin{pmatrix} -0.1087 & 0.5971 & -0.1086 & 0.4081 \\ 0.3679 & 0.1321 & 0.1321 & 0.3679 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 10,\ \beta_1 = 93$	$\beta_0 = 10,\ \beta_1 = 96$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.1074 & 0.6330 & -0.1075 & 0.4220 \\ 0.3676 & 0.1324 & 0.1324 & 0.3676 \end{pmatrix}$	$\begin{pmatrix} -0.1041 & 0.7818 & -0.1041 & 0.4055 \\ 0.3092 & 0.1908 & 0.1908 & 0.3092 \end{pmatrix}$
$\boldsymbol{\beta}$	$\beta_0 = 10,\ \beta_1 = 98$	$\beta_0 = 10,\ \beta_1 = 100$
$\boldsymbol{x}$ $\boldsymbol{w}$	$\begin{pmatrix} -0.1020 & 0.5326 & -0.1020 & 0.4335 \\ 0.4111 & 0.0889 & 0.0889 & 0.4111 \end{pmatrix}$	$\begin{pmatrix} -1.0000 & 0.6989 & -0.0999 & 0.4254 \\ 0.3487 & 0.1513 & 0.1513 & 0.3487 \end{pmatrix}$

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 23, No. 1, 2025 (New Series), pp 55–78
https://www.ssca.org.in/journal

Discrete Type I Half-Logistic Weibull Distribution and its Properties

M. Girish Babu[1] **and K. Jayakumar**[2]
[1]*Department of Statistics, Government Arts and Science College, Kozhikode, Kerala, India-673 018*
[2]*Department of Statistics, University of Calicut, Kerala, India-673 635*

Received: 09 November 2023; Revised: 24 March 2024; Accepted: 11 April 2024

Abstract

In this paper, we develop a discrete version of the type I half-logistic family of distributions. Several members of this family such as, discrete type I half-logistic version of uniform, Lomax, exponential, Fréchet and Weibull distributions are derived. Statistical properties of one of the members of this family, namely, the discrete type I half-logistic Weibull distribution, is studied in detail. The parameters of this distribution are estimated using the maximum likelihood method and a simulation study is conducted to evaluate the consistency of the method. Three data applications are illustrated to show the flexibility for fitting the proposed models to real-life data sets.

Key words: Data modeling; Discrete distributions; Hazard rate function; Order statistics; Weibull distribution.

AMS Subject Classifications: 60E05, 62E10.

1. Introduction

The logistic function is one of the oldest growth functions in the literature and is used to describe both population and organic growth. Various theoretical, methodological and applied issues relating to the logistic model are discussed in Balakrishnan (1991). Often in scientific enquiry we may come across observations which are discrete in nature. In a reliability study or life testing of equipment, it is difficult to quantify the length of life of the equipment on a continuous scale. In survival analysis, we may record the number of days of survival for lung cancer patients since therapy, or the times from remission to relapse are also usually recorded in number of days. But in some analysis, often the interest lies not only in counts but in changes in counts from a given origin, in such situation the variable of interest can take either zero, positive or negative value. The conventional discrete distributions such as, geometric, Poisson, binomial and negative binomial have wide but limited applicability in reliability, failure time modeling, etc. Thus, there is interest in developing new discrete

Corresponding Author: K. Jayakumar
Email: jkumar19@rediffmail.com

family of distributions based on the well known continuous distributions. Among these, the discrete Weibull distribution of Nakagawa and Osaki (1975) is the most popular one.

Recently, discretization of the continuous distributions has attracted the researchers' attention and several forms of discrete lifetime distributions are being established in the literature. Some of the recent works on discretization of continuous distributions are the discrete Laplace distribution by Inusah and Kozubowski (2006), the discrete half-normal distribution by Kemp (2008), the discrete Burr and discrete Pareto distributions by Krishna and Pundir (2009), the discrete generalized exponential distribution by Gomez-Deniz (2010) and the discrete gamma distribution by Chakraborty and Chakravarty (2012).

Other notable works in this direction are the discrete additive Weibull distribution by Bebbington *et al.* (2012), the discrete inverse Weibull distribution by Jazi *et al.* (2010), the discrete generalized exponential distribution by Nekoukhou *et al.* (2012), discrete reduced modified Weibull distribution by Almalki and Nadarajah (2014), the discrete Lindley distribution by Bakouch *et al.* (2014), the discrete Logistic distribution by Chakraborty and Chakravarty (2016), the discrete log-logistic distribution by Para and Jan (2016), the discrete Weibull geometric distribution by Jayakumar and Babu (2018), the truncated discrete Mittag-Leffler distribution by Jayakumar and Sankaran (2018), discrete additive Weibull geometric distribution by Jayakumar and Babu (2019), the discrete Pareto type(IV) model by Ghosh (2020), among others. There are several methods available in the literature to discretize a continuous random variable, for more details, see Chakraborty (2015). This paper discusses the formation of distribution that are more appropriate to modeling discrete failure data in varying failure rate shape.

Let Y is discretized as $Y = \lfloor X \rfloor$, the largest integer less than or equal to X. Using the survival function $S_X(y)$, the discrete version of the random variable X can be derived by

$$P(Y = y) = P(X \geq y) - P(X \geq y + 1) = S_X(y) - S_X(y + 1); \quad y = 0, 1, 2, \ldots . \tag{1}$$

The cumulative distribution function (cdf) of Y is given by,

$$\begin{aligned} F(y) = P(Y \leq y) &= P(Y < y) + P(Y = y) \\ &= 1 - S_X(y) + P(Y = y) \\ &= 1 - S_X(y + 1) = P(X \leq y + 1). \end{aligned}$$

Now, for evaluating the values of the cdf when the value of y can be integer or fractional the following general formula can be used

$$F(y) = P(X \leq \lfloor y \rfloor + 1),$$

where $\lfloor y \rfloor$ denotes the floor of y, i.e., the largest integer less or equal to y. In this paper we take y as $\lfloor y \rfloor$.

The discretization of a continuous distribution given in (1) retains the similar functional form of the survival function, so that many reliability characteristics remains unchanged. This motivated to use this technique of generating discretized version of continuous distribution. The half-logistic probability models are obtained as the models of the absolute value of the standard logistic models.

Recently Chipepa *et al.* (2022) introduced a generalized class of distributions named as exponentiated half-logistic generalized - G power series distribution by combining the exponentiated half-logistic generalized class of distributions and power series distribution. The hazard rate of this distribution exhibits increasing, decreasing, bathtub, bathtub followed by upside down bathtub, J and inverse-J shapes. One member of this family called type I half-logistic-G (TIHL-G) family of distributions was already studied by Cordeiro *et al.* (2016). This family is defined by

$$F(x;\lambda,\Theta)=\int_0^{-\log[1-G(x;\Theta)]}\frac{2\lambda e^{-\lambda t}}{(1+e^{-\lambda t})^2}dt=\frac{1-[1-G(x;\Theta)]^{\lambda}}{1+[1-G(x;\Theta)]^{\lambda}}, \tag{2}$$

where $G(x;\Theta)$ is the baseline *cdf* with parameter vector Θ and $\lambda>0$ is an additional shape parameter. When $\lambda=1$, this family becomes half-logistic-G (HL-G) family of distributions. The probability density function (pdf) of (2) is given by

$$f(x;\lambda,\Theta)=\frac{2\lambda g(x;\Theta)[1-G(x;\Theta)]^{\lambda-1}}{\Big[1+[1-G(x;\Theta)]^{\lambda}\Big]^2}, \tag{3}$$

where $g(x;\Theta)$ is the baseline pdf. Also, the survival function and hazard rate function (hrf) are respectively given by

$$S(x;\lambda,\Theta)=\frac{2[1-G(x;\Theta)]^{\lambda}}{1+[1-G(x;\Theta)]^{\lambda}}, \tag{4}$$

and

$$h(x;\lambda,\Theta)=\frac{\lambda g(x;\Theta)}{\Big[1-G(x;\Theta)\Big]\Big[1+[1-G(x;\Theta)]^{\lambda}\Big]}. \tag{5}$$

Different choices of $G(x;\Theta)$ in (2) leads to special models of this family of distributions. Some of the models are type I half-logistic normal, type I half-logistic gamma and type I half-logistic Fréchet discussed in Cordeiro *et al.* (2016). The objective of this paper is to introduce a discrete version of this family and study their mathematical properties.

The paper is organized as follows. In Section 2, we introduce a discrete type I half-logistic family of distributions. Some members of this family are introduced in Section 3. Section 4 discusses the construction of the type I half-logistic Weibull distribution and in Section 5, the maximum likelihood estimation of unknown parameters are discussed and a simulation study to asses the performance of the MLEs of the model parameters is also presented. Applications of this new discrete distribution for modeling real data sets are discussed in Section 6, and conclusions and future works are presented in Section 7.

2. Discretization of type I half-logistic family of distributions

Let X be a continuous random variable belonging to TIHL-G family of distributions with cdf given in (2). Let Y be the discrete analogue of X derived using the survival function (4) and by using the expression (1) as follows :

$$P(Y=y)=\frac{2\Big[\bar{G}^{\lambda}(y;\Theta)-\bar{G}^{\lambda}(y+1;\Theta)\Big]}{\Big[1+\bar{G}^{\lambda}(y;\Theta)\Big]\Big[1+\bar{G}^{\lambda}(y+1;\Theta)\Big]} \tag{6}$$

where $Y = \lfloor X \rfloor$, the largest integer less than or equal to X and $\bar{G}(y;\Theta) = 1 - G(y;\Theta)$. The corresponding cdf is given by

$$F(y) = \frac{1 - \bar{G}^{\lambda}(y+1;\Theta)}{1 + \bar{G}^{\lambda}(y+1;\Theta)}, \tag{7}$$

survival function is

$$S(y) = 1 - P(Y \leq y) = \frac{2\bar{G}^{\lambda}(y+1;\Theta)}{1 + \bar{G}^{\lambda}(y+1;\Theta)}, \tag{8}$$

and hazard rate function is

$$h(y) = \frac{P(Y=y)}{P(Y \geq y)} = \frac{\bar{G}^{\lambda}(y;\Theta) - \bar{G}^{\lambda}(y+1;\Theta)}{\bar{G}^{\lambda}(y+1;\Theta)\Big[1 + \bar{G}^{\lambda}(y;\Theta)\Big]}. \tag{9}$$

The reverse hazard rate is

$$h^{*}(y) = \frac{P(Y=y)}{P(Y \leq y)} = \frac{2\Big[\bar{G}^{\lambda}(y;\Theta) - \bar{G}^{\lambda}(y+1;\Theta)\Big]}{\Big[1 + \bar{G}^{\lambda}(y;\Theta)\Big]\Big[1 + \bar{G}^{\lambda}(y+1;\Theta)\Big]}. \tag{10}$$

The second rate of failure is given by

$$h^{**}(y) = log\left[\frac{S(y)}{S(y+1)}\right] = log\left[\frac{\bar{G}^{\lambda}(y+1;\Theta)}{\bar{G}^{\lambda}(y+2;\Theta)}\right] + log\left[\frac{1 + \bar{G}^{\lambda}(y+2;\Theta)}{1 + \bar{G}^{\lambda}(y+1;\Theta)}\right]. \tag{11}$$

2.1. Quantile function

The quantile function of the discrete type I half-logistic G family of distributions, say $Q(u)$, defined by $F(Q(u)) = u$, where $u \in (0,1)$ is given by

$$Q(u) = \left\lceil G^{-1}\left[1 - \left(\frac{1-u}{1+u}\right)^{\frac{1}{\lambda}}\right] - 1 \right\rceil, \tag{12}$$

where $\lceil . \rceil$ denotes the ceiling value.

In particular, the median $= \left\lceil G^{-1}\left[1 - \left(\frac{1}{3}\right)^{\frac{1}{\lambda}}\right] - 1 \right\rceil$.

2.2. Probability generating function

The probability generating function (pgf) of discrete TIHL-G family of distributions is given by

$$P_Y(s) = E(s^Y) = 1 + 2(s-1)\sum_{y=1}^{\infty} \frac{s^{y-1}\bar{G}^{\lambda}(y+1;\Theta)}{1 + \bar{G}^{\lambda}(y+1;\Theta)}.$$

Then mean and variance are respectively,

$$E(Y) = \sum_{y=1}^{\infty} \frac{2\bar{G}^{\lambda}(y+1;\Theta)}{1 + \bar{G}^{\lambda}(y+1;\Theta)},$$

and

$$V(Y)=\sum_{y=1}^{\infty}\frac{(2y-1)\bar{G}^{\lambda}(y+1;\Theta)}{1+\bar{G}^{\lambda}(y+1;\Theta)}-\left[\sum_{y=1}^{\infty}\frac{2\bar{G}^{\lambda}(y+1;\Theta)}{1+\bar{G}^{\lambda}(y+1;\Theta)}\right]^{2}.$$

Also, the recurrence relation for generating probabilities, is

$$P_Y(y+1;\Theta,\lambda)=\frac{\left[\bar{G}^{\lambda}(y+1;\Theta)-\bar{G}^{\lambda}(y+2;\Theta)\right]\left[1+\bar{G}^{\lambda}(y;\Theta)\right]}{\left[\bar{G}^{\lambda}(y;\Theta)-\bar{G}^{\lambda}(y+1;\Theta)\right]\left[1+\bar{G}^{\lambda}(y+2;\Theta)\right]}P_Y(y;\Theta,\lambda).$$

Different choices of $G(y;\Theta)$ in (6) will give new family of discrete probability distributions. In the next section, we discuss some discrete probability models obtained from this family.

3. Some members of Discrete Type I Half Logistic- General (DTIHL-G) family

3.1. Discrete type I half-logistic uniform distribution

Let $X\sim U(0,\alpha)$ with cdf $G(x;\alpha)=\frac{x}{\alpha},0<x<\alpha$. Then the pmf, cdf and survival function of the discrete type I half-logistic uniform distribution are respectively,

$$P(Y=y)=\frac{2\alpha^{\lambda}\left[(\alpha-y)^{\lambda}-(\alpha-(y+1))^{\lambda}\right]}{\left[\alpha^{\lambda}+(\alpha-y)^{\lambda}\right]\left[\alpha^{\lambda}+(\alpha-(y+1))^{\lambda}\right]};\ y=0,1,...,\alpha-1,$$

$$F(y;\alpha,\lambda)=\frac{\alpha^{\lambda}-\left[\alpha-(y+1)\right]^{\lambda}}{\alpha^{\lambda}+\left[\alpha-(y+1)\right]^{\lambda}},$$

and

$$S(y;\alpha,\lambda)=\frac{2\left[\alpha^{\lambda}-(y+1)\right]^{\lambda}}{\alpha^{\lambda}+\left[\alpha^{\lambda}-(y+1)\right]^{\lambda}}.$$

3.2. Discrete type I half-logistic Lomax distribution

Let X follow the Lomax distribution with shape parameter $\alpha>0$ and scale parameter $\beta>0$ with cdf $G(x;\alpha,\beta)=1-(1+\beta x)^{-\alpha}$. Then the pmf, cdf and survival function of the discrete type I half-logistic Lomax distribution are given by

$$P(Y=y)=\frac{2\left[(1+\beta y)^{-\alpha\lambda}-(1+\beta(y+1))^{-\alpha\lambda}\right]}{\left[1+(1+\beta y)^{-\alpha\lambda}\right]\left[1+(1+\beta(y+1))^{-\alpha\lambda}\right]};\ y=0,1,...\ ,$$

$$F(y;\alpha,\lambda)=\frac{1-(1+\beta y)^{-\alpha\lambda}}{1+(1+\beta y)^{-\alpha\lambda}},$$

and

$$S(y;\alpha,\lambda)=\frac{2(1+\beta(y+1))^{-\alpha\lambda}}{1+(1+\beta(y+1))^{-\alpha\lambda}}.$$

When $\lambda=1$, this distribution becomes discrete half-logistic Lomax distribution.

3.3. Discrete type I half-logistic exponential distribution

Let X follow the exponential distribution with parameter $\alpha > 0$ with cdf $G(x;\alpha) = 1 - e^{-\alpha x}$. By taking $q = e^{-\alpha}, 0 < q < 1$, we get the pmf, cdf and survival function of the discrete type I half-logistic exponential distribution are respectively,

$$P(Y = y) = \frac{2\left[q^{\lambda y} - q^{\lambda(y+1)}\right]}{\left[1 + q^{\lambda y}\right]\left[q^{\lambda(y+1)}\right]};\ y = 0, 1, \ldots ,$$

$$F(y;\alpha,\lambda) = \frac{1 - q^{\lambda y}}{1 + q^{\lambda y}},$$

and

$$S(y;\alpha,\lambda) = \frac{2q^{\lambda(y+1)}}{1 + q^{\lambda(y+1)}}.$$

3.4. Discrete type I half-logistic Fréchet distribution

Let X follow the Fréchet distribution with scale parameter $\alpha > 0$ and shape parameter $\beta > 0$ with cdf $G(x;\alpha,\beta) = e^{-(\alpha/x)^{\beta}}$. By taking $q = e^{-\alpha^{\beta}}, 0 < q < 1$, we get the pmf, cdf and survival function of the discrete type I half-logistic Fréchet distribution as, respectively

$$P(Y = y) = \frac{2\left[(1 - q^{y^{-\beta}})^{\lambda} - (1 - q^{(y+1)^{-\beta}})^{\lambda}\right]}{\left[1 + (1 - q^{y^{-\beta}})^{\lambda}\right]\left[1 + (1 - q^{(y+1)^{-\beta}})^{\lambda}\right]};\ y = 0, 1, \ldots ,$$

$$F(y;q,\beta,\lambda) = \frac{1 - (1 - q^{(y+1)^{-\beta}})^{\lambda}}{1 + (1 - q^{(y+1)^{-\beta}})^{\lambda}},$$

and

$$S(y;q,\beta,\lambda) = \frac{2(1 - q^{(y+1)^{-\beta}})^{\lambda}}{1 + (1 - q^{(y+1)^{-\beta}})^{\lambda}}.$$

In a similar way, by considering different choices of $G(y;\Theta)$ in (6), we can develop several discrete probability distributions. In the next section we study in detail the discrete type I half-logistic Weibull distribution.

4. Discrete type I half-logistic Weibull distribution

Let X follow the Weibull distribution with parameters $\alpha > 0$ and $\beta > 0$ with cdf and survival function are respectively $G(x;\alpha,\beta) = 1 - e^{-\alpha x^{\beta}}$ and $\bar{G}(x;\alpha,\beta) = e^{-\alpha x^{\beta}}$. By taking $q = e^{-\alpha}, 0 < q < 1$ and using (6), the pmf of the new distribution is given by

$$P(Y = y) = \frac{2(q^{\lambda y^{\beta}} - q^{\lambda(y+1)^{\beta}})}{(1 + q^{\lambda y^{\beta}})(1 + q^{\lambda(y+1)^{\beta}})}. \tag{13}$$

We call this distribution the discrete type I half-logistic Weibull (DTIHLW) distribution with parameters q, λ and β and is denoted by DTIHLW(q,λ,β).
When $\beta = 1$, the pmf becomes,

$$P(Y = y) = \frac{2q^{\lambda y}(1 - q^{\lambda})}{(1 + q^{\lambda y})(1 + q^{\lambda(y+1)})},$$

which is the discrete type I half-logistic exponential distribution.

5. Structural properties of DTIHLW(q, λ, β)

Figure 1 shows the shape of pmf of DTIHLW(q, λ, β) distribution for various selection of the parameter values.

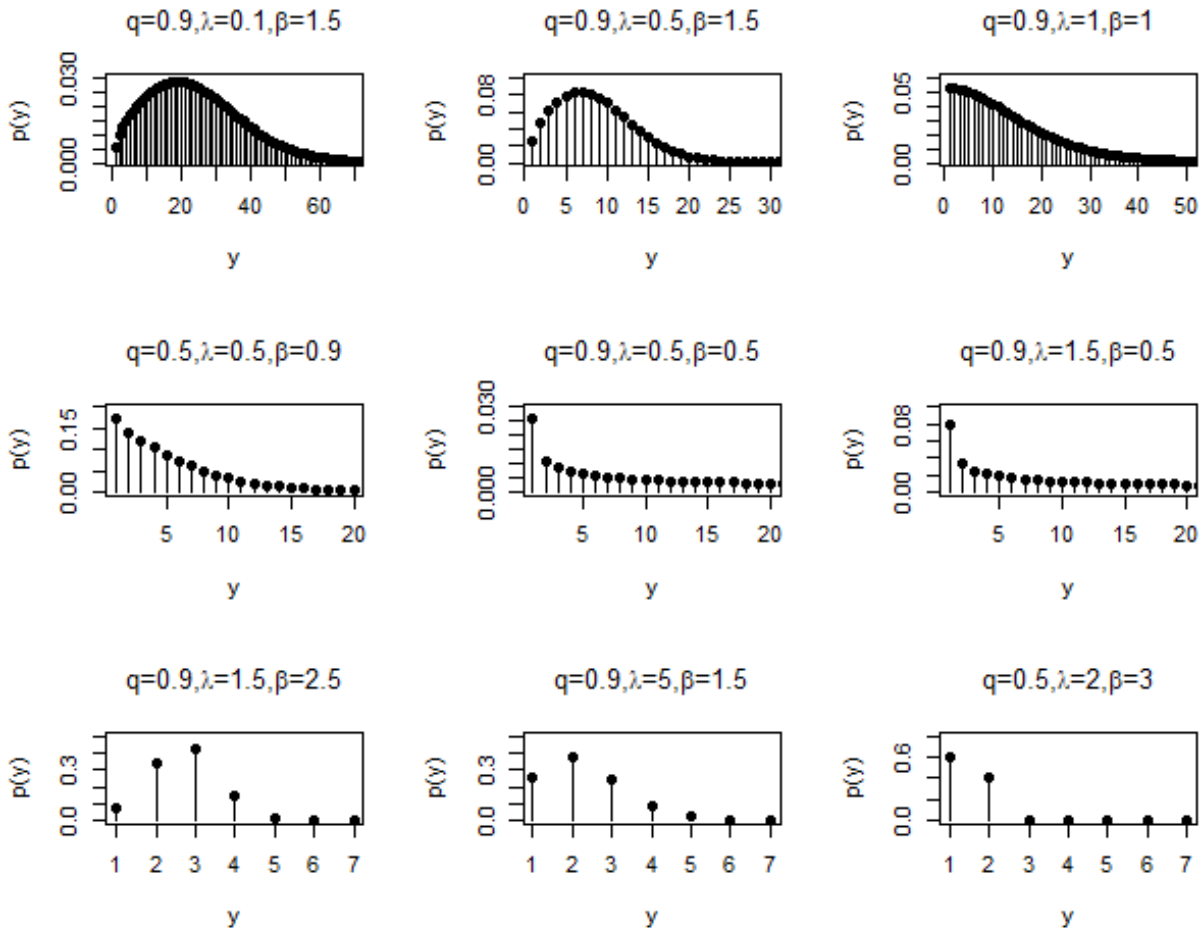


Figure 1: Shapes of the pmf of DTIHLW(q, λ, β) for various parameter values.

Theorem 1: The pmf of DTIHLW(q, λ, β) distribution is log-concave for $\beta \leq 1$.

Proof: From Kus *et al.* (2019), a distribution with pmf $p(y)$ is log-concave if

$$[p(y+1)]^2 > p(y)p(y+2), \tag{14}$$

for all $y \geq 0$.
Under $q \in (0,1)$, $\lambda > 0$ and $\beta \leq 1$ we have

$$\frac{[q^{\lambda(y+1)^\beta} - q^{\lambda(y+2)^\beta}]^2}{(1+q^{\lambda(y+1)^\beta})(1+q^{\lambda(y+2)^\beta})} - \frac{[q^{\lambda y^\beta} - q^{\lambda(y+1)^\beta}][q^{\lambda(y+2)^\beta} - q^{\lambda(y+3)^\beta}]}{(1+q^{\lambda y^\beta})(1+q^{\lambda(y+3)^\beta})} > 0,$$

for all $y \geq 0$. Thus (14) is satisfied by the pmf (13). □

5.1. Cumulative distribution function, survival and hazard rate functions of DTIHLW distribution

The cdf of DTIHLW(q, λ, β) distribution is obtained as

$$F(y; q, \lambda, \beta) = P(Y \leq y) = 1 - P(Y \geq y + 1) = \frac{1 - q^{\lambda(y+1)^{\beta}}}{1 + q^{\lambda(y+1)^{\beta}}}, \tag{15}$$

where $y = 0, 1, 2, ...; 0 < q < 1, \lambda > 0$ and $\beta > 0$.
In particular,

$$F(0) = \frac{1 - q^{\lambda}}{1 + q^{\lambda}},$$

and the proportion of positive values,

$$1 - F(0) = \frac{2q^{\lambda}}{1 + q^{\lambda}}.$$

Also,

$$P(a < Y \leq b) = \frac{1 - q^{\lambda(b+1)^{\beta}}}{1 + q^{\lambda(b+1)^{\beta}}} - \frac{1 - q^{\lambda(a+1)^{\beta}}}{1 + q^{\lambda(a+1)^{\beta}}}.$$

The survival function of DTIHLW(q, λ, β) is given by

$$S(y; q, \lambda, \beta) = P(Y > y) = 1 - P(Y \leq y) = \frac{2q^{\lambda(y+1)^{\beta}}}{1 + q^{\lambda(y+1)^{\beta}}}. \tag{16}$$

The hazard rate function is given by,

$$h(y; q, \lambda, \beta) = \frac{P(Y = y)}{P(Y \geq y)} = \frac{1 - q^{\lambda[(y+1)^{\beta} - y^{\beta}]}}{1 + q^{\lambda(y+1)^{\beta}}}, \tag{17}$$

provided $P(Y \geq y) > 0$. Here note that,

$$\lim_{y \to 0} h(y; q, \lambda, \beta) = \frac{1 - q^{\lambda}}{1 + q^{\lambda}}.$$

Also, when $\lambda > 0$ and $\beta > 1$,

$$\lim_{y \to \infty} h(y; q, \lambda, \beta) = 1,$$

when $\lambda > 0$ and $\beta < 1$,

$$\lim_{y \to \infty} h(y; q, \lambda, \beta) = 0,$$

and when $\lambda > 0$ and $\beta = 1$,

$$\lim_{y \to \infty} h(y; q, \lambda, \beta) = 1 - q^{\lambda}.$$

Figure 2 shows the shape of the hrf of DTIHLW(q, λ, β) for various choices of parameter values. The cumulative hazard function, $H(y; q, \lambda, \beta)$, is given by

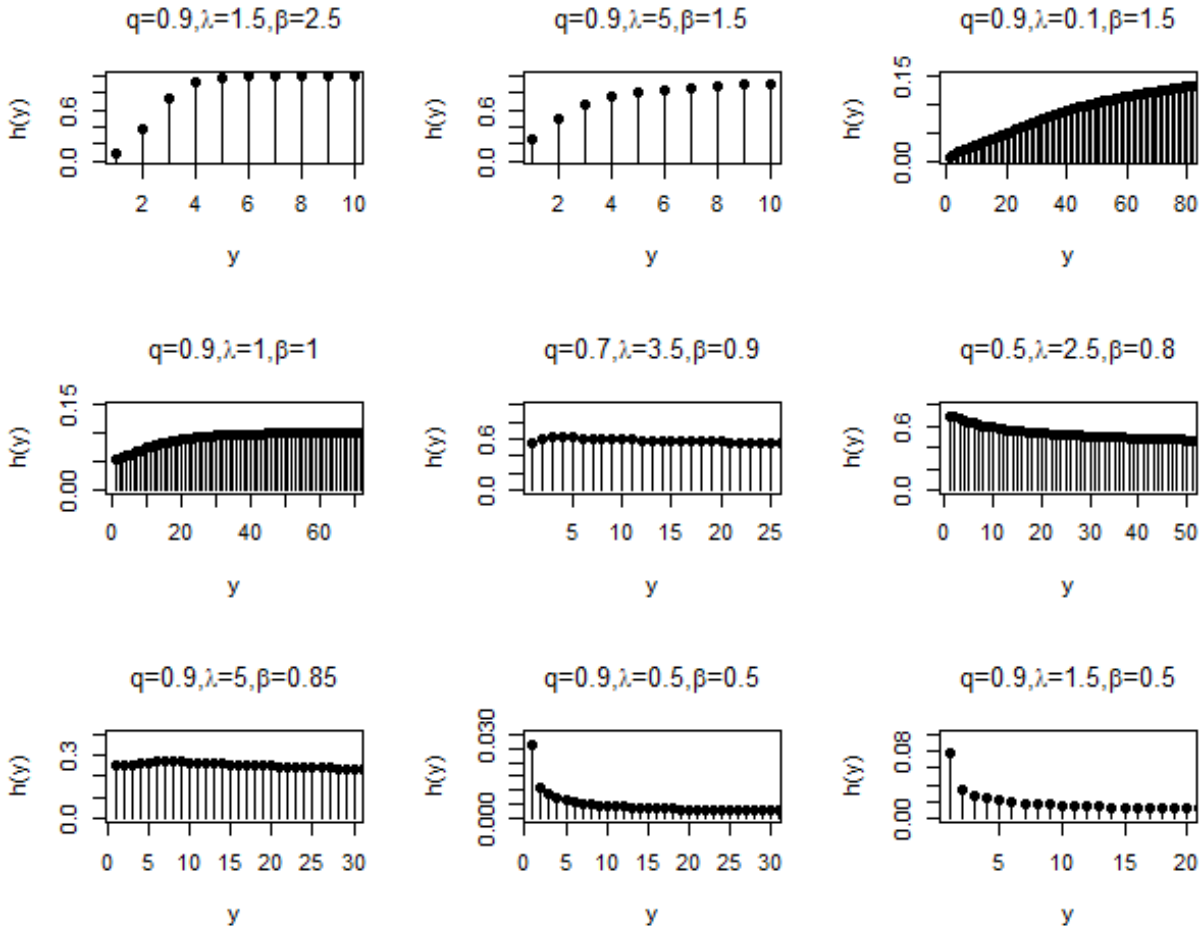


Figure 2: Shapes of the hrf of DTIHLW(q, λ, β) for various parameter values.

$$H(y; q, \lambda, \beta) = \sum_{t=0}^{y} h(t) = \sum_{t=0}^{y} \frac{1 - q^{\lambda[(t+1)^\beta - t^\beta]}}{1 + q^{\lambda(t+1)^\beta}}. \tag{18}$$

The mean residual life (MRL) function (see Jayakumar and Babu (2018)) is given by

$$L(y) = E[(Y - y)/Y \geq y] = \sum_{j \geq y} \prod_{i=y}^{j} \big(1 - h(i)\big) = \sum_{j \geq y} \prod_{i=y}^{j} \frac{1 + q^{-\lambda y^\beta}}{1 + q^{-\lambda(y+1)^\beta}}; \quad y = 0, 1, 2, \ldots . \tag{19}$$

Another expression for MRL by Roy and Gupta (1999) is given by

$$\mu(y) = E[(Y - y)/Y > y] = 1 + L(y+1) = 1 + \sum_{j \geq y+1} \prod_{i=y+1}^{j} \frac{1 + q^{-\lambda y^\beta}}{1 + q^{-\lambda(y+1)^\beta}}; \quad y = 0, 1, 2, \ldots . \tag{20}$$

When $y = 0$, then the MRL function is equal to the mean of the lifetime distribution, that is, $L(0) = \mu$. Thus, we have,

$$\mu(0) = \frac{\mu}{1 - p(0)} = \frac{\mu(1 + q^\lambda)}{2q^\lambda}. \tag{21}$$

Also, the reverse hazard rate function is given by

$$h^*(y) = P(Y = y/Y \leq y) = \frac{2(q^{\lambda y^\beta} - q^{\lambda(y+1)^\beta})}{(1+q^{\lambda y^\beta})(1-q^{\lambda(y+1)^\beta})}. \tag{22}$$

The following Figure 3 shows the change of reverse hrf for given parameter values. The

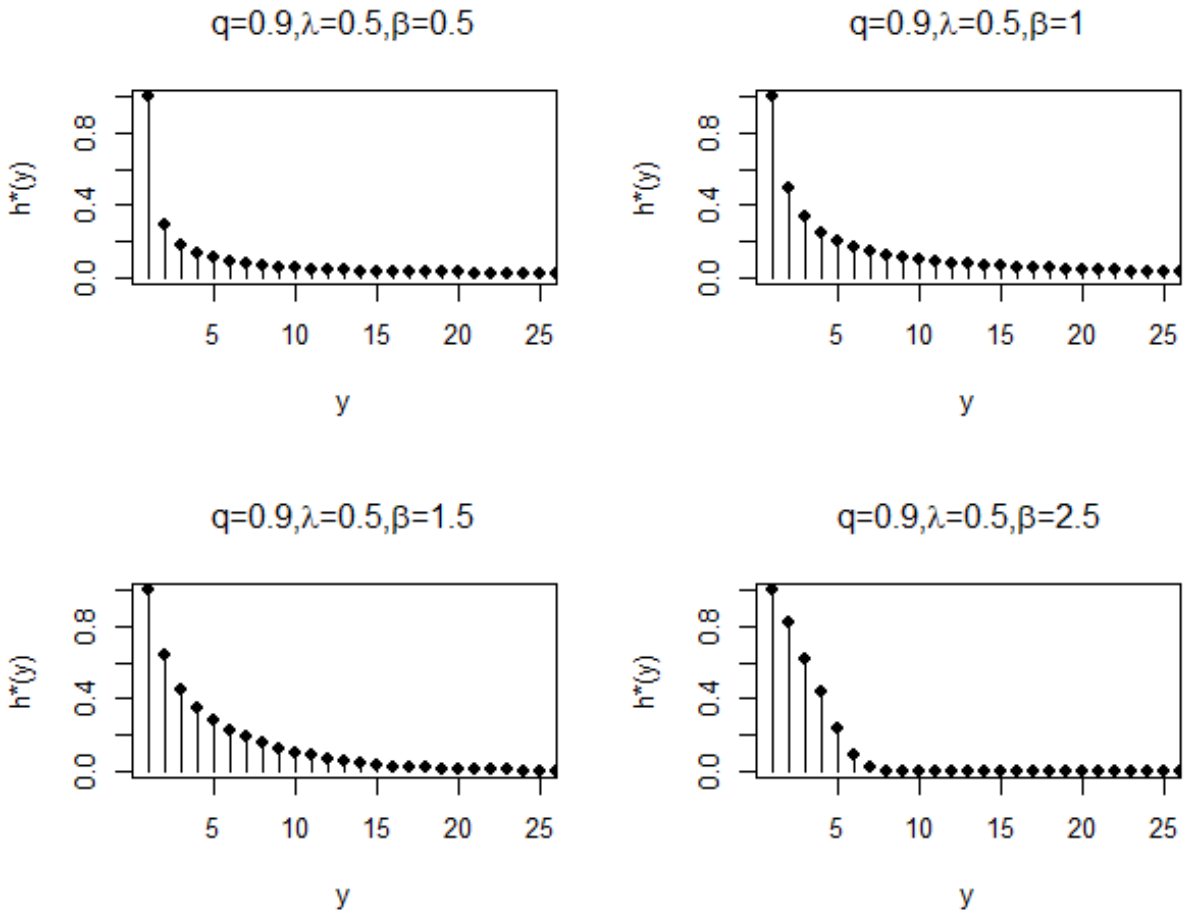


Figure 3: Reverse hrf of DTIHLW(q, λ, β) for various parameter values.

second rate of failure is given by

$$h^{**}(y) = \log\left\{\frac{S(y)}{S(y+1)}\right\} = \log\left\{\frac{q^{\lambda[(y+1)^\beta-(y+2)^\beta]}\left(1+q^{\lambda(y+2)^\beta}\right)}{1+q^{\lambda(y+2)^\beta}}\right\}. \tag{23}$$

5.2. Quantiles

The point y_u is known as the u^{th} quantile of a discrete random variable Y, if it satisfies $P(Y \leq y_u) \geq u$ and $P(Y \geq y_u) \geq 1-u$, see Rohatgi and Saleh (2001). Then we have the following theorem.

Theorem 2: The u^{th} quantile $\phi(u)$ of DTIHLW(q, λ, β) is given by,

$$\phi(u) = \lceil y_u \rceil = \left\lceil \left[ln\left(\frac{1-u}{1+u}\right) \Big/ \lambda \; ln(q) \right]^{\frac{1}{\beta}} - 1 \right\rceil, \tag{24}$$

where $\lceil y_u \rceil$ denotes the smallest integer greater than or equal to y_u.

Proof: Here first we assume that, $P(Y \leq y_u) \geq u$.
That is,

$$\begin{aligned} \frac{1 - q^{\lambda(y_u+1)^\beta}}{1 + q^{\lambda(y_u+1)^\beta}} &\geq u \\ &\Rightarrow \; 1 - q^{\lambda(y_u+1)^\beta} \geq u(1 + q^{\lambda(y_u+1)^\beta}) \\ &\Rightarrow \; \left[\frac{ln\left(\frac{1-u}{1+u}\right)}{\lambda \; ln(q)}\right]^{\frac{1}{\beta}} \leq y_u + 1 \\ &\Rightarrow \; y_u \geq \left[\frac{ln\left(\frac{1-u}{1+u}\right)}{\lambda \; ln(q)}\right]^{\frac{1}{\beta}} - 1, \end{aligned} \tag{25}$$

since $ln(q) < 0$.
Similarly, $P(Y \geq y_u) \geq 1 - u$ gives,

$$y_u \leq \left[\frac{ln\left(\frac{1-u}{1+u}\right)}{\lambda \; ln(q)}\right]^{\frac{1}{\beta}}. \tag{26}$$

From (25) and (26) we get,

$$\left[\frac{ln\left(\frac{1-u}{1+u}\right)}{\lambda \; ln(q)}\right]^{\frac{1}{\beta}} - 1 < y_u \leq \left[\frac{ln\left(\frac{1-u}{1+u}\right)}{\lambda \; ln(q)}\right]^{\frac{1}{\beta}}$$

Hence, $\phi(u)$ is an integer given by,

$$\phi(u) = \lceil y_u \rceil = \left\lceil \left[ln\left(\frac{1-u}{1+u}\right) \Big/ \lambda \; ln(q) \right]^{\frac{1}{\beta}} - 1 \right\rceil,$$

This completes the proof. □

Let U be a random number drawn from a uniform distribution on $(0, 1)$, then a random number Y following DTIHLW(q, λ, β) distribution is obtained by using the expression (24). In particular, the median is given by,

$$\phi\left(\frac{1}{2}\right) = \lceil y_{\frac{1}{2}} \rceil = \left\lceil \left[\frac{-1.099}{\lambda \; ln(q)}\right]^{\frac{1}{\beta}} - 1 \right\rceil.$$

5.3. Probability generating function of DTIHLW(q, λ, β)

The pgf of DTIHLW(q, λ, β) distribution is

$$P_Y(s) = 1 + 2(s-1)\sum_{y=1}^{\infty} \frac{s^{y-1} q^{\lambda(y+1)^\beta}}{1 + q^{\lambda(y+1)^\beta}}. \tag{27}$$

Then the mean is

$$E(Y) = \sum_{y=1}^{\infty} \frac{2q^{\lambda(y+1)^\beta}}{1 + q^{\lambda(y+1)^\beta}}, \tag{28}$$

and the variance is

$$V(Y) = \sum_{y=1}^{\infty} \frac{2(2y-1)q^{\lambda(y+1)^\beta}}{1 + q^{\lambda(y+1)^\beta}} - \left[\sum_{y=1}^{\infty} \frac{2q^{\lambda(y+1)^\beta}}{1 + q^{\lambda(y+1)^\beta}}\right]^2. \tag{29}$$

5.4. Moments

The r^{th} moment about origin of DTIHLW(q, λ, β) is given by

$$\mu_r\prime = E(Y^r) = 2\sum_{y=0}^{\infty} \frac{y^r (q^{\lambda y^\beta} - q^{\lambda(y+1)^\beta})}{(1 + q^{\lambda y^\beta})(1 + q^{\lambda(y+1)^\beta})}. \tag{30}$$

For given values of the parameters, (30) can be numerically computed using R-programming. Table 1 shows the raw and central moments, skewness, and kurtosis for the given values of q, λ and β.

5.5. Order statistics

Let $Y_1, Y_2, ..., Y_n$ be n random samples taken from DTIHLW(q, λ, β) and let $Y_{(1)}$,$Y_{(2)}$,..., $Y_{(n)}$ denote the corresponding order statistics. Then the cdf for the k^{th} order statistic, say $Z = Y_{(k)}$, is given by

$$F_Z(z) = \sum_{j=k}^{n} \binom{n}{j} F^j(z)[1 - F(z)]^{n-j}. \tag{31}$$

Using the binomial expansion for $[1 - F(z)]^{n-j}$, we get

$$\begin{aligned} F_Z(z) &= \sum_{j=k}^{n}\sum_{i=0}^{n-j} \binom{n}{j}\binom{n-j}{i}(-1)^i [F(z)]^{i+j} \\ &= \sum_{j=k}^{n}\sum_{i=0}^{n-j} \binom{n}{j}\binom{n-j}{i}(-1)^i \left[\frac{1 - q^{\lambda(z+1)^\beta}}{1 + q^{\lambda(z+1)^\beta}}\right]^{i+j}. \end{aligned} \tag{32}$$

Table 1: Moments, skewness and kurtosis for $q = 0.5, \lambda = 1.0$ and various choices of β.

Parameter	Raw moments	Central moments	Skewness	Kurtosis
$\beta = 0.5$	$\mu'_1 = 6.42$ $\mu'_2 = 190.36$ $\mu'_3 = 12507.7$ $\mu'_4 = 1482469$	$\mu_2 = 149.2$ $\mu_3 = 9370.9$ $\mu_4 = 1203285$	5.14	54.09
$\beta = 1.0$	$\mu'_1 = 1.53$ $\mu'_2 = 5.15$ $\mu'_3 = 23.98$ $\mu'_4 = 143.84$	$\mu_2 = 2.81$ $\mu_3 = 7.50$ $\mu_4 = 53.02$	1.59	6.69
$\beta = 1.5$	$\mu'_1 = 0.98$ $\mu'_2 = 1.74$ $\mu'_3 = 3.75$ $\mu'_4 = 9.56$	$\mu_2 = 0.78$ $\mu_3 = 0.53$ $\mu_4 = 2.11$	0.76	3.44
$\beta = 2$	$\mu'_1 = 0.79$ $\mu'_2 = 1.04$ $\mu'_3 = 1.57$ $\mu'_4 = 2.69$	$\mu_2 = 0.42$ $\mu_3 = 0.09$ $\mu_4 = 0.47$	0.32	2.69
$\beta = 2.5$	$\mu'_1 = 0.71$ $\mu'_2 = 0.78$ $\mu'_3 = 0.94$ $\mu'_4 = 1.25$	$\mu_2 = 0.29$ $\mu_3 = -0.02$ $\mu_4 = 0.19$	-0.107	2.42
$\beta = 3$	$\mu'_1 = 0.67$ $\mu'_2 = 0.69$ $\mu'_3 = 0.72$ $\mu'_4 = 0.78$	$\mu_2 = 0.24$ $\mu_3 = -0.06$ $\mu_4 = 0.10$	-0.54	1.82

The pmf of the k^{th} order statistics is obtained as

$$\begin{aligned} f_Z(z) &= F_Z(z) - F_Z(z-1) \\ &= \sum_{j=k}^{n} \sum_{i=0}^{n-j} \binom{n}{j} \binom{n-j}{i} (-1)^i \\ &\quad \left(\left[\frac{1 - q^{\lambda(z+1)^\beta}}{1 + q^{\lambda(z+1)^\beta}} \right]^{i+j} - \left[\frac{1 - q^{\lambda z^\beta}}{1 + q^{\lambda z^\beta}} \right]^{i+j} \right) \\ &= \sum_{j=k}^{n} \sum_{i=0}^{n-j} \binom{n}{j} \binom{n-j}{i} (-1)^i \\ &\quad \frac{\left[(1 - q^{\lambda(z+1)^\beta})(1 + q^{\lambda z^\beta}) \right]^{i+j} - \left[(1 + q^{\lambda(z+1)^\beta})(1 - q^{\lambda z^\beta}) \right]^{i+j}}{\left[(1 + q^{\lambda(z+1)^\beta})(1 + q^{\lambda z^\beta}) \right]^{i+j}}. \end{aligned} \tag{33}$$

5.6. Infinite divisibility

From Steutel and van Harn (2004), we have the following result.

Lemma 1: If $p_y = P(Y = y), y \in Z_+$, is infinitely divisible, then we have $p_y \leq e^{-1} = 0.3679$, for all $y \in N$.

From the above Lemma, using (13), we have arrived the condition that DTIHLW distribution is infinitely divisible for a given q, λ and β if it satisfies

$$q^{\lambda y^\beta} \leq q^{\lambda(y+1)^\beta} + \frac{1}{2e}(1 + q^{\lambda y^\beta})(1 + q^{\lambda(y+1)^\beta}).$$

But we can show that $p_y > 0.3679$ for some values of $y \in N, \lambda, \beta$ and q. We take $\lambda = 1.5, \beta = 2$ and $q = 0.5$, then we have $p_1 = 0.4920 > 0.3679$. This shows that the DTIHLW distribution is not infinitely divisible.

5.7. Stress-strength parameter

The stress-strength parameter R is a measure of component reliability. Let the random variable Y be the strength of a component which is subjected to a random stress Z. The estimation of R when Y and Z are independently and identically distributed (iid) has been discussed in the literature by many authors. For a detailed study, one can see Kotz *et al.* (2003). In discrete case, the stress-strength model is defined as,

$$R = P(Y > Z) = \sum_{y=0}^{\infty} p_Y(y)\ F_Z(y), \tag{34}$$

where, $p_Y(y)$ and $F_Z(y)$ are the pmf and cdf of the independent discrete random variables Y and Z, respectively. The stress-strength models are useful in various fields such as medicine, engineering, and psychology. Let $Y \sim DTIHLW(\theta_1)$ and $Z \sim DTIHLW(\theta_2)$, where $\theta_1 = (q_1, \lambda_1, \beta_1)^T$ and $\theta_2 = (q_2, \lambda_2, \beta_2)^T$. Then, using (13) and (15), we have,

$$R = \sum_{y=0}^{\infty} \frac{2\left[q_1^{\lambda_1 y^{\beta_1}} - q_1^{\lambda_1(y+1)y^{\beta_1}}\right]\left[1 - q_2^{\lambda_2(y+1)^{\beta_2}}\right]}{\left[1 + q_1^{\lambda_1 y^{\beta_1}}\right]\left[1 + q_1^{\lambda_1(y+1)^{\beta_1}}\right]\left[1 + q_2^{\lambda_2(y+1)^{\beta_2}}\right]}. \tag{35}$$

The stress strength reliability parameter for different parameter values are numerically computed and presented in Table 2. We see that the value of stress-strength parameter is decreasing when β_1 increases and increasing when β_2 increases.

5.8. Likelihood function of DTIHLW distribution

Consider a random sample $(y_1, y_2, ..., y_n)$ of size n, from the DTIHLW(q, λ, β). Then, the likelihood function is given by,

$$L = \frac{2^n \prod_{i=1}^n (q^{\lambda y_i^\beta} - q^{\lambda(y_i+1)^\beta})}{\prod_{i=1}^n (1 + q^{\lambda y_i^\beta}) \prod_{i=1}^n (1 + q^{\lambda(y_i+1)^\beta})}. \tag{36}$$

The log-likelihood function is,

$$\begin{aligned} ln(L) &= n\ ln(2) + \sum_{i=1}^n ln(q^{\lambda y_i^\beta} - q^{\lambda(y_i+1)^\beta}) \\ &\quad - \sum_{i=1}^n ln(1 + q^{\lambda y_i^\beta}) - \sum_{i=1}^n ln(1 + q^{\lambda(y_i+1)^\beta}). \end{aligned} \tag{37}$$

Table 2: Value of stress-strength parameter (R) for various choices of parameters.

$q_1 = 0.5, q_2 = 0.5$				
$\lambda_1 = 0.5, \lambda_2 = 0.5$				
$\beta_1 \rightarrow$ $\beta_2 \downarrow$	0.5	1.0	1.5	2.0
0.5	0.5236	0.3268	0.2742	0.2528
1.0	0.7374	0.5576	0.4348	0.3744
1.5	0.7989	0.7091	0.6062	0.5253
2.0	0.8257	0.7799	0.7213	0.6592
$\lambda_1 = 0.5, \lambda_2 = 1.5$				
$\beta_1 \rightarrow$ $\beta_2 \downarrow$	0.5	1.0	1.5	2.0
0.5	0.8333	0.7337	0.6743	0.6424
1.0	0.8889	0.8569	0.8193	0.7860
1.5	0.9029	0.8928	0.8796	0.8645
2.0	0.9082	0.9054	0.9016	0.8971
$\lambda_1 = 1, \lambda_2 = 0.5$				
$\beta_1 \rightarrow$ $\beta_2 \downarrow$	0.5	1.0	1.5	2.0
0.5	0.3477	0.2549	0.2325	0.2236
1.0	0.5363	0.3837	0.3224	0.2969
1.5	0.6242	0.5129	0.4342	0.3916
2.0	0.6679	0.6013	0.5392	0.4946

The likelihood equations are the following

$$\begin{aligned}\frac{\partial ln(L)}{\partial q} &= \sum_{i=1}^{n} \frac{y_i^{\beta} q^{\lambda y_i^{\beta}-1} - (y_i+1)^{\beta} q^{\lambda (y_i+1)^{\beta}-1}}{q^{\lambda y_i^{\beta}} - q^{\lambda (y_i+1)^{\beta}}} \\ &\quad - \sum_{i=1}^{n} \frac{y_i^{\beta} q^{\lambda y_i^{\beta}-1}}{1+q^{\lambda y_i^{\beta}}} - \sum_{i=1}^{n} \frac{(y_i+1)^{\beta} q^{\lambda (y_i+1)^{\beta}-1}}{1+q^{\lambda (y_1+1)^{\beta}}} = 0, \end{aligned} \tag{38}$$

$$\begin{aligned}\frac{\partial ln(L)}{\partial \lambda} &= \sum_{i=1}^{n} \frac{y_i^{\beta} q^{\lambda y_i^{\beta}} - (y_i+1)^{\beta} q^{\lambda (y_i+1)^{\beta}}}{q^{\lambda y_i^{\beta}} - q^{\lambda (y_i+1)^{\beta}}} \\ &\quad - \sum_{i=1}^{n} \frac{y_i^{\beta} q^{\lambda y_i^{\beta}}}{1+q^{\lambda y_i^{\beta}}} - \sum_{i=1}^{n} \frac{(y_i+1)^{\beta} q^{\lambda (y_i+1)^{\beta}}}{1+q^{\lambda (y_i+1)^{\beta}}} = 0, \end{aligned} \tag{39}$$

and

$$\begin{aligned}\frac{\partial ln(L)}{\partial \beta} &= \sum_{i=1}^{n} \frac{ln(y_i) y_i^{\beta} q^{\lambda y_i^{\beta}} - ln(y_i+1)(y_i+1)^{\beta} q^{\lambda (y_i+1)^{\beta}}}{q^{\lambda y_i^{\beta}} - q^{\lambda (y_i+1)^{\beta}}} \\ &\quad - \sum_{i=1}^{n} \frac{ln(y_i) y_i^{\beta} q^{\lambda y_i^{\beta}}}{1+q^{\lambda y_i^{\beta}}} - \sum_{i=1}^{n} \frac{ln(y_i+1)(y_i+1)^{\beta} q^{\lambda (y_i+1)^{\beta}}}{1+q^{\lambda (y_i+1)^{\beta}}} = 0. \end{aligned} \tag{40}$$

These equations do not have explicit solutions and their solutions must be obtained numerically by using statistical software like *nlm* or *optim* package in R programming.

We compute the maximized unrestricted and restricted log-likelihood ratio (LR) test statistic for testing on some DTIHLW submodels. The LR test statistic can be used to check whether DTIHLW distribution for a given data set is statistically superior to the submodels. For example, $H_0 : \beta = 1$ versus $H_1 : \beta \neq 1$ is equivalent to compare the DTIHLW distribution and DTIHLE distribution. Here the LR test statistic reduces to $\omega = 2[l(\hat{q}, \hat{\lambda}, \hat{\beta}) - l(\hat{q}', \hat{\lambda}', 1)]$, where $(\hat{q}, \hat{\lambda}, \hat{\beta})$ and $(\hat{q}', \hat{\lambda}')$ are the MLEs under H_1 and H_0, respectively. The test statistic ω is asymptotically (as $n \to \infty$) distributed as $\chi^2_{(k)}$, where k is the length of the parameter vector of interest. The LR test rejects H_0 if $\omega > \chi^2_{(k,\alpha)}$ where $\chi^2_{(k,\alpha)}$ denotes the upper $(1-\alpha)100\%$ quantile of the $\chi^2_{(k)}$ distribution.

5.9. Simulation study

This section demonstrates the performance of the MLEs of the model parameters of DTIHLW distribution using Monte Carlo simulation for various sample sizes and for selected parameter values. The algorithm for the simulation study are as follows:

Step 1. Input the value of replication (N);
Step 2. Specify the sample size n and the values of the parameters q, λ and β;
Step 3. Generate u_i from $U(0,1),\ i = 1, 2, ..., n$;
Step 4. Obtain the random observations from the DTIHLW distribution using (24);
Step 5. Compute the MLEs of the three parameters;
Step 6. Repeat steps 3 to 5, N times;
Step 7. Compute the parameter estimate, standard error of estimate, average bias, mean square error (MSE) and coverage probability (CP) for each parameter.

Here the expected value of the estimator is

$E(\hat{\theta}) = \frac{1}{N}\sum_{i=1}^{N}\hat{\theta}_i$, $E(SE(\hat{\theta})) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(-\frac{\partial^2 \log(L)}{\partial\theta_i^2}\right)}$,

Average Bias $= \frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}_i - \theta)$, $MSE(\hat{\theta}) = \frac{1}{N}\sum_{i=1}^{N}(\hat{\theta}_i - \theta)^2$ and

CP = Probability of $\theta_i \in \left(\hat{\theta}_i \pm 1.96\sqrt{-\frac{\partial^2 \log(L)}{\partial\theta_i^2}}\right)$.

We take random samples of size n=50, 100, 200 and 500 respectively. The MLEs of the parameter vector $\theta = (q, \lambda, \beta)^T$ are determined by maximizing the log-likelihood function given in (37) by using the *optim* package in **R** programming based on each generated samples. This simulation is repeated 1000 times and the average estimate and its standard error, average bias, MSE and CP are computed and presented in Table 3. From Table 3, it can be seen that, as sample size increases the estimates of bias and MSE are decreases. Also note that the CP values are quite closer to the 95% nominal level.

6. Applications

In order to check the use of DTIHLW distribution for real life data modeling, we consider three data sets. The first data set is continuous measurement of flood peaks (in m^3/s) of the Wheaton River near Carcross in Yukon Territory, Canada for the years 1958-1984. This data was analyzed by Choulakian and Stephens (2001). The data are as follows: 1.7 2.2 14.4 1.1 0.4 20.6 5.3 0.7 13.0 12.0 9.3 1.4 18.7 8.5 25.5 11.6 14.1 22.1 1.1 2.5 14.4 1.7 37.6 0.6 2.2 39.0 0.3 15.0 11.0 7.3 22.9 1.7 0.1 1.1 0.6 9.0 1.7 7.0 20.1 0.4 14.1 9.9 10.4 10.7

Table 3: The MLE, standard error, average bias, MSE and CP for given parameters.

Parameter(θ)	Samples(n)	$E(\hat{\theta})(E(SE(\hat{\theta})))$	Average bias	MSE	CP
$q = 0.5$	50	0.368(0.223)	-0.121	0.115	87.5
	100	0.412(0.192)	-0.108	0.101	88.1
	200	0.432(0.115)	-0.097	0.093	90.2
	500	0.486(0.102)	-0.066	0.088	92.8
$\lambda = 1.5$	50	1.733(0.196)	0.228	0.114	91.6
	100	1.692(0.188)	0.119	0.091	91.9
	200	1.544(0.094)	0.105	0.077	92.7
	500	1.539(0.023)	0.099	0.061	94.1
$\beta = 0.9$	50	0.633(0.251)	-0.129	0.025	92.8
	100	0.701(0.190)	-0.116	0.021	93.5
	200	0.876(0.022)	-0.064	0.009	94.2
	500	0.899(0.013)	-0.027	0.003	94.9
$q = 0.8$	50	0.759(0.142)	-0.183	0.226	91.2
	100	0.773(0.136)	-0.086	0.149	93.4
	200	0.792(0.084)	-0.043	0.063	93.8
	500	0.803(0.048)	0.008	0.041	94.7
$\lambda = 1.0$	50	0.836(0.362)	-0.161	0.447	90.7
	100	0.881(0.143)	-0.133	0.219	91.3
	200	0.934(0.081)	-0.096	0.124	93.3
	500	0.962(0.016)	-0.077	0.019	93.9
$\beta = 1.5$	50	1.543(0.116)	0.039	0.118	94.1
	100	1.514(0.108)	0.026	0.103	94.4
	200	1.503(0.099)	0.021	0.081	95.2
	500	1.501(0.086)	0.018	0.011	95.8

30.0 3.6 5.6 30.8 13.3 4.2 25.5 3.4 11.9 21.5 27.6 36.4 2.7 64.0 1.5 2.5 27.4 1.0 27.1 20.2 16.8 5.3 9.7 27.5 2.5 27.0 1.9 2.8. Since the data set is continuous, here first we discretize the data by considering the floor value (y) and fitted the new distribution for the y values.

The second data set is the daily ozone level measurements (in ppm x 1000) taken from Nadarajah (2008) and are as follows: 7 115 79 31 9 8 45 61 23 28 19 23 35 59 21 23 32 48 22 44 28 4 7 65 24 13 18 11 27 44 21 73 12 1 10 110 23 28 36 30 85 89 20 80 41 6 97 122 32 135 34 21 82 73 16 14 23 52 168 24 18 39 20 45 13 14 71 108 9 18 11 29 16 21 46 16 37 63 44 13 12 59 84 7 20 64 118 36 37 50 76 23 13 39 85 14 49 9 96 30 32 16 78 14 64 78 91 18 40 35 47 20 77 66 97 11.

The third data set is from Eliwa *et al.* (2021) which represents the daily new deaths due to COVID-19 in China from 23 January to 28 March, 2019. The data are: 8 16 15 24 26 26 38 43 46 45 57 64 65 73 73 86 89 97 108 97 146 121 143 142 105 98 136 114 118 109 97 150 71 52 29 44 47 35 42 31 38 31 30 28 27 22 17 22 11 7 13 10 14 13 11 8 3 7 6 9 7 4 6 5 3 5.

We fit DTIHLW(q, λ, β) distribution for the three data sets. The fit of the data sets are compared with six competitive models, respectively, type I half-logistic exponential (DTIHLE) distribution, a sub model of the proposed distribution, discrete Weibull geometric

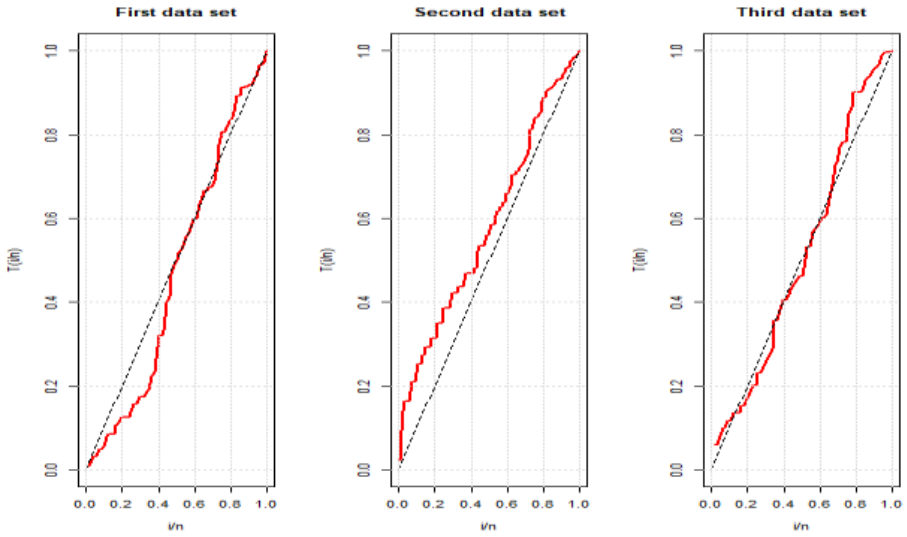


Figure 4: The TTT plots of the three data sets.

(DWG) distribution of Jayakumar and Babu (2018), exponentiated discrete Weibull (EDW) distribution of Nekoukhou and Bidram (2015), discrete modified Weibull (DMW) distribution of Nooghabi *et al.* (2011), discrete logistic (DLOG) distribution of Chakraborty and Chakravarty (2016), discrete Weibull (DW) distribution of Nakagawa and Osaki (1975).

Descriptive statistics of the three data sets are shown in Table 4. The Total Time on Test (TTT) plot of the three data sets are shown in Figure 4.

Table 4: Descriptive statistics for the three data sets

Data	Samples(n)	Mean	SD	Min.	Max.	Skewness	Kurtosis
First set	72	12.204	12.297	0.1	64.0	1.304	3.189
Second set	116	42.129	32.988	1	168	1.242	1.290
Third set	66	49.742	43.873	3	150	0.837	2.450

The values of the log-likelihood function $(-\log L)$, the statistics Kolmogorov-Smirnov (K-S), Akaike Information Criterion (AIC), Akaike Information Criterion with correction(CAIC) and Bayesian Information Criterion(BIC) are calculated for the seven distributions in order to verify which distribution fits better to these data. The better distribution corresponds to smaller $K-S$, $-\log L$, AIC, $CAIC$, BIC values and high p value. Here, AIC=$-2\log L+2k$, CAIC=$-2\log L+(\frac{2kn}{n-k-1})$ and BIC=$-2\log L+k\log n$ where, L is the likelihood function evaluated at the maximum likelihood estimates, k is the number of parameters and n is the sample size.

The values in Table 5 shows that the $DTIHLW$ distribution leads to a better fit compared to the other six models. Figure 5, shows the fitted pdf and cdf with the empirical distribution of the first data set. The LR test statistic to test the hypothesis $H_0 : \beta = 1$ versus $H_0 : \beta \neq 1$ for the first data set is $\omega = 7.264 > 3.841$ with p value 0.0070. So we reject the null hypothesis.

Table 5: Parameter estimates and goodness of fit for the first data set

Model	ML estimates	-log L	AIC	CAIC	BIC	K-S	p-value
DTIHLW	$\hat{q} = 0.758$ $\hat{\lambda} = 0.794$ $\hat{\beta} = 0.765$	251.99	509.97	510.32	516.80	0.109	0.351
EDW	$\hat{q} = 0.866$ $\hat{\lambda} = 0.831$ $\hat{\beta} = 1.089$	252.24	510.48	510.83	517.31	0.125	0.208
DWG	$\hat{q} = 0.123$ $\hat{\lambda} = 0.900$ $\hat{\beta} = 0.912$	252.25	510.49	510.85	517.33	0.136	0.140
DMW	$\hat{q} = 0.917$ $\hat{\lambda} = 0.870$ $\hat{\beta} = 1.016$	253.53	513.06	513.41	519.89	0.137	0.133
DTIHLE	$\hat{q} = 0.121$ $\hat{\lambda} = 0.051$	255.62	515.24	515.42	519.79	0.187	0.013
DW	$\hat{q} = 0.779$ $\hat{\beta} = 0.630$	257.52	519.04	519.21	523.59	0.159	0.051
DLOG	$\hat{q} = 0.664$ $\hat{\lambda} = 9.382$	280.19	564.37	564.55	568.93	0.279	$2.7\text{x}10^{-5}$

Table 6: Parameter estimates and goodness of fit for the second data set

Model	ML estimates	-log L	AIC	CAIC	BIC	K-S	p-value
DTIHLW	$\hat{q} = 0.954$ $\hat{\lambda} = 0.428$ $\hat{\beta} = 1.133$	545.24	1096.49	1096.70	1104.75	0.079	0.464
EDW	$\hat{q} = 0.939$ $\hat{\lambda} = 0.878$ $\hat{\beta} = 1.666$	548.47	1102.93	1103.15	1111.19	0.148	0.013
DWG	$\hat{q} = 0.162$ $\hat{\lambda} = 0.966$ $\hat{\beta} = 0.879$	560.95	1127.91	1128.12	1136.17	0.207	$9.3\text{x}10^{-5}$
DMW	$\hat{q} = 0.978$ $\hat{\lambda} = 0.828$ $\hat{\beta} = 1.010$	551.36	1108.71	1108.93	1116.98	0.104	0.159
DTIHLE	$\hat{q} = 0.101$ $\hat{\lambda} = 0.014$	548.59	1101.18	1101.28	1106.68	0.105	0.153
DW	$\hat{q} = 0.989$ $\hat{\beta} = 1.158$	547.65	1099.31	1099.41	1104.81	0.100	0.193
DLOG	$\hat{q} = 0.946$ $\hat{\lambda} = 38.115$	567.74	1139.48	1139.59	1144.99	0.134	0.032

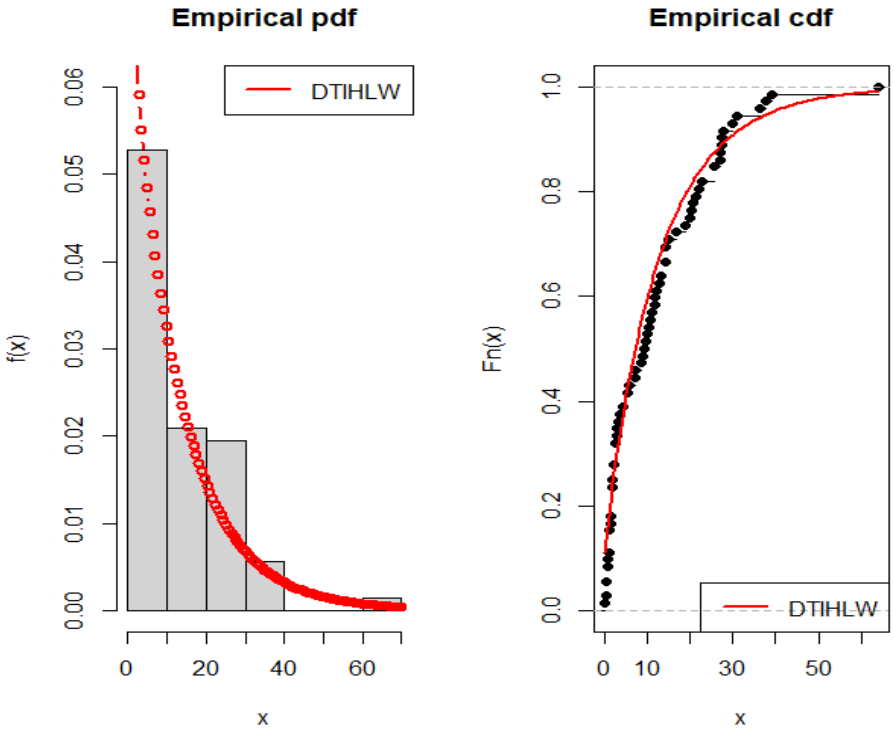


Figure 5: Fitted pdf and cdf plots for the first data set

The values in Table 6 indicates that the DTIHLW distribution leads to a better fit compared to the other six models. Figure 6, shows the fitted pdf and cdf with the empirical distribution of the second data set. The LR test statistic to test the hypothesis $H_0 : \beta = 1$ versus $H_0 : \beta \neq 1$ for the second data set is $\omega = 6.7 > 3.841$ with p value 0.0096. So we reject the null hypothesis. The values in Table 7 indicates that the DTIHLW distribution leads to a better fit compared to the other six models. Figure 7, shows the fitted pdf and cdf with the empirical distribution of the second data set. The LR test statistic to test the hypothesis $H_0 : \beta = 1$ versus $H_0 : \beta \neq 1$ for the third data set is $\omega = 42.3 > 3.841$ with p value $8.02x10^{-11}$. So we reject the null hypothesis.

7. Conclusion and future works

The discrete version of the Type I half logistic distributions was introduced. Several members of this family such as discrete type I half-logistic uniform, discrete type I half-logistic Lomax, discrete type I half-logistic exponential, discrete type I half-logistic Fréchet and discrete type I half-logistic Weibull distributions were specified. Some properties of the discrete type I half-logistic Weibull distribution were studied. The three parameters of the new distribution were estimated using maximum likelihood method and a simulation study was conducted to check the performance of the method. Three real data applications shows that this model is suitable for modeling discrete data. Since the likelihood equations of the present distribution are highly non-linear equations and it is difficult to study the existence and uniqueness of the MLE's of parameters, so we propose further studies in this direction as future work.

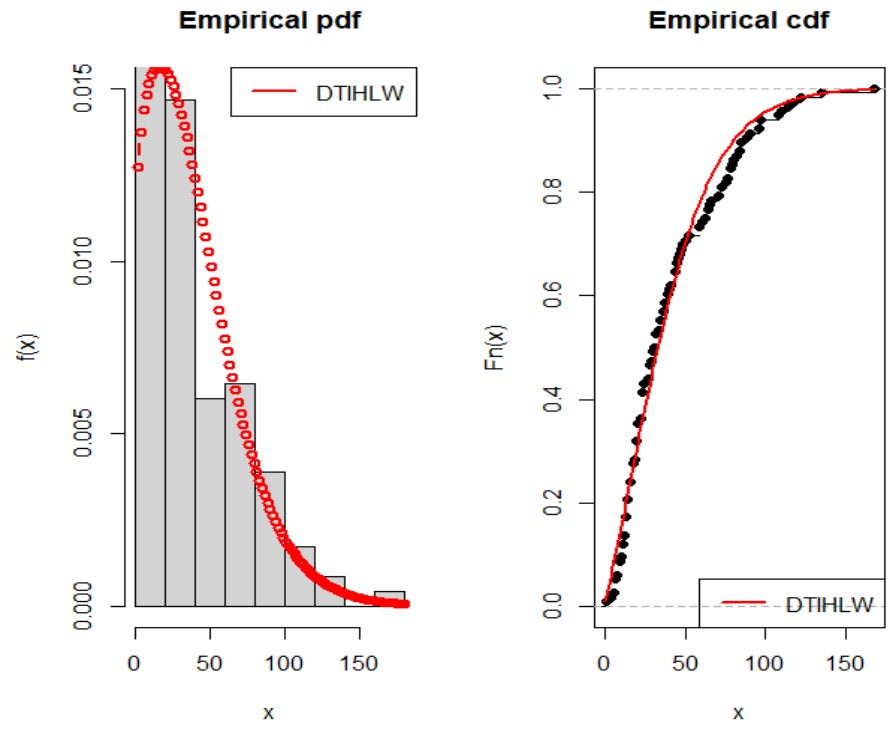


Figure 6: Fitted pdf and cdf plots for the second data set.

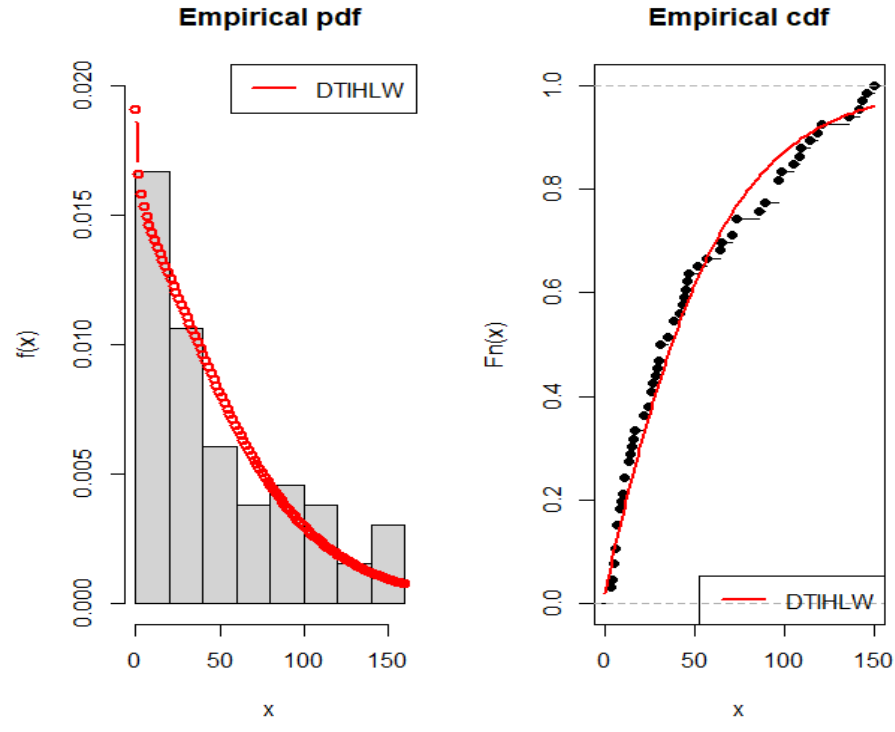


Figure 7: Fitted pdf and cdf plots for the third data set.

Table 7: Parameter estimates and goodness of fit for the third data set.

Model	ML estimates	-log L	AIC	CAIC	BIC	K-S	p-value
DTIHLW	$\hat{q} = 0.798$ $\hat{\lambda} = 0.169$ $\hat{\beta} = 0.922$	324.82	655.64	656.02	662.21	0.091	0.644
EDW	$\hat{q} = 0.899$ $\hat{\lambda} = 0.702$ $\hat{\beta} = 1.716$	325.96	657.92	658.31	664.49	0.114	0.357
DWG	$\hat{q} = 0.645$ $\hat{\lambda} = 0.412$ $\hat{\beta} = 0.739$	1137.19	2280.37	2280.76	2286.94	0.568	$2.2\text{x}10^{-16}$
DMW	$\hat{q} = 0.877$ $\hat{\lambda} = 0.437$ $\hat{\beta} = 1.001$	352.59	711.18	711.57	717.74	0.256	$3.5\text{x}10^{-4}$
DTIHLE	$\hat{q} = 0.305$ $\hat{\lambda} = 0.042$	345.95	695.89	696.08	700.27	0.255	$3.6\text{x}10^{-4}$
DW	$\hat{q} = 0.784$ $\hat{\beta} = 0.531$	357.11	718.22	718.41	722.59	0.391	$3.5x10^{-9}$
DLOG	$\hat{q} = 0.971$ $\hat{\lambda} = 4.885$	366.55	737.10	737.29	741.41	0.494	$2.2x10^{-14}$

Acknowledgements

The authors would like to express their thanks to the Editor and the anonymous reviewer for their constructive comments and suggestions, which significantly improved the presentation of the article.

Conflict of interest

The authors do not have any financial or non-financial conflict of interest to declare for the research work included in this article.

References

Almalki, S. J. and Nadarajah, S. (2014). A new discrete modified Weibull distribution. *IEEE Transactions on Reliability*, **63**, 68–80.

Bakouch, H. S., Jazi, M. A., and Nadarajah, S. (2014). A new discrete distribution. *Statistics*, **48**, 200–240.

Balakrishnan, N. (1991). *Handbook of the Logistic Distribution*. CRC Press.

Bebbington, M., Lai, C. D., Wellington, M., and Zitikis, R. (2012). The discrete additive Weibull distribution: A bathtub-shaped hazard for discontinuous failure data. *Reliability Engineering and System Safety*, **106**, 37–44.

Chakraborty, S. (2015). Generating discrete analogues of continuous probability distributions - a survey of methods and constructions. *Journal of Statistical Distributions and Applications*, **2**, 1–30.

Chakraborty, S. and Chakravarty, D. (2012). Discrete gamma distribution: properties and parameter estimation. *Communications in Statistics - Theory and Methods*, **41**, 3301–3324.

Chakraborty, S. and Chakravarty, D. (2016). A new discrete probability distribution with integer support on $(-\infty, \infty)$. *Communications in Statistics - Theory and Methods*, **45**, 492–505.

Chipepa, F., Charmunorwa, S., Oluyede, B., Makubate, B., and Zidana, C. (2022). The exponentiated half logistic-generalized-G power series class of distributions: Properties and applications. *Journal of Probability and Statistical Science*, **20**, 21–40.

Choulakian, V. and Stephens, M. A. (2001). Goodness-of-fit tests for the generalized Pareto distribution. *Technometrics*, **43**, 478–484.

Cordeiro, G. M., Alizadeh, M., and Diniz Marinho, P. R. (2016). The type I half-logistic family of distributions. *Journal of Statistical Computation and Simulation*, **86**, 707–728.

Eliwa, M., El-Morshedy, M., and Ali., S. (2021). Exponentiated odd Chen-G family of distributions: statistical properties, Bayesian and non-Bayesian estimation with applications. *Journal of Applied Statistics*, **48**, 1948–1974.

Ghosh, I. (2020). A new discrete Pareto type (IV) model: theory, properties and applications. *Journal of Statistical Distributions and Applications*, **7**, 1–17.

Gomez-Deniz, E. (2010). Another generalization of the geometric distribution. *Test*, **19**, 399–415.

Inusah, S. and Kozubowski, T. J. (2006). A discrete analogue of the Laplace distribution. *Journal of Statistical Planning and Inference*, **136**, 1090–1102.

Jayakumar, K. and Babu, M. G. (2018). Discrete Weibull geometric distribution and its properties. *Communications in Statistics - Theory and Methods*, **47**, 1767–1783.

Jayakumar, K. and Babu, M. G. (2019). Discrete additive Weibull geometric distribution. *Journal of Statistical Theory and Applications*, **18**, 33–45.

Jayakumar, K. and Sankaran, K. K. (2018). A generalisation of discrete Weibull distribution. *Communications in Statistics - Theory and Methods*, **47**, 6064–6078.

Jazi, M. A., Lai, C. D., and Alamatsaz, M. H. (2010). A discrete inverse Weibull distribution and estimation of its parameters. *Statistical Methodology*, **7**, 121–132.

Kemp, A. W. (2008). *Advances in Mathematical and Statistical Modelling*, chapter 3, pages 353–360. Springer.

Kotz, S., Lumelskii, Y., and Pensky, M. (2003). *The Stress-strength Model and its Generalizations*. World Scientific.

Krishna, H. and Pundir, P. S. (2009). Discrete Burr and discrete Pareto distributions. *Statistical Methodology*, **6**, 177–188.

Kus, C., Akdogan, Y., Asgharzadeh, A., Kinaci, I., and Karakaya, K. (2019). Binomial-discrete Lindley distribution. *Communications Faculty of Sciences University of Ankara Series A1 Mathematics and Statistics*, **68**, 401–411.

Nadarajah, S. (2008). A truncated inverted beta distribution with application to air pollution data. *Stochastic Environmental Research and Risk Assessment*, **22**, 285–289.

Nakagawa, T. and Osaki, S. (1975). The discrete Weibull distribution. *IEEE Transactions on Reliability*, **24**, 300–301.

Nekoukhou, V., Alamatsaz, M. H., and Bidram, H. (2012). A discrete analogue of the generalized exponential distribution. *Communications in Statistics - Theory and Methods*, **41**, 2000–2013.

Nekoukhou, V. and Bidram, H. (2015). The exponentiated discrete Weibull distribution. *SORT*, **39**, 127–146.

Nooghabi, M. S., Roknabadi, A. H. R., and Borzadaran, G. R. M. (2011). Discrete modified Weibull distribution. *Metron*, **69**, 207–222.

Para, B. A. and Jan, T. R. (2016). Discrete version of log-logistic distribution and its applications in genetics. *International Journal of Modern Mathematical Sciences*, **14**, 407–422.

Rohatgi, V. K. and Saleh, E. A. K. (2001). *An Introduction to Probability and Statistics*. John Wiley & Sons.

Roy, D. and Gupta, R. P. (1999). Characterizations and model selections through reliability measures in the discrete case. *Statistics and Probability Letters*, **43**, 197–206.

Steutel, F. W. and van Harn, K. (2004). *Infinite Divisibility of Probability Distributions on the Real Line*. Marcel Dekker Inc.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 23, No. 1, 2025 (New Series), pp 79–88
https://www.ssca.org.in/journal

Bulk Queuing Model with Reneging of Customers and their Retention

Shejal Gupta[1], **Pradeep K. Joshi**[2] **and K. N. Rajeshwari**[3]
[1]*Medi-Caps University, Indore-452001, Madhya Pradesh, India*
[2]*Department of Mathematics, IPS Academy, Indore-452001, Madhya Pradesh, India*
[3]*School of Mathematics, DAVV, Indore-452001, Madhya Pradesh, India*

Received: 24 May, 2023; Revised: 22 March, 2024; Accepted: 16 April, 2024

Abstract

In the current competitive scenario, customer satisfaction is a key aspect for any organization. This paper deals with the concept of customer reneging in the system. Due to improper quality of service, customers get dissatisfied, which represents queuing with feedback. Unsatisfied customers, after taking partial service, again put efforts into getting service in case of feedback. A single server Markovian feedback bulk queuing model $M^b/M/1$ (where b is the fixed batch size) is considered with the reneging of customers and their retention. The steady-state solution and various system performance measures are established. Sensitivity analysis of parameters is also performed and the effect on the size of the system is compared with the variation in the probability of retention, which shows that the higher the retention of customers, the larger the queue size in the system. MATLAB software is used to show the results graphically. Some particular cases for the proposed model are also examined.

Key words: Customer retention; Feedback; Bulk queuing model; Steady state solution; Performance measure.

AMS Subject Classifications: 62K05, 05B05

1. Introduction

Nowadays, in this competitive era, businesses and organizations can flourish only if customers are satisfied. The quality of the product as well as quick service by the servers are the demands of every customer. Inefficiency in fulfilling these demands leads to customer dissatisfaction, which results in monetary losses for businesses. Thus, customer satisfaction is the measure of success for any business and reflects the degree to which the organization is able to meet the customer's expectations. The customer enters the system for service, but due to poor quality of service, leaves the system before completion of the service. This process is termed reneging, and the customers who leave the system are called reneged customers. Customer retention is the biggest challenge for organizations, as customer impatience is the

Corresponding Author: Pradeep K. Joshi
Email: pradeepkjoshi2@gmail.com

major cause of this problem. Thus, by incorporating various strategies, an unsatisfied customer is convinced to remain in the system, which is termed retained customers. Customer impatience is categorized into three types:

- Balking is when a customer decides not to join the queue after seeing its size.
- Reneging is when a customer joins the queue for service but leaves the queue after waiting for a long time.
- Jockeying is when a customer switches between the parallel queues because they think that by doing so, they might get quick service.

Haight (1957) and Haight (1959) studied the concept of customer impatience and reneging in queuing theory. The concept of reneging and balking was also studied by Ancker and Gafarian (1963) in the M/M/1/N queuing system and obtained its steady-state solution. Abou-El-Ata and Hariri (1992) studied a multiple-channel truncated queue with balking and reneging and established the steady-state solution and various system performance measures of the proposed queuing model. Choudhury and Medhi (2011) analyzed the Markovian multi-server queuing model with balking and reneging, in which explicit closed forms were presented. Two abandonment scenarios with impatient customers in a single server Markovian queue were studied by Kapodistria (2011) In the first scenario, an existing customer becomes impatient and performs synchronized abandonments, and the customer is excluded from taking service in the second scenario. This work is then extended by him to a multi-server Markovian queue under the second abandonment scenario as well.

Kumar and Sharma (2012a) and Kumar and Sharma (2012b) developed an M/M/1/N queuing model with the reneging of customers and their retention and obtained the steady-state solution and various performance measures of the proposed model. They extended this work and developed an M/M/1/N queuing model using the concept of balking and retention of reneged customers. So, balking is another added concept that they used in their research. VijayaLaxmi and Jyothsna (2013) studied the optimization of reneging and balking queues with vacation interruption under N-policy.

Kumar and Sharma (2013) incorporated the notion of balking and reneging of customers with their retention in the M/M/1 feedback queuing model and developed a steady-state solution. VijayaLaxmi and Kassahun (2018) studied a multi-server Markovian queue with working vacations, reneging of customers, and discouraged arrivals and obtained the steady state and steady probabilities of the system. Kumar and Sharma (2021) discussed a Markovian queuing system with multiple heterogeneous servers, reneging, and retention of reneging customers. They performed transient analysis using a probability-generating function and important performance measures, including the average retention rate. Also, the steady-state solution of the model is obtained. Rimmy and Indra (2022) described the effect of balking and reneging on a two-dimensional state queuing model with multiple servers. They derived the time-dependent probabilities by using Laplace transformations and obtained some measurable outcomes of the system.

In our study, the work of Kumar and Sharma (2013) is extended. They investigated the single server infinite capacity Markovian feedback M/M/I queueing model with retention of reneged customers and balking. In their analysis, they considered a single-server

feedback queueing model where one server serves all of the customers who arrive under the presumption that the retention of reneged customers and balking. In this paper web have extended this work to a single server Markovian feedback Mb/M/1 bulk queueing model. The limitations of a single server M/M/1 model are overcome by taking the bulk queueing model into consideration, because many organizations frequently encounter the arrival of customers in batches in real-world settings. In that situation, our study will assist in quickly and successfully resolving their issues. The overhead associated with processing individual requests is reduced in our work by handling requests in batches, which results in greater resource utilization.We also obtained the steady-state solution of the proposed model. Further, several system performance measures and particular cases of the proposed queuing models are obtained.

The issue of batch arrivals is not addressed in the extensive literature that has been published since 2013, which focuses primarily on a single server queueing model with finite and infinite capacity and some assumption-based research on jockeying, reneging, and balking or on their combinations.

In our study, we took into account a single server Markovian feedback bulk queuing model where customers arrive in predetermined fixed batch size. The bulk queueing model outperforms the preceding single server M/M/1 queueing model by allowing numerous requests to come simultaneously as a batch rather than one at a time. This is accomplished by establishing a fixed batch size. Therefore, in real-world situations, this bulk queueing strategy will boost customer retention, which raises the total number of customers using the system. So, our study plays a pivotal role in the field of queueing theory.

2. Model description

In the study, we consider the single-server Markovian feedback bulk queuing model $M^b/M/1$ (where b is the fixed batch size of the arrival of the customer) with reneging of the customer. Customers join the system in a Poisson manner with the arrival rate λ and get the service exponentially with the service rate. Due to the concept of reneging, customers join the queue for service and leave the queue after waiting because the queue is too long. Feedback customers are those unsatisfied customers who re-join the system for another regular service after the completion of the previous service.

Let the parameter ξ of reneging time be exponentially distributed. It is found that by incorporating some strategies and schemes, a reneged customer can be convinced to be retained in the system for the service. Let q be the probability with which reneged customers are retained in the system, the probability of non-retention of customers be $p(= 1 - q))$, n be the number of units in the system, $P_n(t)$ be the transient state probability of having n customers in the system at time t, and P_n be the steady state probability of having n customers in the system.

The differential-difference equations of the bulk queuing model $M^b/M/1$ given by Medhi (2001) are:

$$\frac{dP_0(t)}{dt} = -\lambda P_0(t) + \mu P_1(t) \tag{1}$$

$$\frac{dP_n(t)}{dt} = -\left[(\lambda+\mu)P_n(t)\right] + \mu P_{n+1}(t) \quad , n < b \quad , \quad n = 1, 2, \cdots, b-1 \tag{2}$$

$$\frac{dP_n(t)}{dt} = -\left[(\lambda+\mu q+(n-1)\xi p)\, P_n(t)\right] + (\mu q + n\xi p)P_{n+1}(t) + \lambda P_{n-b}(t), n \geq b \tag{3}$$

Equations (1) and (2) were considered from Medhi (2001) and we expanded them to generate equation (3) under the assumptions that the queueing model is a bulk queueing model with a fixed batch size b and reneging and number of customers n are greater than or equal to the batch size b.

In steady state, $\lim_{t\to\infty} P_n(t) = P_n$ and hence $\frac{\mathrm{dP}_n(t)}{dt} = 0$ as $t \to \infty$ and thus equations (1), (2) and (3) gives the difference equations of the model

$$0 = -\lambda P_0 + \mu P_1 \tag{4}$$

$$0 = -\left[(\lambda+\mu)P_n\right] + \mu P_{n+1} \quad , n < b \quad , \quad n = 1, 2, \cdots, b-1 \tag{5}$$

$$0 = -\left[(\lambda+\mu q+(n-1)\xi p)\, P_n\right] + (\mu q + n\xi p)P_{n+1} + \lambda P_{n-b}, n \geq b \tag{6}$$

Using equation (4), we get

$$P_1 = \frac{\lambda P_0}{\mu} \tag{7}$$

For $n = 1$, equation (5) yields,$(\lambda+\mu)P_1 = \mu P_2$ $i.e;$ $P_2 = \frac{(\lambda+\mu)}{\mu} P_1$

$$i.e;\ P_2 = \frac{\lambda(\lambda+\mu)}{\mu^2} P_0$$

For $n = 2$, equation (5) yields, $P_3 = \frac{\lambda(\lambda+\mu)^2}{\mu^3} P_0$

On solving iteratively, we get

$$P_n = \frac{\lambda(\lambda+\mu)^{n-1}}{\mu^n} P_0 \quad , \quad 1 \leq n \leq b \tag{8}$$

For $n > b$, put $n = b$ in equation (6)

$$\left[(\lambda+\mu q+(b-1)\xi p)\, P_b\right] = (\mu q + b\xi p)P_{b+1} + \lambda P_0$$

$$P_{b+1} = \frac{\left[(\lambda+\mu q+(b-1)\xi p)\, P_b\right] - \lambda P_0}{(\mu q + b\xi p)}$$

Put the value of P_b for $n = b$ from equation (8), we get

$$P_{b+1} = \frac{\lambda[(\lambda+\mu q+(b-1)\xi p)(\lambda+\mu)^{b-1}-\mu^b]}{\mu^b(\mu q+b\xi p)}P_0$$

Similarly for $n > b$, the steady state probabilities $P_n; n > b+1$ are obtained as

$$P_n = \prod_{k=b+1}^{n} \frac{\lambda[\{\lambda+\mu q+(b-1)\xi p\}(\lambda+\mu)^{b-1}-\mu^b]}{\mu^b(\mu q+kb\xi p)}P_0 \tag{9}$$

For finding the value of P_0 , normalization condition $\sum_{n=0}^{\infty} P_n = 1$ is used and the values of $P_n; n \geq 1$

$$\left[1+\sum_{n=1}^{b}\frac{\lambda(\lambda+\mu)^{n-1}}{\mu^n}+\sum_{n=b+1}^{\infty}\prod_{k=b+1}^{n}\frac{\lambda\{(\lambda+\mu q+(b-1)\xi p)(\lambda+\mu)^{b-1}-\mu^b\}}{\mu^b(\mu q+kb\xi p)}P_0\right]=1$$

where

$$P_0 = \frac{1}{1+\sum_{n=1}^{b}\frac{\lambda(\lambda+\mu)^{n-1}}{\mu^n}+\sum_{n=b+1}^{\infty}\prod_{k=b+1}^{n}\frac{\lambda\{(\lambda+\mu q+(b-1)\xi p)(\lambda+\mu)^{b-1}-\mu^b\}}{\mu^b(\mu q+kb\xi p)}} \tag{10}$$

The steady state probabilities exist if

$$\left[1+\sum_{n=1}^{b}\frac{\lambda(\lambda+\mu)^{n-1}}{\mu^n}+\sum_{n=b+1}^{\infty}\prod_{k=b+1}^{n}\frac{\lambda\{(\lambda+\mu q+(b-1)\xi p)(\lambda+\mu)^{b-1}-\mu^b\}}{\mu^b(\mu q+kb\xi p)}\right]<\infty$$

3. System performance measures

Now, we derive some common performance measures from the proposed single-server Mb/M/1feedback bulk queuing model, which are useful for investigating the behavior of the system.

3.1. The expected number of customers waiting in the system(L_s)

$$\begin{aligned} L_s &= \sum_{n=0}^{\infty} nP_n \\ &= \left[\sum_{n=1}^{b}\frac{n\lambda(\lambda+\mu)^{n-1}}{\mu^n}+\sum_{n=b+1}^{\infty}n\Big(\prod_{k=b+1}^{\infty}\frac{\lambda\{(\lambda+\mu q+(b-1)\xi p)(\lambda+\mu)^{b-1}-\mu^b\}}{\mu^b(\mu q+kb\xi p)}\Big)\right]P_0 \end{aligned} \tag{11}$$

3.2. The expected number of customers waiting in the queue(L_q)

$$\begin{aligned} L_q &= L_s - \frac{\lambda}{\mu} \\ &= \left[\sum_{n=1}^{b}\frac{n\lambda(\lambda+\mu)^{n-1}}{\mu^n}+\sum_{n=b+1}^{\infty}n\Big(\prod_{k=b+1}^{\infty}\frac{\lambda\{(\lambda+\mu q+(b-1)\xi p)(\lambda+\mu)^{b-1}-\mu^b\}}{\mu^b(\mu q+kb\xi p)}\Big)\right]P_0-\frac{\lambda}{\mu} \end{aligned} \tag{12}$$

3.3. The expected waiting time of the customer in the system(W_s)

$$W_s = \frac{L_s}{\lambda \mathrm{b}} = \frac{1}{\lambda \mathrm{b}} \left[\sum_{n=1}^{b} \frac{n\lambda(\lambda+\mu)^{n-1}}{\mu^n} + \sum_{n=b+1}^{\infty} n\Big(\prod_{k=b+1}^{\infty} \frac{\lambda\{(\lambda+\mu q+(b-1)\xi p)(\lambda+\mu)^{b-1}-\mu^b\}}{\mu^b(\mu q+kb\xi p)}\Big) \right] P_0 \tag{13}$$

3.4. The expected waiting time of the customer in the queue(W_q)

$$W_q = W_s - \frac{1}{\mu} = \left[\sum_{n=1}^{b} \frac{n\lambda(\lambda+\mu)^{n-1}}{\mu^n} + \sum_{n=b+1}^{\infty} n\Big(\prod_{k=b+1}^{\infty} \frac{\lambda\{(\lambda+\mu q+(b-1)\xi p)(\lambda+\mu)^{b-1}-\mu^b\}}{\mu^b(\mu q+kb\xi p)}\Big)x \right] P_0 - \frac{1}{\mu} \tag{14}$$

4. Sensitivity analysis

Sensitivity analysis evaluates the responsiveness of a model to the changes in various controllable parameters. In this section, we evaluate the sensitivity of the proposed model for different values of various parameters.

For a fixed value of n and for different values of λ, μ, ξ, q, we calculate the variations in the expected number of customers waiting in the system(L_s) by using equation (11) and discuss their effects graphically.

Case I. Effect on the size of the system with the variation in arrival rate

For $n = 4$, $\lambda = 2.0, 2.1, 2.2, 2.3, 2.4$, $\mu = 3$, $\xi = 0.1$, $q = 0.6$, p=0.4, b=3, we substitute these values in (11), we have

Table 1: Effect on the size of the system with the variation in Arrival rate

S.No.	λ	P_0	L_s
1.	2.0	0.115	2.85
2.	2.1	0.103	2.92
3.	2.2	0.093	2.97
4.	2.3	0.084	3.03
5.	2.4	0.076	3.08

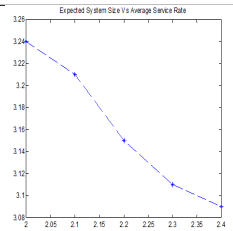


Figure 1
X- Axis : Average arrival rate,
Y-Axis: Expected system size

From Table 1 and Figure 1 above, we observe that the size of the system is directly proportional to the arrival rate, *i.e.* more the arrival of customers, larger the size of the system and vice-versa.

Case II. Effect on the size of the system with the variation in service rate

For $n = 4$, $\lambda = 2$, $\mu = 2.0, 2.1, 2.2, 2.3, 2.4$, $\xi = 0.1$, $q = 0.6$, $p = 0.4$, $b = 3$, we substitute these values in (11), we have

Table 2: Effect on the size of the system with the variation in service rate

S.No.	μ	P_0	L_s
1.	2.0	0.046	3.24
2.	2.1	0.052	3.21
3.	2.2	0.058	3.15
4.	2.3	0.064	3.11
5.	2.4	0.071	3.09

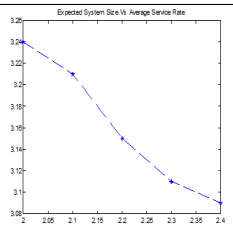


Figure 2
X- Axis : Average service,
Y-Axis: Expected system size

From Table 2 and Figure 2 above, we observe that as the average service rate increases, the size of the system decreases.

Case III. Effect on the size of the system with the variation in average reneging rate

For $n = 4$, $\lambda = 2$, $\mu = 3$, $\xi = 0.01, 0.02, 0.03, 0.04, 0.05$, $q = 0.6$, $p = 0.4$, $b = 3$ we substitute these values in (11), we have

Table 3 : Effect on the size of the system with the variation in average reneging rate

S.No.	ξ	P_0	L_s
1.	0.01	0.102	2.98
2.	0.02	0.104	2.97
3.	0.03	0.105	2.94
4.	0.04	0.107	2.94
5.	0.05	0.108	2.91

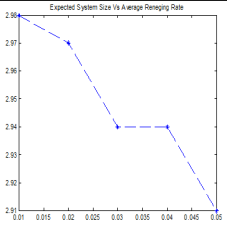


Figure 3
X- Axis : Average reneging rate,
Y-Axis: Expected system size

From Table 3 and Figure 3 above, we observe that as the average reneging rate increases, the size of the system decreases.

Case IV. Effect on the size of the system with the variation in retention probability

For $n = 4$, $\lambda = 2$, $\mu = 3$, $\xi = 0.1$, q=0.1, 0.2, 0.3, 0.4, 0.5, $b = 3$ we substitute these values in (11), we have

Table 4 : Effect on the size of the system with the variation in retention probability

S.No.	**q**	**p(=1-q)**	P_0	L_s
1.	0.1	0.9	0.120	2.63
2.	0.2	0.8	0.122	2.76
3.	0.3	0.7	0.120	2.80
4.	0.4	0.6	0.118	2.82
5.	0.5	0.5	0.116	2.83

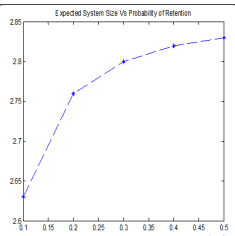


Figure 4
X- Axis : Probability of retention,
Y-Axis: Expected system size

From Table 4 and Figure 4 above, it is observed that the higher the retention of customers from reneging, the larger the size of the system.

5. Particular cases of the model

In this section, some particular cases of the proposed model are derived.

5.1. When retention probability of reneged customers is zero

If retention of reneged customers is zero, then $q = 1 - p = 0$. In this case, proposed model becomes $\mathbf{M^b/M/1}$ feedback bulk queuing model with reneging and we get

$$P_n = \frac{\lambda(\lambda+\mu)^{n-1}}{\mu^n} P_0 \quad , \quad 1 \le n \le b \tag{15}$$

$$P_n = \prod_{k=b+1}^{n} \frac{\lambda[(\lambda+(b-1)\xi p)(\lambda+\mu)^{b-1} - \mu^b]}{\mu^b(kb\xi p)} P_0, \ n > b \tag{16}$$

where $P_0 = \frac{1}{1+\sum_{n=1}^{b} \frac{\lambda(\lambda+\mu)^{n-1}}{\mu^n} + \sum_{n=b+1}^{\infty} \prod_{k=b+1}^{n} \frac{\lambda\{(\lambda+(b-1)\xi p)(\lambda+\mu)^{b-1}-\mu^b\}}{\mu^b(\mathrm{kb\xi p})}}$.

5.2. When no reneging in the system

If there is no reneging in the system, then $\xi = 0$. In this case proposed model reduces to simple $\mathrm{M^b/M/1}$ queue model and we get

$$P_n = \frac{\lambda(\lambda+\mu)^{n-1}}{\mu^n} P_0 \quad , \quad 1 \le n \le b$$

$$P_n = \prod_{k=b+1}^{n} \frac{\lambda[(\lambda+\mu q)(\lambda+\mu)^{b-1} - \mu^b]}{\mu^b(\mu q)} P_0, \ n > b \tag{17}$$

where $P_0 = \frac{1}{1+\sum_{n=1}^{b} \frac{\lambda(\lambda+\mu)^{n-1}}{\mu^n} + \sum_{n=b+1}^{\infty} \prod_{k=b+1}^{n} \frac{\lambda\{(\lambda+\mu q)(\lambda+\mu)^{b-1}-\mu^b\}}{\mu^b(\mu\ \mathrm{q})}}$.

5.3. When the system is of finite capacity

If system capacity is finite, say N, then proposed model reduces to $M^b/M/1/N$ feedback queuing model with retention of reneged customers and

$$P_n = \frac{\lambda(\lambda+\mu)^{n-1}}{\mu^n} P_0 \quad , \quad 1 \le n \le b$$

$$P_n = \prod_{k=b+1}^{n} \frac{\lambda[(\lambda+\mu q+(b-1)\xi p)(\lambda+\mu)^{b-1}-\mu^b]}{\mu^b(\mu q+kb\xi p)} P_0, \ b+1 \le n \le N \tag{18}$$

where $P_0 = \frac{1}{1+\sum_{n=1}^{b} \frac{\lambda(\lambda+\mu)^{n-1}}{\mu^n} + \sum_{n=b+1}^{N} \prod_{k=b+1}^{n} \frac{\lambda\{(\lambda+\mu q+(b-1)\xi p)(\lambda+\mu)^{b-1}-\mu^b\}}{\mu^b(\mu q+kb\xi p)}}$

6. Conclusion

In this paper, a single-server $M^b/M/1$ feedback bulk queuing model with reneged customers and their retention is discussed. The steady-state solution and various system performance measures are also derived for the proposed model. The sensitivity analysis of the proposed model is performed, and the effect of variation in the retention probability on the size of the system is discussed. From the results obtained, we concluded that the higher the retention of customers, the larger the size of the system. Thus, the study suggests any organization employ more strategies to retain customers for maximum profit.However, under some unusual circumstances, like epidemics or catastrophic events, this conclusion may not be true since customer retention will decrease due to impatience if the arrival of customers in batches is exponentially increasing and rises to be extremely large. Numerical results are analyzed by graphical representation using MATLAB software. Further, some particular cases of the proposed model are also discussed, and for different cases, we obtained some more queuing models with feedback. These extensions of models and their comparisons can be explored in future work.

Acknowledgements

The authors express their gratefulness to the reviewer and chief editor for giving suggestions that led to considerable improvement in the research paper.

Conflict of interest

The authors do not have any financial or non-financial conflict of interest to declare for the research work included in this article.

References

Abou-El-Ata, M. O. and Hariri, A. M. A. (1992). The M/M/c/N queue with balking and reneging. *Computers and Operations Research*, **19**, 713–716.

Ancker, J. C. J. and Gafarian, A. V. (1963). Some queuing problems with balking and reneging i. *Operations Research*, **11**, 88–100.

Choudhury, A. and Medhi, P. (2011). Balking and reneging in multiservermarkovianqueuing systems. *International Journal of Mathematics in Operational Research*, **3**, 377–394.

Haight, F. A. (1957). Queuing with balking. *Biometrika*, **44**, 360–369.

Haight, F. A. (1959). Queuing with reneging. *Metrika*, **2**, 186–197.

Kapodistria, S. (2011). The M/M/1 queue with synchronized abandonments. *Queuing Systems*, **68**, 79–109.

Kumar, R. and Sharma, S. (2021). Transient analysis of a markovian queuing model with multiple-heterogeneous servers and customers' impatience. *Operations Research*, **58**, 540–556.

Kumar, R. and Sharma, S. K. (2012a). M/M/1/N queuing system with retention of reneged customers. *Pakistan Journal of Statistics and Operation Research*, **8**, 859–866.

Kumar, R. and Sharma, S. K. (2012b). M/M/1/N queuing system with retention of reneged customers and balking. *American Journal of Operational Research*, **2**, 1–5.

Kumar, R. and Sharma, S. K. (2013). M/M/1 feedback queuing models with retention of reneged customers and balking. *American Journal of Operational Research*, **3**, 1–6.

Medhi, J. (2001). *Stochastic Processes*. Second Edition, New Age International(p) LTD, Math, New Delhi.

Rimmy, S. and Indra (2022). Transient analysis for a multiple vacations queuing model with impatient customers. *Journal of Scientific Research of The Banaras Hindu University*, **66**, 393–403.

VijayaLaxmi, P. and Jyothsna, V. G. K. (2013). Optimization of balking and reneging queue with vacation interruption under n-policy. *Journal of Optimization*, **3**, 1–9.

VijayaLaxmi, P. and Kassahun, T. W. (2018). Analysis of a multi-server markovian queue with working vacations and impatience of customers. *International Journal of Research in Engineering, IT and Social Sciences*, **8**, 10–24.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 23, No. 1, 2025 (New Series), pp 89–122
https://www.ssca.org.in/journal

Inference Techniques, Properties, and Applications of the T-Marshall-Olkin X Family of Distributions

Meenu Jose[1] and Lishamol Tomy[2]
[1]*Department of Statistics, Carmel College(Autonomous) Mala,Thrissur, Kerala, India*
[2]*Department of Statistics, Deva Matha College, Kuravilangad, Kerala, 686633, India*

Received: 12 March 2024; Revised: 15 April 2024; Accepted: 24 April 2024

Abstract

In this research, we study and introduce a new family of continuous distributions known as the T-Marshall-Olkin X family. We present some special models and investigate the asymptotic distributions of order statistics of the family half-logistic-Marshall-Olkin X family, which is explored in depth as a specific instance. The half-logistic-Marshall-Olkin Lomax distribution is one unique model in this family that is explored in depth. We list a few of the new distribution's mathematical properties. We use the maximum likelihood method to estimate the model's parameters. The bias and mean square error of the maximum likelihood estimators are examined in a simulation study that is given. Testing the importance of a distribution parameter is done using the likelihood ratio test with a simulation study. The potentiality and flexibility of the new family are illustrated by using two practical data sets.

Key words: T-X family; Marshall-Olkin; Moments; Maximum likelihood estimation; Likelihood ratio test; Applications.

AMS Subject Classifications: 62K05, 05B05

1. Introduction

The statistical literature is rich with different kinds univariate distributions and is still growing rapidly. The classical distributions have various limitations in modelling real-life data. This persuades the statistical researcher to develop methods for generating new classes of distributions starting with a base line distribution.

Marshall and Olkin (1997) proposed a flexible family of distributions by introducing a new shape parameter to the existing family of distributions called the Marshall-Olkin family of distributions. The cumulative density function (CDF) of the Marshall-Olkin (MO) family is given respectively, by $G(x) = \frac{F(x)}{c+(1-c)F(x)}, c > 0, x \in \mathbb{R}$, where $F(x)$ is the baseline CDF. This approach produces a stable distribution with broad field behaviour in probability density function (PDF) and hazard rate function (HRF) compared to the baseline distribution. It

Corresponding Author: Meenu Jose
Email: meenusgc@gmail.com

provides a flexible framework for modelling a variety of circumstances and is useful in areas such as reliability, finance, simulation studies, health research, and engineering. Some MO families of distributions are MO-extended Lomax by Ghitany *et al.* (2007), MO-extended Lindley by Ghitany *et al.* (2012), MO-Fréchet by Krishna *et al.* (2013), MO-exponential Weibull by Pogány *et al.* (2015), MO-generalized exponential by Ristić and Kundu (2015), MO-Ikum by Tomy and Gillariose (2018), MO modified Lindley by Gillariose *et al.* (2020), MO Gumbel-Lomax by Nwezza and Ugwuowo (2020) MO-Lindley-Log-logistic by Moakofi *et al.* (2021), MO alpha power inverse exponential by Basheer (2022), MO Inverse log-logistic by Aako *et al.* (2022), MO Extended Gumbel Type-II by Willayat *et al.* (2022), MO extended unit-Gompertz by Opone *et al.* (2022), MO Exponentiated Dagum by Sherwani *et al.* (2023), MO Extended Generalized Exponential by Innocent *et al.* (2023), MO Pranav by Alsultan (2023), MO Exponentiated Fréchet by Niyoyunguruza *et al.* (2023), MO Chris-Jerry by Obulezi *et al.* (2023), MO Pareto type-I by Aldahlan *et al.* (2023), MO Cosine Topp-Leon by Osi *et al.* (2024a), MO Bilal by Irhad *et al.* (2024).

Alzaatreh *et al.* (2013) introduced a powerful method to generate new families of distributions called the transformed-transformer method, and the family is called the T-X family of distributions. This approach extends the beta-G by Eugene *et al.* (2002) and Kumaraswamy-G by Cordeiro and de Castro (2011) families by using any continuous distribution for a random variable T on $[a, b]$. The CDF of the T-X family of distributions is given by $R(x) = \int_a^{W[G(x)]} j(t)dt$, where j(t) is the PDF of a random variable T, $T \in [a, b]$ for $-\infty < a < b < \infty$ and $W[G(x)]$ is a function of the baseline CDF of a random variable X and satisfies three conditions, namely

- $W[G(x)] \in [a, b]$.
- $W[G(x)]$ is differentiable and monotonically non decreasing.
- $W[G(x)] \to a$ as $x \to -\infty$ and $W[G(x)] \to b$ as $x \to \infty$.

Numerous research papers on the T-X family have been published in the literature. The Weibull-Pareto distribution by Alzaatreh *et al.* (2013), Kumaraswamy-Geometric Distribution by Akinsete *et al.* (2014), McDonald quasi Lindley distribution by Merovci *et al.* (2015), Kumaraswamy -Weibull geometric distribution by Rasekhi *et al.* (2018), generalized odd inverted exponential generated family of distributions by Chesneau and Djibrila (2019), Weibull Burr X-G family of distribution by Ishaq *et al.* (2019), weighted odd Weibull generated family of distributions by Mi *et al.* (2021), exponentiated odd Chen-G family of distributions by Eliwa *et al.* (2021), generalized odd linear exponential family of distributions by Jamal *et al.* (2022), Rayleigh-Exponentiated Odd Generalized-Pareto distribution by Yahaya and Doguwa (2022), MO odd power generalized Weibull distribution by Chipepa *et al.* (2022), New Generalized Logarithmic-X family of distributions by Shah *et al.* (2023), New Generalized Odd Fréchet-Exponentiated-G family of distribution by Sadiq *et al.* (2023), new generalized exponentiated Fréchet–Weibull distribution by Klakattawi *et al.* (2023), MO Topp-Leone Half-Logistic-G family of distributions by Sengweni *et al.* (2023), exponentiated Cosine Topp-Leone Generalized family of distributions by Osi *et al.* (2024b) and others are a few examples. A review paper by Tomy *et al.* (2019) provides a detailed account of the T-X family of distributions.

Nowadays, there is a trend toward combining various families of distributions to increase the flexibility and properties of new distributions. Some of them are the beta MO family by Alizadeh *et al.* (2015a), Kumaraswamy MO family by Alizadeh *et al.* (2015b), generalized MO Kumaraswamy-G family by Handique and Chakraborty (2015a), MO-Kumaraswamy-G family by Handique and Chakraborty (2015b), T-transmuted X family by Moolath and Jayakumar (2017), MO Zubair-G family by Nasiru and Abubakari (2022), MO Weibull–Burr XII family by Alsadat *et al.* (2023), type II exponentiated half logistic-MO-G family by Oluyede and Gabanakgosi (2023), new generalized exponentiated Fréchet–Weibull family by Klakattawi *et al.* (2023), new Topp-Leone Kumaraswamy MO generated family by Atchadé *et al.* (2024). The new idea is based on both the MO and T-X families of distributions, combining the MO and T-X families of distributions. The motivations for introducing this new family of distributions are:

1. To generate a new family of distributions that have the properties contained in the MO and T-X families of distributions.

2. The new family of distributions is more adaptable to real-life data than models with same number of parameters and baseline distribution.

3. The desirable characteristics and adaptability provided by this new family of distributions, particularly in terms of the forms of the density and hazard rate functions, have inspired us to create this model, as it proves beneficial for real-life data analysis.

In this chapter, we propose a new extension of the T-X family by considering MO as baseline distribution called the T-Marshall-Olkin X family of distributions. The proposed distribution is well-suited to both biomedical and survival datasets. This study demonstrates that the novel extension of the Lomax distribution provides a better match to the datasets than other well-known distributions (see Section 8). The chapter unfolds as follows: In Section 2, we introduce a new family of distributions called "T-Marshall-Olkin X family" and study its properties. In Section 3, some members of T-Marshall-Olkin X family are identified. The mathematical properties of one of the member of T-Marshall-Olkin X family called, half logistic-Marshall-Olkin X family of distributions are studied in Section 4. In Section 5, we study the half logistic-Marshall-Olkin Lomax distribution and its properties. The maximum likelihood estimator of the unknown parameters with simulation study are discussed in Section 6. The analysis of two real data sets has been presented and illustrating the modelling potential of half logistic-Marshall-Olkin Lomax distribution in Section 8. Finally, the conclusion of the paper appears in Section 9.

2. T-Marshall-Olkin X family of distributions

The CDF of a T-X family of distributions is defined as

$$R(x) = \int_{a}^{W[G(x)]} j(t)dt. \tag{1}$$

Let $[W(G(x)] = -\log(1-G(x))$ and the random variable T be defined on $(0,\infty)$. Then the CDF becomes

$$R(x) = \int_0^{-\log(1-G(x))} j(t)dt. \tag{2}$$

As a special case, we assume $G(x)$ is a MO family of distributions.

Then

$$W(G(x)] = -\log\Big\{1 - \frac{F(x)}{c+(1-c)F(x)}\Big\} = -\log\Big\{\frac{c(1-F(x))}{c+(1-c)F(x)}\Big\}.$$

From Equation (2), the CDF of the new family is

$$R(x) = \int_0^{-\log\left\{\frac{c(1-F(x))}{c+(1-c)F(x)}\right\}} j(t)dt = J\Big\{-ln\{\frac{c(1-F(x))}{c+(1-c)F(x)}\}\Big\}. \tag{3}$$

When considering X as a continuous random variable, the probability density function (PDF) can be generated as follows:

$$\begin{aligned} r(x) &= \frac{d}{dx}J\Big\{-\log\{\frac{c(1-F(x))}{c+(1-c)F(x)}\}\Big\} \\ &= j\Big\{-\log\{\frac{c(1-F(x))}{c+(1-c)F(x)}\}\Big\}\frac{f(x)}{(1-F(x))(c+(1-c)F(x))}; \quad x \in \mathbb{R}. \end{aligned} \tag{4}$$

The corresponding HRF can be found using the formula

$$h_r(x) = \frac{j\Big\{-\log\{\frac{c(1-F(x))}{c+(1-c)F(x)}\}\Big\}f(x)}{[1-F(x)][c+(1-c)F(x)][1-J\{-\log\{\frac{c(1-F(x))}{c+(1-c)F(x)}\}\}]}. \tag{5}$$

The shapes of the PDF and HRF can be enumerated analytically. The critical points of the density function are the roots of the equation:

$$\begin{aligned} \frac{\partial \log[r(x)]}{\partial x} &= \frac{j'\Big\{-\log\{\frac{c(1-F(x))}{c+(1-c)F(x)}\}\Big\}}{j\Big\{-\log\{\frac{c[1-F(x)]}{c+(1-c)F(x)}\}\Big\}}\frac{f(x)}{[1-F(x)][c+(1-c)F(x)]} + \frac{f'(x)}{f(x)} \\ &\quad + \frac{f(x)}{1-F(x)} - \frac{(1-c)f(x)}{c+(1-c)F(x)} = 0. \end{aligned} \tag{6}$$

Equation (6) may have more than one root. If the root of Equation (6) is x = x_0, then it corresponds to a local maximum if $\frac{\partial^2 \log[r(x)]}{\partial x^2} < 0$, a local minimum if $\frac{\partial^2 \log[r(x)]}{\partial x^2} > 0$, and a point of inflection if $\frac{\partial^2 \log[r(x)]}{\partial x^2} = 0$

Similarly, the critical points of $h_r(x)$ are the roots of the equation

$$\frac{\partial \log[h_r(x)]}{\partial x} = \frac{j'\Big\{-\log\{\frac{c(1-F(x))}{c+(1-c)F(x)}\}\Big\}}{j\Big\{-\log\{\frac{c[1-F(x)]}{c+(1-c)F(x)}\}\Big\}} \frac{f(x)}{[1-F(x)][c+(1-c)F(x)]} \frac{f'(x)}{f(x)} + \frac{f(x)}{1-F(x)}$$
$$- \frac{(1-c)f(x)}{c+(1-c)F(x)} + \frac{j\Big\{-\log\{\frac{c[1-F(x)]}{c+(1-c)F(x)}\}\Big\}}{1-J\Big\{-\log\{\frac{c[1-F(x)]}{c+(1-c)F(x)}\}\Big\}} \frac{f(x)}{[1-F(x)][c+(1-c)F(x)]}$$
$$= 0. \tag{7}$$

Equation (7) may have more than one root. If the root of Equation (7) is x = x_0, then it corresponds to a local maximum if $\frac{\partial^2 \log[h_r(x)]}{\partial x^2} < 0$, a local minimum if $\frac{\partial^2 \log[h_r(x)]}{\partial x^2} > 0$, and a point of inflection if $\frac{\partial^2 \log[h_r(x)]}{\partial x^2} = 0$.
Some remarks on the T-Marshall-Olkin X family of distributions:

1. The T-Marshall-Olkin X family of distributions CDF and PDF, which are given in equations Equation (3) and Equation (4), can be as $R(x) = J\Big\{-\log\Big\{1-\frac{F(x)}{c+(1-c)F(x)}\Big\}\Big\} = J(H_g(x))$ and $r(x) = h_g(x)j(H_g(x))$ where $h(x)$ and $H(x)$ are HRF and cumulative HRF of the random variable X with CDF $\Big\{\frac{F(x)}{c+(1-c)F(x)}\Big\}$, ie, the Marshall-Olkin distribution. Hence, the T-Marshall-Olkin X family of distributions can be considered as a family of distributions arising from a weighted hazard function.

2. The random variable T which follows the PDF $j(t)$ and the random variable X following PDF $r(x)$ are related in the following way: $X = F^{-1}\Big\{\frac{c(1-e^{-T})}{1-(1-c)(1-e^{-T})}\Big\}$. This inverse function provides an easy way to simulate the random variable from T-Marshall-Olkin X family of distribution by initially simulating the random variable T and subsequently figuring out $X = F^{-1}\Big\{\frac{c(1-e^{-T})}{1-(1-c)(1-e^{-T})}\Big\}$, which has the CDF $R(x)$. Thus, $E(X) = E\Big\{F^{-1}\Big\{\frac{c(1-e^{-T})}{1-(1-c)(1-e^{-T})}\Big\}\Big\}$.
The quantile function, $Q_r(u)$, $0 < u < 1$, for the T-Marshall-Olkin X family of distribution likely to be obtained by
$$Q_r(u) = F^{-1}\Big\{\frac{c(1-e^{-J^{-1}(u)})}{1-(1-c)(1-e^{-J^{-1}(u)})}\Big\}.$$

3. If X is a discrete random variable with probability mass function (PMF) $f(x)$. Then the PMF of the T-Marshall-Olkin X family of discrete distributions can be exhibited as
$$r(x) = R(x)-R(x-1) = J\Big\{-\log\{\frac{c(1-F(x))}{c+(1-c)F(x)}\}\Big\}-J\Big\{-\log\{\frac{c(1-F(x-1))}{c+(1-c)F(x-1)}\}\Big\}.$$
In this article, the situation in which X is a continuous random variable will be covered.

4. when $c = 1$, the T-Marshall-Olkin X family of distributions reduces to the T-X family of distributions.

3. Some members of T-Marshall-Olkin X family of distributions

Several families of distributions can be derived from the T-Marshall-Olkin X family for different choices of $j(t)$. For various T distributions, Table 1 lists a few members of the T-Marshall-Olkin X family.

Some characteristics of the T-Marshall-Olkin X family for various T distributions will be dealt in the remaining areas of this section.

3.1. Exponential-Marshall-Olkin X family of distributions

In the instance when the random variable T follows the exponential distribution with parameter λ then $j(t) = \lambda e^{-\lambda t};\ t > 0,\ \lambda > 0$. Based on the Equation (4), the PDF of the exponential-Marshall-Olkin X family is.

$$r(x) = \lambda\left\{\frac{c(1-F(x))}{c+(1-c)F(x)}\right\}^{\lambda}\frac{f(x)}{(1-F(x))(c+(1-c)F(x))};\quad c,\lambda > 0. \tag{8}$$

The CDF of the exponential distribution is $J(t) = 1 - e^{-\lambda x}$ and from Equation (3) the CDF of the exponential-Marshall-Olkin X family is

$$R(x) = 1 - \left\{\frac{c(1-F(x))}{c+(1-c)F(x)}\right\}^{\lambda}. \tag{9}$$

The corresponding HRF is illustrated as

$$h_r(x) = \frac{\lambda f(x)}{(1-F(x))(c+(1-c)F(x))} = \frac{\lambda h_f(x)}{c+(1-c)F(x)} = \lambda h_g(x), \tag{10}$$

where $h_f(x)$ and $h_g(x)$ are the HRF of the distribution with PDF $f(x)$ and $g(x)$.
Thus

$$\lim_{x\to-\infty} h_r(x) = \lim_{x\to-\infty}\frac{\lambda h_f(x)}{c} = \lim_{x\to-\infty}\lambda h_g(x)$$
$$\lim_{x\to\infty} h_r(x) = \lim_{x\to\infty}\lambda h_f(x) = \lim_{x\to\infty}\lambda h_g(x).$$

It follows from Equation (10) that

$$\frac{\lambda h_f(x)}{c} \le h_r(x) \le \lambda h_f(x) \qquad (-\infty < x < \infty, \lambda \le c)$$

$$\lambda h_f(x) \le h_r(x) \le \frac{\lambda h_f(x)}{c} \qquad (-\infty < x < \infty, \lambda \ge c).$$

Again, Equation (10) shows that $\frac{h_r(x)}{h_f(x)}$ is increasing in x for$c \ge 1$ and dereasing for $0 < c \le 1$.
Some unique instances of exponential-Marshall-Olkin X family are illustrated below

1. When $c = 1$, the exponential-Marshall-Olkin X family reduces to exponential-X family of distribution.

Table 1: Some members of T-Marshall-Olkin X family of distributions for different T distributions

Name	The density of T	The density of the family $r(x)$
Exponential	$\lambda e^{-\lambda x}$	$\lambda\left\{\frac{c(1-F(x))}{c+(1-c)F(x)}\right\}^{\lambda}\frac{f(x)}{(1-F(x))(c+(1-c)F(x))};\quad c,\lambda>0$
Half logistic	$\frac{2\lambda e^{-\lambda x}}{(1+e^{-\lambda x})^2}$	$\frac{2\lambda\left\{\frac{c(1-F(x))}{c+(1-c)F(x)}\right\}^{\lambda}}{\left\{1+\left\{\frac{c(1-F(x))}{c+(1-c)F(x)}\right\}^{\lambda}\right\}^2}\frac{f(x)}{(1-F(x))(c+(1-c)F(x))};\quad c,\lambda>0$
Half Normal	$\frac{1}{\sigma}\left(\frac{2}{\pi}\right)^{\frac{1}{2}}e^{-\frac{t^2}{2\sigma^2}}$	$\frac{1}{\sigma}\left(\frac{2}{\pi}\right)^{\frac{1}{2}}e^{-\frac{\left\{-\log\left\{\frac{c(1-F(x))}{c+(1-c)F(x)}\right\}\right\}^2}{2\sigma^2}}\frac{f(x)}{(1-F(x))(c+(1-c)F(x))};\quad \sigma>0$
Type 2 Gumbel	$abx^{(-a-1)}e^{(-bx^{-a})}$	$ab\left\{-\log\left\{\frac{c(1-F(x))}{c+(1-c)F(x)}\right\}\right\}^{(-a-1)}e^{\left\{-b\left\{-\log\left\{\frac{c(1-F(x))}{c+(1-c)F(x)}\right\}\right\}^{-a}\right\}}\frac{f(x)}{(1-F(x))(c+(1-c)F(x))};\quad a,b,c>0$
Gamma	$\frac{1}{\Gamma(\alpha)\beta^{\alpha}}x^{\alpha-1}e^{\frac{-x}{\beta}}$	$\frac{1}{\Gamma(\alpha)\beta^{\alpha}}\left\{-\log\left\{\frac{c(1-F(x))}{c+(1-c)F(x)}\right\}\right\}^{\alpha-1}\left\{\frac{c(1-F(x))}{c+(1-c)F(x)}\right\}^{1/\beta}\frac{f(x)}{(1-F(x))(c+(1-c)F(x))};\quad \beta,\alpha>0$
Weibull	$\frac{a}{\beta}\left(\frac{x}{\beta}\right)^{a-1}e^{(-\frac{x}{\beta})^a}$	$\frac{a}{\beta}\left\{\frac{-\log\left\{\frac{c(1-F(x))}{c+(1-c)F(x)}\right\}}{\beta}\right\}^{(a-1)}e^{\left\{-\left\{\frac{-\log\{\frac{c(1-F(x))}{c+(1-c)F(x)}\}}{\beta}\right\}^{a}\right\}}\frac{f(x)}{(1-F(x))(c+(1-c)F(x))};\quad a,c,\beta>0$
Lomax	$\frac{\alpha}{\theta}\left[1+\frac{x}{\theta}\right]^{-(\alpha+1)}$	$\frac{\alpha}{\theta}\left\{1-\frac{\log\left\{\frac{c(1-F(x))}{c+(1-c)F(x)}\right\}}{\theta}\right\}^{-(\alpha+1)}\frac{f(x)}{(1-F(x))(c+(1-c)F(x))};\ c,\alpha,\theta>0$
Lindley	$\frac{\theta^2}{1+\theta}(1+x)e^{-\theta x}$	$\frac{\theta^2}{1+\theta}\left\{1-\log\left\{\frac{c(1-F(x))}{c+(1-c)F(x)}\right\}\right\}\left\{\frac{c(1-F(x))}{c+(1-c)F(x)}\right\}^{\theta}\frac{f(x)}{(1-F(x))(c+(1-c)F(x))};\ c,\theta>0$
Exponentiated Exponential	$\alpha\lambda\ (1-e^{-\lambda x})^{\alpha-1}\ e^{-\lambda x}$	$\alpha\lambda\left\{1-\left\{\frac{c(1-F(x))}{c+(1-c)F(x)}\right\}^{\lambda}\right\}^{\alpha-1}\left\{\frac{c(1-F(x))}{c+(1-c)F(x)}\right\}^{\lambda}\frac{f(x)}{(1-F(x))(c+(1-c)F(x))};\ c,\alpha,\lambda>0$
Akash	$\frac{\theta^3}{\theta^2+2}(1+x^2)e^{-\theta x}$	$\frac{\theta^3}{\theta^2+2}\left\{1+\left\{-\log\left\{\frac{c(1-F(x))}{c+(1-c)F(x)}\right\}\right\}^2\right\}\left\{\frac{c(1-F(x))}{c+(1-c)F(x)}\right\}^{\theta}\frac{f(x)}{(1-F(x))(c+(1-c)F(x))};\ c,\theta>0$
Rayleigh	$\frac{x}{\sigma^2}e^{-\frac{x^2}{2\sigma^2}}$	$\frac{\left\{-\log\left\{\frac{c(1-F(x))}{c+(1-c)F(x)}\right\}\right\}}{\sigma^2}e^{-\frac{\left\{-\log\left\{\frac{c(1-F(x))}{c+(1-c)F(x)}\right\}\right\}^2}{2\sigma^2}}\frac{f(x)}{(1-F(x))(c+(1-c)F(x))};\quad c,\sigma>0$

Table 1:(Continued)

Name	The density of T	The density of the family $r(x)$
Half Normal	$\frac{1}{\sigma}(\frac{2}{\pi})^{\frac{1}{2}}e^{-\frac{t^2}{2\sigma^2}}$	$\frac{1}{\sigma}(\frac{2}{\pi})^{\frac{1}{2}}e^{-\frac{\left\{-\log\left\{\frac{c(1-F(x))}{c+(1-c)F(x)}\right\}\right\}^2}{2\sigma^2}}\frac{f(x)}{(1-F(x))(c+(1-c)F(x))}$; $\sigma>0$
Lomax	$\frac{\alpha}{\theta}[1+\frac{x}{\theta}]^{-(\alpha+1)}$	$\frac{\alpha}{\theta}\left\{1-\frac{\log\left\{\frac{c(1-F(x))}{c+(1-c)F(x)}\right\}}{\theta}\right\}^{-(\alpha+1)}\frac{f(x)}{(1-F(x))(c+(1-c)F(x))}$; $c,\alpha,\theta>0$
Power Cauchy	$2\pi^{-1}(\alpha/\sigma)(x/\sigma)^{\alpha-1}[1+(x/\sigma)^{2\alpha}]^{-1}$	$\frac{2\pi^{-1}\alpha}{\sigma}\left\{\frac{\left\{-\log\left\{\frac{c(1-F(x))}{c+(1-c)F(x)}\right\}\right\}}{\sigma}\right\}^{\alpha-1}\left\{1+\left\{\frac{\left\{-\log\left\{\frac{c(1-F(x))}{c+(1-c)F(x)}\right\}\right\}}{\sigma}\right\}^{2\alpha}\right\}^{-1}$ $\frac{f(x)}{(1-F(x))(c+(1-c)F(x))}$; $c,\alpha,\sigma>0$

2. When $\lambda = 1$, the exponential-Marshall-Olkin X family reduces to Marshall-Olkin X family of distribution.

3. When $\lambda = c = 1$, the exponential-Marshall-Olkin X family reduces to a distribution with PDF $f(x)$.

3.2. Half-logistic-Marshall-Olkin X family of distributions

In the instance when the random variable T follows the half-logistic distribution with parameter λ then $j(t) = \frac{2\lambda e^{-\lambda t}}{(1+e^{-\lambda t})^2};\ t > 0,\ \lambda > 0$. Based on the Equation (4), the PDF of the half-logistic-Marshall-Olkin X (HLMO-X) family is

$$r(x) = \frac{2\lambda\{\frac{c(1-F(x))}{c+(1-c)F(x)}\}^{\lambda}}{\left\{1+\{\frac{c(1-F(x))}{c+(1-c)F(x)}\}^{\lambda}\right\}^2}\frac{f(x)}{(1-F(x))(c+(1-c)F(x))};\quad c,\lambda > 0. \tag{11}$$

When $c = 1$, the HLMO-X family reduces to half-logistic-X family of distributions. The CDF of the half-logistic distribution is $J(t) = \frac{1-e^{-\lambda t}}{1+e^{-\lambda t}}$ and hence from Equation (3) the CDF of the HLMO-X family is

$$R(x) = \frac{1-\left\{\frac{c(1-F(x))}{c+(1-c)F(x)}\right\}^{\lambda}}{1+\left\{\frac{c(1-F(x))}{c+(1-c)F(x)}\right\}^{\lambda}}. \tag{12}$$

The corresponding HRF is given by

$$\begin{aligned} h_r(x) &= \frac{\lambda}{\left\{1+\{\frac{c(1-F(x))}{c+(1-c)F(x)}\}^{\lambda}\right\}}\frac{h_f(x)}{(c+(1-c)F(x))} \\ &= \frac{\lambda h_g(x)}{\left\{1+\{\frac{c(1-F(x))}{c+(1-c)F(x)}\}^{\lambda}\right\}}, \end{aligned} \tag{13}$$

where $h_f(x)$ and $h_g(x)$are the HRF of a distribution with PDF $f(x)$ and $g(x)$. Thus

$$\lim_{x\to-\infty} h_r(x) = \lim_{x\to-\infty}\frac{\lambda h_f(x)}{2c} = \lim_{x\to-\infty}\frac{\lambda h_g(x)}{2}$$

$$\lim_{x\to\infty} h_r(x) = \lim_{x\to\infty}\lambda h_f(x) = \lim_{x\to\infty}\lambda h_g(x).$$

It follows from Equation (13) that

$$\frac{\lambda h_f(x)}{2c} \le h_r(x) \le \lambda h_f(x) \qquad (-\infty < x < \infty, \lambda \le 2c)$$

$$\lambda h_f(x) \le h_r(x) \le \frac{\lambda h_f(x)}{2c} \qquad (-\infty < x < \infty, \lambda \ge 2c).$$

Again, Equation (13) shows that $\frac{h_r(x)}{h_f(x)}$ is increasing in x for $c \leq 1$ and decreasing for $0 < c \geq 1$.

The quantile function, $Q_r(u)$, $0 < u < 1$, is given by

$$Q_r(u) = F^{-1}\Bigg\{\frac{c\Big[1 - \big[\frac{1-u}{1+u}\big]^{1/\lambda}\Big]}{1 - [1-c]\Big[1 - \big[\frac{1-u}{1+u}\big]^{1/\lambda}\Big]}\Bigg\}. \tag{14}$$

To generate a random variable from HLMO-X first generate a $U \sim U(0,1)$ then use

$$X = F^{-1}\Bigg\{\frac{c\Big[1 - \big[\frac{1-u}{1+u}\big]^{1/\lambda}\Big]}{1 - [1-c]\Big[1 - \big[\frac{1-u}{1+u}\big]^{1/\lambda}\Big]}\Bigg\}.$$

Another approach to simulate the HLMO-X random variable is to simulate the half-Logistic random variable T and then calculate $X = F^{-1}\Big\{\frac{c(1-e^{-T})}{1-(1-c)(1-e^{-T})}\Big\}$
The p^{th} quantile for HLMO-X family can be obtained as

$$Q_r(p) = F^{-1}\Bigg\{\frac{c\Big[1 - \big[\frac{1-p}{1+p}\big]^{1/\lambda}\Big]}{1 - [1-c]\Big[1 - \big[\frac{1-p}{1+p}\big]^{1/\lambda}\Big]}\Bigg\}.$$

4. Properties of HLMO-X family of distributions

This section is devoted to some important properties of HLMO-X family of distributions.

4.1. Some valuable expansions

Here we provide linear representations for the CDF and PDF of the HLMO-X family of distributions. If c $\in (0,1)$, by applying the generalized binomial expansion in Equation (12), we are getting the following result.

$$R(x) = -1 + 2\Big\{\sum_{j=0}^{\infty}\sum_{k=0}^{\infty}\sum_{l=0}^{k}(-1)^{j+k+l}c^{\lambda j}[1-c]^k\binom{-\lambda j}{k}\binom{\lambda j + k}{l}(F(x))^l\Big\}.$$

By swapping the indices k and l in the sum symbol,

$$R(x) = -1 + 2\Big\{\sum_{j=0}^{\infty}\sum_{l=0}^{\infty}\sum_{k=l}^{\infty}(-1)^{j+k+l}c^{\lambda j}[1-c]^k\binom{-\lambda j}{k}\binom{\lambda j + k}{l}(F(x))^l\Big\}.$$

and then

$$R(x) = \sum_{l=0}^{\infty} b_l[F(x)]^l, \tag{15}$$

where $a_l = 2\sum_{j=0}^{\infty}\sum_{k=l}^{\infty}(-1)^{j+k+l}c^{\lambda j}[1-c]^k\binom{-\lambda j}{k}\binom{\lambda j+k}{l}$, $b_0 = -1 + a_0$and, for $l \geq 1$, $b_l = a_l$. That is, the PDF of X can be expressed as a mixture of exponentiated-F ("exp-F" for short) densities

$$r(x) = \sum_{l=0}^{\infty} b_{l+1}h_{l+1}(x), \tag{16}$$

where $h_{l+1}(x) = (l+1)[F(x)]^l(f(x))$ represents the PDF of exp-F distribution with $(l+1)$ as the power parameter. Therefore, using Equation (16), several mathematical properties of the new distribution are able to be readily derived from those of the exp-F distribution. For instance, the ordinary and incomplete moments as well as the moment generating function of X can be derived from those quantities of the exp-F distribution.

4.2. Moments, generating functions and mean deviation

Let $Y_{l+1}(l > 0)$ be a random variable with power parameter $l+1$ and PDF h_{l+1}. The n^{th} raw moment of X, that is n^{th} raw moment of HLMO-X family of distribution follows from Equation (16) as

$$\acute{\mu_n} = E(X^n) = \sum_{l=0}^{\infty} b_{l+1}E(Y_{l+1}^n). \tag{17}$$

Another formula for $\acute{\mu_n}$ follows from (17) as

$$\acute{\mu_n} = E(X^n) = \sum_{l=0}^{\infty}(l+1)b_{l+1}w_{n,l}, \tag{18}$$

where $w_{n,l} = \int_0^1 Q_F(u)^n u^l du$, $Q_F(u)$ is the quantile function with CDF $F(x)$.

The m^{th} central moment of X by using $\acute{\mu_n}$ in Equation (18) is given by

$$\mu_m = E(X - \acute{\mu_1})^m = \sum_{n=0}^{m}\binom{m}{n}(-\acute{\mu_1})^{m-n}\acute{\mu_n}. \tag{19}$$

The n^{th} incomplete moment of X is described by $m_n(y) = \int_{-\infty}^{y} x^n r(x)$. So $m_n(y)$ follows as

$$m_n(y) = \sum_{l=0}^{\infty}(l+1)b_{l+1}\int_o^{F(y)} Q_F(u)^n u^l du, \tag{20}$$

For most F distributions, the integral can be calculated at least numerically.

For the moment generating function (MGF) $M(t)$ of X, we propose two formulas. The first formula comes from Equation (16) as

$$M(t) = \sum_{l=0}^{\infty} b_{l+1}M_{l+1}(t), \tag{21}$$

where $M_{l+1}(t)$ represented as the MGF of exp-F distribution with power parameter $(l+1)$. The second formula comes from Equation (21) as

$$M(t) = \sum_{l=0}^{\infty}(l+1)b_{l+1}\tau(t,l), \tag{22}$$

Where $\tau(t,l) = \int_0^1 exp[tQ_F(u)]u^l du$.

The mean deviation about the mean($\delta_1 = E(|X - \acute{\mu_1}|)$ and about the median ($\delta_2 = E(|X - M|)$of X are given by

$$\delta_1 = 2\acute{\mu_1}R(\acute{\mu_1}) - 2m_1(\acute{\mu_1}) \tag{23}$$

and

$$\delta_2 = \acute{\mu_1} - 2m_1(M), \tag{24}$$

where $M = Q_r(0.5)$ is the median of X ,$\acute{\mu_1} = E(X)$, $R(\acute{\mu_1})$ is simply calculated from Equation (12) and $m_1(y)$ is the first incomplete moment given by Equation (20) with n = 1 that is,

$$m_1(y) = \sum_{l=0}^{\infty}(l+1)b_{l+1}\rho(y,l), \tag{25}$$

where $\rho(y,l) = \int_o^{F(y)} Q_F(u)u^l du$ can be computed numerically. Other formulae for $m_1(y)$ is

$$m_1(y) = \sum_{l=0}^{\infty} b_{l+1}j_{l+1}(y), \tag{26}$$

where $j_{l+1}(y) = \int_{-\infty}^{y} xh_{l+1}(x)dx$ is the key quantity needed to compute the first incomplete moment of the exp-F distribution. The equations Equation (25) and Equation (26) may be applied to construct Bonferroni and Lorenz curves that are useful in reliability, economics, insurance, demography, and medicine. For a given probability π the Bonferroni and Lorenz curves is defined by $B(\pi) = m_1(q)/(\pi\acute{\mu_1})$ and $L(\pi) = m_1(q)/\acute{\mu_1}$ respectively, where $q = Q(\pi)$ is the quantile function of X at π.

4.3. Order statistics

Assume that $X_1, X_2..., X_n$ is a random sample drawn from HLMO-X family of distribution and $X_{1:n}, X_{2:n}, ..., X_{n:n}$ is the corresponding order statistic. Then the PDF $f_{i:n}(x)$ of the i^{th} order statistic, let's say $X_{i:n}$, is provided by

$$\begin{aligned} f_{i:n}(x) &= \frac{n!}{(i-1)!\,(n-i)!}r(x)R^{i-1}(x)[1-R(x)]^{n-i} \\ &= \frac{n!}{(i-1)!\,(n-i)!}\sum_{j=0}^{n-i}(-1)^j\binom{n-i}{j}r(x)[R(x)]^{i+j-1}. \end{aligned} \tag{27}$$

using Equation (16) and Equation (17) we can get

$$f_{i:n}(x) = \frac{n!}{(i-1)!}\sum_{j=0}^{n-i}\frac{(-1)^j}{(n-i-j)!\,j!}\Big[\sum_{k=0}^{\infty} b_{k+1}(k+1)[F(x)]^k f(x)\Big]\Big[\sum_{l=0}^{\infty} b_l[F(x)]^l\Big]^{i+j-1}.$$

Then we use power series expansion raised to a positive integer by Gradshteyn and Ryzhik (2014)

$$f_{i:n}(x) = \sum_{k,l=0}^{\infty} m_{k,l} h_{k+l+1}(x). \tag{28}$$

where h_{k+l+1} represents the exp-F density function with k+l+1 as its parameter, $m_{k,l} = \frac{n!(k+1)b_{k+1}}{(i-1)!(k+l+1)}\sum_{j=0}^{n-i}\frac{(-1)^j r_{j+i-1,l}}{(n-i-j)!j!}$, b_l is defined in Equation (16), the quantities $r_{j+i-1,l}$ are obtained recursively from $r_{j+i-1,0} = b_0^{j+i-1}$ and (for $l \geq 1$) $r_{j+i-1,l} = (lb_0)^{-1}\sum_{m=1}^{l}[m(i+j)-l]b_m r_{j+i-1,l-m}$. Equation (28) allows us to obtain the ordinary and incomplete moments, generating function and mean deviations of $X_{i:n}$.

4.4. Asymptotic distributions of sample extremes

A CDF R is said to belong to the domain of maximal (minimal) attraction of a non degenerate CDF $H(H^\star)$, denoted by $R \in D_{max}(H)(R \in D_{min}(H^\star))$, if there exist normalizing constants a_n and $b_n > 0$ ($a_n^\star$ and $b_n^\star > 0$) such that $R_{n:n}(a_n + b_n x) = P(X_{n:n} \leq a_n + b_n x) \to H(x)(R_{1:n}(a_n^\star + b_n^\star x) = P(X_{1:n} \leq a_n^\star + b_n^\star x) \to H(x))$ for all continuity points of $H(H^\star)$, where $H^\star(x) = 1 - H(-x)$.
As it is widely known, see (Arnold *et al.* (2008), p. 210, 213), that H belongs to any of the following types:

$$(i)\ H_1(x,\alpha) = e^{-x^{-\alpha}},\quad x > 0, \alpha > 0.$$
$$(ii)\ H_2(x,\alpha) = e^{-(-x)^{\alpha}},\quad x < 0, \alpha > 0.$$
$$(iii)\ H_3(x,\alpha) = e^{-e^{-x}},\quad -\infty < x < \infty.$$

Lemma 1: (See Arnold *et al.* (2008), p. 218)
(i) $F \in D_{max}(H)$ if and only if $n\overline{F}(a_n + b_n x) \to -\log H(x)$
(ii) $F \in D_{min}(H^\star)$ if and only if $nF(a_n^\star + b_n^\star x) \to -\log[1 - H^\star(x)]$.

Theorem 1: For any CDF F, we have
(i) $R \in D_{max}(H)$ if and only if $G \in D_{max}(H)$
(ii) $R \in D_{max}(H)$ if and only if $F \in D_{max}(H)$.
More specifically, we have
(1) $G \in D_{max}(H_1(x;\alpha))$ if and only if $R \in D_{max}(H_1((2)^{-1/\alpha\lambda}x;\alpha\lambda))$
Also $F \in D_{max}(H_1(x;\alpha))$ if and only if $R \in D_{max}(H_1(2c^\lambda)^{-1/\alpha\lambda}x;\alpha\lambda))$
(2) $G \in D_{max}(H_2(x;\alpha))$ if and only if $R \in D_{max}(H_2((2)^{1/\alpha\lambda}x;\alpha\lambda))$
Also $F \in D_{max}(H_2(x;\alpha))$ if and only if $R \in D_{max}(H_2(2c^\lambda)^{1/\alpha\lambda}x;\alpha\lambda))$
(3) $G \in D_{max}(H_3(x))$ if and only if $R \in D_{max}(H_3(x\lambda - \log 2))$
Also $F \in D_{max}(H_3(x))$ if and only if $R \in D_{max}(H_3(x\lambda - \log 2c^\lambda)$
If a_n and $b_n > 0$ are the appropriate normalizing constants for the weak convergence of the upper extremes according to G(or F) in the three cases mentioned above, then $a_{\varphi(n;\lambda)}$ and $b_{\varphi(n;\lambda)} > 0$ are the appropriate normalizing constants for the weak convergence of the upper extremes according to R, where $\varphi(n;b) = [n^{1/b}]$ and $[\mu]$ indicates the integer part of μ.

Proof. If $G \in D_{max}(H)$, with appropriate normalizing constants a_n and $b_n > 0$, then by applying (i) of Lemma 1, as $n \to \infty$,

$$\varphi(n;\lambda)(1 - G(a_{\varphi(n;\lambda)} + b_{\varphi(n;\lambda)}x)) \to -\log H(x),$$

which implies $n(1 - G(a_{\varphi(n;\lambda)} + b_{\varphi(n;\lambda)}x))^\lambda \to [-\log H(x)]^\lambda$. Instead, we have $1 - G(a_n + b_n x) \to 0$, for all values of x for which $-\log H(x)$ is finite. This implies that $(1 - G(a_{\varphi(n;\lambda)} + b_{\varphi(n;\lambda)}x) \to 0$, for all values of x for which $-\log H(x)$ is finite. Thus,

$$\begin{aligned} n[1 - R(a_{\varphi(n;\lambda)} + b_{\varphi(n;\lambda)}x);\lambda] &= n\left\{\frac{2[\bar{G}(a_{\varphi(n;\lambda)} + b_{\varphi(n;\lambda)}x)]^\lambda}{1 + [\bar{G}(a_{\varphi(n;\lambda)} + b_{\varphi(n;\lambda)}x)]^\lambda}\right\} \\ &\sim 2n[\bar{G}(a_{\varphi(n;\lambda)} + b_{\varphi(n;\lambda)}x)]^\lambda \to 2[-\log H(x)]^\lambda. \end{aligned}$$

and also noting that
$2(-\log H_1(x;\alpha))^\lambda = -\log(H_1((2)^{-1/\alpha\lambda}x;\alpha\lambda))$;
$2(-\log H_2(x;\alpha))^\lambda = -\log(H_2((2)^{1/\alpha\lambda}x;\alpha\lambda))$;
$2(-\log H_3(x))^\lambda = -\log(H_3(x\lambda - \log 2))$.
Moving on to the converse claim, let us assume that for a given $\lambda > 0$ we have $R \in D_{max}(H)$, with $\hat{a}_n$ and $\hat{b}_n > 0$ are the normalizing constants based on R. From (i) of Lemma 1, we then have

$$n[1 - R(\hat{a}_n + \hat{b}_n x);\lambda] \to -\log H(x),$$

as $n \to \infty$, which implies $1 - R(\hat{a}_n + \hat{b}_n x;\lambda) \to 0$, that is, $G(\hat{a}_n + \hat{b}_n x \to 1)$, as $n \to \infty$, for all values of x for which $-\log H(x)$ is finite. Thus,

$$n[1 - R(\hat{a}_n + \hat{b}_n x);\lambda] = n\left\{\frac{2[\bar{G}(\hat{a}_n + \hat{b}_n x)]^\lambda}{1 + \bar{G}(\hat{a}_n + \hat{b}_n x)^\lambda}\right\} \sim 2n[\bar{G}(\hat{a}_n + \hat{b}_n x)]^\lambda.$$

From this we get, $2n[\bar{G}(\hat{a}_n + \hat{b}_n x)]^\lambda \to -\log H(x)$ or equivalently, $\varphi(n;\lambda)(1 - G(\hat{a}_n + \hat{b}_n x)) \to \frac{[-\log H(x)]^{1/\lambda}}{2}$. Since the last convergence holds for all subsequence of n and specifically holds for the subsequence $\acute{n} = \varphi(n;1/\lambda) = [n^\lambda]$, where $\varphi(\acute{n};\lambda) = [[n^\lambda]^{1/\lambda}] \sim n$, we get $n(1 - G(\tilde{a}_n + \tilde{b}_n x)) \to \frac{[-\log H(x)]^{1/\lambda}}{2}$, where $\tilde{a}_n = \hat{a}_{[n^\lambda]}$ and $\tilde{b}_n = \hat{b}_{[n^\lambda]}$. Thus, we get the expected result (notice that the theorem's converse portion holds true for the normalizing constants $\tilde{a}_n$ and $\tilde{b}_n$, that is $R(\hat{a}_n + \hat{b}_n x) \in D_{max}(H)$. Implies $G(\tilde{a}_n + \tilde{b}_n x) = G(\hat{a}_{[n^\lambda]} + \hat{b}_{[n^\lambda]}x) \in D_{max}(H))$, Hence, the given theorem is proved for the part (i) scenario. The proof of theorem for the part (ii) scenario follows by similar manner by using Lemma 1, Part (i).This completes the proof.

5. Half-logistic-Marshal-Olkin Lomax distribution

Let X be a random variable following the Lomax (L) distribution with parameters α and θ then $f(x) = \frac{\alpha}{\theta}[1 + \frac{x}{\theta}]^{-(\alpha+1)};\ x > 0,\ \alpha, \theta > 0$. The PDF of half-logistic-Marshall-Olkin Lomax (HLMOL) distribution using Equation (11) is defined as

$$r(x) = \frac{2\lambda\alpha c^\lambda}{\theta} \frac{[(1 + \frac{x}{\theta})^\alpha + c - 1]^{\lambda-1}[1 + \frac{x}{\theta}]^{\alpha-1}}{\left[[(1 + \frac{x}{\theta})^\alpha + c - 1]^\lambda + c^\lambda\right]^2};\ x > 0,\ c, \lambda, \alpha, \theta > 0, \tag{29}$$

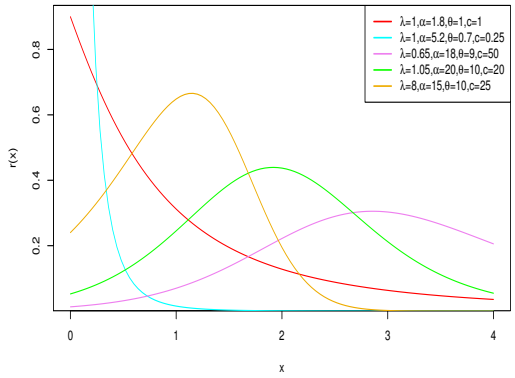


Figure 1: PDF of HLMOL for various values of α, θ, λ and c

where c, λ, α and θ are location, location scale, scale, and shape parameters, respectively. Hereafter, a random variable X with a PDF in Equation (29) will be denoted by $X \sim$ HLMOL$(c, \lambda, \alpha, \theta)$. The CDF of the Lomax distribution is $\mathrm{F(x)} = 1 - [1 + \frac{x}{\theta}]^{-\alpha}$ and hence from Equation (12) the CDF of the HLMOL distribution is

$$R(x) = \frac{[(1 + \frac{x}{\theta})^{\alpha} + c - 1]^{\lambda} - c^{\lambda}}{[(1 + \frac{x}{\theta})^{\alpha} + c - 1]^{\lambda} + c^{\lambda}}. \tag{30}$$

In the form of graphical representations, Figure 1 displays a few plots of $r(x)$ for selected values of the parameter c, λ, α and θ. These plots demonstrate that the PDF has good shape flexibility. It can be reversed J-shape, left-skewed, right-skewed, or symmetric.

Some unique cases of HLMOL distribution:

1. When $c = 1$, the HLMOL distribution reduces to half-logistic Lomax by Anwar and Zahoor (2018) distribution.
2. When $\lambda = 1$, the HLMOL distribution reduces to Marshall-Olkin half-logistic Lomax distribution.
3. When $\lambda = 1$ and c=0.5, the HLMOL distribution reduces to Lomax distribution.

In lifetime analysis, the HRF is a useful function. Therefore, the HRF of $X \sim$ HLMOL$(c, \lambda, \alpha, \theta)$ is given by

$$h_r(x) = \frac{\frac{\lambda\alpha}{\theta}[(1 + \frac{x}{\theta})^{\alpha} + c - 1]^{\lambda-1}[1 + \frac{x}{\theta}]^{\alpha-1}}{\left[[(1 + \frac{x}{\theta})^{\alpha} + c - 1]^{\lambda} + c^{\lambda}\right]}. \tag{31}$$

Figure 2 displays the graphs of $h_r(x)$ for selected values of the parameters α, θ, λ and c. It can be upside down bathtub and decreasing.

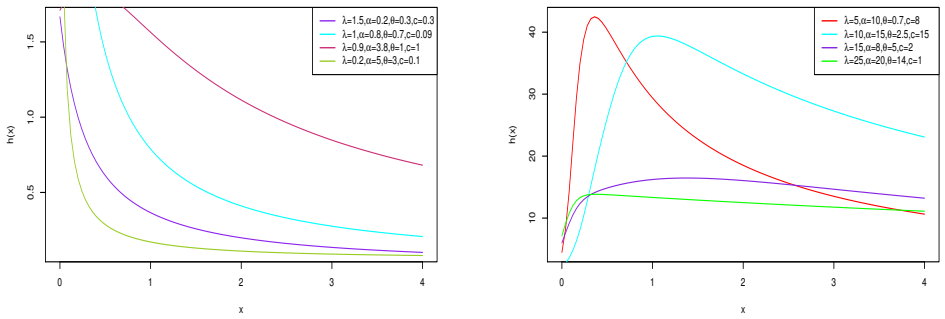


Figure 2: HRF of HLMOL for various values of α, θ, λ and c

By using Equation (14) the quantile function, $Q_r(u)$, $0 < u < 1$, is given by

$$Q_r(u) = \theta\Big\{\Big[\frac{[\frac{1-u}{1+u}]^{1/\lambda}}{1-[1-c][1-[\frac{1-u}{1+u}]^{1/\lambda}]}\Big]^{-1/\alpha} - 1\Big\}.$$

To generate a random variable from HLMOL, first generate a $U \sim U(0,1)$ then use

$$X = \theta\Big\{\Big[\frac{[\frac{1-u}{1+u}]^{1/\lambda}}{1-[1-c][1-[\frac{1-u}{1+u}]^{1/\lambda}]}\Big]^{-1/\alpha} - 1\Big\}.$$

Another approach to simulate the HLMOL random variable is by simulating the half-logistic random variable T and then calculate

$$X = \theta\Big\{\Big[\frac{[e^{-T}]^{1/\lambda}}{1-[1-c][1-[e^{-T}]^{1/\lambda}]}\Big]^{-1/\alpha} - 1\Big\}.$$

The p^{th} quantile for HLMOL distribution can be obtained as

$$Q_r(p) = \theta\Big\{\Big[\frac{[\frac{1-p}{1+p}]^{1/\lambda}}{1-[1-c][1-[\frac{1-p}{1+p}]^{1/\lambda}]}\Big]^{-1/\alpha} - 1\Big\}.$$

If p =1/2, that is median of HLMOL is given by

$$M = \theta\Big\{\Big[\frac{[\frac{1}{3}]^{1/\lambda}}{1-[1-c][1-[\frac{1}{3}]^{1/\lambda}]}\Big]^{-1/\alpha} - 1\Big\}.$$

5.1. Linear representation

By using Equations (16) and (17) the linear representation of CDF and PDF of HLMOL distribution is given by

$$R(x) = \sum_{l=0}^{\infty} b_l[1-[1+\frac{x}{\theta}]^{-\alpha}]^l, \tag{32}$$

Where $a_l = 2\sum_{j=0}^{\infty}\sum_{k=l}^{\infty}(-1)^{j+k+l}c^{\lambda j}[1-c]^k\binom{-\lambda j}{k}\binom{\lambda j+k}{l}$, $b_0 = -1 + a_0$and, for $l \geq 1$,$b_l = a_l$

$$\begin{aligned} r(x) &= \sum_{l=0}^{\infty}\sum_{j=0}^{l}(l+1)b_{l+1}(-1)^j\binom{l}{j}\frac{\alpha}{\theta}[1+\frac{x}{\theta}]^{-(\alpha+\alpha j+1)} \\ &= \sum_{l=0}^{\infty}\sum_{j=0}^{l}(l+1)b_{l+1}(-1)^j\binom{l}{j}\frac{\alpha}{\alpha+\alpha j}L(x;\alpha+\alpha j,\theta), \end{aligned} \tag{33}$$

Where $L(x;\alpha+\alpha j,\theta)$ denoted the Lomax PDF with parameter θ and $\alpha+\alpha j$. So the PDF of HLMOL is simply an infinite linear combination of Lomax distribution. Thus, some mathematical properties of the new distribution can be obtained straightly from those Lomax distribution properties based on Equation (33).

5.2. Moments and generating functions

The n^{th} raw moment of X is obtained from Equation (20)

$$\acute{\mu_n} = E(X^n) = \sum_{l=0}^{\infty}\sum_{i=0}^{n}(l+1)b_{l+1}\theta^n(-1)^i\binom{n}{i}\beta(l+1,1+\frac{i-n}{\alpha}), \qquad n<\alpha.$$

If n=1, That is the mean of HLMOL distribution is given by

$$\acute{\mu_1} = E(X) = \sum_{l=0}^{\infty}(l+1)b_{l+1}\theta[\beta(l+1,1-1/\alpha)-\beta(l+1,1)], \qquad \alpha>1.$$

If n=2

$$\begin{aligned} \acute{\mu_2} &= E(X^2) \\ &= \sum_{l=0}^{\infty}(l+1)b_{l+1}\theta^2[\beta(l+1,1-2/\alpha)-2\beta(l+1,1-1/\alpha)+\beta(l+1,1)], \quad \alpha>2. \end{aligned}$$

The m^{th} central moment of X by using $\acute{\mu_n}$ in Equation (19) is given by

$$\mu_m = E(X-\acute{\mu_1})^m = \sum_{n=0}^{m}\binom{m}{n}(-\acute{\mu_1})^{m-n}\acute{\mu_n}.$$

If m=2, That is the variance of HLMOL distribution is given by

$$\begin{aligned} \mu_2 &= \sum_{l=0}^{\infty}(l+1)b_{l+1}\theta^2[\beta(l+1,1-2/\alpha)-2\beta(l+1,1-1/\alpha)+\beta(l+1,1)] \\ &\quad -\Big[\sum_{l=0}^{\infty}(l+1)b_{l+1}\theta[\beta(l+1,1-1/\alpha)-\beta(l+1,1)]\Big]^2, \qquad \alpha>2. \end{aligned}$$

Then, the moment measure of skewness $S = \frac{\mu_3^2}{\mu_2^3}$ and moment measure of kurtosis $K = \frac{\mu_4}{\mu_2^2}$ can be calculated from the second, third and fourth central moments.

The n^{th} incomplete moment of X is defined by using Equation (20) is given by

$$m_n(y) = \sum_{l=0}^{\infty}\sum_{i=0}^{n}(l+1)b_{l+1}\theta^n(-1)^i\binom{n}{i}\beta_{F(y)}(l+1, 1+\frac{i-n}{\alpha}), \qquad n < \alpha.$$

By using Equation (22) the MGF of X is given by

$$\begin{aligned} M(t) &= \sum_{l=0}^{\infty}(l+1)b_{l+1}\int_0^1\sum_{n=0}^{\infty}\frac{(t\theta)^n}{n!}[(1-u)^{-1/\alpha}-1]^n u^l du \\ &= \sum_{n=0}^{\infty}\frac{t^n}{n!}\acute{\mu_n}, \end{aligned}$$

where $\acute{\mu_n}$ is the n^{th} raw moment of the HLMOL distribution. The Bonferroni and the Lorenz curve are given by

$$B(\pi) = \frac{\sum_{l=0}^{\infty}(l+1)b_{l+1}[\beta_{F(q)}(l+1, 1-1/\alpha) - \beta_{F(q)}(l+1,1)]}{\pi\sum_{l=0}^{\infty}(l+1)b_{l+1}[\beta(l+1,1-1/\alpha)-\beta(l+1,1)]}, \quad \alpha > 1.$$

$$L(\pi)) = \frac{\sum_{l=0}^{\infty}(l+1)b_{l+1}[\beta_{F(q)}(l+1, 1-1/\alpha) - \beta_{F(q)}(l+1,1)]}{\sum_{l=0}^{\infty}(l+1)b_{l+1}[\beta(l+1,1-1/\alpha)-\beta(l+1,1)]}, \quad \alpha > 1, \tag{34}$$

where $q = Q(\pi)$ is the quantile function of X at π.

Table 2 gives the mean, variance, third raw moment, skewness and kurtosis of HLMOL distribution for different choices of parameter values. For fixed λ and c, the mean and variance of the HLMOL distribution are increasing functions of θ and α. Also the distribution of the HLMOL distribution tends to be skewed more to the right as θ and α decreases. For fixed λ, θ and α, the HLMOL distribution can be platykurtic, mesokurtic and leptokurtic as c increases. Also the distribution of the HLMOL distribution tends to be skewed more to the left as c increases. That is, the HLMOL is positively and negatively skewed, platykurtic, mesokurtic and leptokurtic distribution.

5.3. Order statistics

Assume that $X_1, X_2..., X_n$ is a random sample drawn from HLMOL distribution and $X_{1:n}, X_{2:n}, ..., X_{n:n}$ is the corresponding order statistic. Then the PDF $f_{i:n}(x)$ of the i^{th} order statistic, let's say $X_{i:n}$, is provided by

$$f_{i:n}(x) = \sum_{k,l=0}^{\infty}\sum_{j=0}^{k+l}m_{k,l}(-1)^j(k+l+1)\binom{k+l}{j}\frac{\alpha}{\alpha+\alpha j}L(x;\alpha+\alpha j,\theta), \tag{35}$$

where $m_{k,l} = \frac{n!(k+1)b_{k+1}}{(i-1)!(k+l+1)}\sum_{j=0}^{n-i}\frac{(-1)^j r_{j+i-1,l}}{(n-i-j)!j!}$, the quantities $r_{j+i-1,l}$ are obtained recursively from $r_{j+i-1,0} = b_0^{j+i-1}$ and (for $l \geq 1$) $r_{j+i-1,l} = (lb_0)^{-1}\sum_{m=1}^{l}[m(i+j)-l]b_m r_{j+i-1,l-m}$ and $L(x;\alpha+\alpha j,\theta)$ denoted the Lomax PDF with parameter θ and $\alpha+\alpha j$. So the PDF of i^{th} order statistic of HLMOL distribution is simply an infinite linear combination of Lomax distribution.

Table 2: $\acute{\mu}_1$, μ_2, μ_3, S and K for various choices of parameters

Parameter	$\acute{\mu}_1$	μ_2	μ_3	S	K
$\lambda = 0.95$ c=0.25 θ=0.7 α=5.2	0.1218	0.0343	0.03572	31.6755	107.1577
$\lambda = 0.95$ c=0.25 θ=0.9 α=6	0.1319	0.0370	0.0342	23.0538	61.2956
$\lambda = 0.95$ c=0.25 θ=10 α=20	0.3904	0.2398	0.3475	8.7539	17.3673
λ =3 c=20 θ=50 α=50	2.3191	1.1223	0.1005	0.0071	2.7521
λ =3 c=74.4263 θ=50 α=50	3.5554	1.5711	-0.5373	0.0744	3
λ =3 c=80 θ=50 α=50	3.6277	1.5920	-0.5798	0.0833	3.0237
$\lambda = 0.6$ c=0.2 θ=0.7 α=7	0.1686	0.0757	0.1411	45.8365	433.6134
$\lambda = 10$ c=1.1 θ=1.1 α=1.1	0.1706	0.0257	0.0091	4.8254	12.1528
$\lambda = 4.9$ c=1 θ=1 α=1	0.3709	0.1731	0.2933	16.5825	62.3333

5.4. Asymptotic distributions of sample extremes

Consider the asymptotic distributions of first order statistic $X_{1:n}$ and n^{th} order statistic $X_{n:n}$. We using the asymptotic results for $X_{1:n}$ and $X_{n:n}$ by Arnold *et al.* (2008) and Theorem 1, we can find the limiting distribution of extreme order statistic.

For HLMOL distribution $R^{-1}(O) = 0$ which is finite and by using L'Hospital's

$$\lim_{\epsilon\to 0+} \frac{R[R^{-1}(0)+\epsilon x]}{R[R^{-1}(0)+\epsilon]} = \lim_{\epsilon\to 0+} x\frac{r[\epsilon x]}{r[x]} = x.$$

Therefore the asymptotic distribution $X_{1:n}$ is of Weibull type with $\alpha = 1$, that is $R \in D_{min}(H_2^\star(x;1)) = 1 - e^{-x}, x > 0$. Here the normalizing constants based on R are given by $a_n^\star = R^{-1}(O) = 0$, $b_n^\star = R^{-1}(1/n) - R^{-1}(O) = \theta\Big\{\Big[\frac{\left[\frac{n-1}{n+1}\right]^{1/\lambda}}{1-[1-c]\left[1-\left[\frac{n-1}{n+1}\right]^{1/\lambda}\right]}\Big]^{-1/\alpha} - 1\Big\}$.

For Lomax distribution $F^{-1}(1) = \infty$, by using L'Hospital's

$$\lim_{t\to\infty} \frac{1-F(tx)}{1-F(t)} = \lim_{t\to\infty} x\frac{f(tx)}{f(t)} = x^{-\alpha}.$$

Therefore the asymptotic distribution of $X_{n:n}$ based on F is Fréchet type. From Theorem 1 the asymptotic distribution of $X_{n:n}$ based on R is Fréchet type, that is $R \in D_{max}(H_1(x;\alpha)) = e^{-x^{-\alpha}}, x > 0, \alpha > 0$. Here the normalizing constants based on R are given by $a_n = 0$, $b_n = R^{-1}(1-1/n) = \theta\Big\{\Big[\frac{\left[\frac{1}{2n-1}\right]^{1/\lambda}}{1-[1-c]\left[1-\left[\frac{1}{2n-1}\right]^{1/\lambda}\right]}\Big]^{-1/\alpha} - 1\Big\}$.

6. Estimation of parameters by maximum likelihood method

Here, we discuss maximum likelihood estimation of HLMO-X family of distribution along with a simulation study of HLMOL. Let $x_1, ..., x_n$ be a sample from $X \sim$ HLMO-X(λ, c, ξ). Let $\Theta = (\lambda, c, \xi)^T$ be the parameter vector and ξ corresponds to the parameter vector of the baseline distribution F, $F(x) = F(x_i;\xi)$, $f(x) = f(x_i;\xi)$. The total log-likelihood function for Θ is given by

$$\begin{aligned}\ell_n = \ell_n(\Theta|x_1,...,x_n) = n\log(2\lambda) + \lambda\sum_{i=1}^n \log\left\{\frac{c[1-F(x_i;\xi)]}{c+(1-c)F(x_i;\xi)}\right\} + \sum_{i=1}^n \log[f(x_i;\xi)] \\ - \sum_{i=1}^n \log[1-F(x_i;\xi)] - 2\sum_{i=1}^n \log\left\{1+\left[\frac{c[1-F(x_i;\xi)]}{c+(1-c)F(x_i;\xi)}\right]^\lambda\right\} \\ - \sum_{i=1}^n \log[c+(1-c)F(x_i;\xi)].\end{aligned}$$

The score function $U_n(\Theta) = (\frac{\partial\ell_n}{\partial\lambda}, \frac{\partial\ell_n}{\partial c}, \frac{\partial\ell_n}{\partial\xi})^T$ has components given by

$$\begin{aligned}\frac{\partial\ell_n}{\partial\lambda} = \frac{n}{\lambda} - 2\sum_{i=1}^n \log\left\{\frac{c[1-F(x_i;\xi)]}{c+(1-c)F(x_i;\xi)}\right\}\frac{[c[1-F(x_i;\xi)]]^\lambda}{[c+(1-c)F(x_i;\xi)]^\lambda + [c[1-F(x_i;\xi)]]^\lambda} \\ + \sum_{i=1}^n \log\left\{\frac{c[1-F(x_i;\xi)]}{c+(1-c)F(x_i;\xi)}\right\},\end{aligned}$$

$$\frac{\partial \ell_n}{\partial c} = \lambda \sum_{i=1}^{n} \frac{F(x_i;\xi)}{c[c+(1-c)F(x_i;\xi)]} - \frac{[1-F(x_i;\xi)]}{[c+(1-c)F(x_i;\xi)]} - 2\lambda \sum_{i=1}^{n} \frac{c^{(\lambda-1)}[1-F(x_i;\xi)]^{\lambda}F(x_i;\xi)}{[c+(1-c)F(x_i;\xi)]\{[c+(1-c)F(x_i;\xi)]^{\lambda}+[c[1-F(x_i;\xi)]]^{\lambda}\}},$$

$$\frac{\partial \ell_n}{\partial \xi} = -\lambda \sum_{i=1}^{n} \frac{F^{(\xi)}(x_i;\xi)}{[1-F(x_i;\xi)][c+(1-c)F(x_i;\xi)]} + \sum_{i=1}^{n} \frac{f^{(\xi)}(x_i;\xi)}{f(x_i;\xi)} - (1-c)\sum_{i=1}^{n} \frac{F^{(\xi)}(x_i;\xi)}{c+(1-c)F(x_i;\xi)} + \sum_{i=1}^{n} \frac{F^{(\xi)}(x_i;\xi)}{1-F(x_i;\xi)} - 2c\sum_{i=1}^{n} \frac{[1-F(x_i;\xi)]^{\lambda-1}F^{(\xi)}(x_i;\xi)}{[c+(1-c)F(x_i;\xi)]\{[c+(1-c)F(x_i;\xi)]^{\lambda}+[c[1-F(x_i;\xi)]]^{\lambda}\}},$$

where $f^{(\xi)}(x_i;\xi) = \frac{\partial f(x_i;\xi)}{\partial \xi}$ and $F^{(\xi)}(x_i;\xi) = \frac{\partial F(x_i;\xi)}{\partial \xi}$. The maximum likelihood estimates (MLEs) of Θ, say $\hat{\Theta} = (\hat{\lambda}, \hat{c}, \hat{\xi})$, are the simultaneous solutions of the following equations: $\frac{\partial \ell_n}{\partial \lambda} = 0$, $\frac{\partial \ell_n}{\partial c} = 0$ and $\frac{\partial \ell_n}{\partial \xi} = 0$. These equations cannot be solved analytically and statistical software can be used to solve them numerically.

6.1. Simulation study

Here we perform a simulation study evaluating the performance of the MLEs presented above for the HLMOL distribution for selected values of the parameters θ, α, λ and c. The simulation experiment was repeated 1000 times each with sample sizes 50, 100, 150, 200 and parameter combinations are

1. λ=1.5, α=4 fixed c=1 and θ=1.
2. α=0.5, θ=0.6 fixed λ=1 and c=0.5.
3. θ= 0.2, c=0.5 fixed λ=1 and α=1.
4. λ=1, θ= 1 fixed c=1 and α=1.
5. λ=0.5, c= 0.2, θ=0.4 and fixed α=1.
6. λ= 0.75, α=0.15, c=0.1, θ= 0.05.

Table 3 presents the average estimates (*AEs*), average bias (*Bias*) and mean square error (*MSE*) values of parameters for different sample sizes. It can be noted that as sample size increases, the *Bias* decay towards zero and *MSE* decreases. That is, the estimators are asymptotically unbiased and consistent. Therefore the maximum likelihood estimation method works quite well to estimate the parameters of the HLMOL distribution.

Table 3: *AE*s, *Bias* and *MSE* of parameters based on 1000 simulations of the HLMOL distribution

	n	Parameter	*AE*s	*Bias*	*MSE*
I	50	λ	1.1136	-0.3864	5.6075
		α	3.8552	-0.1448	0.7885
	100	λ	1.3263	-0.1736	4.0699
		α	3.9348	-0.0652	0.5723
	150	λ	1.4484	-0.0516	1.8639
		α	3.9805	-0.0195	0.2621
	200	λ	1.5022	0.0022	0.0059
		α	4.0006	0.0006	0.0008
II	50	α	0.5389	0.0388	0.0182
		θ	0.7409	0.1409	0.0205
	100	α	0.5196	0.0196	0.0069
		θ	0.6635	0.0635	0.0131
	150	α	0.5113	0.0113	0.0043
		θ	0.6367	0.0367	0.0120
	200	α	0.5080	0.0079	0.0033
		θ	0.6256	0.0256	0.0117
III	50	c	0.5013	0.0013	0.0002
		θ	0.2051	0.0051	0.0019
	100	c	0.5002	0.0002	0.0001
		θ	0.2017	0.0017	0.0010
	150	c	0.5001	6.5753e-05	0.0001
		θ	0.2016	0.0016	0.0007
	200	c	0.4999	-1.2853e-05	1.2186e-05
		θ	0.2001	5.4980e-05	7.6654e-05
IV	50	λ	1.1002	0.1002	0.1468
		θ	1.2458	0.2458	0.8529
	100	λ	1.0464	0.0464	0.0350
		θ	1.1082	0.1082	0.1621
	150	λ	1.0325	0.0325	0.0224
		θ	1.0732	0.0732	0.1109
	200	λ	1.0233	0.0233	0.0149
		θ	1.0489	0.0489	0.0635

Continued on the next page

Table 3:(Continued)

	n	Parameter	*AE*s	*Bias*	*MSE*
V	50	λ	0.5263	0.0263	0.0105
		c	0.2272	0.0272	0.0096
		θ	0.4128	0.0128	0.0023
	100	λ	0.5132	0.0132	0.0043
		c	0.2136	0.0136	0.0040
		θ	0.4067	0.0067	0.0010
	150	λ	0.5085	0.0085	0.0028
		c	0.2076	0.0076	0.0024
		θ	0.4035	0.0035	0.0006
	200	λ	0.5032	0.0032	0.0020
		c	0.2049	0.0049	0.0017
		θ	0.4022	0.0022	0.0004
VI	50	λ	0.4659	-0.2840	0.8821
		α	0.1225	-0.0274	0.0375
		c	0.0669	-0.0331	0.0938
		θ	0.2106	0.1606	1.5323
	100	λ	0.8875	0.1375	0.5642
		α	0.1976	0.04764	0.0187
		c	0.0882	-0.0118	0.0074
		θ	0.1965	0.1465	0.2604
	150	λ	0.8732	0.1232	0.4111
		α	0.1873	0.0373	0.0147
		c	0.0885	-0.0114	0.0061
		θ	0.1566	0.1066	0.0987
	200	λ	0.8688	0.1188	0.2939
		α	0.1763	0.0263	0.0119
		c	0.0901	-0.0099	0.0048
		θ	0.1357	0.0857	0.0739

7. Test to compare HLMOL with Lomax and Half-logistic-Lomax distributions

Since Lomax (L), half-logistic-Lomax (HLL) by Anwar and Zahoor (2018) and HLMOL distributions are nested models, To distinguish between them, the likelihood ratio (LR) test is employed. For the nested models, the LR statistic is

$$LR = -2\left\{\frac{likelihood\ under\ the\ null\ hypothesis}{likelihood\ under\ the\ alternative\ hypothesis}\right\}.$$

This statistic is asymptotically (as $n \to \infty$) distributed as chi-square distribution with m degrees of freedom (df), where m is the number of additional parameters .

When c is equal to 1, the HLMOL distribution becomes the HLL distribution. So, in order to compare the HLMOL with the HLL distribution, we test the null hypothesis that $H_0 : c = 1\ against\ H_1 : c \neq 1$, and the corresponding LR statistic asymptotically (as $n \to$

∞) distributed as chi-square distribution with 1 df. To investigate how well the test statistic performed for the above hypothesis , we conducted a simulation study.The simulation experiment was performed 1000 times, with sample sizes of 100, 250, and 500 with different parameter combinations. From the HLMOL distribution, a random sample is created, and the test is then run with a 5% level of significance . Calculating the proportion of times the null hypothesis H_0 is rejected requires running the simulation 1000 times for each set of parameter combinations. In order to estimate the test's power, we look at the proportion of times that H_0 is rejected. Table 4 provides the proportions for the 5% level of significance.

The findings in Table 4 show that, for fixed c, θ and α the power of the tests increases as a function of λ. Additionally, given a fixed value of θ, α and λ the tests' power is a diminishing function of c. In general, as sample sizes grow, power grows as well

Table 4: The proportion of times (out of 1000) that the H_0 is rejected at 5% level of significance.

Parameter value				n=100	n=250	n=500
c	θ	α	λ			
			0.15	0.979	0.989	0.993
		0.9	1.25	0.984	0.991	0.999
	0.05		2	0.987	0.993	0.999
			0.15	0.969	0.971	0.999
		1.25	1.25	0.972	0.993	1
			2	0.986	0.988	1
0.1			0.15	0.96	0.976	0.998
		0.9	1.25	0.971	0.991	1
	0.5		2	0.987	0.993	1
			0.15	0.973	0.981	0.986
		1.25	1.25	0.984	0.985	1
			2	0.989	0.994	1
			0.15	0.848	0.886	0.904
		0.9	1.25	0.924	0.935	0.946
	0.05		2	0.902	0.945	0.985
			0.15	0.907	0.952	0.954
		1.25	1.25	0.924	0.956	0.983
			2	0.972	0.980	0.983
0.25			0.15	0.893	0.899	0.95
		0.9	1.25	0.9	0.912	0.954
	0.5		2	0.911	0.921	0.962
			0.15	0.87	0.901	0.915
		1.25	1.25	0.907	0.927	0.939
			2	0.916	0.949	0.966

Similarly, When λ is equal to 1 and c=0.5, the HLMOL distribution becomes the L distribution. So, in order to compare the HLMOL with the L distribution, we test the null hypothesis that
$H_0 : \lambda = 1,\ c = 0.5$ against $H_1 : \lambda \neq 1,\ c \neq 0.5$, and the corresponding LR statistic asymptotically (as n $\to \infty$) distributed as chi-square distribution with 2 DF. To investigate

how well the test statistic performed for the above hypothesis, we conducted a simulation study. The simulation experiment was performed 1000 times, with sample sizes of 100, 250, and 500 with different combination of parameters. From the HLMOL distribution, a random sample is created, and the test is then run with a 5% level of significance . Calculating the proportion of times the null hypothesis H_0 is rejected requires running the simulation 1000 times for each set of combination of parameters. In order to estimate the test's power, we look at the proportion of times that H_0 is rejected. Table 5 provides the proportions for the 5% level of significance.

Table 5: The proportion of times (out of 1000) that the H_0 is rejected at 5% level of significance.

Parameter value				n=100	n=250	n=500
c	θ	α	λ			
0.1			0.2	0.986	0.998	1
		0.15	1	0.987	0.999	1
	0.05		2	0.991	1	1
			0.2	0.998	0.999	1
		1.5	1	0.972	0.973	0.986
			2	0.975	0.982	0.994
0.1			0.2	0.989	0.998	1
		0.15	1	0.993	0.999	1
	0.5		2	0.998	1	1
			0.2	0.996	0.997	1
		1.5	1	0.997	0.997	1
			2	0.997	1	1
			0.2	0.961	0.986	0.993
		0.15	1	0.982	0.988	0.994
	0.05		2	0.982	0.985	0.997
			0.2	0.997	0.999	1
		1.5	1	0.914	0.95	0.998
			2	0.95	0.981	1
0.25			0.2	0.97	0.984	0.99
		0.15	1	0.979	0.985	0.991
	0.5		2	0.98	0.985	0.991
			0.2	0.987	0.992	0.999
		1.5	1	0.988	0.994	1
			2	0.993	0.994	1

The results in Table 5 demonstrate that, for fixed c, θ and α, the power of the tests generally increases as a function of λ. Additionally, the power of the tests is a decreasing function of c for a certain value of λ, θ and α. In general, power increases as sample size increase.

8. Applications

Under this head, we exhibit the importance of the proposed family. We fit the HLMOL distribution to two data sets and compare this distribution with four other models, namely: Kumaraswamy-generalized Lomax (Kw-GL) distribution by Shams (2013), Weibull Lomax (WL) distribution by Tahir *et al.* (2015), HLL and L distribution. The MLEs of the parameters of the models are calculated and goodness-of-fit statistics for the models are compared. The measures including the Akaike information criterion (AIC), Bayesian information criterion (BIC) and Kolmogorov-Smirnov (K-S) statistic with p-value (p-V). Additionally, we employ the LR test to compare the HLMOL distribution with the L and HLL distributions.

8.1. The secondary reactor pumps data set

This data represents the time period between secondary reactor pump failures. The data was originally discussed in Suprawhardana and Prayoto (1999). and was previously used by Bebbington *et al.* (2007). Following are the time between failures for 23 secondary reactor pumps.
{2.160, 0.150, 4.082, 0.746, 0.358, 0.199, 0.402, 0.101, 0.605, 0.954, 1.359, 0.273, 0.491, 3.465, 0.070, 6.560, 1.060, 0.062, 4.992, 0.614, 5.320, 0.347, 1.921}

The necessary numerical summaries for the five fits using the secondary reactor pumps data set includes the estimated log-likelihood function ($\hat{\ell}$), AIC, BIC and K-S with p-V are provided in Tables 6 and 7. Additionally, Table 8 provides two LR statistics based on data set from secondary reactor pumps along with (p-V).

Table 6: Estimated values, log-likelihood, AIC and BIC for the secondary reactor pumps data set

Distribution	Estimates	$-ln(L)$	AIC	BIC
HLMOL	$\boldsymbol{\hat{\lambda} = 0.5250}$ $\boldsymbol{\hat{\alpha} = 8442.2096}$ $\boldsymbol{\hat{c} = 0.1662}$ $\boldsymbol{\hat{\theta} = 9025.7431}$	**31.862**	**67.7242**	**69.9952**
Kw-GL	$\hat{a} = 0.8085$ $\hat{b} = 185.7834$ $\hat{\lambda} = 297.5083$ $\hat{\alpha} = 0.3337$	32.51709	73.03418	77.57616
WL	$\hat{a} = 7.2122$ $\hat{b} = 0.8163$ $\hat{\beta} = 12.6936$ $\hat{\alpha} = 0.8239$	32.51238	73.02476	77.56674
HLL	$\hat{\lambda}$= 0.6797 $\hat{\alpha} = 2.3802$ $\hat{\theta} = 0.7796$	32.64682	71.29364	74.70013
L	$\hat{\alpha} = 2.2425$ $\hat{\theta}$= 2.1699	32.4952	68.9903	71.2613

Table 7: K-S with p-V for the secondary reactor pumps data set

Distributions	K-S	p-V
HLMOL	**0.0954**	**0.9718**
Kw-GL	0.1186	0.8654
WL	0.1176	0.8717
HLL	0.096283	0.9695
L	0.099734	0.9589

Table 8: The values of LR statistic for different hypothesis and data sets

Models	Hypothesis	Secondary reactor pumps data set		
		LR	df	p-V
HLMOL vs. L	H_0 : λ= 1,c=0.5 vs. H_1 : H_0 is false	7.2334	2	0.0269
HLMOL vs. HLL	H_0 : c=1 vs. H_1 : H_0 is false	14.6222	1	< 0.001

Figure 3 display the total time test (TTT) plot for the secondary reactor pumps data set, and Figure 4 display the graphs of estimated PDF and CDF of the considered distributions for secondary reactor pumps data set.

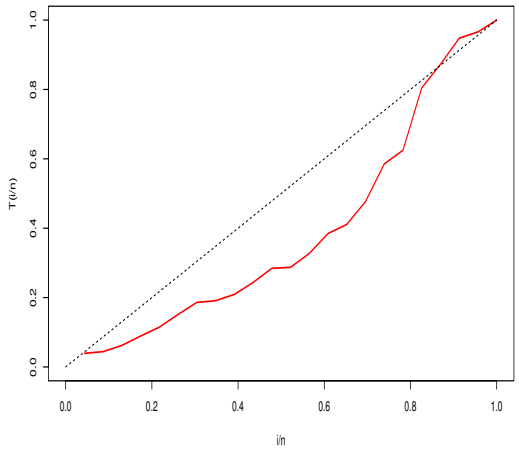


Figure 3: TTT-plot for the secondary reactor pumps data set

8.2. Bladder cancer patients data set

The data set was given by Almheidat *et al.* (2015). It is corresponding to remission times (months) of a random sample of 128 bladder cancer patients. The data are as given below
{0.080, 0.200, 0.400, 0.500, 0.510, 0.810, 0.900, 1.050, 1.190, 1.260, 1.350, 1.400,1.460, 1.760, 2.020, 2.020, 2.070, 2.090, 2.230, 2.260, 2.460, 2.540, 2.620, 2.640, 2.690, 2.690, 2.750,

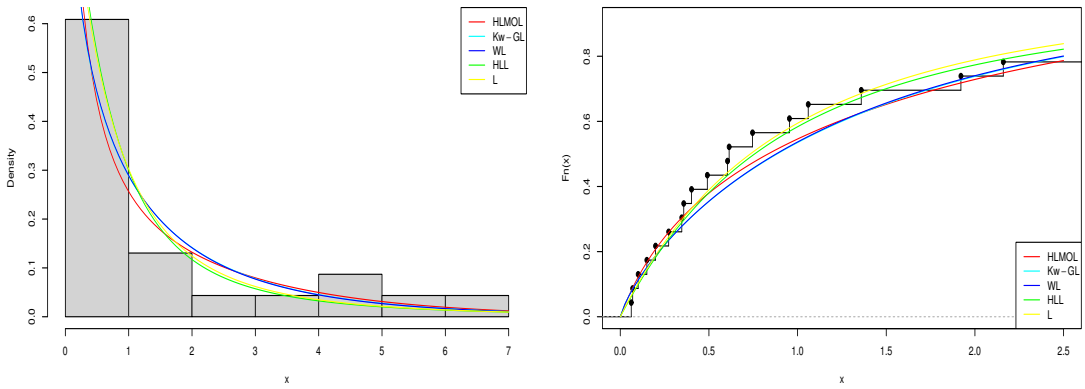


Figure 4: Estimated PDF and CDF for the HLMOL, Kw-GL, WL, HLL and L distributions for secondary reactor pumps data set

2.830, 2.870, 3.020, 3.250, 3.310, 3.360, 3.360, 3.480, 3.520, 3.570, 3.640, 3.700, 3.820, 3.880, 4.180, 4.230, 4.260, 4.330, 4.340, 4.400, 4.500, 4.510, 4.870, 4.980, 5.060, 5.090, 5.170, 5.320, 5.320, 5.340, 5.410, 5.410, 5.490,5.620, 5.710, 5.850, 6.250, 6.540, 6.760, 6.930, 6.940, 6.970, 7.090, 7.260, 7.280, 7.320, 7.390, 7.590, 7.620, 7.630, 7.660, 7.870, 7.930, 8.260, 8.370, 8.530, 8.650,8.660, 9.020, 9.220, 9.470, 9.740, 10.06, 10.34, 10.66, 10.75, 11.25, 11.64, 11.79, 11.98, 12.02, 12.03, 12.07, 12.63, 13.11, 13.29, 13.80, 14.24, 14.76, 14.77, 14.83, 15.96, 16.62, 17.12, 17.14, 17.36, 18.10, 19.13, 20.28, 21.73, 22.69, 3.63, 25.74, 25.82, 26.31, 32.15, 34.26, 36.66, 43.01, 46.12, 79.05}

The necessary numerical summaries for the five fits using the bladder cancer patients data set includes $\hat{\ell}$, AIC, BIC and K-S with p-V are provided in Tables 9 and 10. Additionally, Table 11 provides two LR statistics based on data set bladder cancer patients along with p-V.

Figure 5 display the TTT-plot for the bladder cancer patients data set, and Figure 6 displays the graphs of estimated PDF and CDF of the considered distributions for bladder cancer patients data sets.

In Tables 6, 7, 9 and 10, the MLEs of the parameters for the fitted distributions along with -log-likelihood, AIC, BIC, K-S with p-V values are given for two distinct data sets. The HLMOL distribution proves to be a superior model than the Kw-GL, WLo, HLL, and L models because it has the lowest values of AIC, BIC, K-S, and the highest p-V of the K-S statistic. Tables 8 and 11 also show the LR statistic values and p-V. In light of these results, we reject the null hypothesis for the aforementioned data sets and come to the conclusion that the HLMOL distribution offers a much more accurate depiction than the L and HLL distributions.

Figures 3 and 5 indicates decreasing HRF for he secondary reactor pumps data set and upside-down bathtub shaped HRF for the bladder cancer patients data set. Therefore, the HLMOL distribution can fit these data sets.

Figures 4 and 6 present a diagrammatic comparison of the closeness of the fitted

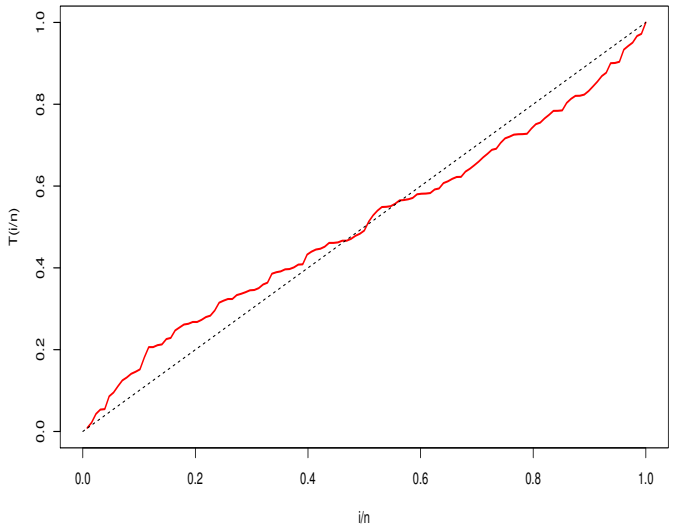


Figure 5: TTT-plot for the bladder cancer patients data set

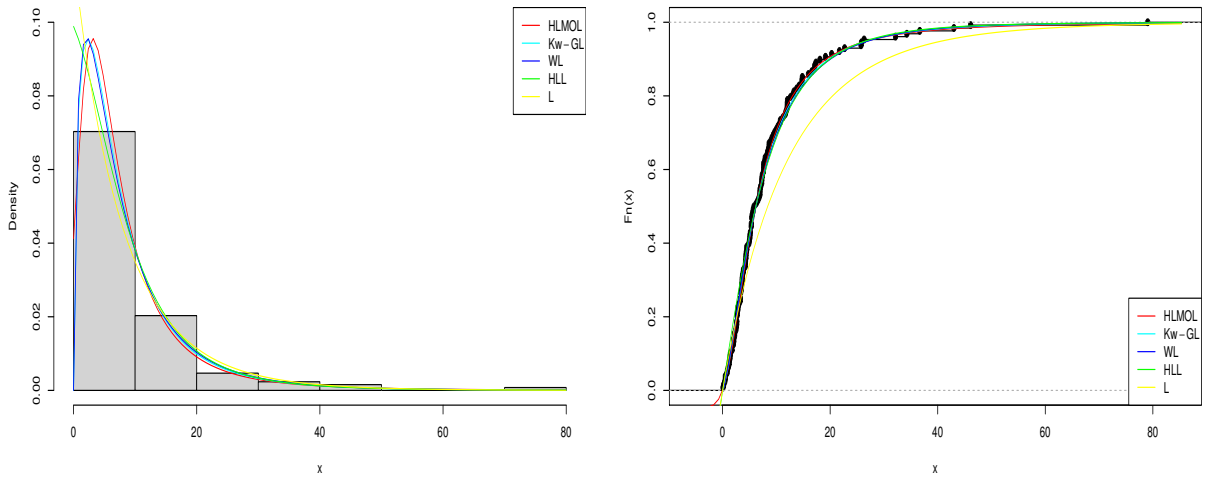


Figure 6: Estimated PDF and CDF for the HLMOL, Kw-GL, WL, HLL and L distributions for bladder cancer patients data sets

Table 9: Estimated values, log-likelihood, AIC and BIC bladder cancer patients data set

Distribution	Estimates	$-ln(L)$	AIC	BIC
HLMOL	$\hat{\boldsymbol{\lambda}} = \mathbf{0.3401}$ $\hat{\boldsymbol{\alpha}} = \mathbf{8.6402}$ $\hat{\boldsymbol{c}} = \mathbf{4.0693}$ $\hat{\boldsymbol{\theta}} = \mathbf{8.7546}$	**406.579**	**821.158**	**832.5661**
Kw-GL	$\hat{a} = 1.5493$ $\hat{b} = 10.3464$ $\hat{\lambda} = 11.5419$ $\hat{\alpha} = 0.4372$	407.3357	822.6713	834.0794
WL	$\hat{a} = 16.3314$ $\hat{b} = 1.5541$ $\hat{\beta} = 5.3873$ $\hat{\alpha} = 0.1607$	407.611	823.222	834.6301
HLL	$\hat{\lambda}= 0.5540$ $\hat{\alpha} = 0.4941$ $\hat{\theta} = 26.6014$	409.4457	824.8915	833.4476
L	$\hat{\alpha} = 13.0380$ $\hat{\theta}$=110.7043	411.5897	827.1794	832.8835

Table 10: K-S with p-V for the secondary reactor pumps data set

Distributions	K-S	p-V
HLMOL	**0.0286**	**0.9999**
Kw-GL	0.0404,	0.985
WL	0.0449	0.9587
HLL	0.0808	0.3738
L	0.1006	0.1498

densities with the observed histogram and CDFs with the empirical CDFs of the data sets. These diagrams demonstrate that the proposed distribution renders a closer fit the above two data sets.

9. Conclusion

In this article, the T-X method was utilized to introduce the T-Marshall Olkin X family of distribution, a novel family of distributions. HLMO-X and one of its members, HLMOL, are investigated in depth as a particular case. The quantile function, moments, incomplete moments, moment generating function, Lorenz curve, Bonferroni curve, skewness, kurtosis, order statistics, and asymptotic distributions of order statistics are some of the structural characteristics are investigated. The maximum likelihood approach, together with simulation analysis, is the technique utilized to estimate the model parameters. The distribution fit between HLL and HLMOL and also between L and HLMOL is tested using the LR test with simulation research. The outcome demonstrates that the HLMOL dis-

Table 11: The values of LR statistic for different hypothesis and data sets

Models	Hypothesis	Bladder cancer patients data set		
		LR	df	p-V
HLMOL vs. L	$H_0 : \lambda = 1, c = 0.5$ vs. $H_1 :$ H_0 is false	709.4762	2	< 0.001
HLMOL vs. HLL	$H_0 : c = 1$ vs. $H_1 :$ H_0 is false	24.6404	1	< 0.001

tribution is superior to the other two. When compared to the Kw-GL, WL, GL, and EL distributions, fitting to two real-world data produce good results in favour of the suggested distribution. As a result, the proposed distribution can be viewed as making a worthwhile contribution to the existing knowledge. Future research will include more generalizations that can be made for both continuous and discrete cases. One such generalization is the exponential-Marshall-Olkin X family of distributions. For evaluating the accuracy of the new models, different inferential investigations will be taken into consideration.

Conflict of interest

The authors do not have any financial or non-financial conflict of interest to declare for the research work included in this article.

References

Aako, O. L., Adewara, J., and Nkemnole, E. (2022). Marshall-Olkin generalized inverse log-logistic distribution: Its properties and applications. *International Journal of Mathematical Sciences and Optimization: Theory and Applications*, **8**, 79–93.

Akinsete, A., Famoye, F., and Lee, C. (2014). The Kumaraswamy-geometric distribution. *Journal of Statistical Distributions and Applications*, **1**, 1–21.

Aldahlan, M. A., Rabie, A. M., Abdelhamid, M., Ahmed, A. H. N., and Afify, A. Z. (2023). The marshall–Olkin Pareto type-I distribution: Properties, inference under complete and censored samples with application to breast cancer data. *Pakistan Journal of Statistics and Operation Research*, **1**, 603–622.

Alizadeh, M., Cordeiro, G. M., Brito, E. d., and B. Demétrio, C. G. (2015a). The beta Marshall-Olkin family of distributions. *Journal of Statistical Distributions and Applications*, **2**, 1–18.

Alizadeh, M., Tahir, M., Cordeiro, G. M., Mansoor, M., Zubair, M., and Hamedani, G. (2015b). The Kumaraswamy Marshal-Olkin family of distributions. *Journal of the Egyptian Mathematical Society*, **23**, 546–557.

Almheidat, M., Famoye, F., and Lee, C. (2015). *Some Generalized Families of Weibull Distribution: Properties and Applications*. PhD thesis, Central Michigan University.

Alsadat, N., Nagarjuna, V. B., Hassan, A. S., Elgarhy, M., Ahmad, H., and Almetwally, E. M. (2023). Marshall–Olkin Weibull–Burr XII distribution with application to physics data. *AIP Advances*, **13**.

Alsultan, R. (2023). The Marshall-Olkin pranav distribution: theory and applications. *Pakistan Journal of Statistics and Operation Research*, **1**, 155–166.

Alzaatreh, A., Lee, C., and Famoye, F. (2013). A new method for generating families of continuous distributions. *Metron*, **71**, 63–79.

Anwar, M. and Zahoor, J. (2018). The half-logistic Lomax distribution for lifetime modeling. *Journal of Probability and Statistics*, **2018**, 1–12.

Arnold, B. C., Balakrishnan, N., and Nagaraja, H. N. (2008). *A First Course in Order Statistics*. SIAM.

Atchadé, M. N., N'bouké, M. A., Djibril, A. M., Al Mutairi, A., Mustafa, M. S., Hussam, E., Alsuhabi, H., and Nassr, S. G. (2024). A new Topp-Leone Kumaraswamy Marshall-Olkin generated family of distributions with applications. *Heliyon*, **10**.

Basheer, A. M. (2022). Marshall–Olkin alpha power inverse exponential distribution: properties and applications. *Annals of Data Science*, **9**, 301–313.

Bebbington, M., Lai, C.-D., and Zitikis, R. (2007). A flexible Weibull extension. *Reliability Engineering & System Safety*, **92**, 719–726.

Chesneau, C. and Djibrila, S. (2019). The generalized odd inverted exponential-G family of distributions: properties and applications. *Eurasian Bulletin of Mathematics*, **2**, 86–110.

Chipepa, F., Moakofi, T., and Oluyede, B. (2022). The Marshall-Olkin-odd power generalized Weibull-G family of distributions with applications of COVID-19 data. *Journal of Probability and Statistical Science*, **20**, 1–20.

Cordeiro, G. M. and de Castro, M. (2011). A new family of generalized distributions. *Journal of Statistical Computation and Simulation*, **81**, 883–898.

Eliwa, M., El-Morshedy, M., and Ali, S. (2021). Exponentiated odd Chen-G family of distributions: statistical properties, Bayesian and non-Bayesian estimation with applications. *Journal of Applied Statistics*, **48**, 1948–1974.

Eugene, N., Lee, C., and Famoye, F. (2002). Beta-normal distribution and its applications. *Communications in Statistics-Theory and methods*, **31**, 497–512.

Ghitany, M., Al-Awadhi, F., and Alkhalfan, L. (2007). Marshall–Olkin extended Lomax distribution and its application to censored data. *Communications in Statistics—Theory and Methods*, **36**, 1855–1866.

Ghitany, M., Al-Mutairi, D., Al-Awadhi, F., and Al-Burais, M. (2012). Marshall-Olkin extended Lindley distribution and its application. *International Journal of Applied Mathematics*, **25**, 709–721.

Gillariose, J., Tomy, L., Jamal, F., and Chesneau, C. (2020). The Marshall-Olkin modified lindley distribution: properties and applications. *Journal of Reliability and Statistical Studies*, **1**, 177–198.

Gradshteyn, I. S. and Ryzhik, I. M. (2014). *Table of Integrals, Series, and Products*. Academic press.

Handique, L. and Chakraborty, S. (2015a). The generalized Marshall-Olkin-Kumaraswamy-G family of distributions. *arXiv preprint arXiv:1510.08401*, **1**.

Handique, L. and Chakraborty, S. (2015b). The Marshall-Olkin-Kumarswamy-G family of distributions. *arXiv preprint arXiv:1509.08108*, **1**.

Innocent, O. O., Femi, A. J., Nuga, O. A., and Adebisi, O. A. (2023). Marshall-Olkin extended generalized exponential distribution: Properties, inference and application to traffic data. *International Journal of Statistics and Probability*, **12**.

İrhad, M. R., Ahammed, E. S. M., Radhakumari, M., and Amer, A.-O. (2024). Marshall-Olkin Bilal distribution with associated minification process and acceptance sampling plans. *Hacettepe Journal of Mathematics and Statistics*, **53**, 1–29.

Ishaq, A. I., Usman, A., Usman, A. A., and Tasi'u, M. (2019). Weibull burr x-generalized family of distributions. *Nigerian Journal of Scientific Research*, **18**, 269–283.

Jamal, F., Handique, L., Ahmed, A. H. N., Khan, S., Shafiq, S., and Marzouk, W. (2022). The generalized odd linear exponential family of distributions with applications to reliability theory. *Mathematical and Computational Applications*, **27**, 55.

Klakattawi, H. S., Khormi, A. A., Baharith, L. A., et al. (2023). The new generalized exponentiated Fréchet–Weibull distribution: Properties, applications, and regression model. *Complexity*, **2023**.

Krishna, E., Jose, K., Alice, T., and Ristić, M. M. (2013). The Marshall-Olkin Fréchet distribution. *Communications in Statistics-Theory and Methods*, **42**, 4091–4107.

Marshall, A. W. and Olkin, I. (1997). A new method for adding a parameter to a family of distributions with application to the exponential and Weibull families. *Biometrika*, **84**, 641–652.

Merovci, F., Elbatal, I., and Puka, L. (2015). The McDonald quasi Lindley distribution and its applications. *Acta Universitatis Apulensis*, **45**, 87–105.

Mi, Z., Hussain, S., and Chesneau, C. (2021). On a special weighted version of the odd weibull-generated class of distributions. *Mathematical and Computational Applications*, **26**, 62.

Moakofi, T., Oluyede, B., and Makubate, B. (2021). Marshall-Olkin Lindley-log-logistic distribution: Model, properties and applications. *Mathematica Slovaca*, **71**, 1269–1290.

Moolath, G. B. and Jayakumar, K. (2017). T-transmuted x family of distributions. *Statistica*, **77**, 251–276.

Nasiru, S. and Abubakari, A. G. (2022). Marshall-Olkin Zubair-G family of distributions. *Pakistan Journal of Statistics and Operation Research*, **1**, 195–210.

Niyoyunguruza, A., Odongo, L. O., Nyarige, E., Habineza, A., and Muse, A. H. (2023). Marshall-Olkin exponentiated Fréchet distribution. *Journal of Data Analysis and Information Processing*, **11**, 262–292.

Nwezza, E. E. and Ugwuowo, F. I. (2020). The Marshall-Olkin Gumbel-Lomax distribution: properties and applications. *Heliyon*, **6**.

Obulezi, O. J., Anabike, I. C., Oyo, O. G., Igbokwe, C., and Etaga, H. (2023). Marshall-Olkin Chris-Jerry distribution and its applications. *International Journal of Innovative Science and Research Technology*, **8**, 522–533.

Oluyede, B. and Gabanakgosi, M. (2023). The type II exponentiated half logistic-Marshall-Olkin-G family of distributions with applications. *Colombian Journal of Statistics/Revista Colombiana de Estadística*, **46**.

Opone, F. C., Akata, I. U., and Altun, E. (2022). The Marshall-Olkin extended unit-Gompertz distribution: its properties, simulations and applications. *Statistica*, **82**, 97–118.

Osi, A., Doguwa, S., Yahaya, A., Zakari, Y., and Usman, A. (2024a). Marshall-Olkin cosine Topp-Leone family of distributions with application to real-life datasets. *Preprint*, **1**.

Osi, A. A., Doguwa, S. I., Abubakar, Y., Zakari, Y., and Abubakar, U. (2024b). Development of exponentiated cosine Topp-Leone generalized family of distributions and its applications to lifetime data. *UMYU Scientifica*, **3**, 157–167.

Pogány, T. K., Saboor, A., and Provost, S. (2015). The Marshall–Olkin exponential Weibull distribution. *Hacettepe Journal of Mathematics and Statistics*, **44**, 1579–1594.

Rasekhi, M., Alizadeh, M., and Hamedani, G. G. (2018). The Kumaraswamy Weibull geometric distribution with applications. *Pakistan Journal of Statistics and Operation Research*, **1**, 347–366.

Ristić, M. M. and Kundu, D. (2015). Marshall-Olkin generalized exponential distribution. *Metron*, **73**, 317–333.

Sadiq, I. A., Doguwa, S., Yahaya, A., and Usman, A. (2023). Development of new generalized odd fréchet-exponentiated-g family of distribution. *UMYU Scientifica*, **2**, 169–178.

Sengweni, W., Oluyede, B., and Makubate, B. (2023). The marshall-Olkin Topp-Leone half-logistic-G family of distributions with applications. *Statistics, Optimization & Information Computing*, **11**, 1001–1026.

Shah, Z., Khan, D. M., Khan, Z., Faiz, N., Hussain, S., Anwar, A., Ahmad, T., and Kim, K.-I. (2023). A new generalized logarithmic–X family of distributions with biomedical data analysis. *Applied Sciences*, **13**, 3668.

Shams, T. M. (2013). The Kumaraswamy-generalized Lomax distribution. *Middle-East Journal of Scientific Research*, **17**, 641–646.

Sherwani, R. A. K., Ashraf, S., Abbas, S., and Aslam, M. (2023). Marshall Olkin exponentiated Dagum distribution: Properties and applications. *Journal of Statistical Theory and Applications*, **22**, 70–97.

Suprawhardana, M. S. and Prayoto, S. (1999). Total time on test plot analysis for mechanical components of the rsg-gas reactor. *Atom Indones*, **25**, 81–90.

Tahir, M. H., Cordeiro, G. M., Mansoor, M., and Zubair, M. (2015). The Weibull-Lomax distribution: properties and applications. *Hacettepe Journal of Mathematics and Statistics*, **44**, 455–474.

Tomy, L. and Gillariose, J. (2018). The Marshall-Olkin ikum distribution. *Biometrics and Biostatistics International Journal*, **7**, 00186.

Tomy, L., Jose, M., and Jose, M. (2019). The TX family of distributions: a retrospect. *Think India Journal*, **22**, 9407–9420.

Willayat, F., Saud, N., Ijaz, M., Silvianita, A., and El-Morshedy, M. (2022). Marshall–Olkin extended Gumbel type-II distribution: properties and applications. *Complexity*, 1–23.

Yahaya, A. and Doguwa, S. (2022). On Rayleigh-exponentiated odd generalized-Pareto distribution with its applications. *Benin Journal of Statistics*, **5**, 89–107.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 23, No. 1, 2025 (New Series), pp 123–137
https://www.ssca.org.in/journal

Competing Risks Models with Multistate and Intermediate States

A. M. Rangoli, A. S. Talawar and R. P. Agadi
Department of Statistics, Karnatak University, Dharwad, India.

Received: 15 July 2023; Revised: 07 December 2023; Accepted: 25 April 2024

Abstract

In the present paper, we are considering competing risks with multistate and intermediate state models. For competing risks analysis, we are considering prostate cancer data for multistate model and simulated data for intermediate state model. We can see from the prostate cancer data, that in the presence of competing risks 1-Kaplan Meier has more cumulative incidence as compared to the Aalen-Johnson estimator. Hence in comparison to Aalen-Johnson, Kaplan-Meier overestimates the cumulative incidence for multistate model. We have generated competing risks data for the intermediate states model via simulation using the case of HIV/AIDS. The simulated data on competing risks in the presence of intermediate states contain 4 transient states and 3 absorbing states. For simulation purpose we have used Weibull, Binomial and Uniform distributions. Simulated data on HIV/AIDS explain the behavior of competing risks model with intermediate states.

Key words: Competing risks; Intermediate state; Kaplan-Meier; Aalen-Johnson estimator; Simulation.

1. Introduction

Competing risk models are survival statistical models that take a wide variety of failure causes into account. Competing risk models analyze the time until some first event occurs at that time. Standard survival analysis often simply takes into account the amount of time until a first occurrence. Individuals in the study are followed from a common initial point until the occurrence of the first event, and such occurrences preclude the observation of the rest of the events and are called competing risks Llopis-Cardona *et al.* (2021). The estimation of competing risk data is more prominent in everyday life. The competing risks generalize survival function from a single combined endpoint to multiple first event types.

Many authors have worked on the multistate and intermediate state models in competing risk settings. Chiang (1968) modeled competing risks with multistate and intermediate states and also illness death models. Andersen *et al.* (2002) considered competing risk models as a special case of the multistate model. They have used the Fine and Gray

Corresponding Author: A. M. Rangoli
Email: ajmr1008@gmail.com

model to estimate the parameters of the covariates. Putter *et al.* (2007) modeled multistate with intermediate states for a breast cancer dataset. Schmoor *et al.* (2013) developed a model of competing risk with multistate models and intermediate state models with MGUS (monoclonal gammopathy of undetermined significance) data inbuilt in R software. Andersen and Keiding (2012) considered competing risks, multistate, and illness death models to illustrate the transition hazard. Schmoor *et al.* (2013) considered multistate models for the hematopoietic stem cell transplantation (HSCT) dataset with different scenarios. Jepsen *et al.* (2015) considered multistate models for disease progression, and the other authors contributed towards the multistate models (Meira-Machado *et al.*, 2009; Gasperoni *et al.*, 2017; Manevski *et al.*, 2022). For the analysis purpose, they have used real data from the patients with cirrhosis disease. Moreno-Betancur *et al.* (2017) considered a multistate model for studying disease-related mortality.

In our present paper, we are considering competing risks in the multistate model and the intermediate state model. Initially, every individual is in state 0 (that is, initial state or alive) at time of origin. An individual stays in the state until the occurrence of the first event. Generally, there is one event of interest modeled by the transition into state 1, and all other first event types are considered competing events. Figure 1 and 2 show the transfer diagram of the multistate and intermediate models respectively.

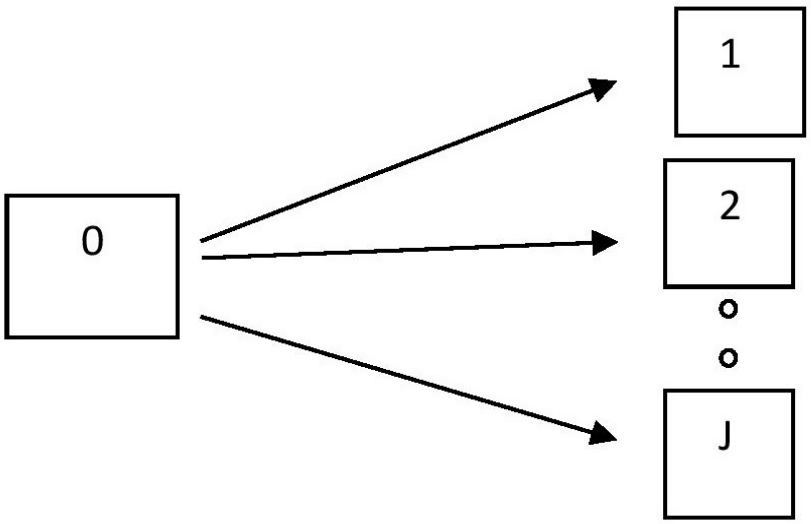


Figure 1: Competing Risks with multistate model.

Figure 1 is the competing risks with multistate model with 8 states and are given by, 0→ Alive, 1→ Cancer, 2→ cardio vascular disease (CVD), 3→ Cerebrovascular accident, 4→ Pulmonary embolus, 5→ Other cancers, 6→ Respiratory disease, 7→Unknown cause. Here state 0 is transient where as J= 1, 2, 3, 4, 5, 6 and 7 are absorbing states.

Figure 2 is transition diagram of the competing risks with intermediate states model and the states are given by, 0→ Human immunodeficiency virus (HIV) Infected, 1→ Acute HIV Infection, 2→ Chronic HIV Infection, 3→ Acquired immune deficiency syndrome (AIDS), 4→ Death due to Acute HIV Infection, 5→ Death due to Chronic HIV Infection, 6→ Death due to AIDS. Here states 0, 1, 2 and 3 are transient where as 4, 5 and 6 are absorbing states.

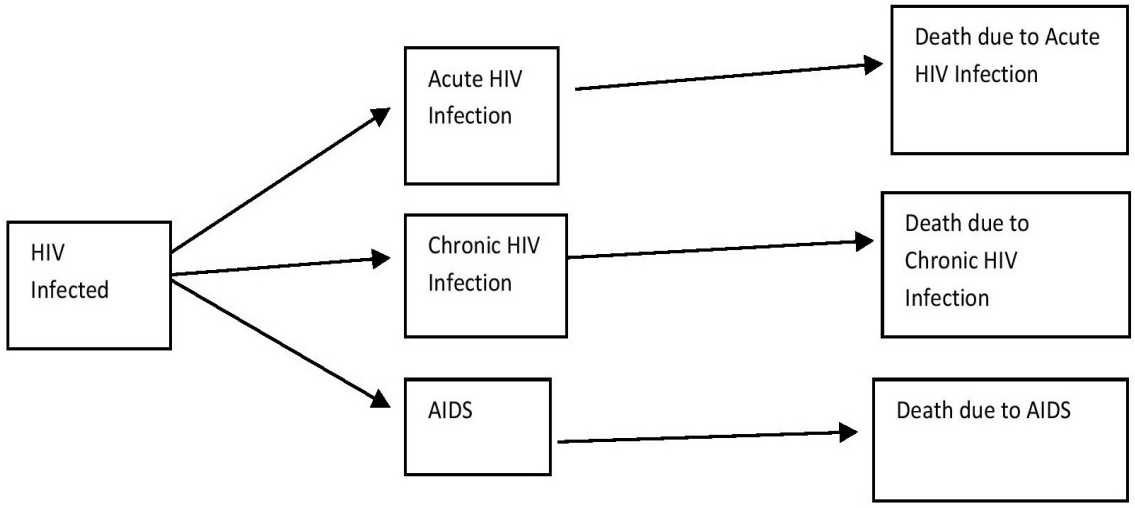


Figure 2: Competing risks with intermediate states.

A multistate model is used to model a process where subjects transition takes place from one state to the next. In the Figure, we can say each box is a state and each arrow is a possible transition. In Figure 1, we can see that there is one transient state and seven absorbing states, and reverse transitions are not allowed. In Figure 2, there are four transient states and three absorbing states.

2. Methodology

For our study, we first consider the secondary dataset on prostate cancer available in Andrews and Herzberg (2012) for modeling competing risks with multistate, as shown in Figure 1. Next, we have generated competing risk data via simulation for a competing risk model with intermediate states (Beyersmann *et al.*, 2011), as shown in Figure 2. For simulation, we assume that the failure time follows the Weibull distribution.

The process is Markovian in nature because present state depends upon immediate previous state but not older. In the competing risks process, $X_t\ ,\ t \geq 0$ denotes the state.

Every individual starts in the initial state 0 at time origin 0

$$p\left(X_0 = 0\right) = 1$$

An individual stays in state 0 $(i, e,\ \ X_t = 0)$ as long as neither competing event happened.

Generally everyone will leave the initial state at some point in time

$$p\left(T \in (0, \infty)\right) = 1$$

where T→ survival time or failure time.

Let T_{0j} be the time from the start point 0 to event j, $j = 1,\ 2, \ldots, J$. Consider $T = min(T_{01}, T_{02}, .., T_{0j})$ the time to the occurrence of the first event.

Observation of the competing risks process $\{X_t,\ t \geq 0\}$ will in general be subject to right censoring (Geskus, 2015). Let C be the right censoring time, the observed data is given as $(\min(T, C),\ 1(T \leq C).X_T)$. The status indicator $1(T \leq C).X_T \in \{1, 2, \ldots, 7\}$ equal 0 if observation was censored.

Let the hazard function be

$$\lambda(t) = \lim_{h\to 0} \tfrac{1}{h} P(t \leq T < t + h/T \geq t) = \tfrac{f(t)}{S(t)}$$

where $f(t)$ be density function and $S(t)$ be survival function,

and the survival function in terms of hazard function can be expressed as,

$$S(t) = e^{-\int_0^t \lambda(u)du}$$

Now the cause specific hazard function be given by,

$$\lambda_j(t) = \lim_{h\to 0} \tfrac{1}{h} p(t \leq T < t + h, J = j\ /T \geq t)$$

$$\lambda(t) = \textstyle\sum_{j=1}^{J} \lambda_j(t)$$

and the cumulative cause specific hazard function be given by

$$\Lambda(t) = \textstyle\int_0^t \lambda(u)du$$

$$\Lambda_j(t) = \textstyle\sum_{k=1}^{K} \frac{Number\ of\ individuals\ observed\ to\ fail\ due\ to\ cause\ j\ at\ t_k}{Number\ of\ individuals\ at\ risk\ just\ prior\ to\ t_k}$$

Now $X_t,\ t \geq 0$ denote the state at time t (Beyersmann *et al.*, 2011). The hazard function given by,

$$\lambda_{0j}(t) = \lim_{h\to 0} \frac{1}{h} p(t \leq T < t + h, J = j\ /T \geq t) \quad j = 1, 2, \ldots, J \tag{1}$$

which means an individual is in state 0 at time t and in small interval h it reaches to state j.

From Figure 2 we can see the intermediate transition and the hazard function are given by,

$$\lambda_{0k} = \lim_{h\to 0} \tfrac{p(t \leq T_{0k} < t+h/T_{0k} > t)}{h}\ ,\ k = 1, 2, 3$$

$$\lambda_{14} = \lim_{h\to 0} \tfrac{p(t-t_{01} \leq T_{14} < t-t_{01}+h/T_{14} > t-t_{01})}{h}$$

$$\lambda_{25} = \lim_{h\to 0} \tfrac{p(t-t_{02} \leq T_{25} < t-t_{02}+h/T_{25} > t-t_{02})}{h}$$

$$\lambda_{36} = \lim_{h\to 0} \tfrac{p(t-t_{03} \leq T_{36} < t-t_{03}+h/T_{36} > t-t_{03})}{h}$$

In many real problems (e.g., those associated with relapse of cancer diseases) the behavior of T_{14} often depends on the characteristic of transition 0 to 1.

Now the overall hazard is given by,

$\lambda(t) = \sum_{j=1}^{J} \lambda_{0j}(t)$

and cumulative hazard function given by,

$$\Lambda_{0j} = \int_0^t \lambda_{0j}(u)du \tag{2}$$

and cause specific Nelson-Aalen estimator of the cumulative hazard is given by,

$$\Lambda_{0j}(t) = \sum_{k=1}^{K} \frac{Number\ of\ individuals\ observed\ type\ j\ event\ at\ t_k}{Number\ of\ individuals\ at\ risk\ \ just\ prior\ to\ t_k}, j = 1, 2, .., J \tag{3}$$

Now for Intermediate multistate models (Geskus, 2015), the transition probability function is given by,

$$p_{ij}(t) = p(X(t + dt) = j/X(t) = i),\ i, j = 0, 1, 2, \ldots J, \quad i \neq j$$

and cumulative hazard function given by,

$$\Lambda_{ij}(t) = \sum_{k=1}^{K} \frac{Number\ of\ individuals\ observed\ i \to j\ transition\ at\ t_k}{Number\ of\ individuals\ at\ risk\ in\ state\ i\ just\ prior\ to\ t_k}, i, j = 0, 1, 2, \ldots J,\ \ i \neq j \tag{4}$$

Chapman-Kolmogorov equation is used to find out the transition probabilities from state i to state j over m steps, which is denoted by P_{ij}^m. It is given by the probability of a chain moving from state i to state j in exactly m steps, where $m \geq 2$ is given by,

$$P_{ij}^m = p(X_{m+n} = j/X_n = i) \tag{5}$$

P_{ij}^m gives the probability that from state i at n^{th} trail and the state j is reached at $(m+n)^{th}$ trial in m steps.

In our model, from Figure 2 we can see that initially an individual in state 0 will reaches to state j, $j = 4,\ 5,\ 6$ in two steps that is first it reaches to state i, $\ i = 1,\ 2,\ 3$ then they move towards state j respectively, that is we can say two step transition.

2.1. The empirical transition matrix

Let $N_{ij}(t)$ be the number of observed direct transitions from state i to state j up to time t. $Y_i(t)$ be the number of individuals under observation in state i just before time t (Geskus, 2015). The Non-diagonal entries of the matrix of cumulative transition hazards $\Lambda(t)$ may be estimated by the Nelson-Aalen estimator.

$$\widehat{\Lambda}_{ij}(t) = \int_0^t \frac{N_{ij}(t)}{Y_i(t)} \tag{6}$$

The transition probabilities are conditional probabilities $p_{ij}(s, t)$ defines as, the probability of being at state j at time t given that the individual was in state i at s and is given by,

$p_{ij}(s,t) = p(X(t) = j/X(s) = i) \quad , s \leq t, \quad i,j = 0,1,2,\ldots J$

Or

$$\widehat{p}(s,t) = \prod_{(s,t]} \left(I + d\widehat{\Lambda}(u)\right) \tag{7}$$

As $\widehat{\Lambda}(u)$ is a matrix of step functions with a finite number of jumps on $(s,t]$, the product integral can be written as a finite matrix product.

$$\widehat{p}(s,t) = \prod_{s<t_k<t} \left(I + d\widehat{\Lambda}\left(t_k\right)\right) \tag{8}$$

where the product is taken over all observed transition times in $(s,t]$. Note that the non-diagonal elements $(i,\ j)$, i$\neq$ j of $I + d\widehat{\Lambda}\left(t_k\right)$ are the number of observed direct $i \rightarrow j$ transitions, divided by the number of individuals under observation in state i just prior to t_k. The diagonal entries of $I + d\widehat{\Lambda}\left(t_k\right)$are such that each row equal to 1.

That is initially it is in state i at time s and reaches to state j at time t. That is next state is depends on current state not the previous states.

The survival function of the waiting time T in the initial state 0 is,

$$p(T > t) = \exp\left(-\int_0^t \lambda_0(u)du\right)$$

Cumulative incidence function (CIF) for cause j at t describes the probability of failing from cause j before time t and is given by

$$F_{0j}(t) = p\left(T \leq t,\ X_T = j\right) = \int_0^t p(T > u-)\ \lambda_{0j}(u)du \tag{9}$$

where $p(T > u-)$ is survival function just before u.

This probability is not proper distribution function because $F_{0j}(t)$ does not go to 1 when t goes to ∞.

Now Aalen–Johnson (AJ) estimator for cumulative incidence function (Beyersmann *et al.*, 2011) is given by,

$$\widehat{F_k^{AJ}}(t) = \sum_{t_{(i)} \leq t} \overline{F}^{PL}\left(t_{(i)}-\right)\ \widehat{\lambda_k}(t_{(i)}) \tag{10}$$

where $\overline{F}^{PL}\left(t_{(i)}-\right)$ is the Kaplan-Meier estimate of the overall cumulative incidence that combine all event types. The overall survival function $S(t)$ within the competing risk scenario is the transition probability that assesses the performance of the process in the initial state at time t, $S(t) = p_{00}(0,t)$ and the CIF for cause j, $F_{0j}(t)$ in the competing risk framework are the transition probabilities $p_{oj}(t)$.

3. Results and discussions

The analysis is carried out using the observed prostate cancer data for multistate model and the simulated data for intermediate state model.

3.1. Prostate cancer dataset

We have considered the analysis of competing risks with a multistate model using prostate cancer data. The data consists of 7 causes of failure, as shown in Figure 1, with 489 patients, and contains right censored observations. The data contains one transient state and seven absorbing states. The summary of the data is given in Table 1. Out of 489 patients, 126 take the transition from state 0 to state 1, which is 25.7% of the patients failure due to cancer with an average failure time of 26.4 months, and 91 take the transition from state 0 to state 2, which is 18.6% of the patients failure due to CVD with an average failure time of 24 months, and so on. And 150 patients who remain in the initial state are called censored; that is, 30.6% of the patients are censored. And we can see that the estimated probability of an individual being in the initial state is 0.2473704, and the transition probability from state 0 to state 1 is 0.2820819, and to state 2 is 0.1895011, and so on. Figure 3 explains the cumulative incidence curve given by Aalen-Johnson (red line) and the 1-Kaplan Meier curve (black line) for the prostate cancer data. We can see that in the presence of competing risks, 1-Kaplan Meier has more cumulative incidence as compared to the Aalen-Johnson estimator, which implies Kaplan-Meier overestimates cumulative incidence in the presence of competing risks (Talawar and Rangoli, 2023). Figure 4 explains the cumulative hazards given by Nelson-Aalen, and it can be seen that transitions from states 0 to 1 (cancer) and 0 to 2 (CVD) have a high cumulative hazard as compared to all transitions.

3.2. Simulation study

We generate competing risks data for the intermediate state model via simulation. The pattern of the competing risks in the presence of intermediate states is shown in Figure 2, which contains four transient states and three absorbing states. Table 2 contains the transition of subjects from one state to another. For simulation purposes, we have used the Weibull distribution with scale parameter 2 and shape parameter 0.5 (Sathian *et al.*, 2018; Okpala and Okoli, 2021).

Algorithm for generating competing risks data with intermediate states.

i. Generate event time T with some specified distribution (like exponential, Weibull).

ii. Run the binomial test for assigning the cause i,e generating failure cause j,

$$j = 1, 2, 3, \ldots, J.$$

iii. Generate censored observation C. In this we have used uniform distribution to generate censored observations.

iv. Now considering min (T, C), and assigning cause if we get T else it will be considered as censored.

v. Repeat the above steps for each intermediate states.

Using this algorithm and R-programming, we have simulated 1000 observations. The simulated sample data is presented in Table 2.

- Initially, the patient ID-1 moves from state 0 to state 1 at a time of 1.5 years, and the same patient will remain in state 1 only; that is, the patient becomes censored due to some reason.
- The patient ID-4 initially moves from state 0 to state 1 in 3 years, and the same patient moves from state 1 to state 4 at time 1.2 years. Therefore, patient ID-4 dies in 4.2 years after acquiring an HIV infection.
- Similarly, considering that patient ID-649 initially moves from state 0 to state 3 in 9.5 years of time and the same patient moves from state 3 to state 6 in 0.3 years, Therefore, patient ID-649 has a total failure time of 9.8 years after acquiring an HIV infection.

Table 3 explains simulated data for the number of patients who move from one state to another. Here we can see 340 patients move from state 0 to state 1 with an average time of 0.96 years, 271 patients move from state 0 to state 2 with an average time of 1.08 years, and 120 patients move from state 0 to 3 with an average time of 1.01 years. Out of 340 patients in state 1, 250 move to state 4 with an average failure time of 1.17 years; out of 271 patients in state 2, 208 move to state 5 with an average failure time of 1.17 years; and out of 120 patients in state 3, 81 move to state 6 with an average failure time of 1.42 years. Hence, an HIV-infected patient dies due to acute HIV infection, chronic HIV infection, and AIDS, with an average failure time of 2.13, 2.25, and 2.43, respectively. Table 4 explains the transition probability matrix, such that the probability of patients being in the initial state only is 0.09207136, the estimated transition probability for states 0 to 1 is 0.06163742, state 2 is 0.02921785, and state 3 is 0.03700813. And from state 0 to state 4 in two-step transitions, the probability is 0.3472133, and the probability of patients being in state 1 only is 0.05215, and so on.

From Figure 5, we can see the cumulative hazards of the intermediate states given by Nelson-Aalen. The plots in the first row are transitions from initial state 0, and the second row plots show intermediate transition states. The cumulative transition hazard for $0 \to 1$, $0 \to 2$, $0 \to 3$ were less as compared to the transition from $1 \to 4$, $2 \to 5$, $3 \to 6$ respectively. From this, we can see that the risk of dying from acute HIV infection, chronic HIV infection, and AIDS is higher. Hence, we can say that intermediate states play an important role in understanding failures from different causes. And also from Figure 5, we can see that after 8 years of time, the hazard becomes constant. Figure 6 gives the cumulative incidence for intermediate states. In this Figure, we can see that cumulative incidence is low for transitions from 0 to 3 and 3 to 6 compared to all other transitions. Figure 7 explains about the transition probability curve, initially the patient moves $0 \to 1$, $0 \to 2$, $0 \to 3$ and then these patients after reaching state 1, 2, 3 they move towards 4, 5, 6 respectively. From Figure 7, we can see that the first three plots initially have an increasing incidence, but soon after reaching approximately 1 year of time they become decreasing because when patients reaches states 1, 2, and 3, they again take a transition towards states 4, 5, and 6,

respectively. The states 1, 2, and 3 are transient states, and they do not reach a high value. But once the patients reach states 4, 5, and 6, the incidence increases because these states are absorbing states.

Table 1: Number of transitions from initial states for prostate cancer data

states	0	1	2	3	4	5	6	7
Causes	Healthy	Cancer	CVD	Cerebrovascular accident	Pulmonary embolus	Other cancer	Respiratory disease	Unknown cause
From 0 to	150	126	91	31	12	24	16	39
Transition Probability	0.24737	0.28208	0.18950	0.065040	0.02455	0.069655	0.03604	0.08575
Average Failure Time in Months		26.4	24	31.3	14.3	25.8	27.6	30.2

Table 2: Simulated sample data.

Id	from	To	Time
1	HIV infected	Acute HIV Infection	1.561531
1	Acute HIV Infection	Death due to Acute HIV Infection	1.442998
2	HIV infected	Acute HIV Infection	0.732673
2	Acute HIV Infection	Death due to Acute HIV Infection	0.054669
3	HIV infected	Chronic HIV Infection	0.183823
3	Chronic HIV Infection	Death due to Chronic HIV Infection	0.198274
17	HIV infected	AIDS	1.15567
17	AIDS	Death due to AIDS	0.541392
177	HIV infected	Chronic HIV Infection	6.824169
177	Chronic HIV Infection	Death due to Chronic HIV Infection	0.819118
207	HIV infected	Acute HIV Infection	0.221875
207	Acute HIV Infection	Death due to Acute HIV Infection	6.772362
300	HIV infected	Acute HIV Infection	0.015122
300	Acute HIV Infection	Censored	8.730437
606	HIV infected	Chronic HIV Infection	7.602475
606	ChronicHIV Infection	Censored	8.308631
649	HIV infected	AIDS	9.468228
649	AIDS	Death due to AIDS	0.297785

Table 3: Number of transition from different states for simulated data

states	0	1	2	3	4	5	6
0	269	340 (0.96)*	271 (1.08)*	120 (1.01)*	0	0	0
1	0	90	0	0	250 (1.17)*	0	0
2	0	0	63	0	0	208 (1.17)*	0
3	0	0	0	39	0	0	81 (1.42)*
4	0	0	0	0	0	0	0
5	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0

* values in parenthesis give the average failure time in years

Table 4: Transition probability matrix with time (0, 9.7) years for simulated data.

	0	1	2	3	4	5	6
0	0.09207136	0.06163742	0.02921785	0.03700813	0.3472133	0.3038753	0.1289767
1	0	0.05215555	0	0	0.9478445	0	0
2	0	0	0.01311025	0	0	0.9868898	0
3	0	0	0	0.04103397	0	0	0.958966
4	0	0	0	0	1	0	0
5	0	0	0	0	0	1	0
6	0	0	0	0	0	0	1

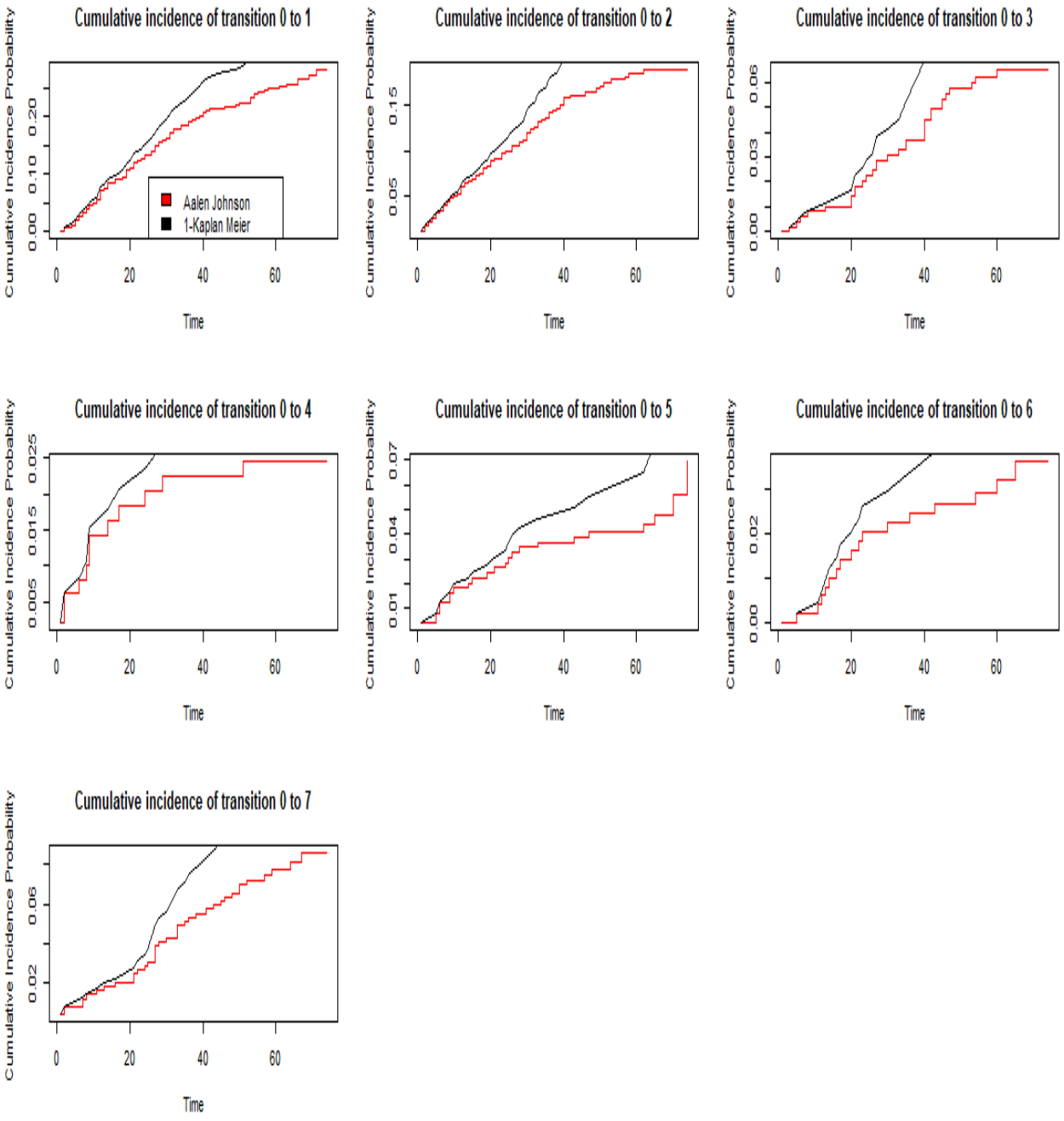


Figure 3: Cumulative incidence curve with 1-KM curve for all transition states for prostate cancer dataset.

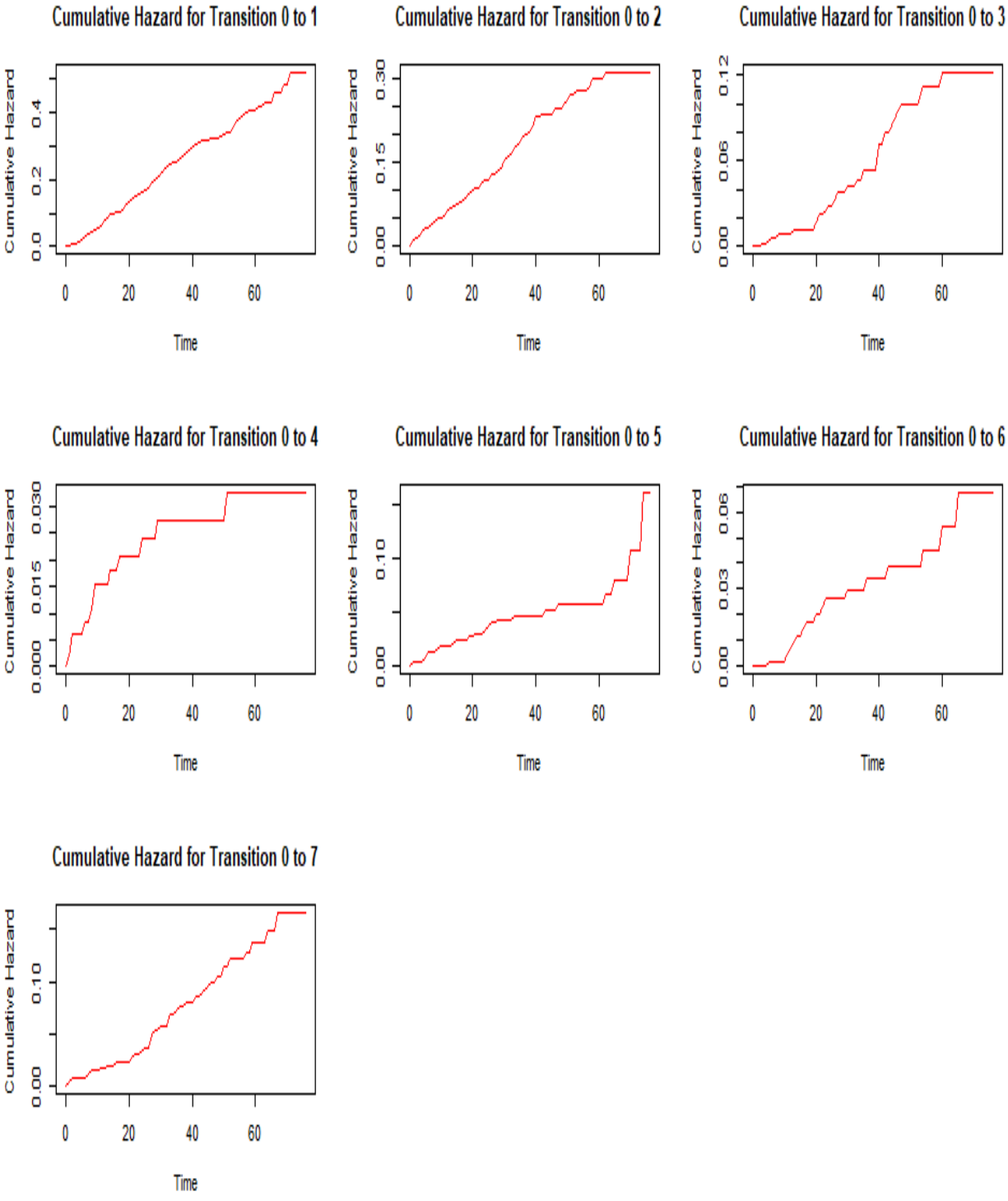


Figure 4: Cumulative hazard curve for all transition states for prostate cancer dataset.

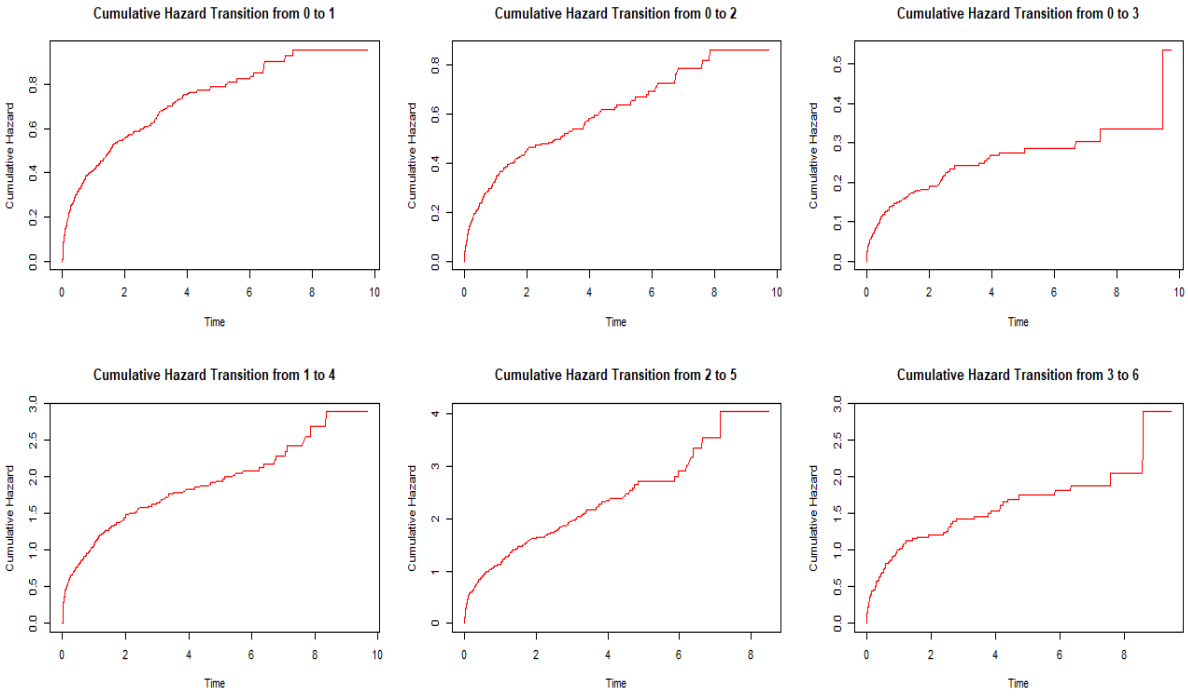


Figure 5: Nelson-Aalen cumulative hazard curve for simulated data.

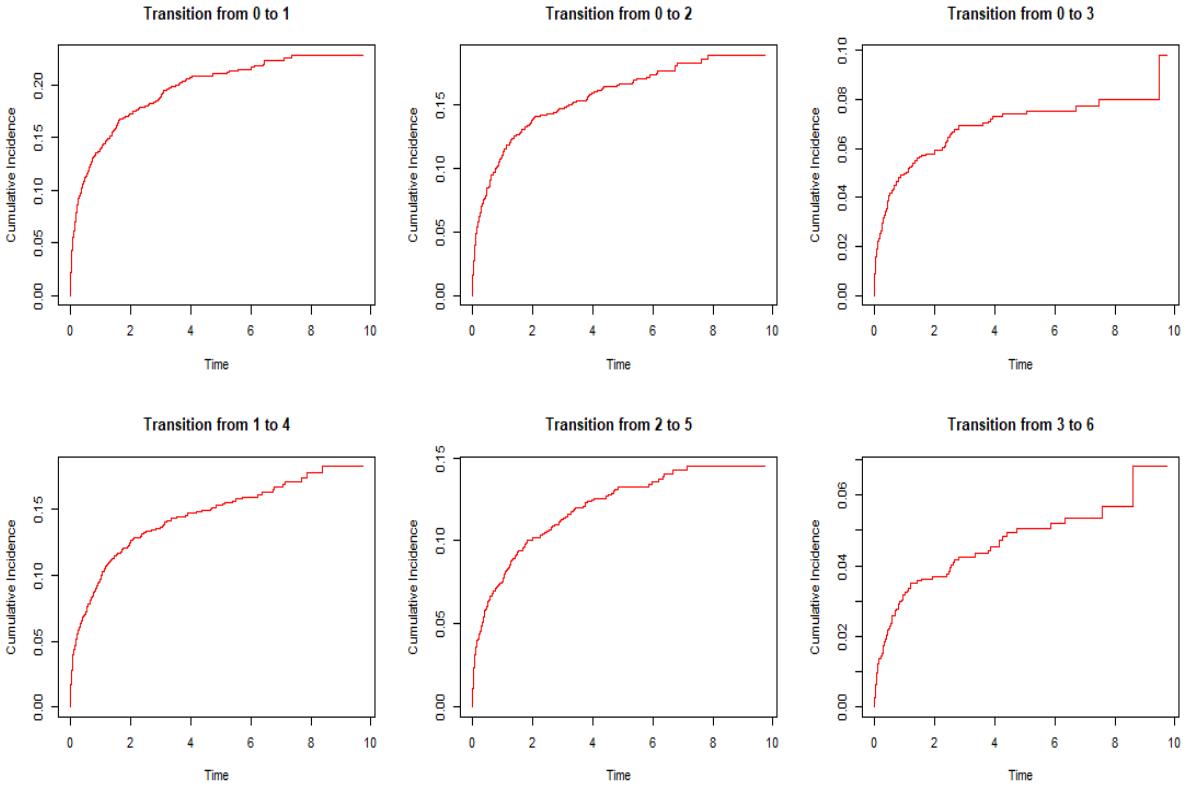


Figure 6: Aalen-Johnson cumulative incidence curve for simulated data.

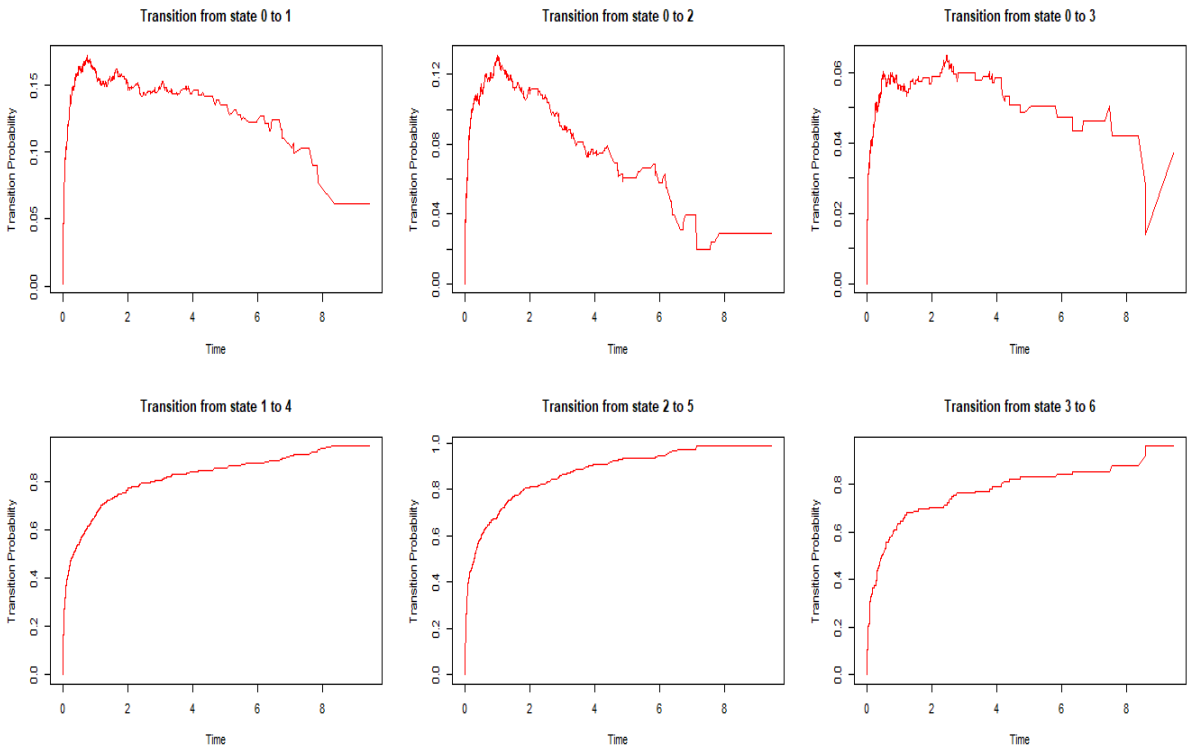


Figure 7: Transition probability curve for simulated data.

4. Conclusion

From the study we see behavior of the competing risks model for multistate and intermediate states and we conclude that, it is better to use intermediate states to clearly understand the reason of failure. The two sets of data such as prostate cancer data and simulated data respectively provide an information to model the competing risks with multistate and intermediate state models. In the competing risks model with intermediate states, we have seen the two step transition occurrence and transition probability curve initially increases and then decreases due to transient states and also we conclude that in the presence of competing risks with multistate model, the Kaplan-Meier overestimates the cumulative incidence.

Acknowledgements

The first author is thankful to Department of Science and Technology, innovation in science pursuit for inspired research (DST-INSPIRE) for financial support (Fellowship/2021/210203).

Conflict of interest

The authors do not have any financial or non-financial conflict of interest to declare for the research work included in this article.

References

Andersen, P. K., Abildstrom, S. Z., and Rosthøj, S. (2002). Competing risks as a multi-state model. *Statistical Methods in Medical Research*, **11**, 203–215.

Andersen, P. K. and Keiding, N. (2012). Interpretability and importance of functionals in competing risks and multistate models. *Statistics in Medicine*, **31**, 1074–1088.

Andrews, D. F. and Herzberg, A. M. (2012). *Data: A Collection of Problems from Many Fields for the Student and Research Worker*. Springer Science and Business Media.

Beyersmann, J., Allignol, A., and Schumacher, M. (2011). *Competing Risks and Multistate Models with R*. Springer Science and Business Media.

Chiang, C. L. (1968). *Introduction to Stochastic Processes in Biostatistics*. John Wiley and Sons.

Gasperoni, F., Ieva, F., Barbati, G., Scagnetto, A., Iorio, A., Sinagra, G., and Di Lenarda, A. (2017). Multi-state modelling of heart failure care path: a population-based investigation from Italy. *PloS one*, **12**, e0179176.

Geskus, R. B. (2015). *Data Analysis with Competing Risks and Intermediate States*. CRC Press.

Jepsen, P., Vilstrup, H., and Andersen, P. K. (2015). The clinical course of cirrhosis: the importance of multistate models and competing risks analysis. *Hepatology*, **62**, 292–302.

Llopis-Cardona, F., Armero, C., and Sanfélix-Gimeno, G. (2021). Reflection on modern methods: Competing risks versus multi-state models. *arXiv preprint arXiv:2104.03671*.

Manevski, D., Putter, H., Pohar Perme, M., Bonneville, E. F., Schetelig, J., and de Wreede, L. C. (2022). Integrating relative survival in multi-state models - a non-parametric approach. *Statistical Methods in Medical Research*, **31**, 997–1012.

Meira-Machado, L., de Uña-Álvarez, J., Cadarso-Suárez, C., and Andersen, P. K. (2009). Multi-state models for the analysis of time-to-event data. *Statistical Methods in Medical Research*, **18**, 195–222.

Moreno-Betancur, M., Sadaoui, H., Piffaretti, C., and Rey, G. (2017). Survival analysis with multiple causes of death. *Epidemiology*, **28**, 12–19.

Okpala, S. T. and Okoli, C. N. (2021). Comparison on performance of the lognormal, log logistic and weibull distribution on survival of hiv patients with opportunistic infections in Anambra State, Nigeria. *European Journal of Statistics and Probability*, **9**, 11–22.

Putter, H., Fiocco, M., and Geskus, R. B. (2007). Tutorial in biostatistics: competing risks and multi-state models. *Statistics in Medicine*, **26**, 2389–2430.

Sathian, B., Sreedharan, J., Asim, M., Menezes, R. G., van Teijlingen, E., and Unnikrishnan, B. (2018). Estimation of the burden of people living with human immunodeficiency virus/acquired immunodeficiency syndrome (hiv/aids) in Kerala state, India. *Nepal Journal of Epidemiology*, **8**, 738.

Schmoor, C., Schumacher, M., Finke, J., and Beyersmann, J. (2013). Competing risks and multistate models. *Clinical Cancer Research*, **19**, 12–21.

Talawar, A. and Rangoli, A. (2023). Comparative analysis of competing risks models using covariates. *International Journal of Agricultural and Statistical Sciences*, **19**, 691–705.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 23, No. 1, 2025 (New Series), pp 139–175
https://www.ssca.org.in/journal

Extinction and Stationary Distribution of a Stochastic $SEII_aI_qHR$ Epidemic Model with Intervention

Tamalendu Das[1], **Tridip Sardar**[2] **and Sourav Rana**[1]
[1]*Department of Statistics, Visva-Bharati University, Santiniketan, West Bengal, India*
[2]*Department of Mathematics, Dinabandhu Andrews College, Kolkata, West Bengal, India*

Received: 31 July 2023; Revised: 25 March 2024; Accepted: 29 April 2024

Abstract

Mathematical and statistical models serve as valuable tools for the analysis and simulation of infectious disease transmission. This study explores the dynamics of Covid-19 through the utilization of a deterministic epidemic model denoted as $SEII_aI_qHR$, incorporating interventions. The investigation focuses on essential aspects such as the positivity, boundedness, existence of various equilibria based on the basic reproduction number (R_0), and asymptotic behavior of solutions around these equilibria in the deterministic model. Recognizing the significance of environmental noise and the involvement of random factors in real-world disease propagation systems, we also develop a stochastic version of the $SEII_aI_qHR$ model to account for the impact of noise. We establish the necessary conditions for the existence and uniqueness of solutions for the system and discuss the ergodic stationary distribution as well as the conditions for system extinction. To validate our analytical findings, we conduct numerical studies. Our results indicate that the rate of intervention and the fraction of the population in quarantine actively influence disease control efforts.

Key words: Stochastic model; Disease intervention; Extinction; Stationary distribution; Sieve bootstrap test.

AMS Subject Classifications: 37H30, 37A50, 60G17, 34A34, 37N25

1. Introduction

Infectious diseases are the leading cause of deaths in the low-income countries (W.H.O., 2020). As of 2019, all communicable diseases together accounted for 36% of all deaths worldwide (W.H.O., 2020). Some example of communicable diseases are SARS, MERS-CoV, COVID-19, Dengue, Malaria, *etc.* Severe acute respiratory syndrome (SARS) is a viral respiratory disease caused by a SARS-associated coronavirus. Burden of SARS outbreak in 2003 in Asian countries is around USD $60 billion (Ding and Zhang, 2022). Middle East Respiratory Syndrome (MERS) is viral respiratory illness and it was first occurred in 2012 in Saudi Arabia. Approximately 35% of MERS cases reported to WHO have died (W.H.O., 2022). Recent outbreak of COVID-19 infection causes around 7 million deaths worldwide (W.H.O.,

Corresponding Author: Sourav Rana
Email: sourav.rana@visva-bharati.ac.in

2023b). Dengue is a viral infection caused by the bite of infected mosquitoes. Around half of the world population are at risk of dengue infection with 100–400 million infections occurring each year (Bhatt *et al.*, 2013). Along with the dengue, as of 2021, around half of the world population at risk of Malaria with around 247 million cases and approximately 0.61 million deaths currently occurring each year (W.H.O., 2023a).

In epidemiology, compartmental SIR type models can provide an overall understanding of the dynamics of infectious diseases. Information like spread dynamics, incidence peak timing, transmission severity, effect of disease control strategies *etc.* can be obtained by studying mathematical models (Cai *et al.*, 2017; Ding and Zhang, 2022; Tang *et al.*, 2020; Li *et al.*, 2020). Classical epidemiological models of communicable diseases are mainly deterministic compartmental systems (Choisy *et al.*, 2007; Wearing *et al.*, 2005). However, disease incidence growth in general random in nature since uncertainty in contact rates (Cai *et al.*, 2013; Allen, 2017). Furthermore, disease incidence also depend on population demographic rates which in-general follows Markovian process therefore, it is related to environmental noise (Cai *et al.*, 2013; Allen, 2017). Thus, stochastic differential equation (SDE) based models can provide more realistic information on disease spread at initial stage of infection (Allen, 2008, 2017; Cai *et al.*, 2013; Mao, 2007; Oksendal, 2013).

Recently, there are few works on infectious diseases can be found in literature based on stochastic differential equations (Cai *et al.*, 2013; Lahrouz and Omari, 2013; Ding and Zhang, 2022; Cai *et al.*, 2017; Rao *et al.*, 2012; Din *et al.*, 2021; Sun *et al.*, 2022; Din *et al.*, 2020; Tuckwell and Williams, 2007). Randomness in these models are incorporated either by adding random noise in the state equations or by considering environmental fluctuations in some model parameters (Allen, 2008, 2017). Cai *et al.* (2013) found that random fluctuations can suppress the disease outbreak that leads some insight on disease control strategies. Lahrouz and Omari (2013) considered a SIRS epidemic model with general incidence rate in a population of varying size. They analytically determined the sufficient conditions for the extinction and the existence of a unique stationary distribution. Ding and Zhang (2022) developed a stochastic SIRS epidemic model with information intervention. Author's determined that the average in time of the second moment of the solutions of the stochastic system is bounded for a relatively small noise. Furthermore, they found that information interaction response rate have a vital role in reducing disease incidence, and as the intensity of the response increases, the number of infected population decreases, which is beneficial for disease control (Ding and Zhang, 2022). Cai *et al.* (2017) considered a stochastic version of SIRS epidemic model with ratio-dependent incidence rate. Author's mathematically derived some results on permanence and extinction of the proposed stochastic epidemic model. Rao *et al.* (2012) determined stability of an epidemic model with diffusion and stochastic perturbation. Din *et al.* (2021) use a stochastic Markovian dynamics approach to describe the spreading of dengue and the threshold of the disease. Some mathematical properties of the stochastic epidemic model are determined.

In this paper, we first develop a deterministic $SEII_aI_qHR$ epidemic model with frequency dependent incidence rate based on the assumption that a susceptible individual may get infection either by contacting a symptomatic or an asymptomatic or an exposed individual. This deterministic model also considered the transmission variability among different transmission rates from symptomatic, asymptomatic and exposed individuals. Furthermore, model also considered the awareness effect (for example spreading awareness program

through media, proper hand sanitization, social distancing, wearing mask, *etc*), and infection (exposed population) quarantine effect. Main objective of this work is to study the effect of stochastic perturbations in the developed deterministic $SEII_aI_qHR$ epidemic model. In particular, we focused on answering the following questions:

- A detailed study of the $SEII_aI_qHR$ epidemic model and its stochastic version. Then comparison between their dynamics based on various factors.
- How the effect of intervention and quarantine effect influenced the dynamics of a disease in presence of environmental fluctuations.

The rest of the paper is presented as follows: In section 2, detailed $SEII_aI_qHR$ model is formulated. In section 3, some basic properties (example: positivity of solution, global stability of the disease-free equilibrium, local stability of the endemic equilibrium, etc) of the deterministic $SEII_aI_qHR$ model are studied. Detailed formulation of the stochastic $SEII_aI_qHR$ model is shown in section 4. We also discussed Euler Maruyama scheme to determine the numerical solution of the stochastic differential equation. Next, we analytically studied the existence and uniqueness of the solution for the SDE model in section 5. Moreover, long term disease extinction, ergodicity of the solution is studied analytically through various mathematical as well statistical concept. In section 6, we numerically studied the deterministic system to support its analytical findings. We further studied the stochastic system and generated various sample paths, average density paths, histograms of densities, stochastic extinction scenario, *etc.* We have replicated the system very large times to adress the role of quarantine population in the trend of infection. Finally, we discuss and conclude our study.

2. The mathematical model

We start with a deterministic compartmental SIR-type model where population is subdivided into seven mutually exclusive sub-classes namely susceptible (S), exposed (E), symptomatic (I), asymptomatic (I_a), quarantined (I_q), hospitalized (H) and recovered (R), respectively. We considered frequency dependent force of infection with the assumption that susceptible can get infection in contact with the symptomatic (I), asymptomatic (I_a), and exposed (E) cases, respectively. However, we also assumed that the probability of infection form the exposed and asymptomatic cases are lesser compared to the symptomatic cases with transmission modification parameters $\eta_1(0 \leq \eta_1 \leq 1)$, and $\eta_2(0 \leq \eta_2 \leq 1)$, respectively. Furthermore, we also considered the effect of some intervention that reduce the transmission rate β by a factor $(1-k)$, where $0 \leq k \leq 1$. In epidemiological point of view, this intervention represents some awareness effect among the susceptible population that reduce the contact with the infected populations (exposed, symptomatic and asymptomatic). The intervention strategies includes the preventive measures such as lock-down, spreading awareness program through media, proper hand sanitization, social distancing, wearing mask, *etc.* which results in slowing down the disease transmission process.

We assume variable human population with recruitment rate Π. The susceptible compartment reduced due to new infection and natural deaths at rate μ_d. Exposed population increased due to new infection coming from the susceptible compartment and reduced due to natural deaths at a rate μ_d. After the incubation period $\frac{1}{\sigma}$, a fractions ρ_1 and ρ_2 of the exposed population become symptomatic and asymptomatic infected and the remaining fraction $(1-\rho_1-\rho_2)$ of the exposed population become quarantined. Symptomatic infected compartment (I) is increased due to inflow of infected population coming from the exposed

class (E) and reduced due to natural death at a rate μ_d and a fraction α become hospitalized. Asymptomatic infected compartment (I_a) increased due to inflow of a fraction ρ_2 of the exposed population after completion of the incubation period $\frac{1}{\sigma}$. This compartment is decreased due to natural recovery and death at rates γ_a and μ_d, respectively. Quarantined compartment (I_q) increased due to those exposed individuals who are quarantined and this compartment is reduced due to hospitalization of symptomatic cases, natural death and recovery at rates α_q, γ_q and μ_d, respectively. Hospitalized compartment (H) is increased by the patient coming from the symptomatic class and quarantined compartments at rates α, and α_q, respectively. This compartment is decreased due to recovery, disease related death, and natural death at rates γ, δ, and μ_d, respectively. Recovered compartment increased due to inflow of individuals coming from asymptomatic, quarantined, and hospitalized compartments, respectively. This population is reduced by natural death at a rate μ_d. Based on all the assumptions our deterministic the epidemic model that represents the rate of change of different disease classes are provided below:

$$\begin{aligned}
\frac{dS}{dt} &= \Pi - (1-k)\frac{\beta S}{N}(I + \eta_1 I_a + \eta_2 E) - \mu_d S,\\
\frac{dE}{dt} &= (1-k)\frac{\beta S}{N}(I + \eta_1 I_a + \eta_2 E) - \sigma E - \mu_d E,\\
\frac{dI}{dt} &= \rho_1 \sigma E - \alpha I - \mu_d I,\\
\frac{dI_a}{dt} &= \rho_2 \sigma E - \gamma_a I_a - \mu_d I_a,\\
\frac{dI_q}{dt} &= (1-\rho_1-\rho_2)\sigma E - (\alpha_q + \gamma_q) I_q - \mu_d I_q,\\
\frac{dH}{dt} &= \alpha I + \alpha_q I_q - (\gamma + \delta) H - \mu_d H,\\
\frac{dR}{dt} &= \gamma_a I_a + \gamma_q I_q + \gamma H - \mu_d R,
\end{aligned} \tag{1}$$

The schematic diagram and the description of the parameters used in the model (1) is presented in Fig. 1 and Table 1 respectively.

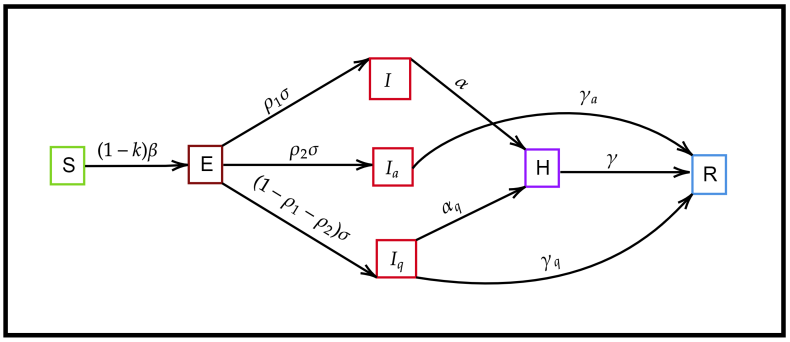


Figure 1: A Flow diagram of the model (1).

Table 1: Description of various parameters used in the model (1).

Parameter	Definitions	Value	Reference
Π	Recruitment rate	10	Din *et al.* (2021)
μ_d	Death rate	0.2	Din *et al.* (2020)
η_1	Modification parameter	0.1002	Senapati *et al.* (2021)
η_2	Modification parameter	(0.1,0.4)	Assumed
k	Strength of intervention	$(0, 0.6544)$	Senapati *et al.* (2021)
β	Rate of disease transmission	1.7399	Senapati *et al.* (2021)
σ	Rate of transition from E to I	0.1923	Li *et al.* (2020)
ρ_1	Fraction of the E move to I	0.3362	Senapati *et al.* (2021)
ρ_2	Fraction of the E move to I_a	0.4204	Senapati *et al.* (2021)
α	Rate of transition from I to H	0.2174	Li *et al.* (2020)
α_q	Rate of transition from I_q to H	0.1429	Senapati *et al.* (2021)
γ_a	Recovery rate of I_a	0.13978	Tang *et al.* (2020)
γ_q	Recovery rate of I_q	0.11624	Tang *et al.* (2020)
γ	Recovery rate of H	0.0701	Senapati *et al.* (2021)
δ	Rate of disease induced death	0.0175	Senapati *et al.* (2021)

3. Analysis

3.1. Model positivity

Theorem 1: The solution to the system (1) remains positive for all time t (≥ 0) given a non-negative initial condition.

Proof: From (1) we can write
$\left.\frac{dS}{dt}\right|_{S=0} = \Pi \geq 0,\ \left.\frac{dE}{dt}\right|_{E=0} = \frac{(1-k)}{N}\beta S(I+\eta_1 I_a) \geq 0,\ \left.\frac{dI}{dt}\right|_{I=0} = \rho_1\sigma E \geq 0,\ \left.\frac{dI_a}{dt}\right|_{I_a=0} = \rho_2\sigma E \geq 0,$

$\left.\frac{dI_q}{dt}\right|_{I_q=0} = (1-\rho_1-\rho_2)\sigma E \geq 0, \left.\frac{dH}{dt}\right|_{H=0} = \alpha I + \alpha_q I_q \geq 0, \left.\frac{dR}{dt}\right|_{R=0} = \gamma_a I_a + \gamma_q I_q + \gamma H \geq 0.$
Consequently, the system (1) is positive at all times when positive initial conditions are given. □

3.2. Boundness

Theorem 2: The system (1) is bounded in the feasible region $\{(S,E,I,I_a,I_q,H,R) \in R_+^7 : N(t) \leq \frac{\Pi}{\mu_d};\ S(t),E(t),I(t),I_a(t),I_q(t),H(t),R(t) \geq 0,\ \text{at any time } t \geq 0\}$.

Proof: We begin by considering the total population density $N(t)$ and utilize the model (1) in the following manner:

$$N(t) = S(t)+E(t)+I(t)+I_a(t)+I_q(t)+H(t)+R(t),$$

$$\frac{dN}{dt} = \Pi - \mu_d N,$$

By using Gronwall's inequality,

$$\begin{aligned} &N(t) = N(0)e^{-\Pi t} + \tfrac{\Pi}{\mu_d}, \quad t \geq 0, \\ &\Rightarrow \lim_{n\to\infty} SupN(t) \leq \tfrac{\Pi}{\mu_d}. \end{aligned} \tag{2}$$

So we can say that the system (1) is bounded in the region $\{(S, E, I, I_a, I_q, H, R) \in R_+^7 : N(t) \leq \frac{\Pi}{\mu_d};\ S(t), E(t), I(t), I_a(t), I_q(t), H(t), R(t) \geq 0,\ \text{at any time } t \geq 0\}$. □

3.3. Local stability of disease-free equilibrium (DFE)

The DFE of the model (1) is given by $E_0(\frac{\Pi}{\mu_d}, 0, 0, 0, 0, 0, 0)$.
The local stability of E_0 in the system (1) can be established using the next generation operator method. Following the notation in Driessche and Watmough (2002), we denote the matrices F and V for the new infection and transition terms, respectively, as follows:

$$F = \begin{bmatrix} \eta_2(1-k)\beta & (1-k)\beta & (1-k)\beta\eta_1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix},$$

$$V = \begin{bmatrix} (\sigma+\mu_d) & 0 & 0 & 0 \\ -\rho_1\sigma & (\alpha+\mu_d) & 0 & 0 \\ -\rho_2\sigma & 0 & (\gamma_a+\mu_d) & 0 \\ -(1-\rho_1-\rho_2)\sigma & 0 & 0 & (\alpha_q+\gamma_q+\mu_d) \end{bmatrix}$$

Therefore, the basic reproduction number, denoted by R_0 (Hethcote, 2000) and calculated as $\rho(FV^{-1})$ where ρ represents the spectral radius, can be expressed as $R_0 = \frac{\eta_2(1-k)\beta}{(\sigma+\mu_d)} + \frac{\rho_1\sigma(1-k)\beta}{(\sigma+\mu_d)(\alpha+\mu_d)} + \frac{(1-k)\beta\eta_1\rho_2\sigma}{(\sigma+\mu_d)(\gamma_a+\mu_d)}$.

By utilizing Theorem 2 from Driessche and Watmough (2002), we can establish the following result.

Lemma 1: The DFE, E_0 of the model (1) is locally-asymptotically stable (LAS) if $R_0 < 1$, and unstable if $R_0 > 1$.

3.4. Global stability of DFE

In order to demonstrate the global stability of E_0 in the model (1), we can rewrite the system as follows:

$$\begin{aligned} \frac{dX}{dt} &= T(X, I') \\ \frac{dI'}{dt} &= G(X, I'), \quad G(X, 0) = 0, \end{aligned} \tag{3}$$

where $X = (S, H, R) \in \mathbb{R}_+^3$ represents the components denoting the number of uninfected individuals, and $I' = (E, I, I_a, I_q) \in \mathbb{R}_+^4$ represents the components denoting the number of infected individuals, including latent, infectious, and other categories. $E_0 = (X^*, 0)$

represents the disease-free equilibrium of the system eqrefEQ:eqn 2.3. For the system (1), the expressions for $T(X, I')$ and $G(X, I')$ are in the Annexure.

From the expression of $G(X, I')$, it is evident that $G(X, 0) = 0$.

To demonstrate the global stability of $\varepsilon_0 = (X^*, 0)$, the following two conditions must be satisfied:

(H1) For $\dfrac{dX}{dt} = T(X, 0)$, X^* is globally asymptotically stable.

(H2) $G(X, I') = AI - \widehat{G}(X, I)$, $\widehat{G}(X, I) \geq 0$ for $(X, I') \in \Omega$,

Here, $A = D_{I'}G(X^*, 0)$ represents an M-matrix, where the off-diagonal elements are non-negative. Additionally, Ω denotes the region in which the model (1) holds biological significance.

Now, we can express the system defined in **(H1)** as follows:

$$\begin{aligned} \frac{dS}{dt} &= \Pi - \mu_d S, \\ \frac{dR}{dt} &= -\mu_d R. \end{aligned} \tag{4}$$

By solving this system of equations analytically, we obtain the following solution: $S(t) = \frac{\Pi}{\mu_d} + e^{-\mu_d t}(S(0) - \frac{\Pi}{\mu_d})$, $R(t) = e^{-\mu_d t}R(0)$. As $t \to \infty$, $S(t) = \frac{\Pi}{\mu_d}$, $R(t) \to 0$. Hence, X^* is globally asymptotically stable for $\dfrac{dX}{dt} = T(X, 0)$.

Therefore, we can conclude that **(H1)** holds for the system (1). Now, the matrices A and $\widehat{G}(X, I)$ for the system (1) are in the Annexure.

It is evident that A is an M-matrix, and since $S(t) \leq N(t)$ holds in Ω, we can conclude that $\widehat{G}(X, I) \geq 0$ for $(X, I) \in \Omega$. Based on the findings presented in Castillo-Chavez *et al.* (2002), the following result can be stated:

Theorem 3: The DFE of the model (1) is globally asymptotically stable in Ω whenever $R_0 < 1$.

3.5. Existence of endemic equilibria

In this section, we establish the existence of the endemic equilibrium for the model (1). Let us denote $k_1 = \sigma + \mu_d$, $k_2 = \alpha + \mu_d$, $k_3 = \gamma_a + \mu_d$, $k_4 = \alpha_q + \gamma_q + \mu_d$, $k_5 = \gamma + \delta + \mu_d$. Let $E_*(S^*, E^*, I^*, I_a^*, I_q^*, H^*, R^*)$ represents any arbitrary endemic equilibrium point (EEP) of the model (1). Further, define $\lambda^* = \frac{(1-k)\beta I^*}{N^*} + \frac{(1-k)\beta\eta_1 I_a^*}{N^*} + \frac{(1-k)\beta\eta_2 E^*}{N^*}$.
So we have E_* in terms of λ^* by solving the equations in (1) at steady-state (see Annexure).

By substituting the E_* expressions into λ^*, we can observe that the non-zero equilibrium of the model (1) satisfies the following linear equation in terms of λ^*: $a_0\lambda^* + a_1 = 0$, where, $a_0 = k_2k_3k_4k_5\mu_d + k_3k_4k_5\mu_d\rho_1\sigma + k_2k_4k_5\mu_d\rho_2\sigma + k_2k_3k_5\mu_d(1-\rho_1-\rho_2)\sigma + k_3k_4\mu_d\alpha\rho_1\sigma +$

$k_2k_3\mu_d\alpha_q(1-\rho_1-\rho_2)\sigma+k_2k_4k_5\gamma_a\rho_2\sigma+k_2k_3k_5\gamma_a(1-\rho_1-\rho_2)\sigma+k_3k_4\gamma\alpha\rho_1\sigma+k_2k_3\alpha_q(1-\rho_1-\rho_2)\sigma$ $a_1 = k_1k_2k_3k_4k_5\mu_d(1-R_0)$. Since $a_0 > 0,\ k_1 > 0,\ k_2 > 0,\ k_3 > 0,\ k_4 > 0,\ k_5 > 0$ and $\mu_d > 0$, it becomes evident that the model (1) possesses a unique endemic equilibrium point (EEP) when $R_0 > 1$. On the other hand, when $R_0 < 1$, there is no positive endemic equilibrium point in the model. Based on the analysis, we can conclude that there is no existence of equilibrium other than the disease-free equilibrium (DFE) when $R_0 < 1$. Additionally, it can be demonstrated that the DFE E_0 of the model (1) is globally asymptotically stable (GAS) when $R_0 < 1$.

From the above discussion we have concluded that,

Theorem 4: The model (1) possesses a unique endemic (positive) equilibrium, denoted as E^*, whenever the basic reproduction number $R_0 > 1$. However, for $R_0 \leq 1$, the model does not have any endemic equilibrium.

3.6. Local stability of endemic equilibrium point (EEP)

The EEP of the model (1) is given by $E_*(S^*, E^*, I^*, I_a^*, I_q^*, H^*, R^*)$ where the expressions are computed analytically in the Annexure.

3.7. Local stability

Theorem 5: The endemic equilibrium E_* exhibits local asymptotic stability if all the roots of the characteristic equation possess negative real parts.

Proof: The Jacobian matrix of the system at E_* is as follows:

$$J_{E_*} = \begin{pmatrix} -P_{11} & -P_{12} & -P_{13} & -P_{14} & P_{15} & P_{16} & P_{17} \\ P_{21} & P_{22} & P_{23} & P_{24} & -P_{25} & -P_{26} & -P_{27} \\ 0 & P_{32} & -P_{33} & 0 & 0 & 0 & 0 \\ 0 & P_{42} & 0 & -P_{44} & 0 & 0 & 0 \\ 0 & P_{52} & 0 & 0 & -P_{55} & 0 & 0 \\ 0 & 0 & P_{63} & 0 & P_{65} & -P_{66} & 0 \\ 0 & 0 & 0 & P_{74} & P_{75} & P_{76} & -P_{77} \end{pmatrix},$$

where, $P_{11} = \frac{\beta(1-k)(N-S^*)}{N^2}(I^* + \eta_1 I_a^* + \eta_2 E^*) + \mu_d,\ P_{12} = \frac{(1-k)\beta S^*}{N^2}(\eta_2 N - I^* - \eta_1 I_a^* - \eta_2 E^*),\ P_{13} = \frac{(1-k)\beta S^*}{N^2}(N - I^* - \eta_1 I_a - \eta_2 E),\ P_{14} = \frac{(1-k)\beta S^*}{N^2}(\eta_1 N - I^* - \eta_1 I_a^* - \eta_2 E^*),$ $P_{15} = P_{16} = P_{17} = \frac{(1-k)\beta S^*}{N^2}(I^* + \eta_1 I_a^* + \eta_2 E^*)\ P_{21} = \frac{\beta(1-k)(N-S^*)}{N^2}(I^* + \eta_1 I_a^* + \eta_2 E^*),$ $P_{22} = \frac{(1-k)\beta S^*}{N^2}(\eta_2 N - I^* - \eta_1 I_a^* - \eta_2 E^*) - \sigma - \mu_d,\ P_{23} = \frac{(1-k)\beta S^*}{N^2}(N - I^* - \eta_1 I_a - \eta_2 E),$ $P_{24} = \frac{(1-k)\beta S^*}{N^2}(\eta_1 N - I^* - \eta_1 I_a^* - \eta_2 E^*),\ P_{25} = P_{26} = P_{27} = \frac{(1-k)\beta S^*}{N^2}(I^* + \eta_1 I_a^* + \eta_2 E^*),$ $P_{32} = \rho_1\sigma,\ P_{33} = (\alpha + \mu_d),\ P_{42} = \rho_2\sigma,\ P_{44} = (\gamma_a + \mu_d),\ P_{52} = \rho_3\sigma,\ P_{55} = (\alpha_q + \gamma_q + \mu_d),$ $P_{63} = \alpha,\ P_{65} = \alpha_q,\ P_{66} = (\gamma + \delta + \mu_d),\ P_{74} = \gamma_a,\ P_{75} = \gamma_q,\ P_{76} = \gamma,\ P_{77} = \mu_d.$

Here the stability of E_* is determined by the presence of negative real roots in the characteristic equation of J_{E_*}. □

Now, the corresponding characteristic equation is a polynomial of degree 7, and analytical computation becomes challenging. Therefore, we will validate Theorem 5 by per-

forming numerical computations.

4. Stochastic model

The role of environmental change in shaping epidemic development has been widely recognized (Oksendal, 2006). The unpredictable nature of human contact introduces inherent randomness into the growth and spread of epidemics, leading to ongoing disruptions in population dynamics (Beddington and May, 1977; Chen *et al.*, 2023). In the study of epidemic dynamics, the utilization of SDE models is often necessary due to their ability to provide a more suitable framework in various scenarios. These models effectively capture the stochastic nature of population fluctuations and account for the dynamical changes resulting from subtle parameter variations. In a recent investigation, Hussain *et al.* (2023) explored a stochastic version of the MERS-CoV epidemic model, focusing on the ergodic stationary distribution and criteria for disease extinction. Concurrently, Shi and Jiang (2023) introduced a stochastic compartmental model for COVID-19, integrating an Ornstein-Uhlenbeck (OU) process into the contact rate. Their analysis included the criteria for stationary distribution and the derivation of the probability density function near quasi-equilibrium. Additionally, the impact of the OU process on the stochastic model's dynamic behavior was examined. Tan *et al.* (2023) delved into a stochastic SIS epidemic model enriched by media coverage. Through the consideration of two threshold quantities, they investigated the stochastic dynamics, illustrating scenarios where disease eradication is certain or persistent with a distinct stationary distribution. Their study also inferred insights based on the intensity of random disturbances. Furthermore, Ullah *et al.* (2023) explored a stochastic epidemic model incorporating vaccination programs. Extinction and persistence conditions were scrutinized, supported by graphical representations to validate analytical findings.

Many real-world stochastic epidemic models are formulated based on their deterministic counterparts, with the deterministic version serving as a foundation for their development (Jiang *et al.*, 2010; Mao *et al.*, 2002; Li *et al.*, 2020; Thomas and Shelemyahu, 1989). Under the assumption that the coefficients of model (1) are influenced by random noise, which can be accurately represented by Brownian motion, the resulting model (1) can be transformed into a SDE in the following manner:

$$\begin{aligned}
dS &= \left[\Pi - \mu_d S - \frac{(1-k)}{N}\beta S(I + \eta_1 I_a + \eta_2 E)\right]dt + \theta_1 S\ dB_1,\\
dE &= \left[\frac{(1-k)}{N}\beta S(I + \eta_1 I_a + \eta_2 E) - \sigma E - \mu_d E\right]dt + \theta_2 E\ dB_2,\\
dI &= \left[\rho_1\sigma E - \alpha I - \mu_d I\right]dt + \theta_3 I\ dB_3,\\
dI_a &= \left[\rho_2\sigma E - \gamma_a I_a - \mu_d I_a\right]dt + \theta_4 I_a\ dB_4,\\
dI_q &= \left[(1-\rho_1-\rho_2)\sigma E - (\alpha_q + \gamma_q) I_q - \mu_d I_q\right]dt + \theta_5 I_q\ dB_5,
\end{aligned} \tag{5}$$

$$dH = \Big[\alpha I + \alpha_q I_q - (\gamma+\delta)H - \mu_d H\Big]dt + \theta_6 H\ dB_6,$$

$$dR = \Big[\gamma_a I_a + \gamma_q I_q + \gamma H - \mu_d R\Big]dt + \theta_7 R\ dB_7,$$

In the model (5), all parameters and state variables are assumed to be non-negative real numbers. The influence of noise is taken into account through the functions $B_i(t),\ i = 1(1)7$, which represent standard Brownian motions, and $\theta_i(>0),\ i = 1(1)7$, which represent the corresponding intensities of the white noise. Additionally, the Brownian motion satisfies the fundamental axiom $B_1(0) = B_2(0) = B_3(0) = B_4(0) = B_5(0) = B_6(0) = B_7(0)$.

Let's define the vector G for the system (5) as $G = [S, E, I, I_a, I_q, H, R]^T$. The transition probability is specified in Table 2. The expectation $E_x[\Delta G]$ and variance $E_x[\Delta G \Delta G^T]$ are defined as follows.

So the Expectation is $E_x[\Delta G] = \sum_{i=1}^{22} P_i(\Delta G)_i =$

$$\begin{bmatrix} \Pi - \mu_d S - \frac{(1-k)}{N}\beta S(I + \eta_1 I_a + \eta_2 E) \\ \frac{(1-k)}{N}\beta S(I + \eta_1 I_a + \eta_2 E) - \sigma E - \mu_d E \\ \rho_1\sigma E - \alpha I - \mu_d I \\ \rho_2\sigma E - \gamma_a I_a - \mu_d I_a \\ (1-\rho_1-\rho_2)\sigma E - (\alpha_q+\gamma_q)I_q - \mu_d I_q \\ \alpha I + \alpha_q I_q - (\gamma+\delta)H - \mu_d H \\ \gamma_a I_a + \gamma_q I_q + \gamma H - \mu_d R \end{bmatrix} \Delta t.$$

Also the variance is given below:

$$E_x[\Delta G \Delta G^T] = \sum_{i=1}^{22} P_i[(\Delta G)_i][(\Delta G)_i]^T = \begin{bmatrix} M_{11} & M_{12} & 0 & 0 & 0 & 0 & 0 \\ M_{21} & M_{22} & M_{23} & M_{24} & M_{25} & 0 & 0 \\ 0 & M_{32} & M_{33} & 0 & 0 & M_{36} & 0 \\ 0 & M_{42} & 0 & M_{44} & 0 & 0 & M_{47} \\ 0 & M_{52} & 0 & 0 & M_{55} & M_{56} & M_{57} \\ 0 & 0 & M_{63} & 0 & M_{65} & M_{66} & M_{67} \\ 0 & 0 & 0 & M_{74} & M_{75} & M_{76} & M_{77} \end{bmatrix} \Delta t,$$

Here,
$M_{11} = P_1 + P_2 + P_3 + P_4 + P_5 = \Pi + \mu_d S + \frac{(1-k)}{N}\beta S\eta_2 E + \frac{(1-k)}{N}\beta S I + \frac{(1-k)}{N}\beta S\eta_1 I_a;$
$M_{12} = M_{21} = -P_3 = -(\frac{(1-k)}{N}\beta S\eta_2 E);$
$M_{22} = P_3 + P_4 + P_5 + P_6 + P_7 = \frac{(1-k)}{N}\beta S\eta_2 E + \frac{(1-k)}{N}\beta S I + \frac{(1-k)}{N}\beta S\eta_1 I_a + \sigma E + \mu_d E;$
$M_{23} = M_{32} = P_8 = \rho_1\sigma E;$
$M_{24} = M_{42} = P_{11} = \rho_2\sigma E;$
$M_{25} = M_{52} = P_{14} = (1-\rho_1-\rho_2)\sigma E;$
$M_{33} = P_8 + P_9 + P_{10} = \rho_1\sigma E + \alpha I + \mu_d I;$
$M_{36} = M_{63} = -P_9 = -\alpha I;$

$M_{44} = P_{11} + P_{12} + P_{13} = \rho_2\sigma E + \gamma_a I_a + \mu_d I_a;$
$M_{47} = M_{74} = -P_{12} = -\gamma_a I_a;$
$M_{55} = P_{14} + P_{15} + P_{16} + P_{17} = (1-\rho_1-\rho_2)\sigma E + \alpha_q I_q + \gamma_q I_q + \mu_d I_q;$
$M_{56} = M_{65} = -P_{15} = -\alpha_q I_q;$
$M_{57} = M_{75} = -P_{16} = -\gamma_q I_q;$
$M_{66} = P_9 + P_{15} + P_{19} + P_{20} + P_{21} = \alpha I + \alpha_q I_q + \gamma H + \delta H + \mu_d H;$
$M_{67} = M_{76} = -P_{19} = -\gamma H;$
$M_{77} = P_{12} + P_{16} + P_{19} + P_{22} = \gamma_a I_a + \gamma_q I_q + \gamma H + \mu_d R;$

Now we define,

$$Drift = \mathcal{C}(\mathcal{G},t) = \frac{E_x[\Delta G]}{\Delta t} = \begin{bmatrix} \Pi - \mu_d S - \frac{(1-k)}{N}\beta S(I+\eta_1 I_a + \eta_2 E) \\ \frac{(1-k)}{N}\beta S(I+\eta_1 I_a+\eta_2 E) - \sigma E - \mu_d E \\ \rho_1\sigma E - \alpha I - \mu_d I \\ \rho_2\sigma E - \gamma_a I_a - \mu_d I_a \\ (1-\rho_1-\rho_2)\sigma E - (\alpha_q+\gamma_q)I_q - \mu_d I_q \\ \alpha I + \alpha_q I_q - (\gamma+\delta)H - \mu_d H \\ \gamma_a I_a + \gamma_q I_q + \gamma H - \mu_d R \end{bmatrix}.$$

Also the diffusion is defined as

$$Diffusion = \mathcal{D}(\mathcal{G},t) = \sqrt{\frac{E_x[\Delta G \Delta G^T]}{\Delta t}} = \sqrt{\begin{bmatrix} M_{11} & M_{12} & 0 & 0 & 0 & 0 & 0 \\ M_{21} & M_{22} & M_{23} & M_{24} & M_{25} & 0 & 0 \\ 0 & M_{32} & M_{33} & 0 & 0 & M_{36} & 0 \\ 0 & M_{42} & 0 & M_{44} & 0 & 0 & M_{47} \\ 0 & M_{52} & 0 & 0 & M_{55} & M_{56} & M_{57} \\ 0 & 0 & M_{63} & 0 & M_{65} & M_{66} & M_{67} \\ 0 & 0 & 0 & M_{74} & M_{75} & M_{76} & M_{77} \end{bmatrix}}.$$

By incorporating the drift and diffusion equations, the SDE for the system can be expressed as follows:

$$d\mathcal{G}(t) = \mathcal{C}(\mathcal{G},t)\ dt + \mathcal{D}(\mathcal{G},t)\ dB(t)$$

i.e.,

$$d\begin{bmatrix} S \\ E \\ I \\ I_a \\ I_q \\ H \\ R \end{bmatrix} = \begin{bmatrix} \Pi - \mu_d S - \frac{(1-k)}{N}\beta S(I+\eta_1 I_a + \eta_2 E) \\ \frac{(1-k)}{N}\beta S(I+\eta_1 I_a+\eta_2 E) - \sigma E - \mu_d E \\ \rho_1\sigma E - \alpha I - \mu_d I \\ \rho_2\sigma E - \gamma_a I_a - \mu_d I_a \\ (1-\rho_1-\rho_2)\sigma E - (\alpha_q+\gamma_q)I_q - \mu_d I_q \\ \alpha I + \alpha_q I_q - (\gamma+\delta)H - \mu_d H \\ \gamma_a I_a + \gamma_q I_q + \gamma H - \mu_d R \end{bmatrix} dt + \sqrt{\begin{bmatrix} M_{11} & M_{12} & 0 & 0 & 0 & 0 & 0 \\ M_{21} & M_{22} & M_{23} & M_{24} & M_{25} & 0 & 0 \\ 0 & M_{32} & M_{33} & 0 & 0 & M_{36} & 0 \\ 0 & M_{42} & 0 & M_{44} & 0 & 0 & M_{47} \\ 0 & M_{52} & 0 & 0 & M_{55} & M_{56} & M_{57} \\ 0 & 0 & M_{63} & 0 & M_{65} & M_{66} & M_{67} \\ 0 & 0 & 0 & M_{74} & M_{75} & M_{76} & M_{77} \end{bmatrix}} dB(t).$$

Table 2: Possible changes in the process of the model

Transition	Probability
$(\Delta G)_1 = [1\,0\,0\,0\,0\,0\,0]^T$	$P_1 = \Pi\ \Delta t$
$(\Delta G)_2 = [-1\,0\,0\,0\,0\,0\,0]^T$	$P_2 = \mu_d S\ \Delta t$
$(\Delta G)_3 = [-1\,1\,0\,0\,0\,0\,0]^T$	$P_3 = \frac{(1-k)}{N}\beta S\eta_2 E\ \Delta t$
$(\Delta G)_4 = [-1\,0\,1\,0\,0\,0\,0]^T$	$P_4 = \frac{(1-k)}{N}\beta S I\ \Delta t$
$(\Delta G)_5 = [-1\,0\,0\,1\,0\,0\,0]^T$	$P_5 = \frac{(1-k)}{N}\beta S\eta_1 I_a\ \Delta t$
$(\Delta G)_6 = [0\ -1\,0\,0\,0\,0\,0]^T$	$P_6 = \sigma E\ \Delta t$
$(\Delta G)_7 = [0\ -1\,0\,0\,0\,0\,0]^T$	$P_7 = \mu_d E\ \Delta t$
$(\Delta G)_8 = [0\,1\,1\,0\,0\,0\,0]^T$	$P_8 = \rho_1\sigma E\ \Delta t$
$(\Delta G)_9 = [0\,0\ -1\,0\,0\,0\,0]^T$	$P_9 = \alpha I\ \Delta t$
$(\Delta G)_{10} = [0\,0\ -1\,0\,0\,0\,0]^T$	$P_{10} = \mu_d I\ \Delta t$
$(\Delta G)_{11} = [0\,1\,0\,1\,0\,0\,0]^T$	$P_{11} = \rho_2\sigma E\ \Delta t$
$(\Delta G)_{12} = [0\,0\,0\,1\,0\,0\,0]^T$	$P_{12} = \gamma_a I_a\ \Delta t$
$(\Delta G)_{13} = [0\,0\,0\ -1\,0\,0\,0]^T$	$P_{13} = \mu_d I_a\ \Delta t$
$(\Delta G)_{14} = [0\,0\,0\ -1\,0\,0\,0]^T$	$P_{14} = (1-\rho_1-\rho_2)\sigma E\ \Delta t$
$(\Delta G)_{15} = [0\,0\,0\,0\ -1\,0\,0]^T$	$P_{15} = \alpha_q I_q\ \Delta t$
$(\Delta G)_{16} = [0\,0\,0\,0\ -1\,0\,0]^T$	$P_{16} = \gamma_q I_q\ \Delta t$
$(\Delta G)_{17} = [0\,0\,0\,0\ -1\,0\,0]^T$	$P_{17} = \mu_d I_q\ \Delta t$
$(\Delta G)_{18} = [0\,0\,1\,0\,0\,0\,0]^T$	$P_{18} = \alpha I\ \Delta t$
$(\Delta G)_{19} = [0\,0\,0\,0\,0\ -1\,0]^T$	$P_{19} = \gamma H\Delta t$
$(\Delta G)_{20} = [0\,0\,0\,0\,0\ -1\,0]^T$	$P_{20} = \delta H\Delta t$
$(\Delta G)_{21} = [0\,0\,0\,0\,0\ -1\,0]^T$	$P_{21} = \mu_d H\Delta t$
$(\Delta G)_{22} = [0\,0\,0\,0\,0\,0\ -1]^T$	$P_{22} = \mu_d R\Delta t$

4.1. Euler Maruyama scheme

In this section, we employ the Euler-Maruyama scheme to obtain the numerical solution of the stochastic differential equation. The model parameters used in the computations are listed in Table 1. The following computational procedure is followed:

$$d\mathcal{G}_n(t) = \mathcal{C}(\mathcal{G}_n, t)\ dt + \mathcal{D}(\mathcal{G}_n, t)dB(t)$$

$$\begin{bmatrix} S^{n+1} \\ E^{n+1} \\ I^{n+1} \\ I_a^{n+1} \\ I_q^{n+1} \\ H^{n+1} \\ R^{n+1} \end{bmatrix} = \begin{bmatrix} S^n \\ E^n \\ I^n \\ I_a^n \\ I_q^n \\ H^n \\ R^n \end{bmatrix} + \begin{bmatrix} \Pi - \mu_d S - \dfrac{(1-k)}{N}\beta S(I + \eta_1 I_a + \eta_2 E) \\ \dfrac{(1-k)}{N}\beta S(I + \eta_1 I_a + \eta_2 E) - \sigma E - \mu_d E \\ \rho_1 \sigma E - \alpha I - \mu_d I \\ \rho_2 \sigma E - \gamma_a I_a - \mu_d I_a \\ (1 - \rho_1 - \rho_2)\sigma E - (\alpha_q + \gamma_q) I_q - \mu_d I_q \\ \alpha I + \alpha_q I_q - (\gamma + \delta) H - \mu_d H \\ \gamma_a I_a + \gamma_q I_q + \gamma H - \mu_d R \end{bmatrix} dt$$

$$+ \sqrt{\begin{bmatrix} M_{11} & M_{12} & 0 & 0 & 0 & 0 & 0 \\ M_{21} & M_{22} & M_{23} & M_{24} & M_{25} & 0 & 0 \\ 0 & M_{32} & M_{33} & 0 & 0 & M_{36} & 0 \\ 0 & M_{42} & 0 & M_{44} & 0 & 0 & M_{47} \\ 0 & M_{52} & 0 & 0 & M_{55} & M_{56} & M_{57} \\ 0 & 0 & M_{63} & 0 & M_{65} & M_{66} & M_{67} \\ 0 & 0 & 0 & M_{74} & M_{75} & M_{76} & M_{77} \end{bmatrix}} \delta B_n$$

5. Parametric perturbation of the model

Let $(\Omega, \mathcal{F}, \{\mathcal{F}_t\}_{t\geq 0}, \mathcal{P})$ be a complete probability space equipped with the filtration $\{\mathcal{F}_t\}_{t\geq 0}$. The filtration is assumed to be increasing and right-continuous, and $\mathcal{F}_0$ contains all $\mathcal{P}$-null sets. Throughout the paper, we denote $a \wedge b$ as the minimum of a and b, $a \vee b$ as the maximum of a and b, and $\langle y(t) \rangle$ as the time average of $y(t)$ defined as $\frac{1}{t}\int_0^t y(s)\ ds$.

5.1. Existence and uniqueness of the global solutions

In order to investigate the dynamic characteristics of the system described by equation (5), the initial step involves verifying the presence of a unique positive solution for this system. This section aims to provide a comprehensive explanation regarding the existence of a unique positive solution to the SDE model represented by equation (5).

Theorem 6: For any initial value $(S(0), E(0), I(0), I_a(0), I_q(0), H(0), R(0)) \in \mathbb{R}_+^7$, there is a

positive solution $(S(t), E(t), I(t), I_a(t), I_q(t), H(t), R(t))$ of the stochastic model (5) for $t \geq 0$ and the solution will maintain in $\mathbb{R}^7_+$ with probability one.

Proof: The constants involved in the equations are locally Lipschitz continuous for the given initial population sizes $(S(0), E(0), I(0), I_a(0), I_q(0), H(0), R(0)) \in \mathbb{R}^7+$ when $t \in [0, \tau e]$, where τ_e is the explosion time (Yanan and Daqing, 2014; Ji and Jiang, 2014). To establish the global nature of the solution, it is necessary to prove that $\tau_e = \infty$ almost surely (a.s.). We select $k_0 \geq 0$ to be sufficiently large such that $S(0), E(0), I(0), I_a(0), I_q(0), H(0)$, and $R(0)$ all fall within the interval $[\frac{1}{k_0}, k_0]$. For each integer $k \geq k_0$, we define the stopping time $\tau_k = \inf\{t \in [0, \tau_e] : \min(S(t), E(t), I(t), I_a(t), I_q(t), H(t), R(t)) \leq \frac{1}{k}$ *or*, $\max(S(t), E(t), I(t), I_a(t), I_q(t), H(t), R(t)) \geq k\}$.

We define $\inf(\phi) = \infty$ for the empty set ϕ according to the given notation. By definition, as k approaches infinity, τ_k increases. We set τ_∞ as the limit of τ_k as k tends to infinity, with $0 \leq \tau_\infty \leq \tau_e$ almost surely (a.s.). By proving that $\tau_\infty = \infty$ almost surely, we can demonstrate that $\tau_e = \infty$, and it follows that $(S(t), E(t), I(t), I_a(t), I_q(t), H(t), R(t)) \in \mathbb{R}^7_+$ a.s. for all $t \geq 0$.

Now, we define a C^2 function $V : \mathbb{R}^7_+ \to \mathbb{R}_+$ such that
$V = V(S(t), E(t), I(t), I_a(t), I_q(t), H(t), R(t)) = S(t) - 1 - \log S(t) + E(t) - 1 - \log E(t)$
$+I(t) - 1 - \log I(t) + I_a(t) - 1 - \log I_a(t) + I_q(t) - 1 - \log I_q(t) + H(t) - 1 - \log H(t) + R(t) -$
$1 - \log R(t)$.

Here the function V is non negative as $y - 1 - \log y \geq 0, \forall\ y \geq 0$. For arbitrary values of $k \geq k_0$ and $T \geq 0$, applying the Itô formula to equation (5) yields the following result.

$dV(S, E, I, I_a, I_q, H, R) = (1-\frac{1}{S})dS+\theta_1(S-1)dB_1(t)+(1-\frac{1}{E})dE+\theta_2(E-1)dB_2(t)+$
$(1-\frac{1}{I})dI+\theta_3(I-1)dB_3(t)+(1-\frac{1}{I_a})dI_a+\theta_4(I_a-1)dB_4(t)+(1-\frac{1}{I_q})dI_q+\theta_5(I_q-1)dB_5(t)+$
$(1-\frac{1}{H})dH+\theta_6(H-1)dB_6(t)+(1-\frac{1}{R})dR+\theta_7(R-1)dB_7(t)$

$= LV(S, E, I, I_a, I_q, H, R)dt+\theta_1(S-1)dB_1(t)+\theta_2(E-1)dB_2(t)+\theta_3(I-1)dB_3(t)+$
$\theta_4(I_a-1)dB_4(t)+\theta_5(I_q-1)dB_5(t)+\theta_6(H-1)dB_6(t)+\theta_7(R-1)dB_7(t)$.

In equation (5), $LH : \mathbb{R}^7_+ \to \mathbb{R}_+$ is defined by the following equation

$$LV(S, E, I, I_a, I_q, H, R) = (1-\tfrac{1}{S})[\Pi - \mu_d S - \frac{(1-k)}{N}\beta S(I+\eta_1 I_a+\eta_2 E)]+\frac{\theta_1^2}{2}+(1-\frac{1}{E})\Big[\frac{(1-k)}{N}\beta S(I+\eta_1 I_a+\eta_2 E)-\sigma E-\mu_d E\Big]+\frac{\theta_2^2}{2}+(1-\frac{1}{I})(\rho_1\sigma E-\alpha I-\mu_d I)+\frac{\theta_3^2}{2}+(1-\frac{1}{I_a})(\rho_2\sigma E-\gamma_a I_a-\mu_d I_a)+\frac{\theta_4^2}{2}+(1-\frac{1}{I_q})((1-\rho_1-\rho_2)\sigma E-(\alpha_q+\gamma_q)I_q-\mu_d I_q)+\frac{\theta_5^2}{2}+(1-\frac{1}{H})(\alpha I+\alpha_q I_q-(\gamma+\delta)H-\mu_d H)+\frac{\theta_6^2}{2}+(1-\frac{1}{R})(\gamma_a I_a+\gamma_q I_q+\gamma H-\mu_d R)+\frac{\theta_7^2}{2}$$

$$\leq \Pi(1-\tfrac{1}{S})+7\mu_d+\tfrac{\beta I}{N}+\eta_1\beta\tfrac{I_a}{N}+\eta_2\beta\tfrac{E}{N}-k\beta\tfrac{I}{N}-k\eta_1\beta\tfrac{I_a}{N}-k\eta_2\beta\tfrac{E}{N}+\sigma+\alpha+\gamma_a+\rho_1\sigma\tfrac{E}{I_q}+\rho_2\sigma\tfrac{E}{I_q}-\alpha_q-\gamma_q-\alpha\tfrac{I}{H}-\alpha_q\tfrac{I_q}{H}+\gamma+\delta-\gamma_a\tfrac{I_a}{R}-\gamma_q\tfrac{I_q}{R}-\gamma\tfrac{H}{R}+\sum_{i=1}^{7}\tfrac{\theta_i^2}{2}$$

$$\leq \Pi+7\mu_d+\sigma+\alpha+\gamma_a+\gamma+\delta-\alpha_q-\gamma_q+\sum_{i=1}^{7}\tfrac{\theta_i^2}{2}=K\ \text{(say)}$$

Here, K is a positive constant that is independent of the variables S, E, I, I_a, I_q, H, R, and the time t. Therefore, $dV(S,E,I,I_a,I_q,H,R) \leq Kdt + \theta_1(S-1)dB_1(t) + \theta_2(E-1)dB_2(t) + \theta_3(I-1)dB_3(t) + \theta_4(I_a-1)dB_4(t) + \theta_5(I_q-1)dB_5(t) + \theta_7(R-1)dB_7(t)$

Integration both sides of above equation from 0 to $\tau_k \wedge T$

$E[V(S(\tau_k \wedge T), E(\tau_k \wedge T), I(\tau_k \wedge T), I_a(\tau_k \wedge T), I_q(\tau_k \wedge T), H(\tau_k \wedge T), R(\tau_k \wedge T))] \leq V(S(0), E(0), I(0), I_a(0),$
$I_q(0), H(0), R(0)) + K(\tau_k \wedge T) + E[\int_0^{\tau_k \wedge T} \theta_1(S-1)dB_1(t) + \theta_2(E-1)dB_2(t) + \theta_3(I-1)dB_3(t) +$
$\theta_4(I_a-1)dB_4(t) + \theta_5(I_q-1)dB_5(t) + \theta_7(R-1)dB_7(t)] \leq V(S(0), E(0), I(0), I_a(0), I_q(0), H(0),$
$R(0)) + KT$

Setting $\Omega_k = \tau_k \leq T$ for $k \geq k_1$ and by $P(\tau_\infty \leq T) > \epsilon,\ P(\Omega_k) \geq \epsilon$.
It is worth noting that for every $w \in \Omega_k$, there exists at least one combination of $S(\tau_k, w)$, $E(\tau_k, w)$, $I(\tau_k, w)$, $I_a(\tau_k, w)$, $I_q(\tau_k, w)$, $H(\tau_k, w)$,$R(\tau_k, w)$ that is equal to either k or $\frac{1}{k}$ and hence $V(S(\tau_k), E(\tau_k), I(\tau_k), I_a(\tau_k), I_q(\tau_k), H(\tau_k),$
$R(\tau_k))$ is not less than $(k - 1 - \log k)$ or $(\frac{1}{k} - 1 + \log k)$.
Consequently, $V(S(\tau_k), E(\tau_k), I(\tau_k), I_a(\tau_k), I_q(\tau_k), H(\tau_k), R(\tau_k)) \geq E[(k - 1 - \log k) \wedge (\frac{1}{k} - 1 + \log k)]$.
Thus, it follows from $P(\tau_\infty \leq T) > \epsilon$ and equation (5) that
$V(S(0), E(0), I(0), I_a(0), I_q(0), H(0), R(0)) + KT \geq E[1_{\Omega_{(w)}} V(S(\tau_k), E(\tau_k), I(\tau_k), I_a(\tau_k),$
$I_q(\tau_k), H(\tau_k), R(\tau_k))] \geq \epsilon[(k - 1 - \log k) \wedge (\frac{1}{k} - 1 + \log k)]$
Here, $1_{\Omega_{(w)}}$ denotes the indicator function of Ω. By letting $k \to \infty$, we arrive at the contradiction $\infty > V(S(0), E(0), I(0), I_a(0), I_q(0), H(0), R(0)) + KT = \infty$. This implies that $\tau_\infty = \infty$ a.s., thereby completing the proof. □

5.2. Extinction of the disease

Next, we will investigate the dynamic behavior of the epidemic model to determine the conditions for long-term disease elimination. We aim to derive the conditions under which the disease will become extinct within the community. This leads us to the following lemma.

Lemma 2 (Strong Law of Large Number, (Lahrouz and Omari, 2013; Din *et al.*, 2020)**:** Let $M = \{M\}_{t \geq 0}$ be continuous and real-valued local martingale, which vanish as $t \to 0$, then $\lim_{t\to\infty} \langle M, M \rangle_t = \infty$, a.s., $\Rightarrow \lim_{t\to\infty} \frac{M_t}{\langle M,M \rangle_t} = 0$, a.s. and also, $\lim_{t\to\infty} sup \frac{\langle M,M \rangle_t}{t} < 0$ a.s., $\Rightarrow \lim_{t\to\infty} \frac{M_t}{t} = 0$, a.s.

Theorem 7: Let $(S(t), E(t), I(t), I_a(t), I_q(t), H(t), R(t))$ represent the solution of system (5) for any initial value $(S(0), E(0), I(0), I_a(0), I_q(0), H(0), R(0)) \in \mathbb{R}^7_+$. If $R^0_E < 1$, then the solution $(S(t), E(t), I(t), I_a(t), I_q(t), H(t), R(t))$ of system (5) satisfies $\lim_{t\to\infty} sup \frac{\ln E(t)}{t} \leq (\sigma + \mu_d + \frac{\theta_2^2}{2})(R^0_E - 1) < 0$ a.s., where $R^0_E = \frac{(1-k)\beta(1+\eta_1+\eta_2)}{(\sigma+\mu_d+\frac{\theta_2^2}{2})}$. So for $R^0_E < 1$ the disease will be eradicated in the long term.

Proof: Applying the $Itô$ formula to the second equation of model (5), we obtain
$d\ln E(t) = \frac{dE(t)}{E(t)} = [\frac{(1-k)}{N}\beta\frac{S}{E}(I + \eta_1 I_a + \eta_2 E) - \sigma - \mu_d - \frac{\theta_2^2}{2}]dt + \theta_2 dB_2(t)$

$$\leq [(1-k)\beta + (1-k)\beta\eta_1 + (1-k)\beta\eta_2 - \sigma - \mu_d - \tfrac{\theta_2^2}{2}]dt + \theta_2 dB_2(t)$$

$$\leq [(1-k)\beta(1+\eta_1+\eta_2) - (\sigma+\mu_d) - \tfrac{\theta_2^2}{2}]dt + \theta_2 dB_2(t)$$

Integrating the above formula from 0 to t on both sides, we obtain
$\ln E(t) - \ln E(0) \leq \int_0^t [(1-k)\beta(1+\eta_1+\eta_2) - (\sigma+\mu_d) - \frac{\theta_2^2}{2}]ds + \int_0^t \theta_2 dB_2(t)$.

According to the strong law of large numbers (Lahrouz and Omari, 2013; Khasminskii, 2011), we have, $\limsup_{t\to\infty} \frac{\theta_2}{t}\int_0^t dB_2(t) = 0$, a.s.
So, $\lim_{t\to\infty} sup\frac{\ln E(t)}{t} \leq \frac{1}{t}\int_0^t [(1-k)\beta(1+\eta_1+\eta_2) - (\sigma+\mu_d) - \frac{\theta_2^2}{2}]ds$

$$\leq \left[(1-k)\beta(1+\eta_1+\eta_2) - (\sigma+\mu_d) - \tfrac{\theta_2^2}{2}\right]$$

$$\leq (\sigma+\mu_d+\tfrac{\theta_2^2}{2})\left[\frac{(1-k)\beta(1+\eta_1+\eta_2)}{(\sigma+\mu_d+\frac{\theta_2^2}{2})} - 1\right]$$

If we choose $R_E^0 = \frac{(1-k)\beta(1+\eta_1+\eta_2)}{(\sigma+\mu_d+\frac{\theta_2^2}{2})}$, it implies $\lim_{t\to\infty} sup\frac{\ln E(t)}{t} \leq (\sigma+\mu_d+\frac{\theta_2^2}{2})[R_E^0 - 1] < 0$ if $R_E^0 < 1$.
Therefore, the above result indicates that

$$\lim_{t\to\infty} E(t) = 0 \text{ a.s.},$$

which implies that the disease will be eradicated. This completes the proof. □

5.3. Ergodic stationary distribution

When a disease spreads rapidly within a population, understanding its long-term dynamics becomes a significant concern for health officials. In order to study and address this issue mathematically, stability analysis tools are commonly utilized. Deterministic models, under certain conditions, can show the existence of an endemic equilibrium and its global asymptotic stability. However, in the context of stochastic systems, the presence of an endemic equilibrium is not guaranteed, posing challenges in predicting the persistence of the disease within the population (Din *et al.*, 2020). In our study, inspired by the work of Khasminskii (2011), we aim to investigate the existence of an ergodic stationary distribution for system (5). This analysis provides insights into the long-term persistence of the disease. The deterministic version of the system (5) can be easily obtained by setting $\theta_i = 0$ for $i = 1$ to 7, resulting in a straightforward conversion. However, it is important to note that the original stochastic model and its deterministic counterpart exhibit significant differences. Moreover, empirical evidence suggests the absence of an endemic disease state in the stochastic system, challenging the applicability of traditional linear stability analysis to assess the disease's sustained presence. Consequently, our research focuses on investigating the stationary distribution of the proposed system (5), specifically exploring the existence

of ergodic stationary components.

Let's consider the assumption that $X(t)$ is a regular time-homogeneous Markov process in $\mathbb{R}^n_+$. Mathematically, it can be represented as $dX(t) = b(X)dt + \sum_{r=1}^{k} \sigma_r dB_r(t)$, where $b(X)$ represents the drift term.
The diffusion matrix is defined as $A(X) = [a_{ij}(x)],\ a_{ij}(x) = \sum_{r=1}^{k} \sigma_r^i(x)\sigma_j^r(x)$ a.s.

Lemma 3: (Din *et al.*, 2020) The Markov process $X(t)$ has a unique stationary distribution $m(\cdot)$ if there exists a bounded domain $U \subseteq \mathbb{R}^d$ with a regular boundary such that the closure $U \in \mathcal{R}^d$ satisfies the following properties:

1. In the open domain U and some of its neighbors, the smallest eigenvalue of the diffusion matrix $A(t)$ is set far from zero.

2. If $x \in R^d U$, the mean time τ at which a path issuing from x reaches the set U is finite, and $sup_{x\in k} E\tau^x < \infty$ for every compact subset. Moreover, if $f(.)$ is a function integrable with respect to the measure π, then $\quad P\Big[\lim_{T\to\infty} \frac{1}{T}\int_0^T f(X_x(t))dt = \int_{R^d} f(x)\pi dx\Big] = 1.$

For future reference, let us define another threshold value $R_0^* = \Big[\frac{\mu_d(1-k)\beta\rho_1\sigma}{\left(\mu_d+\frac{\theta_1^2}{2}\right)\left(\sigma+\mu_d+\frac{\theta_2^2}{2}\right)\left(\alpha+\mu_d+\frac{\theta_3^2}{2}\right)}\Big]$.

Theorem 8: If $R_0^* > 1$, then a solution $(S(t), E(t), I(t), I_a(t), I_q(t), H(t), R(t))$, of system (5) is ergodic. Moreover, $\exists$ a unique stationary distribution $\pi(.)$.

Proof: First, we will demonstrate that the second condition of Lemma 3 is satisfied. To accomplish this, we will construct a non-negative C^2 function $\overline{V} : \mathbb{R}^7_+ \to \mathbb{R}_+$ such that it satisfies the following properties:
$\overline{V} = N(t) - c_1 \ln S(t) - c_2 \ln E(t) - c_3 \ln I(t)$, with $c_i \geq 0,\ i = 1(1)3$. Applying $Ito\hat{o}'s$ formula (Mao, 1997), we obtain

$$L\overline{V} = (\Pi - \mu_d N - \delta H) - c_1\Big[\frac{\Pi}{S} - \mu_d - \frac{(1-k)}{N}\beta(I + \eta_1 I_a + \eta_2 E) - \frac{\theta_1^2}{2}\Big] - c_2\Big[\frac{(1-k)}{N}\beta\frac{S}{E}(I + \eta_1 I_a + \eta_2 E) - \sigma - \mu_d - \frac{\theta_2^2}{2}\Big] - c_3\Big[\rho_1\sigma\frac{E}{I} - \alpha - \mu_d - \frac{\theta_3^2}{2}\Big]$$

$$= \Pi - \mu_d N - \delta H - c_1\tfrac{\Pi}{S} + c_1\mu_d + c_1(1-k)\beta\tfrac{I}{N} + c_1(1-k)\beta\eta_1\tfrac{I_a}{N} + c_1(1-k)\beta\eta_2\tfrac{E}{N} + c_1\tfrac{\theta_1^2}{2} - c_2(1-k)\beta\tfrac{SI}{NE} - c_2(1-k)\beta\eta_1\tfrac{SI_a}{NE} - c_2(1-k)\beta\eta_2\tfrac{SE}{NE} + c_2(\sigma+\mu_d) + c_2\tfrac{\theta_2^2}{2} - c_3\rho_1\sigma\tfrac{E}{I} + c_3(\alpha+\mu_d) + c_3\tfrac{\theta_3^2}{2}$$

$$\leq -\Big[\mu_d N + c_1\tfrac{\Pi}{S} + c_2(1-k)\beta\tfrac{SI}{NE} + c_3\rho_1\sigma\tfrac{E}{I}\Big] + \Pi + c_1(\mu_d + \tfrac{\theta_1^2}{2}) + c_2(\sigma + \mu_d + \tfrac{\theta_2^2}{2}) + c_3(\alpha + \mu_d + \tfrac{\theta_3^2}{2}) + c_1\Big[(1-k)\beta\tfrac{I}{N} + (1-k)\beta\eta_1\tfrac{I_a}{N} + (1-k)\beta\eta_2\tfrac{E}{N}\Big]$$

$$= -4\Big[\mu_d N c_1\tfrac{\Pi}{S}c_2(1-k)\beta\tfrac{SI}{NE}c_3\rho_1\sigma\tfrac{E}{I}\Big]^{\frac{1}{4}} + \Pi + c_1(\mu_d+\tfrac{\theta_1^2}{2}) + c_2(\sigma+\mu_d+\tfrac{\theta_2^2}{2}) + c_3(\alpha+\mu_d+\tfrac{\theta_3^2}{2}) + c_1\Big[(1-k)\beta\tfrac{I}{N} + (1-k)\beta\eta_1\tfrac{I_a}{N} + (1-k)\beta\eta_2\tfrac{E}{N}\Big]$$

$$= -4\Big[\mu_d(1-k)\beta\rho_1\sigma\Pi c_1c_2c_3\Big]^{\frac{1}{4}} + \Pi + c_1(\mu_d + \tfrac{\theta_1^2}{2}) + c_2(\sigma + \mu_d + \tfrac{\theta_2^2}{2}) + c_3(\alpha + \mu_d + \tfrac{\theta_3^2}{2}) + c_1\Big[(1-k)\beta\tfrac{I}{N} + (1-k)\beta\eta_1\tfrac{I_a}{N} + (1-k)\beta\eta_2\tfrac{E}{N}\Big].$$

Now we assume that, $\Pi = c_1(\mu_d + \frac{\theta_1^2}{2}) = c_2(\sigma + \mu_d + \frac{\theta_2^2}{2}) = c_3(\alpha + \mu_d + \frac{\theta_3^2}{2})$ where, $c_1 = \frac{\Pi}{\left(\mu_d+\frac{\theta_1^2}{2}\right)}, c_2 = \frac{\Pi}{\left(\sigma+\mu_d+\frac{\theta_2^2}{2}\right)}$ and $c_3 = \frac{\Pi}{\left(\alpha+\mu_d+\frac{\theta_3^2}{2}\right)}$.

So, $L\overline{V} \leq -4\Big[\frac{\mu_d(1-k)\beta\rho_1\sigma\Pi^4}{\left(\mu_d+\frac{\theta_1^2}{2}\right)\left(\sigma+\mu_d+\frac{\theta_2^2}{2}\right)\left(\alpha+\mu_d+\frac{\theta_3^2}{2}\right)}\Big]^{\frac{1}{4}} + 4\Pi + c_1\Big[(1-k)\beta\frac{I}{N} + (1-k)\beta\eta_1\frac{I_a}{N} + (1-k)\beta\eta_2\frac{E}{N}\Big] \leq -4\Pi\Big[(R_0^*)^{\frac{1}{4}} - 1\Big] + c_1\Big[(1-k)\beta\frac{I}{N} + (1-k)\beta\eta_1\frac{I_a}{N} + (1-k)\beta\eta_2\frac{E}{N}\Big]$
where, $R_0^* = \Big[\frac{\mu_d(1-k)\beta\rho_1\sigma}{\left(\mu_d+\frac{\theta_1^2}{2}\right)\left(\sigma+\mu_d+\frac{\theta_2^2}{2}\right)\left(\alpha+\mu_d+\frac{\theta_3^2}{2}\right)}\Big]$.

We define another function of the form:

$V = c_4\Big[N(t) - c_1\ln S(t) - c_2\ln E(t) - c_3\ln I(t)\Big] - \ln S(t) - \ln E(t) - \ln I(t) - \ln I_a(t) - \ln I_q(t) - \ln H(t) - \ln R(t) + N(t)$, where, $c_4 > 0$ represents a constant that will be determined later.
Therefore, $V = c_4\overline{V} - \ln S(t) - \ln E(t) - \ln I(t) - \ln I_a(t) - \ln I_q(t) - \ln H(t) - \ln R(t) + N(t)$.

According to Lemma 3 and the continuity of $\overline{V}(S, E, I, I_a, I_q, H, R)$, we can conclude that $\overline{V}(S, E, I, I_a, I_q, H, R)$ has a unique minimum value around $(S_0, E_0, I_0, I_{a_0}, I_{q_0}, H_0, R_0)$ in the interior of $\mathbb{R}_+^7$. Now we define a non-negative C^2 function $V : \mathbb{R}_+^7 \to \mathbb{R}_+$ as $V = \overline{V}(S, E, I, I_a, I_q, H, R) - \overline{V}(S_0, E_0, I_0, I_{a_0}, I_{q_0}, H_0, R_0)$.

Applying $It\hat{o}'s$ formula to V, we obtain

$$\begin{aligned}
LV &= c_4L\overline{V} - L\ln S(t) - L\ln E(t) - L\ln I(t) - L\ln I_a(t) - L\ln H(t) - L\ln R(t) + LN(t)\\
&\leq c_4\Big\{-4\Pi\Big[(R_0^*)^{\frac{1}{4}} - 1\Big] + c_1\Big[(1-k)\beta\frac{I}{N} + (1-k)\beta\eta_1\frac{I_a}{N} + (1-k)\beta\eta_2\frac{E}{N}\Big]\Big\} - \Big[\frac{\Pi}{S} - \mu_d\\
&\quad - \frac{(1-k)}{N}\beta(I + \eta_1 I_a + \eta_2 E) - \frac{\theta_1^2}{2}\Big] - \Big[\frac{(1-k)}{N}\beta\frac{S}{E}(I + \eta_1 I_a + \eta_2 E) - \sigma - \mu_d - \frac{\theta_2^2}{2}\Big] -\\
&\quad \Big[\rho_1\sigma\frac{E}{I} - \alpha - \mu_d - \frac{\theta_3^2}{2}\Big] - \Big[\rho_2\sigma\frac{E}{I_a} - \gamma_a - \mu_d - \frac{\theta_4^2}{2}\Big] - \Big[\alpha\frac{I}{H} + \alpha_q\frac{I_q}{H} - \gamma - \delta - \mu_d - \frac{\theta_6^2}{2}\Big] -\\
&\quad \Big[\gamma_a\frac{I_a}{R} + \gamma_q\frac{I_q}{R} + \gamma\frac{H}{R} - \mu_d - \frac{\theta_7^2}{2}\Big] + \Pi - \mu_d N - \delta H\\
&\leq -c_4c_5 + c_1c_4(1-k)\beta\tfrac{I}{N} + c_1c_4(1-k)\beta\eta_1\tfrac{I_a}{N} + c_1c_4(1-k)\beta\eta_2\tfrac{E}{N} - \tfrac{\Pi}{S} + \mu_d + (1-k)\beta\tfrac{I}{N} + (1-k)\beta\eta_1\tfrac{I_a}{N} + (1-k)\beta\eta_2\tfrac{E}{N} + \tfrac{\theta_1^2}{2} - (1-k)\beta\tfrac{SI}{NE} - (1-k)\beta\eta_1\tfrac{SI_a}{NE} - (1-k)\beta\eta_2\tfrac{S}{N} + \sigma + \mu_d + \tfrac{\theta_2^2}{2} - \rho_1\sigma\tfrac{E}{I} + \alpha + \mu_d + \tfrac{\theta_3^2}{2} - \rho_2\sigma\tfrac{E}{I_a} + \gamma_a + \mu_d + \tfrac{\theta_4^2}{2} - \alpha\tfrac{I}{H} - \alpha_q\tfrac{I_q}{H} + \gamma + \delta + \mu_d + \tfrac{\theta_6^2}{2} - \gamma_a\tfrac{I_a}{R} - \gamma_q\tfrac{I_q}{R} - \gamma\tfrac{H}{R} + \mu_d + \tfrac{\theta_7^2}{2} + \Pi - \mu_d N - \delta H
\end{aligned}$$

where, $c_5 = \Pi\Big[\Big(R_0^*\Big)^{\frac{1}{4}} - 1\Big] > 0$.

So, $LV \leq -c_4c_5 + (c_1c_4+1)(1-k)\beta(1+\eta_1+\eta_2) - \frac{\Pi}{S} - (1-k)\beta\frac{SI}{NE} - (1-k)\beta\eta_1\frac{SI_a}{NE} + 6\mu_d + \sigma + \alpha + \gamma_a + \gamma + \delta - \rho_1\sigma\frac{E}{I} - \rho_2\sigma\frac{E}{I_a} - \alpha\frac{I}{H} - \alpha_q\frac{I_q}{H} - \gamma_a\frac{I_a}{R} - \gamma_q\frac{I_q}{R} - \gamma\frac{H}{R} + \Pi - \mu_d N - \delta H + \frac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2}$.

We define a set as follow:

$D = \Big\{\epsilon_1 \leq S \leq \frac{1}{\epsilon_2},\ \epsilon_1 \leq E \leq \frac{1}{\epsilon_2},\ \epsilon_1 \leq I \leq \frac{1}{\epsilon_2},\ \epsilon_1 \leq I_a \leq \frac{1}{\epsilon_2},\ \epsilon_1 \leq I_q \leq \frac{1}{\epsilon_2},\ \epsilon_1 \leq H \leq \frac{1}{\epsilon_2},\ \epsilon_1 \leq R \leq \frac{1}{\epsilon_2}\Big\}$

where $\epsilon_i > 0,\ i = 1, 2$ are constants, which are very small and will be determined later. We can divide $\mathbb{R}^7_+ \setminus D$ into the following sixteen domains:

$$D_1 = \{(S, E, I, I_a, I_q, H, R) \in R^7_+, 0 < S < \epsilon_1\},$$
$$D_2 = \{(S, E, I, I_a, I_q, H, R) \in R^7_+, 0 < E < \epsilon_2, S > \epsilon_1\};$$
$$D_3 = \{(S, E, I, I_a, I_q, H, R) \in R^7_+, E > \epsilon_1, I < \epsilon_2\},$$
$$D_4 = \{(S, E, I, I_a, I_q, H, R) \in R^7_+, E > \epsilon_1, I_a < \epsilon_2\},$$
$$D_5 = \{(S, E, I, I_a, I_q, H, R) \in R^7_+, I > \epsilon_1, 0 < H < \epsilon_2\},$$
$$D_6 = \{(S, E, I, I_a, I_q, H, R) \in R^7_+, I_q > \epsilon_1, 0 < H < \epsilon_2\},$$
$$D_7 = \{(S, E, I, I_a, I_q, H, R) \in R^7_+, I_a > \epsilon_1, 0 < R < \epsilon_2\},$$
$$D_8 = \{(S, E, I, I_a, I_q, H, R) \in R^7_+, I_q > \epsilon_1, 0 < R < \epsilon_2\},$$
$$D_9 = \{(S, E, I, I_a, I_q, H, R) \in R^7_+, H > \epsilon_1, 0 < R < \epsilon_2\},$$
$$D_{10} = \{(S, E, I, I_a, I_q, H, R) \in R^7_+, S > \tfrac{1}{\epsilon_2}\},$$
$$D_{11} = \{(S, E, I, I_a, I_q, H, R) \in R^7_+, E > \tfrac{1}{\epsilon_2}\},$$
$$D_{12} = \{(S, E, I, I_a, I_q, H, R) \in R^7_+, I > \tfrac{1}{\epsilon_2}\},$$
$$D_{13} = \{(S, E, I, I_a, I_q, H, R) \in R^7_+, I_a > \tfrac{1}{\epsilon_2}\},$$
$$D_{14} = \{(S, E, I, I_a, I_q, H, R) \in R^7_+, I_q > \tfrac{1}{\epsilon_2}\},$$
$$D_{15} = \{(S, E, I, I_a, I_q, H, R) \in R^7_+, H > \tfrac{1}{\epsilon_2}\},$$
$$D_{16} = \{(S, E, I, I_a, I_q, H, R) \in R^7_+, R > \tfrac{1}{\epsilon_2}\}.$$

For all the above cases, it can be observed that there exists a positive constant $c > 0$ such that

$$LV(S, E, I, I_a, I_q, H, R) < -c,\ \forall\ (S, E, I, I_a, I_q, H, R) \in R^7_+ \setminus D.$$ (see Annexure for detail)

Let $(S, E, I, I_a, I_q, H, R) = x \in R^7_+ \setminus D$, the time τ^x at which a trajectory starting from x reaches to the set D, $\tau^n = inf\{t : |(X(t)| = n\}$ and $\tau^n(t) = min\{\tau^x, t, \tau^n\}$.

By integrating LV from 0 to $\tau^n(t)$ and using expectations, as well as applying Dynkin's formula, we have reached the conclusion that

$EV(S(\tau^n(t)), E(\tau^n(t)), I(\tau^n(t)), I_a(\tau^n(t)), I_q(\tau^n(t)), H(\tau^n(t)), R(\tau^n(t))) - V(x)$
$= E\int_0^{\tau^n(t)} LV(S(u), E(u), I(u), I_a(u), I_q(u), H(u), R(u))du$
$\leq E\int_0^{\tau^n(t)} -cdu = -cE\tau^n(t)$. By utilizing the fact that the function $V(x)$ is non-negative, we can deduce that $E\tau^n(t) \leq \frac{V(x)}{c}$.

Thus, $P(\tau_e = \infty) = 1$, which implies that the model (5) is regular. Applying the well-known Fatou's lemma, we obtain $E\tau^n(t) \leq \frac{V(x)}{c} < \infty$.

Obviously, $sup_{x\in K} E\tau^x < \infty$ where $K \subset R^7_+$. So the second condition of Lemma 3 is

satisfied. Moreover, the diffusion matrix for system (5) takes the form

$$B = \begin{bmatrix} \theta_1^2 S^2 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \theta_2^2 E^2 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \theta_3^2 I^2 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \theta_4^2 I_a^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \theta_5^2 I_q^2 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \theta_6^2 H^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \theta_7^2 R^2 \end{bmatrix}$$

$M = min_{(S,E,I,I_a,I_q,H,R)\in D}\{\theta_1^2 S^2, \theta_2^2 E^2, \theta_3^2 I^2, \theta_4^2 I_a^2, \theta_5^2 I_q^2, \theta_6^2 H^2, \theta_7^2 R^2\}$, we can obtain

$\sum_{i,j=1}^{7} a_{ij}(S,E,I,I_a,I_q,H,R)\xi_i\xi_j = \theta_1^2 S^2\xi_1^2 + \theta_2^2 E^2\xi_2^2 + \theta_3^2 I^2\xi_3^2 + \theta_4^2 I_a^2\xi_4^2 + \theta_5^2 I_q^2\xi_5^2 + \theta_6^2 H^2\xi_6^2 + \theta_7^2 R^2\xi_7^2 > M|\xi^2|$

where $\xi = (\xi_1, \xi_2, \xi_3, \xi_4, \xi_5, \xi_6, \xi_7) \in R_+^7$. Thus, the first condition of Lemma 3 is satisfied. It follows from Lemma 3 that the proposed stochastic model is ergodic with a unique stationary distribution.

□

6. Numerical simulations

In this section, we perform numerical simulations using R programming to support our analytical findings. We have taken most of the parameter values from Table 1 and demonstrated the system dynamics for both R_0 greater and less than 1. For the parameter $k = 0$ *i.e.* under no intervention, it is observed that $R_0 = 1.68(> 1)$ which implies the disease persist in the deterministic system (1). Similarly, for the parameter $k = 0.6544$ *i.e.* with intervention effect, it is observed that $R_0 = 0.5805(< 1)$ which implies the disease will die out from the deterministic system (1).

First, we have plotted the relationship $F = \frac{\beta S(I+\eta_1 I_a+\eta_2 E)}{N}$ with respect to a) S, I, b) S, H and c) I, H respectively in Figure 2(a),(b) and (c). It is observed that curve (a) exhibits a quadratic shape, curve (b) follows a sigmoidal pattern, and curve (c) shows a linear relationship. Figure 2(a) illustrates the significant dependence of F on the infection I. The three-dimensional representation reveals that for a fixed I, the shape remains relatively stable concerning S. However, altering I while keeping S constant leads to a rapid increase or decrease in the shape of F, consequently resulting in a swift change in disease propagation within the system. Moving to Figure 2(b), an initial rapid increase in F is observed due to sudden changes in S, albeit with less intensity compared to the previous scenario. However, gradual increments in S result in a slower evolution of F, leading to an initial rapid disease propagation that gradually diminishes as the susceptible population increases. Finally, Figure 2(c) depicts a gradual yet consistent rise in disease propagation as the infection rate increases within the system. This indicates that the different compartments have varying and complex impacts on the spread of new infections. For the above-mentioned parameter values together with $\eta_2 = 0.2$, we have drawn a time series diagram to visualize these two scenarios in Figure 3(a), and (b) for two different values of control parameters $k = 0$ and $k = 0.6544$ respectively. Here it is clear that all the compartments go towards a stable equilibrium. So in Figure 3(a), the susceptible population S (green) goes to stable equilibrium density approximately 29.67, the exposed E (purple), infected I (red), asymptomatic I_a (black), quarantine I_q (pink), hospitalised H (yellow) and recovery population R (light blue) goes

to stable equilibrium density approximately $(10.36, 1.6, 2.47, 1.06, 1.74, 2.95)$ respectively. It also supports Theorem 5, as $R_0 > 1$. Similarly, in Figure 3(b), the susceptible population S (green) goes to stable equilibrium density at 50, rest of the compartment dies out as time goes. It also supports Theorem 3, as $R_0 < 1$ and the DFE is $E_0(50, 0, 0, 0, 0, 0, 0)$.

Next, we have simulated the stochastic version of the model (5) through the Euler Maruyama method. To simulate the path of $S(t)$, $E(t)$, $I(t)$, $I_a(t)$, $I_q(t)$, $H(t)$ and $R(t)$ for the model (5), we fixed the initial values $(S(0), E(0), I(0), I_a(0), I_q(0), H(0), R(0)) = (40, 30, 10, 30, 12, 15, 8)$ throughout the stochastic simulation unless it stated in the figure caption. The parameter values are taken from Table 1 with $k = 0$ and intensity parameters $\theta_1 = 0.3, \theta_5 = 0.2, \theta_7 = 0.1$. In Figure 4(a), we consider the other intensity parameters $\theta_2 = 0.2, \theta_3 = 0.1, \theta_4 = 0.3, \theta_6 = 0.2$ and generated the stochastic densities for $S(t)$ (green), $E(t)$ (purple), $I(t)$ (red), $I_a(t)$ (blue), $I_q(t)$ (black), $H(t)$ (cyan) and $R(t)$ (violet). We further generated the stochastic densities corresponding to $\theta_2 = 0.4, \theta_3 = 0.4, \theta_4 = 0.3, \theta_6 = 0.4$ in Figure 4(b) and $\theta_2 = 0.4, \theta_3 = 0.4, \theta_4 = 0.6, \theta_6 = 0.4$ in Figure 4(c). In a similar way, we have also simulated the scenario in the presence of and high ($k = 0.6544$) and moderate interventions ($k = 0.4$). For high intervention we have generated the stochastic densities corresponding to $\theta_2 = 0.2, \theta_3 = 0.1, \theta_4 = 0.3, \theta_6 = 0.2$ in Figure 5(a); $\theta_2 = 0.4, \theta_3 = 0.4, \theta_4 = 0.3, \theta_6 = 0.4$ in Figure 5(b) and $\theta_2 = 0.4, \theta_3 = 0.4, \theta_4 = 0.6, \theta_6 = 0.4$ in Figure 5(c). Similarly, for low intervention we have generated the stochastic densities in Figure 6(a)-(c). We observed that all the Figures 4(a)-(c), Figures 5(a)-(c) and Figures 6(a)-(c) are stochastically bounded and have positive, unique solution converges in probability (Theorem 6). Figures 7(a)-(f) represents four different sample path and their average path of $S(t)$, $E(t)$, $I(t)$, $I_a(t)$, $I_q(t)$, $H(t)$ and $R(t)$ respectively for the stochastic model (5). The parameters are taken from Figure 4 with $\sigma_1 = 0.3, \sigma_2 = 0.2, \sigma_3 = 0.1, \sigma_4 = 0.3, \sigma_5 = 0.2, \sigma_6 = 0.2, \sigma_7 = 0.1$ *i.e.* without the presence of intervention. In Figure 7(a) (*i.e.* stochastic densities with respect to S), we observed that the one sample path have decreasing flow, others and the average density path (black) shows stable trend. Similarly, in Figure 7(b) (*i.e.*, stochastic densities with respect to E) and Figure 7(c) (*i.e.*, stochastic densities with respect to I), we observed that almost all the sample path shows a stable type of path, as does the average path (black). Figure 7(d) (*i.e.*, stochastic densities with respect to I_a) and Figure 7(e) (*i.e.*, stochastic densities with respect to I_q) shows mixed types of sample path with a larger variation and the average path (black) also reveals a stable type scenario. Various stochastic densities with respect to H and its average path also shows a stable scenario (not shown here). Although, in Figure 7(f) (*i.e.* stochastic densities with respect to R), we observed that the one sample path goes to extinction, others and the average density path (black) shows stochastic oscillating, implies the complex dynamical behavior of the system. Here it reveals there is no extinction scenario on the average run (see Figure 7(a) -(f)), although some downward trend in sample paths is observed in $S(t), I(t), I_q(t)$ and $R(t)$. Next, we generate the figures of average sample path in the presence of intervention (*i.e.* $k = 0.6544$). Following the ideas of Figure 7, we have generated Figure 8 when $R_0 < 1$. In Figure 8(a)-(b), we observed stable scenario in the sample paths as well as average path. However, a downward trend is observed in the average path (see Figure 8(c)-(f)) and in certain extent the result shows a similar behavior like the deterministic system in long run.

To get more detail on the distribution of the densities of various compartments, we have drawn histograms (see Figure 9(a-f)) of the densities at the time point 150 for 5000 runs of the system (5). The parameters are taken from Fig. 4. Here, we have observed that

some sample path shows extinction due to stochastic fluctuation in the I_a, I_q, R population. The average densities lies in the approximate range $(30, 100)$, $(20, 60)$, $(15, 26)$, $(20, 60)$, $(10, 32)$ and $(10, 28)$ for S, E, I, I_a, I_q and R respectively. Similarly, various histograms of the densities (see Figure 10(a)-(f)) at the time point 150 shows I_a, I_1, R compartments have the chance to extinct in the present scenario, although it is possible to have more probability of extinction for a large time point instead of 150 as we have already observed a sharp downtrend in the various compartments in the average run. Histograms were also calculated at time point 100 to provide enhanced understanding of the temporal dynamics (not shown here). An extinction scenario may occur for I_a at a frequency lower than that of I_q, H, and R. The S compartment exhibits a distribution with a long right tail. Furthermore, the distributions of E, I, and I_a are leptokurtic, while that of I_q is positively skewed. Moreover, we have studied the stochastic extinction of the exposed compartment (see Theorem 7) and plot R_E^0 with respect to the parameters k and θ_2. Other parameters are taken from Table 1 with $\eta_2 = 0.1$. We have drawn two heat map diagrams by varying disease transmission rate (β). In Fig. 11(a), we consider a low value of $\beta = 0.74$ and observed that moderate value of control (k) leads to $R_E^0 < 1$. Consequently its easy to control the disease in a long time. Similarly, Fig. 11(b), we consider a moderate value of $\beta = 1.74$ and observed that large value of control (k) needed to make $R_E^0 < 1$. Consequently its no so easy to control the disease in a long time as more area has R_E^0 value greater than one. Two different sample path are drawn (see Fig. 12(a),(b)) for the parameter set same as Fig. 11(a) with $k = 0.6544$ and $\theta_1 = 0.3,\ \theta_2 = 0.7,\ \theta_3 = 0.4,\ \theta_4 = 0.6,\ \theta_5 = 0.2,\ \theta_6 = 0.4,\ \theta_7 = 0.1$. We have computed the value of R_E^0 (< 1) and observed that both sample path leads to extinction.

6.1. Role of quarantine proportion to the trend of infection

Here we have numerically studied the impact of the fraction of quarantine population $\rho_3 = 1 - \rho_1 - \rho_2$ to the model (5) in terms of disease propagation. We defined a new infection term $I_{dis} = I + I_a + I_q$ and studied its long term behaviour with respect to the parameter ρ_3. We simulate the model (5) for two different values at $\rho_3 = 0.25$, $\rho_3 = 0.5$ and find the time series of I, I_a, I_q. We repeat the process for 5000 times and compute the average values *i.e.* $I^{av}, I_a^{av}, I_q^{av}$. After that we compute $I_{dis} = I^{av} + I_a^{av} + I_q^{av}$ to observe the flow of infection in the system. The quantity I_{dis} is simulated for $\rho_3 = 0.25, 0.25$ and plotted in Fig. 13(a). The time series plot $I_{dis}(t)$ for a lower value of $\rho = 0.25$ is presented in green colour and for a relatively higher value of $\rho = 0.5$ is presented in red colour. Now following Noguchi *et al.* (2011) we have performed robust sieve bootstrap approaches for linear trend detection for the generated $I_{dis}(t)$ data. As we found the p-value is very small (< 0.01) in both the case, we tried to fit linear regression models to check the slope of the trend line. The slope of green line is 0.002578 whereas for the red line its 0.003069. So comparing the slope we can say that in long term on average the disease for stochastic system with high value of ρ_3 leads to rapid fall of disease compare to the low one. In this context, it is to be noted that the first difference of $I_{dis}(t)$ *i.e.* $D(I_{dis}(t))$ is stationary (see Fig. 13(b)) in both the case due to Augmented Dickey-Fuller (ADF) test with p-value less than 0.01. Although $I_{dis}(t)$ is not stationary for both the case due to ADF test with p-values 0.8812 and 0.3716 respectively.

7. Discussion and conclusion

The World Health Organization (W.H.O., 2020) states that infectious diseases are the main reason for death in nations with low incomes. Furthermore, according to a recent report, 36% of all deaths worldwide in 2019 were attributable to communicable diseases (W.H.O., 2020). COVID-19 is a rapidly spreading infectious disease that could pose a worldwide threat. Mathematical and statistical models are useful tools for forecasting the pattern, duration, effects of different interventions, and other aspects of disease outbreaks. The present study aimed to develop an intervention-based, deterministic $SEII_aI_qHR$ epidemic model to study the dynamics of the most recent COVID-19 outbreak. Moreover, the model includes the intervention parameter k, which takes into consideration factors like vaccinations, social distancing policies, lockdowns, and other intervention tactics. Symptomatic, asymptomatic, and exposed compartments contribute to the spread of new infections. The disease circulates among the symptomatic, asymptomatic, and quarantine populations in proportions represented by the variables ρ_1, ρ_2, and $(1-\rho_1-\rho_2)$, respectively. We explored the positive invariance and boundedness of every forward solution of the model. Furthermore, using the basic reproduction number (R_0), we explore the local and global stability of the unique disease-free equilibrium of the model. In addition, we also studied the existence and local stability of the endemic equilibrium of the model. The deterministic model offers a general understanding of the spread of disease, but it ignores uncertain variables like immigration, human behavior, the effects of the climate, and other random factors. Therefore, we developed a stochastic version of the $SEII_aI_qHR$ model with a frequency-dependent force of infection and intervention to study the dynamics of the disease transmission in the context of changing environmental and population factors. Moreover, we calculated the transition probabilities to investigate the drift and diffusion components of the SDE while developing the stochastic $SEII_aI_qHR$ model. We then discussed some fundamental properties of the model, including the existence of a unique positive global solution with probability one, which shows that the problem is well-behaved. We also analytically found that the criteria $R_E^0 < 1$ leads to disease extinction in the long term. Additionally, we found the ergodic stationary distribution and the extinction conditions of the disease by constructing an appropriate Lyapunov function and using the Iô formula. Finally, we validated the theoretical findings by generating several numerical solutions to the models. Furthermore, we numerically determined the relationship between the disease transfer function F and various disease compartments of the model (5). Our findings suggest the possibility of three different types of scenarios, e.g., linear, sigmoidal, and quadratic. Furthermore, for two different scenarios, $R_0 < 1$ (stability of the DFE) and $R_0 > 1$ (stability of the EE), we generated time-series diagrams of densities by varying the control parameter k. In addition, to visualize different sample paths, we simulated the SDE model by varying the intervention strength and intensity parameters. The results of our study indicate that the disease does not extinct in the majority of cases. However, the average density of the sample path in the presence of intervention shows a decline in average for the disease compartments compare to without intervention scenario. We have drawn multiple histograms and compared those in two distinct scenarios to see how the densities of various compartments are distributed at a given time. In order to observe the extension scenario, we additionally display the R_E^0 heat map in the (k, θ_2) plane. To calculate R_E^0, two distinct values of disease transmission—low and high—are used. It is noted that the values roughly fall between $(0, 2.5)$ and $(0, 5.5)$, respectively. Lastly, our numerical analysis has demonstrated the positive impact of quarantine

proportions on the infection trend.

In conclusion, The study has mainly two aspects: (1) To study the deterministic aspects of the model and observe the disease propagation and impact of intervention. (2) To formulate the stochastic version of the model and observe the impact of noise, intervention and quarantine proportion in the disease propagation, extinction and ergodic stationary distribution. Here we found that as the intensity of intervention increases, the number of infected patients decreases. This means that intervention plays important roles in the outbreak of sudden infectious diseases. For example, media reports can be used to provide the public with information about the current situation of the epidemic and the effective prevention and control measures proposed by experts. Outbreaks of infectious diseases have led to a dramatic increase in interventions like media, self protection, containment zone, *etc.*, which in turn can help raise awareness and change their behaviors for better implementation of mitigation measures. People will adopt relatively conservative behaviors to reduce the possibility of infection, and individual behavior can effectively delay the peak period of infectious disease outbreaks and reduce the severity of infectious disease outbreaks. However, a part of this study only focuses on the qualitative analysis of the stochastic models. The estimation of some key parameters and studying the distribution of intervention scenario will be an interesting study for the future work.

Acknowledgements

We would like to thank the editor and an anonymous referee for pointing out some important issues which have undoubtedly enhanced the quality of the work.

Conflict of interest

The authors do not have any financial or non-financial conflict of interest to declare for the research work included in this article.

References

Allen, L. J. (2008). An introduction to stochastic epidemic models. *Mathematical Epidemiology*, , 81–130.

Allen, L. J. (2017). A primer on stochastic epidemic models: Formulation, numerical simulation, and analysis. *Infectious Disease Modelling*, **2(2)**, 128–142.

Beddington, J. and May, R. (1977). Harvesting natural populations in a randomly fluctuating environment. *Science*, **197**, 463–465.

Bhatt, S., Gething, P. W., Brady, O. J., Messina, J. P., Farlow, A. W., Moyes, C. L., Drake, J. M., Brownstein, J. S., Hoen, A. G., Sankoh, O., and Others (2013). The global distribution and burden of dengue. *Nature*, **496**, 504–507.

Cai, Y., Kang, Y., and Wang, W. (2017). A stochastic sirs epidemic model with nonlinear incidence rate. *Applied Mathematics and Computation*, **305**, 221–240.

Cai, Y., Wang, X., Wang, W., and Zhao, M. (2013). Stochastic dynamics of an sirs epidemic model with ratio-dependent incidence rate. In *Abstract and Applied Analysis*, volume 2013. Hindawi.

Castillo-Chavez, C., Feng, Z., and Huang, W. (2002). *On the Computation of R0 and Its Role in Global Stability. In: Castillo-Chavez C, van den Driessche P, Kirschner D, Yakubu A.-A, Editors. Mathematical Approaches for Emerging and Reemerging Infection Diseases: An Introduction.* Springer, New York.

Chen, H., Tan, X., Wang, J., Qin, W., and Luo, W. (2023). Stochastic dynamics of a virus variant epidemic model with double inoculations. *Mathematics*, **11**, Doi:10.3390/math11071712.

Choisy, M., Guégan, J.-F., and Rohani, P. (2007). Mathematical modeling of infectious diseases dynamics. In M, T., editor, *Encyclopedia of Infectious Diseases: Modern Methodologies*, volume 379, pages 379–404. Chichester, USA: John Wiley & Sons.

Din, A., Khan, A., and Baleanu, D. (2020). Stationary distribution and extinction of stochastic coronavirus (covid-19) epidemic model. *Chaos Solitons Fractals*, **139**, 110036.

Din, A., Khan, T., Li, Y., Tahir, H., Khan, A., and Khan, W. A. (2021). Mathematical analysis of dengue stochastic epidemic model. *Results in Physics*, **20**, 103719.

Ding, T. and Zhang, T. (2022). Asymptotic behavior of the solutions for a stochastic sirs model with information intervention. *Mathematical Biosciences and Engineering*, **19**, 6940–6961.

Driessche, V. and Watmough, J. (2002). Reproduction numbers and sub-threshold endemic equilibria for compartmental models of disease transmission. *Mathematical Biosciences*, **180**, 29–48.

Hethcote, H. (2000). The mathematics of infectious diseases. *SIAM review*, **42**, 599–653.

Hussain, S., Tunç, O., ur Rahman, G., Khan, H., and Nadia, E. (2023). Mathematical analysis of stochastic epidemic model of mers-corona & application of ergodic theory. *Mathematics and Computers in Simulation*, **207**, 130–150.

Ji, C. and Jiang, D. (2014). Threshold behaviour of a stochastic sir model. *Applied Mathematical Model*, **38**, 5067–5079.

Jiang, D., Ji, C., Shi, N., and Yu, J. (2010). The long time behavior of di sir epidemic model with stochastic perturbation. *Mathematical Analysis and Applications*, **372**, 162–180.

Khasminskii, R. (2011). *Stochastic Stability of Differential Equations.* Springer Science and Business Media, New York.

Lahrouz, A. and Omari, L. (2013). Extinction and stationary distribution of a stochastic sirs epidemic model with non-linear incidence. *Statistics and Probability Letters*, **83**, 960–968.

Li, Q., Guan, X., Wu, P., and Others (2020). Early transmission dynamics in wuhan, china, of novel coronavirus-infected pneumonia. *The New England Journal of Medicine*, , https://doi.org/10.1056/NEJMoa2001316.

Mao, X. (1997). *Stochastic Differential Equations and Applications.* Horwood, Chichester.

Mao, X. (2007). *Stochastic Differential Equations and Applications.* Elsevier.

Mao, X., Marion, G., and Renshaw, E. (2002). Environmental brownian noise suppresses explosions in population dynamics. *Stochastic Processes and Their Applications*, **97**, 95–110.

Noguchi, K., Gel, Y., and Duguay, C. (2011). Bootstrap-based tests for trends in hydrological time series,with application to ice phenology data. *Journal of Hydrology*, **410**, 150–61.

Oksendal, B. (2006). Stochastic differential equations: An introduction with applications. *Journal of American Statistical Association*, **51**, 1721–1732.

Oksendal, B. (2013). *Stochastic Differential Equations: An Introduction with Applications*. Springer Science & Business Media.

Rao, F., Wang, W., and Li, Z. (2012). Stability analysis of an epidemic model with diffusion and stochastic perturbation. *Communications in Nonlinear Science and Numerical Simulation*, **17**, 2551–2563.

Senapati, A., Rana, S., Das, T., and Chattopadhyay, J. (2021). Impact of intervention on the spread of covid-19 in india: A model based study. *Journal of Theoretical Biology*, **523**, 110711.

Shi, Z. and Jiang, D. (2023). Dynamics and density function of a stochastic covid-19 epidemic model with ornstein–uhlenbeck process. *Nonlinear Dynamics*, **111**, 18559–18584.

Sun, J., Gao, M., and Jiang, D. (2022). Threshold dynamics and the density function of the stochastic coronavirus epidemic model. *Fractal and Fractional*, **6**, 245.

Tan, Y., Cai, Y., Wang, X., Peng, Z., Wang, K., Yao, R., and Wang, W. (2023). Stochastic dynamics of an sis epidemiological model with media coverage. *Mathematics and Computers in Simulation*, **204**, 1–27.

Tang, B., Wang, X., Li, Q., and Others (2020). Estimation of the transmission risk of the 2019-ncov and its implication for public health interventions. *Journal of Clinical Medicine*, **9**, 462.

Thomas, C. and Shelemyahu, Z. (1989). Introduction to stochastic differential equations. *Journal of the American Statistical Association*, **84**, 1104.

Tuckwell, H. C. and Williams, R. J. (2007). Some properties of a simple stochastic epidemic model of sir type. *Mathematical Biosciences*, **208**, 76–97.

Ullah, R., Al Mdallal, Q., Khan, T., Ullah, R., Al Alwan, B., Faiz, F., and Zhu, Q. (2023). The dynamics of novel corona virus disease via stochastic epidemiological model with vaccination. *Scientific Reports*, **13**, 3805.

Wearing, H. J., Rohani, P., and Keeling, M. J. (2005). Appropriate models for the management of infectious diseases. *PLoS Medicine*, **2**, e174.

W.H.O. (2020). The top 10 causes of death. `https://www.who.int/news-room/fact-sheets/detail/the-top-10-causes-of-death`. Published on: 09-12-2020.

W.H.O. (2022). Middle east respiratory syndrome coronavirus (mers-cov). `https://www.who.int/news-room/fact-sheets/detail/middle-east-respiratory-syndrome-coronavirus-(mers-cov)`. Published on: 05-08-2022.

W.H.O. (2023a). Malaria. `https://www.who.int/news-room/fact-sheets/detail/malaria`. Published on: 29-03-2023.

W.H.O. (2023b). Who coronavirus (covid-19) dashboard. `https://covid19.who.int/`. Published on: 27-05-2023.

Yanan, Z. and Daqing, J. (2014). The threshold of a stochastic sis epidemic model with vaccination. *Applied Mathematics and Computation*, **243**, 718–727.

ANNEXURE

Expression of $T(X, I')$, $G(X, I')$, A and $\hat{G}(X, I)$ used in section 3.4.

$$T(X, I') = \begin{bmatrix} \Pi - (1-k)\frac{\beta S}{N}(I + \eta_1 I_a + \eta_2 E) - \mu_d S \\ \\ \gamma_a I_a + \gamma_q I_q + \gamma H - \mu_d R \end{bmatrix},$$

$$G(X, I') = \begin{bmatrix} (1-k)\frac{\beta S}{N}(I + \eta_1 I_a + \eta_2 E) - \sigma E - \mu_d E \\ \\ \rho_1 \sigma E - \alpha I - \mu_d I \\ \\ \rho_2 \sigma E - \gamma_a I_a - \mu_d I_a \\ \\ (1 - \rho_1 - \rho_2)\sigma E - (\alpha_q + \gamma_q) I_q - \mu_d I_q \\ \\ \alpha I + \alpha_q I_q - (\gamma + \delta) H - \mu_d H \end{bmatrix}.$$

$$A = \begin{bmatrix} -(\mu_d + \sigma) + (1-k)\beta\eta_2 & (1-k)\beta & (1-k)\beta\eta_1 & 0 & 0 \\ \\ \rho_1 \sigma & -(\alpha + \mu_d) & 0 & 0 & 0 \\ \\ \rho_2 \sigma & 0 & -(\gamma_a + \mu_d) & 0 & 0 \\ \\ (1 - \rho_1 - \rho_2)\sigma & 0 & 0 & -(\alpha_q + \gamma_q + \mu_d) & 0 \\ \\ 0 & \alpha & 0 & \alpha_q & -(\gamma + \delta + \mu_d) \end{bmatrix},$$

$$\hat{G}(X, I) = \begin{bmatrix} (1-k)\beta\eta_2 E(1 - \frac{S}{N}) + (1-k)\beta I(1 - \frac{S}{N}) + (1-k)\beta\eta_1 I_a(1 - \frac{S}{N}) \\ \\ 0 \\ \\ 0 \\ \\ 0 \end{bmatrix}.$$

Calculations used in section 3.5.

$$S^* = \frac{\Pi}{\lambda^* + \mu_d},\ E^* = \frac{\lambda^* S^*}{k_1},\ I^* = \frac{\rho_1\sigma\lambda^* S^*}{k_1k_2},\ I_a^* = \frac{\rho_2\sigma\lambda^* S^*}{k_1k_3},\ I_q^* = \frac{(1-\rho_1-\rho_2)\sigma\lambda^* S^*}{k_1k_4},\ H^* = \frac{\alpha\rho_1\sigma\lambda^* S^*}{k_1k_2k_5} + \frac{\alpha_q(1-\rho_1-\rho_2)\sigma\lambda^* S^*}{k_1k_4k_5},\ R^* = \frac{\gamma_a\rho_2\sigma\lambda^* S^*}{\mu_d k_1k_3} + \frac{\gamma_q(1-\rho_1-\rho_2)\sigma\lambda^* S^*}{\mu_d k_1k_4} + \frac{\gamma\alpha\rho_1\sigma\lambda^* S^*}{\mu_d k_1k_2k_5} + \frac{\gamma\alpha_q(1-\rho_1-\rho_2)\sigma\lambda^* S^*}{\mu_d k_1k_4k_5}.$$

Calculations used in section 3.6. From the model (1), we have

$$I^* = \frac{\rho_1\sigma E^*}{\alpha+\mu_d},\ I_a^* = \frac{\rho_2\sigma E^*}{\gamma_a+\mu_d},\ I_q^* = \frac{(1-\rho_1-\rho_2)\sigma E^*}{(\alpha_q+\gamma_q+\mu_d)},\ H^* = \frac{1}{\gamma+\delta+\mu_d}\Big(\frac{\alpha\rho_1\sigma}{\alpha+\mu_d} + \frac{\alpha_q\rho_2\sigma}{\gamma_a+\mu_d}\Big)E^*,$$
$$R^* = \Big[\frac{\gamma_a\rho_2\sigma}{\mu_d(\gamma_a+\mu_d)} + \frac{\gamma_q\sigma(1-\rho_1-\rho_2)}{\mu_d(\alpha_q+\gamma_q+\mu_d)} + \frac{\gamma}{\mu_d(\gamma+\delta+\mu_d)}\Big(\frac{\alpha\rho_1\sigma}{\alpha+\mu_d} + \frac{\alpha_q\rho_2\sigma}{\gamma_a+\mu_d}\Big)\Big]E^*.$$

$$I^* = \frac{\rho_1\sigma E^*}{\alpha+\mu_d},\ I_a^* = \frac{\rho_2\sigma E^*}{\gamma_a+\mu_d},\ I_q^* = \frac{(1-\rho_1-\rho_2)\sigma E^*}{(\alpha_q+\gamma_q+\mu_d)},\ H^* = \Big(\frac{\frac{\alpha\rho_1\sigma}{\alpha+\mu_d}+\frac{\alpha_q\rho_2\sigma}{\gamma_a+\mu_d}}{\gamma+\delta+\mu_d}\Big)E^*,$$
$$R^* = \frac{\frac{\gamma_a\rho_2\sigma}{\gamma_a+\mu_d}+\frac{\gamma_q\sigma(1-\rho_1-\rho_2)}{(\alpha_q+\gamma_q+\mu_d)}+\gamma\big(\frac{\frac{\alpha\rho_1\sigma}{\alpha+\mu_d}+\frac{\alpha_q\rho_2\sigma}{\gamma_a+\mu_d}}{\gamma+\delta+\mu_d}\big)}{\mu_d}E^*$$
$$N = S + E + I + I_a + I_q + H + R$$
$$N = S + E\Big(1 + \frac{\rho_1\sigma}{\alpha+\mu_d} + \frac{\rho_2\sigma}{\gamma_a+\mu_d} + \frac{(1-\rho_1-\rho_2)\sigma}{(\alpha_q+\gamma_q+\mu_d)} + \frac{\frac{\alpha\rho_1\sigma}{\alpha+\mu_d}+\frac{\alpha_q\rho_2\sigma}{\gamma_a+\mu_d}}{\gamma+\delta+\mu_d} + \frac{\frac{\gamma_a\rho_2\sigma}{\gamma_a+\mu_d}+\frac{\gamma_q\sigma(1-\rho_1-\rho_2)}{(\alpha_q+\gamma_q+\mu_d)}+\gamma\big(\frac{\frac{\alpha\rho_1\sigma}{\alpha+\mu_d}+\frac{\alpha_q\rho_2\sigma}{\gamma_a+\mu_d}}{\gamma+\delta+\mu_d}\big)}{\mu_d}\Big)$$
$$N = S + m_1E,$$

where, $m_1 = \Big(1 + \frac{\rho_1\sigma}{\alpha+\mu_d} + \frac{\rho_2\sigma}{\gamma_a+\mu_d} + \frac{(1-\rho_1-\rho_2)\sigma}{(\alpha_q+\gamma_q+\mu_d)} + \frac{\frac{\alpha\rho_1\sigma}{\alpha+\mu_d}+\frac{\alpha_q\rho_2\sigma}{\gamma_a+\mu_d}}{\gamma+\delta+\mu_d} + \frac{\frac{\gamma_a\rho_2\sigma}{\gamma_a+\mu_d}+\frac{\gamma_q\sigma(1-\rho_1-\rho_2)}{(\alpha_q+\gamma_q+\mu_d)}+\gamma\big(\frac{\frac{\alpha\rho_1\sigma}{\alpha+\mu_d}+\frac{\alpha_q\rho_2\sigma}{\gamma_a+\mu_d}}{\gamma+\delta+\mu_d}\big)}{\mu_d}\Big)$

$$\Rightarrow \frac{(1-k)\beta S}{S+m_1E}(I + \eta_1 I_a + \eta_2 E) = E(\sigma+\mu_d)$$
$$\Rightarrow \frac{(1-k)\beta S}{S+m_1E}\Big(\frac{\rho_1\sigma}{\alpha+\mu_d} + \eta_1\frac{\rho_2\sigma}{\gamma_a+\mu_d} + \eta_2\Big) = (\sigma+\mu_d)$$
$$\Rightarrow (1-k)\beta S\Big(\frac{\rho_1\sigma}{\alpha+\mu_d} + \eta_1\frac{\rho_2\sigma}{\gamma_a+\mu_d} + \eta_2\Big) = (S+m_1E)(\sigma+\mu_d)$$
$$\Rightarrow S\Big[(1-k)\beta\Big(\frac{\rho_1\sigma}{\alpha+\mu_d} + \eta_1\frac{\rho_2\sigma}{\gamma_a+\mu_d} + \eta_2\Big) - (\sigma+\mu_d)\Big] = m_1(\sigma+\mu_d)E$$
$$\Rightarrow S^* = \frac{m_1(\sigma+\mu_d)}{\big[(1-k)\beta\big(\frac{\rho_1\sigma}{\alpha+\mu_d}+\eta_1\frac{\rho_2\sigma}{\gamma_a+\mu_d}+\eta_2\big)-(\sigma+\mu_d)\big]}E^*$$
$$\text{Now, } N = \Big(\frac{m_1(\sigma+\mu_d)}{\big[(1-k)\beta\big(\frac{\rho_1\sigma}{\alpha+\mu_d}+\eta_1\frac{\rho_2\sigma}{\gamma_a+\mu_d}+\eta_2\big)-(\sigma+\mu_d)\big]} + m_1\Big)E$$
$$\Rightarrow N = m_2E; \text{where, } m_2 = \Big(\frac{m_1(\sigma+\mu_d)}{\big[(1-k)\beta\big(\frac{\rho_1\sigma}{\alpha+\mu_d}+\eta_1\frac{\rho_2\sigma}{\gamma_a+\mu_d}+\eta_2\big)-(\sigma+\mu_d)\big]} + m_1\Big)$$
$$\text{Again, } \Pi = S\Big[\mu_d + \frac{(1-k)\beta}{m_2}\Big(\frac{\rho_1\sigma}{\alpha+\mu_d} + \eta_1\frac{\rho_2\sigma}{\gamma_a+\mu_d} + \eta_2\Big)\Big]$$
$$\Rightarrow E^* = \frac{\Pi}{\big[\mu_d+\frac{(1-k)\beta}{m_2}\big(\frac{\rho_1\sigma}{\alpha+\mu_d}+\eta_1\frac{\rho_2\sigma}{\gamma_a+\mu_d}+\eta_2\big)\big]}\frac{\big[(1-k)\beta\big(\frac{\rho_1\sigma}{\alpha+\mu_d}+\eta_1\frac{\rho_2\sigma}{\gamma_a+\mu_d}+\eta_2\big)-(\sigma+\mu_d)\big]}{m_1(\sigma+\mu_d)}$$

Proof of $LV < 0$ for $(S, E, I, I_a, I_q, H, R) \in D_i,\ i = 1(1)16$ used in Theorem 8.

Case I: $(S, E, I, I_a, I_q, H, R) \in D_1$

$$LV \le -c_4c_5 + (c_1c_4+1)(1-k)\beta(1+\eta_1+\eta_2) - \frac{\Pi}{S} - (1-k)\beta\frac{SI}{NE} - (1-k)\beta\eta_1\frac{SI_a}{NE} + 6\mu_d + \sigma + \alpha + \gamma_a + \gamma + \delta - \rho_1\sigma\frac{E}{I} - \rho_2\sigma\frac{E}{I_a} - \alpha\frac{I}{H} - \alpha_q\frac{I_q}{H} - \gamma_a\frac{I_a}{R} - \gamma_q\frac{I_q}{R} - \gamma\frac{H}{R} + \Pi - \mu_dN - \delta H + \frac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2}$$
$$\le (c_1c_4+1)(1-k)\beta(1+\eta_1+\eta_2) + 6\mu_d + \sigma + \alpha + \gamma_a + \gamma + \delta + \Pi + \frac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2} - \frac{\Pi}{S}$$

$$\leq (c_1c_4+1)(1-k)\beta(1+\eta_1+\eta_2)+6\mu_d+\sigma+\alpha+\gamma_a+\gamma+\delta+\Pi+\tfrac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2}-\tfrac{\Pi}{\epsilon_1}$$

Let $\epsilon_1 > 0$ be as sufficiently small so that, $(c_1c_4+1)(1-k)\beta(1+\eta_1+\eta_2)+6\mu_d+\sigma+\alpha+\gamma_a+\gamma+\delta+\Pi+\frac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2}-\frac{\Pi}{\epsilon_1}<0$.

In such case, we have $LV < 0$.

Case II: $(S, E, I, I_a, I_q, H, R) \in D_2$

$$LV \leq -c_4c_5+(c_1c_4+1)(1-k)\beta(1+\eta_1+\eta_2)-\tfrac{\Pi}{S}-(1-k)\beta\tfrac{SI}{NE}-(1-k)\beta\eta_1\tfrac{SI_a}{NE}+6\mu_d+\sigma+\alpha+\gamma_a+\gamma+\delta-\rho_1\sigma\tfrac{E}{I}-\rho_2\sigma\tfrac{E}{I_a}-\alpha\tfrac{I}{H}-\alpha_q\tfrac{I_q}{H}-\gamma_a\tfrac{I_a}{R}-\gamma_q\tfrac{I_q}{R}-\gamma\tfrac{H}{R}+\Pi-\mu_dN-\delta H+\tfrac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2}$$
$$\leq (c_1c_4+1)(1-k)\beta(1+\eta_1+\eta_2)+6\mu_d+\sigma+\alpha+\gamma_a+\gamma+\delta+\Pi+\tfrac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2}-(1-k)\beta\tfrac{S}{E}-(1-k)\beta\eta_1\tfrac{S}{E}$$
$$\leq (c_1c_4+1)(1-k)\beta(1+\eta_1+\eta_2)+6\mu_d+\sigma+\alpha+\gamma_a+\gamma+\delta+\Pi+\tfrac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2}-(1-k)\beta\tfrac{\epsilon_1}{\epsilon_2}-(1-k)\beta\eta_1\tfrac{\epsilon_1}{\epsilon_2}$$

Let $\epsilon_1 > \epsilon_2^2$, very small, such that $(c_1c_4+1)(1-k)\beta(1+\eta_1+\eta_2)+6\mu_d+\sigma+\alpha+\gamma_a+\gamma+\delta+\Pi+\frac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2}-(1-k)\beta\frac{\epsilon_1}{\epsilon_2}-(1-k)\beta\eta_1\frac{\epsilon_1}{\epsilon_2}<0$.

In such case, we have $LV < 0$.

Case III: $(S, E, I, I_a, I_q, H, R) \in D_3$

$$LV \leq -c_4c_5+(c_1c_4+1)(1-k)\beta(1+\eta_1+\eta_2)-\tfrac{\Pi}{S}-(1-k)\beta\tfrac{SI}{NE}-(1-k)\beta\eta_1\tfrac{SI_a}{NE}+6\mu_d+\sigma+\alpha+\gamma_a+\gamma+\delta-\rho_1\sigma\tfrac{E}{I}-\rho_2\sigma\tfrac{E}{I_a}-\alpha\tfrac{I}{H}-\alpha_q\tfrac{I_q}{H}-\gamma_a\tfrac{I_a}{R}-\gamma_q\tfrac{I_q}{R}-\gamma\tfrac{H}{R}+\Pi-\mu_dN-\delta H+\tfrac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2}$$
$$\leq (c_1c_4+1)(1-k)\beta(1+\eta_1+\eta_2)+6\mu_d+\sigma+\alpha+\gamma_a+\gamma+\delta+\Pi+\tfrac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2}-\rho_1\sigma\tfrac{E}{I}$$
$$\leq (c_1c_4+1)(1-k)\beta(1+\eta_1+\eta_2)+6\mu_d+\sigma+\alpha+\gamma_a+\gamma+\delta+\Pi+\tfrac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2}-\rho_1\sigma\tfrac{\epsilon_1}{\epsilon_2}$$

Let $\epsilon_1 > \epsilon_2^2$, very small, such that $(c_1c_4+1)(1-k)\beta(1+\eta_1+\eta_2)+6\mu_d+\sigma+\alpha+\gamma_a+\gamma+\delta+\Pi+\frac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2}-\rho_1\sigma\frac{\epsilon_1}{\epsilon_2}<0$.

In such case, we have $LV < 0$.

Case IV: $(S, E, I, I_a, I_q, H, R) \in D_4$

$$LV \leq -c_4c_5+(c_1c_4+1)(1-k)\beta(1+\eta_1+\eta_2)-\tfrac{\Pi}{S}-(1-k)\beta\tfrac{SI}{NE}-(1-k)\beta\eta_1\tfrac{SI_a}{NE}+6\mu_d+\sigma+\alpha+\gamma_a+\gamma+\delta-\rho_1\sigma\tfrac{E}{I}-\rho_2\sigma\tfrac{E}{I_a}-\alpha\tfrac{I}{H}-\alpha_q\tfrac{I_q}{H}-\gamma_a\tfrac{I_a}{R}-\gamma_q\tfrac{I_q}{R}-\gamma\tfrac{H}{R}+\Pi-\mu_dN-\delta H+\tfrac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2}$$
$$\leq (c_1c_4+1)(1-k)\beta(1+\eta_1+\eta_2)+6\mu_d+\sigma+\alpha+\gamma_a+\gamma+\delta+\Pi+\tfrac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2}-\rho_2\sigma\tfrac{E}{I_a}$$
$$\leq (c_1c_4+1)(1-k)\beta(1+\eta_1+\eta_2)+6\mu_d+\sigma+\alpha+\gamma_a+\gamma+\delta+\Pi+\tfrac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2}-\rho_2\sigma\tfrac{\epsilon_1}{\epsilon_2}$$

Let $\epsilon_1 = \epsilon_2^2$, choose$\epsilon_1 > 0$ small enough such that, $(c_1c_4+1)(1-k)\beta(1+\eta_1+\eta_2)+6\mu_d+\sigma+\alpha+\gamma_a+\gamma+\delta+\Pi+\frac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2}-\rho_2\sigma\frac{\epsilon_1}{\epsilon_2}<0$.

For this case, we get $LV < 0$.

Case V: $(S, E, I, I_a, I_q, H, R) \in D_5$

$$LV \leq -c_4c_5+(c_1c_4+1)(1-k)\beta(1+\eta_1+\eta_2)-\tfrac{\Pi}{S}-(1-k)\beta\tfrac{SI}{NE}-(1-k)\beta\eta_1\tfrac{SI_a}{NE}+6\mu_d+\sigma+\alpha+\gamma_a+\gamma+\delta-\rho_1\sigma\tfrac{E}{I}-\rho_2\sigma\tfrac{E}{I_a}-\alpha\tfrac{I}{H}-\alpha_q\tfrac{I_q}{H}-\gamma_a\tfrac{I_a}{R}-\gamma_q\tfrac{I_q}{R}-\gamma\tfrac{H}{R}+\Pi-\mu_dN-\delta H+\tfrac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2}$$
$$\leq (c_1c_4+1)(1-k)\beta(1+\eta_1+\eta_2)+6\mu_d+\sigma+\alpha+\gamma_a+\gamma+\delta+\Pi+\tfrac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2}-\alpha\tfrac{I}{H}$$
$$\leq (c_1c_4+1)(1-k)\beta(1+\eta_1+\eta_2)+6\mu_d+\sigma+\alpha+\gamma_a+\gamma+\delta+\Pi+\tfrac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2}-\alpha\tfrac{\epsilon_1}{\epsilon_2}$$

Let $\epsilon_1 = \epsilon_2^2$ be as sufficiently small so that, $(c_1c_4+1)(1-k)\beta(1+\eta_1+\eta_2)+6\mu_d+\sigma+\alpha+\gamma_a+\gamma+\delta+\Pi+\frac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2}-\alpha\frac{\epsilon_1}{\epsilon_2}<0$. Here we get $LV<0$.

Case VI: $(S,E,I,I_a,I_q,H,R)\in D_6$

$$\begin{aligned}LV &\leq -c_4c_5+(c_1c_4+1)(1-k)\beta(1+\eta_1+\eta_2)-\tfrac{\Pi}{S}-(1-k)\beta\tfrac{SI}{NE}-(1-k)\beta\eta_1\tfrac{SI_a}{NE}+6\mu_d+\sigma+\\ &\alpha+\gamma_a+\gamma+\delta-\rho_1\sigma\tfrac{E}{I}-\rho_2\sigma\tfrac{E}{I_a}-\alpha\tfrac{I}{H}-\alpha_q\tfrac{I_q}{H}-\gamma_a\tfrac{I_a}{R}-\gamma_q\tfrac{I_q}{R}-\gamma\tfrac{H}{R}+\Pi-\mu_dN-\delta H+\tfrac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2}\\ &\leq (c_1c_4+1)(1-k)\beta(1+\eta_1+\eta_2)+6\mu_d+\sigma+\alpha+\gamma_a+\gamma+\delta+\Pi+\tfrac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2}-\alpha_q\tfrac{I_q}{H}\\ &\leq (c_1c_4+1)(1-k)\beta(1+\eta_1+\eta_2)+6\mu_d+\sigma+\alpha+\gamma_a+\gamma+\delta+\Pi+\tfrac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2}-\alpha_q\tfrac{\epsilon_1}{\epsilon_2}\end{aligned}$$

Let $\epsilon_1=\epsilon_2^2$ be as sufficiently small so that, $(c_1c_4+1)(1-k)\beta(1+\eta_1+\eta_2)+6\mu_d+\sigma+\alpha+\gamma_a+\gamma+\delta+\Pi+\frac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2}-\alpha_q\frac{\epsilon_1}{\epsilon_2}<0$.

Therefore, we have $LV<0$.

Case VII: $(S,E,I,I_a,I_q,H,R)\in D_7$

$$\begin{aligned}LV &\leq -c_4c_5+(c_1c_4+1)(1-k)\beta(1+\eta_1+\eta_2)-\tfrac{\Pi}{S}-(1-k)\beta\tfrac{SI}{NE}-(1-k)\beta\eta_1\tfrac{SI_a}{NE}+6\mu_d+\sigma+\\ &\alpha+\gamma_a+\gamma+\delta-\rho_1\sigma\tfrac{E}{I}-\rho_2\sigma\tfrac{E}{I_a}-\alpha\tfrac{I}{H}-\alpha_q\tfrac{I_q}{H}-\gamma_a\tfrac{I_a}{R}-\gamma_q\tfrac{I_q}{R}-\gamma\tfrac{H}{R}+\Pi-\mu_dN-\delta H+\tfrac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2}\\ &\leq (c_1c_4+1)(1-k)\beta(1+\eta_1+\eta_2)+6\mu_d+\sigma+\alpha+\gamma_a+\gamma+\delta+\Pi+\tfrac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2}-\gamma_a\tfrac{I_a}{R}\\ &\leq (c_1c_4+1)(1-k)\beta(1+\eta_1+\eta_2)+6\mu_d+\sigma+\alpha+\gamma_a+\gamma+\delta+\Pi+\tfrac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2}-\gamma_a\tfrac{\epsilon_1}{\epsilon_2}\end{aligned}$$

Let $\epsilon_1=\epsilon_2^2$, choose$\epsilon_1>0$ small enough such that, $(c_1c_4+1)(1-k)\beta(1+\eta_1+\eta_2)+6\mu_d+\sigma+\alpha+\gamma_a+\gamma+\delta+\Pi+\frac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2}-\gamma_a\frac{\epsilon_1}{\epsilon_2}<0$.

In such case, we have $LV<0$.

Case VIII: $(S,E,I,I_a,I_q,H,R)\in D_8$

$$\begin{aligned}LV &\leq -c_4c_5+(c_1c_4+1)(1-k)\beta(1+\eta_1+\eta_2)-\tfrac{\Pi}{S}-(1-k)\beta\tfrac{SI}{NE}-(1-k)\beta\eta_1\tfrac{SI_a}{NE}+6\mu_d+\sigma+\\ &\alpha+\gamma_a+\gamma+\delta-\rho_1\sigma\tfrac{E}{I}-\rho_2\sigma\tfrac{E}{I_a}-\alpha\tfrac{I}{H}-\alpha_q\tfrac{I_q}{H}-\gamma_a\tfrac{I_a}{R}-\gamma_q\tfrac{I_q}{R}-\gamma\tfrac{H}{R}+\Pi-\mu_dN-\delta H+\tfrac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2}\\ &\leq (c_1c_4+1)(1-k)\beta(1+\eta_1+\eta_2)+6\mu_d+\sigma+\alpha+\gamma_a+\gamma+\delta+\Pi+\tfrac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2}-\gamma_q\tfrac{I_q}{R}\\ &\leq (c_1c_4+1)(1-k)\beta(1+\eta_1+\eta_2)+6\mu_d+\sigma+\alpha+\gamma_a+\gamma+\delta+\Pi+\tfrac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2}-\gamma_q\tfrac{\epsilon_1}{\epsilon_2}\end{aligned}$$

Let $\epsilon_1>\epsilon_2^2$, very small, such that $(c_1c_4+1)(1-k)\beta(1+\eta_1+\eta_2)+6\mu_d+\sigma+\alpha+\gamma_a+\gamma+\delta+\Pi+\frac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2}-\gamma_q\frac{\epsilon_1}{\epsilon_2}<0$.

In such case, we have $LV<0$.

Case IX: $(S,E,I,I_a,I_q,H,R)\in D_9$

$$\begin{aligned}LV &\leq -c_4c_5+(c_1c_4+1)(1-k)\beta(1+\eta_1+\eta_2)-\tfrac{\Pi}{S}-(1-k)\beta\tfrac{SI}{NE}-(1-k)\beta\eta_1\tfrac{SI_a}{NE}+6\mu_d+\sigma+\\ &\alpha+\gamma_a+\gamma+\delta-\rho_1\sigma\tfrac{E}{I}-\rho_2\sigma\tfrac{E}{I_a}-\alpha\tfrac{I}{H}-\alpha_q\tfrac{I_q}{H}-\gamma_a\tfrac{I_a}{R}-\gamma_q\tfrac{I_q}{R}-\gamma\tfrac{H}{R}+\Pi-\mu_dN-\delta H+\tfrac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2}\\ &\leq (c_1c_4+1)(1-k)\beta(1+\eta_1+\eta_2)+6\mu_d+\sigma+\alpha+\gamma_a+\gamma+\delta+\Pi+\tfrac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2}-\gamma\tfrac{H}{R}\\ &\leq (c_1c_4+1)(1-k)\beta(1+\eta_1+\eta_2)+6\mu_d+\sigma+\alpha+\gamma_a+\gamma+\delta+\Pi+\tfrac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2}-\gamma\tfrac{\epsilon_1}{\epsilon_2}\end{aligned}$$

Let $\epsilon_1=\epsilon_2^2$, choose $\epsilon_1>0$ small enough such that, $(c_1c_4+1)(1-k)\beta(1+\eta_1+\eta_2)+6\mu_d+\sigma+\alpha+\gamma_a+\gamma+\delta+\Pi+\frac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2}-\gamma\frac{\epsilon_1}{\epsilon_2}<0$.

For this case, we have $LV<0$.

Case X: $(S, E, I, I_a, I_q, H, R) \in D_{10}$

$$LV \leq -c_4c_5 + (c_1c_4+1)(1-k)\beta(1+\eta_1+\eta_2) - \frac{\Pi}{S} - (1-k)\beta\frac{SI}{NE} - (1-k)\beta\eta_1\frac{SI_a}{NE} + 6\mu_d + \sigma + \alpha + \gamma_a + \gamma + \delta - \rho_1\sigma\frac{E}{I} - \rho_2\sigma\frac{E}{I_a} - \alpha\frac{I}{H} - \alpha_q\frac{I_q}{H} - \gamma_a\frac{I_a}{R} - \gamma_q\frac{I_q}{R} - \gamma\frac{H}{R} + \Pi - \mu_d N - \delta H + \frac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2}$$
$$\leq (c_1c_4+1)(1-k)\beta(1+\eta_1+\eta_2) + 6\mu_d + \sigma + \alpha + \gamma_a + \gamma + \delta + \Pi + \frac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2} - \mu_d N$$
$$\leq (c_1c_4+1)(1-k)\beta(1+\eta_1+\eta_2) + 6\mu_d + \sigma + \alpha + \gamma_a + \gamma + \delta + \Pi + \frac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2} - \frac{\mu_d}{\epsilon_2}$$

Let $\epsilon_2 > 0$ be as sufficiently small so that, $(c_1c_4+1)(1-k)\beta(1+\eta_1+\eta_2) + 6\mu_d + \sigma + \alpha + \gamma_a + \gamma + \delta + \Pi + \frac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2} - \frac{\mu_d}{\epsilon_2} < 0$.

In such case, we have $LV < 0$.

Case XI: $(S, E, I, I_a, I_q, H, R) \in D_{11}$

$$LV \leq -c_4c_5 + (c_1c_4+1)(1-k)\beta(1+\eta_1+\eta_2) - \frac{\Pi}{S} - (1-k)\beta\frac{SI}{NE} - (1-k)\beta\eta_1\frac{SI_a}{NE} + 6\mu_d + \sigma + \alpha + \gamma_a + \gamma + \delta - \rho_1\sigma\frac{E}{I} - \rho_2\sigma\frac{E}{I_a} - \alpha\frac{I}{H} - \alpha_q\frac{I_q}{H} - \gamma_a\frac{I_a}{R} - \gamma_q\frac{I_q}{R} - \gamma\frac{H}{R} + \Pi - \mu_d N - \delta H + \frac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2}$$
$$\leq (c_1c_4+1)(1-k)\beta(1+\eta_1+\eta_2) + 6\mu_d + \sigma + \alpha + \gamma_a + \gamma + \delta + \Pi + \frac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2} - \mu_d N$$
$$\leq (c_1c_4+1)(1-k)\beta(1+\eta_1+\eta_2) + 6\mu_d + \sigma + \alpha + \gamma_a + \gamma + \delta + \Pi + \frac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2} - \frac{\mu_d}{\epsilon_2}$$

Again choosing $\epsilon_2 > 0$ be as sufficiently small so that, $(c_1c_4+1)(1-k)\beta(1+\eta_1+\eta_2) + 6\mu_d + \sigma + \alpha + \gamma_a + \gamma + \delta + \Pi + \frac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2} - \frac{\mu_d}{\epsilon_2} < 0$.

In such case, we have $LV < 0$.

Case XII: $(S, E, I, I_a, I_q, H, R) \in D_{12}$

$$LV \leq -c_4c_5 + (c_1c_4+1)(1-k)\beta(1+\eta_1+\eta_2) - \frac{\Pi}{S} - (1-k)\beta\frac{SI}{NE} - (1-k)\beta\eta_1\frac{SI_a}{NE} + 6\mu_d + \sigma + \alpha + \gamma_a + \gamma + \delta - \rho_1\sigma\frac{E}{I} - \rho_2\sigma\frac{E}{I_a} - \alpha\frac{I}{H} - \alpha_q\frac{I_q}{H} - \gamma_a\frac{I_a}{R} - \gamma_q\frac{I_q}{R} - \gamma\frac{H}{R} + \Pi - \mu_d N - \delta H + \frac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2}$$
$$\leq (c_1c_4+1)(1-k)\beta(1+\eta_1+\eta_2) + 6\mu_d + \sigma + \alpha + \gamma_a + \gamma + \delta + \Pi + \frac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2} - \mu_d N$$
$$\leq (c_1c_4+1)(1-k)\beta(1+\eta_1+\eta_2) + 6\mu_d + \sigma + \alpha + \gamma_a + \gamma + \delta + \Pi + \frac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2} - \frac{\mu_d}{\epsilon_2}$$

Again choosing $\epsilon_2 > 0$ be as sufficiently small so that, $(c_1c_4+1)(1-k)\beta(1+\eta_1+\eta_2) + 6\mu_d + \sigma + \alpha + \gamma_a + \gamma + \delta + \Pi + \frac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2} - \frac{\mu_d}{\epsilon_2} < 0$.

In such case, we have $LV < 0$.

Case XIII: $(S, E, I, I_a, I_q, H, R) \in D_{13}$

$$LV \leq -c_4c_5 + (c_1c_4+1)(1-k)\beta(1+\eta_1+\eta_2) - \frac{\Pi}{S} - (1-k)\beta\frac{SI}{NE} - (1-k)\beta\eta_1\frac{SI_a}{NE} + 6\mu_d + \sigma + \alpha + \gamma_a + \gamma + \delta - \rho_1\sigma\frac{E}{I} - \rho_2\sigma\frac{E}{I_a} - \alpha\frac{I}{H} - \alpha_q\frac{I_q}{H} - \gamma_a\frac{I_a}{R} - \gamma_q\frac{I_q}{R} - \gamma\frac{H}{R} + \Pi - \mu_d N - \delta H + \frac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2}$$
$$\leq (c_1c_4+1)(1-k)\beta(1+\eta_1+\eta_2) + 6\mu_d + \sigma + \alpha + \gamma_a + \gamma + \delta + \Pi + \frac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2} - \mu_d N$$
$$\leq (c_1c_4+1)(1-k)\beta(1+\eta_1+\eta_2) + 6\mu_d + \sigma + \alpha + \gamma_a + \gamma + \delta + \Pi + \frac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2} - \frac{\mu_d}{\epsilon_2}$$

Again choosing $\epsilon_2 > 0$ be as sufficiently small so that, $(c_1c_4+1)(1-k)\beta(1+\eta_1+\eta_2) + 6\mu_d + \sigma + \alpha + \gamma_a + \gamma + \delta + \Pi + \frac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2} - \frac{\mu_d}{\epsilon_2} < 0$.

In such case, we have $LV < 0$.

Case XIV: $(S, E, I, I_a, I_q, H, R) \in D_{14}$

$$LV \leq -c_4c_5 + (c_1c_4+1)(1-k)\beta(1+\eta_1+\eta_2) - \frac{\Pi}{S} - (1-k)\beta\frac{SI}{NE} - (1-k)\beta\eta_1\frac{SI_a}{NE} + 6\mu_d + \sigma + \alpha + \gamma_a + \gamma + \delta - \rho_1\sigma\frac{E}{I} - \rho_2\sigma\frac{E}{I_a} - \alpha\frac{I}{H} - \alpha_q\frac{I_q}{H} - \gamma_a\frac{I_a}{R} - \gamma_q\frac{I_q}{R} - \gamma\frac{H}{R} + \Pi - \mu_d N - \delta H + \frac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2}$$

$$\leq (c_1c_4+1)(1-k)\beta(1+\eta_1+\eta_2)+6\mu_d+\sigma+\alpha+\gamma_a+\gamma+\delta+\Pi+\tfrac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2}-\mu_d N$$
$$\leq (c_1c_4+1)(1-k)\beta(1+\eta_1+\eta_2)+6\mu_d+\sigma+\alpha+\gamma_a+\gamma+\delta+\Pi+\tfrac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2}-\tfrac{\mu_d}{\epsilon_2}$$

Again choosing $\epsilon_2 > 0$ be as sufficiently small so that, $(c_1c_4 + 1)(1 - k)\beta(1 + \eta_1 + \eta_2) + 6\mu_d + \sigma + \alpha + \gamma_a + \gamma + \delta + \Pi + \frac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2} - \frac{\mu_d}{\epsilon_2} < 0$.

Here we get $LV < 0$.

Case XV: $(S, E, I, I_a, I_q, H, R) \in D_{15}$

$$LV \leq -c_4c_5+(c_1c_4+1)(1-k)\beta(1+\eta_1+\eta_2)-\tfrac{\Pi}{S}-(1-k)\beta\tfrac{SI}{NE}-(1-k)\beta\eta_1\tfrac{SI_a}{NE}+6\mu_d+\sigma+\alpha+\gamma_a+\gamma+\delta-\rho_1\sigma\tfrac{E}{I}-\rho_2\sigma\tfrac{E}{I_a}-\alpha\tfrac{I}{H}-\alpha_q\tfrac{I_q}{H}-\gamma_a\tfrac{I_a}{R}-\gamma_q\tfrac{I_q}{R}-\gamma\tfrac{H}{R}+\Pi-\mu_d N-\delta H+\tfrac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2}$$
$$\leq (c_1c_4+1)(1-k)\beta(1+\eta_1+\eta_2)+6\mu_d+\sigma+\alpha+\gamma_a+\gamma+\delta+\Pi+\tfrac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2}-\mu_d N-\delta H$$
$$\leq (c_1c_4+1)(1-k)\beta(1+\eta_1+\eta_2)+6\mu_d+\sigma+\alpha+\gamma_a+\gamma+\delta+\Pi+\tfrac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2}-\tfrac{\mu_d}{\epsilon_2}-\tfrac{\delta}{\epsilon_2}$$

Again choosing $\epsilon_2 > 0$ be as sufficiently small so that, $(c_1c_4 + 1)(1 - k)\beta(1 + \eta_1 + \eta_2) + 6\mu_d + \sigma + \alpha + \gamma_a + \gamma + \delta + \Pi + \frac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2} - \frac{\mu_d}{\epsilon_2} - \frac{\delta}{\epsilon_2} < 0$.

In such case, we have $LV < 0$.

Case XVI: $(S, E, I, I_a, I_q, H, R) \in D_{16}$

$$LV \leq -c_4c_5+(c_1c_4+1)(1-k)\beta(1+\eta_1+\eta_2)-\tfrac{\Pi}{S}-(1-k)\beta\tfrac{SI}{NE}-(1-k)\beta\eta_1\tfrac{SI_a}{NE}+6\mu_d+\sigma+\alpha+\gamma_a+\gamma+\delta-\rho_1\sigma\tfrac{E}{I}-\rho_2\sigma\tfrac{E}{I_a}-\alpha\tfrac{I}{H}-\alpha_q\tfrac{I_q}{H}-\gamma_a\tfrac{I_a}{R}-\gamma_q\tfrac{I_q}{R}-\gamma\tfrac{H}{R}+\Pi-\mu_d N-\delta H+\tfrac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2}$$
$$\leq (c_1c_4+1)(1-k)\beta(1+\eta_1+\eta_2)+6\mu_d+\sigma+\alpha+\gamma_a+\gamma+\delta+\Pi+\tfrac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2}-\mu_d N$$
$$\leq (c_1c_4+1)(1-k)\beta(1+\eta_1+\eta_2)+6\mu_d+\sigma+\alpha+\gamma_a+\gamma+\delta+\Pi+\tfrac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2}-\tfrac{\mu_d}{\epsilon_2}$$

Again choosing $\epsilon_2 > 0$ be as sufficiently small so that, $(c_1c_4 + 1)(1 - k)\beta(1 + \eta_1 + \eta_2) + 6\mu_d + \sigma + \alpha + \gamma_a + \gamma + \delta + \Pi + \frac{\theta_1^2+\theta_2^2+\theta_3^2+\theta_4^2+\theta_6^2+\theta_7^2}{2} - \frac{\mu_d}{\epsilon_2} < 0$.

In such case, we have $LV < 0$.

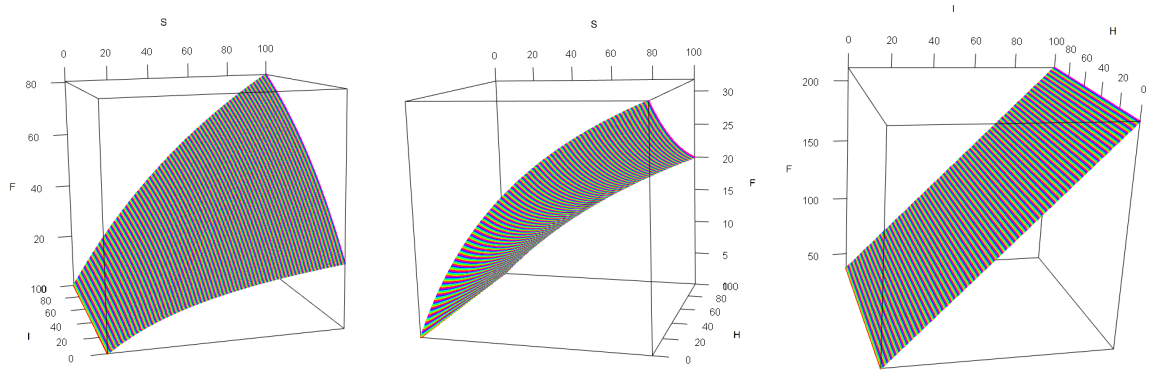


Figure 2: The relationship between $F = \frac{\beta S(I+\eta_1 I_a+\eta_2 E)}{N}$ and (a) S, I [upper left panel] (b) S, H [upper right panel] and (c) I, H [lower panel]. Figure (a) depicts a quadratic shape, while Figure (b) illustrates a sigmoidal form, and Figure (c) exhibits a linear shape. The other parameters are $\eta_2 = 0.4$ and the same from Table 1.

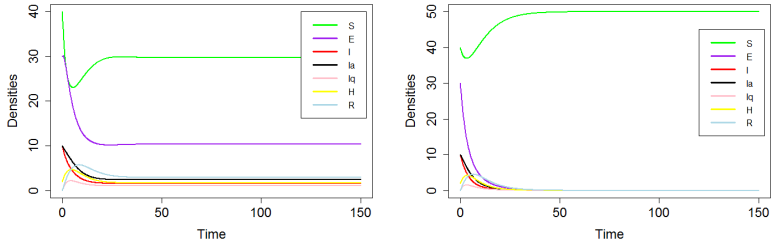


Figure 3: The time series plot of the model (1) for (a) $k = 0$ and (b) $k = 0.6544$. The other parameters are same as in Table 1 with $\eta_2 = 0.2$.

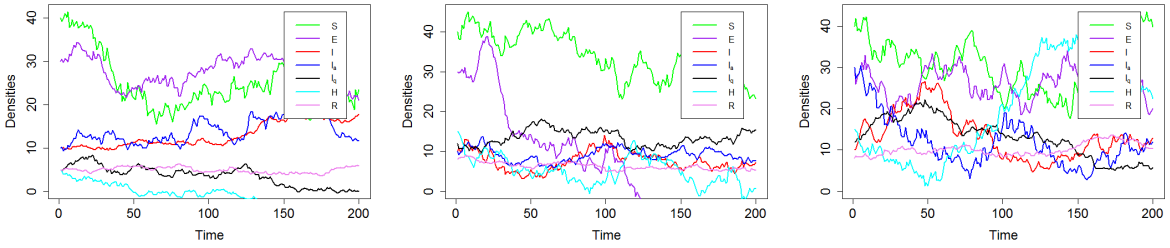


Figure 4: The path $S(t)$, $E(t)$, $I(t)$, $I_a(t)$, $I_q(t)$, $H(t)$ and $R(t)$ for the stochastic model (5) with initial values $(S(0), E(0), I(0), I_a(0), I_q(0), H(0), R(0)) = (40, 30, 10, 30, 12, 15, 8)$. The parameters are taken from Table 1, $\theta_1 = 0.3, \theta_5 = 0.2, \theta_7 = 0.1$, $k = 0$ with a) $\theta_2 = 0.2, \theta_3 = 0.1, \theta_4 = 0.3, \theta_6 = 0.2$; b) $\theta_2 = 0.4, \theta_3 = 0.4, \theta_4 = 0.3, \theta_6 = 0.4$ and c) $\theta_2 = 0.4, \theta_3 = 0.4, \theta_4 = 0.6, \theta_6 = 0.4$.

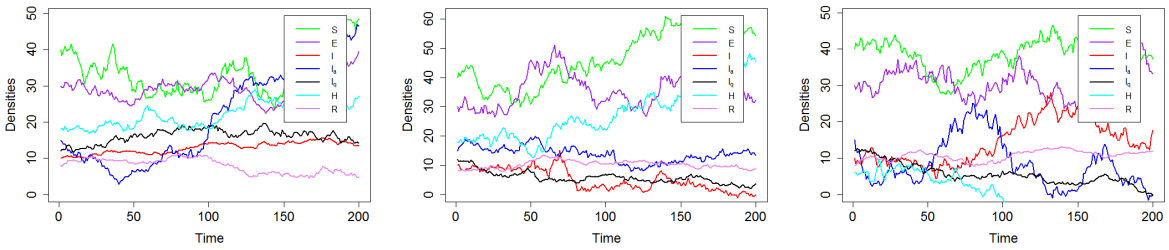


Figure 5: The path $S(t)$, $E(t)$, $I(t)$, $I_a(t)$, $I_q(t)$, $H(t)$ and $R(t)$ for the stochastic model (5) with initial values $(S(0), E(0), I(0), I_a(0), I_q(0), H(0), R(0)) = (40, 30, 10, 15, 12, 18, 8)$. The parameters are taken from Table 1, $\theta_1 = 0.3, \theta_5 = 0.2, \theta_7 = 0.1$ and $k = 0.6544$ with a)$\theta_2 = 0.2, \theta_3 = 0.1, \theta_4 = 0.3, \theta_6 = 0.2$; b) $\theta_2 = 0.4, \theta_3 = 0.4, \theta_4 = 0.3, \theta_6 = 0.4$ and c) $\theta_2 = 0.4, \theta_3 = 0.4, \theta_4 = 0.6, \theta_6 = 0.4$.

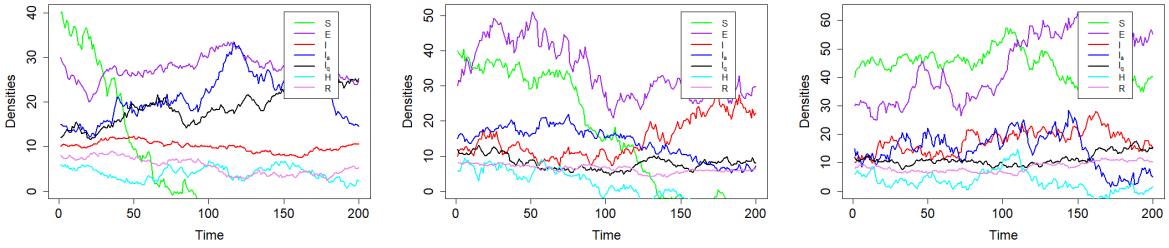


Figure 6: The path $S(t)$, $E(t)$, $I(t)$, $I_a(t)$, $I_q(t)$, $H(t)$ and $R(t)$ for the stochastic model (5) with initial values $(S(0), E(0), I(0), I_a(0), I_q(0), H(0), R(0)) = (40, 30, 10, 15, 12, 18, 8)$. The parameters are taken from Table 1, $\theta_1 = 0.3, \theta_5 = 0.2, \theta_7 = 0.1$ and $k = 0.4$ with a)$\theta_2 = 0.2, \theta_3 = 0.1, \theta_4 = 0.3, \theta_6 = 0.2$; b) $\theta_2 = 0.4, \theta_3 = 0.4, \theta_4 = 0.3, \theta_6 = 0.4$ and c) $\theta_2 = 0.4, \theta_3 = 0.4, \theta_4 = 0.6, \theta_6 = 0.4$.

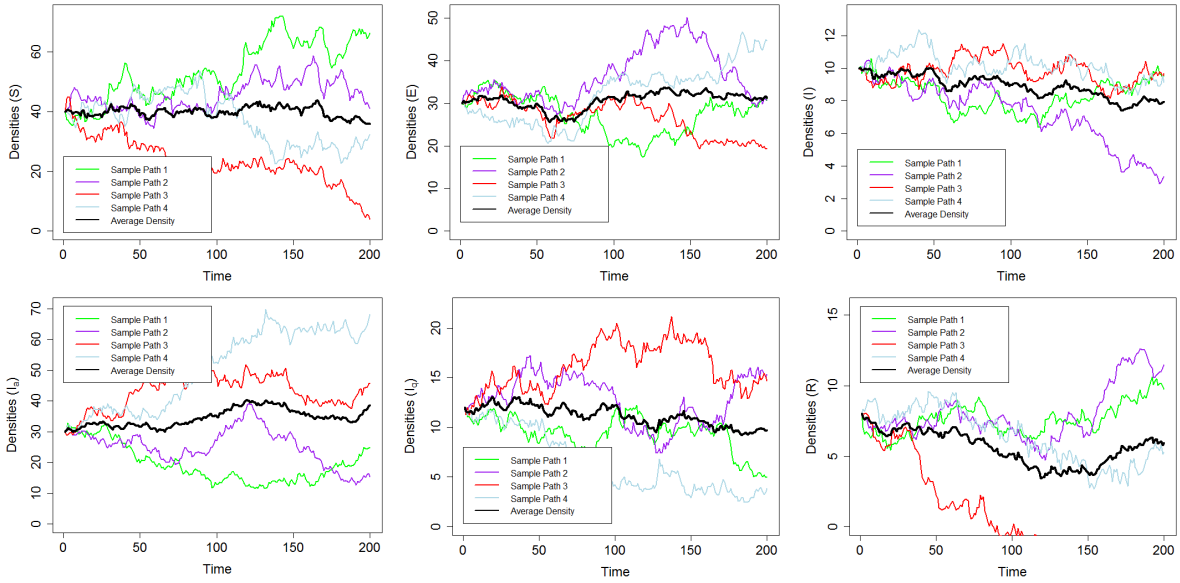


Figure 7: The four different sample paths and their average path of $S(t)$, $E(t)$, $I(t)$, $I_a(t)$, $I_q(t)$, $H(t)$ and $R(t)$ for the stochastic model (5). The parameters are taken from Fig. 4 with $\theta_1 = 0.3, \theta_2 = 0.2, \theta_3 = 0.1, \theta_4 = 0.3, \theta_5 = 0.2, \theta_6 = 0.2, \theta_7 = 0.1$ and $k = 0$.

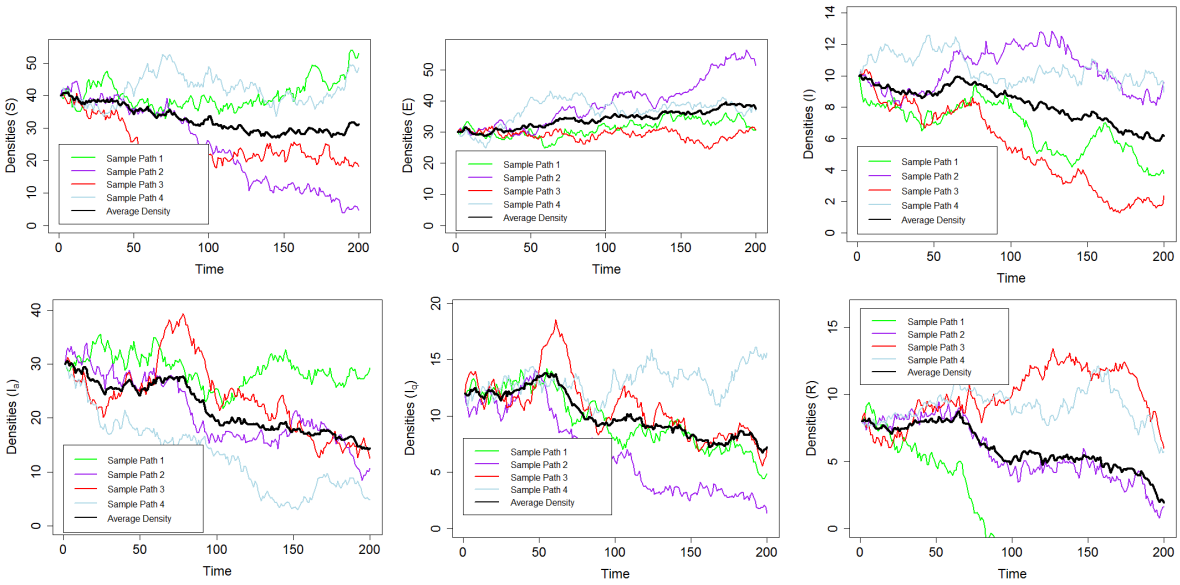


Figure 8: The four different sample paths and their average path of $S(t)$, $E(t)$, $I(t)$, $I_a(t)$, $I_q(t)$, $H(t)$ and $R(t)$ for the stochastic model (5). The parameters are taken from Fig. 5 with $\theta_1 = 0.3, \theta_2 = 0.2, \theta_3 = 0.1, \theta_4 = 0.3, \theta_5 = 0.2, \theta_6 = 0.2, \theta_7 = 0.1$ and $k = 0.6544$.

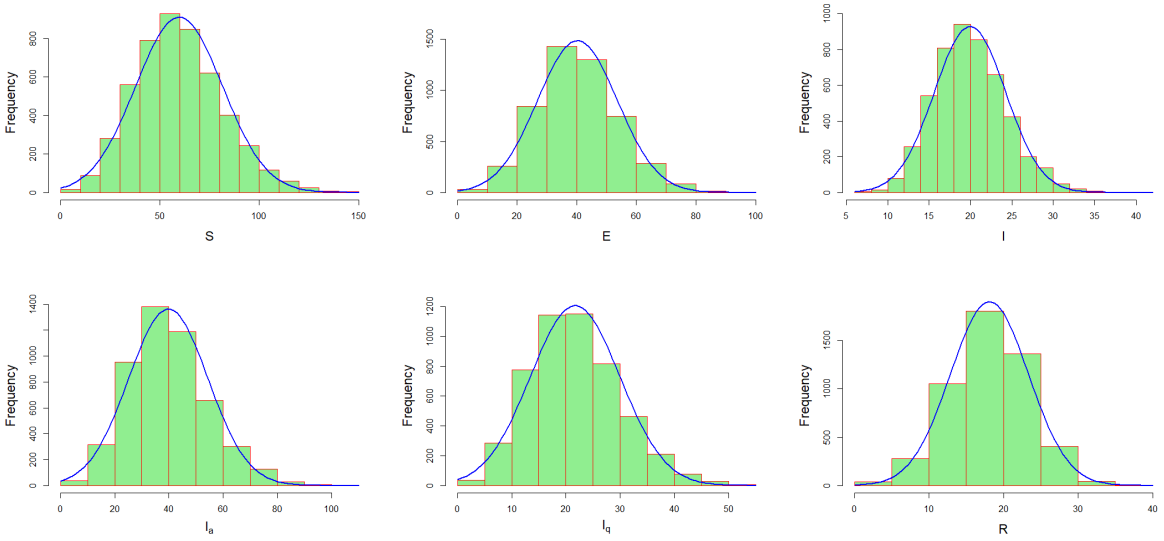


Figure 9: Histogram of the densities at the time point 150 of the system (5). The parameters are taken from Fig. 4 with $\theta_1 = 0.3, \theta_2 = 0.2, \theta_3 = 0.1, \theta_4 = 0.3, \theta_5 = 0.2, \theta_6 = 0.2, \theta_7 = 0.1$.

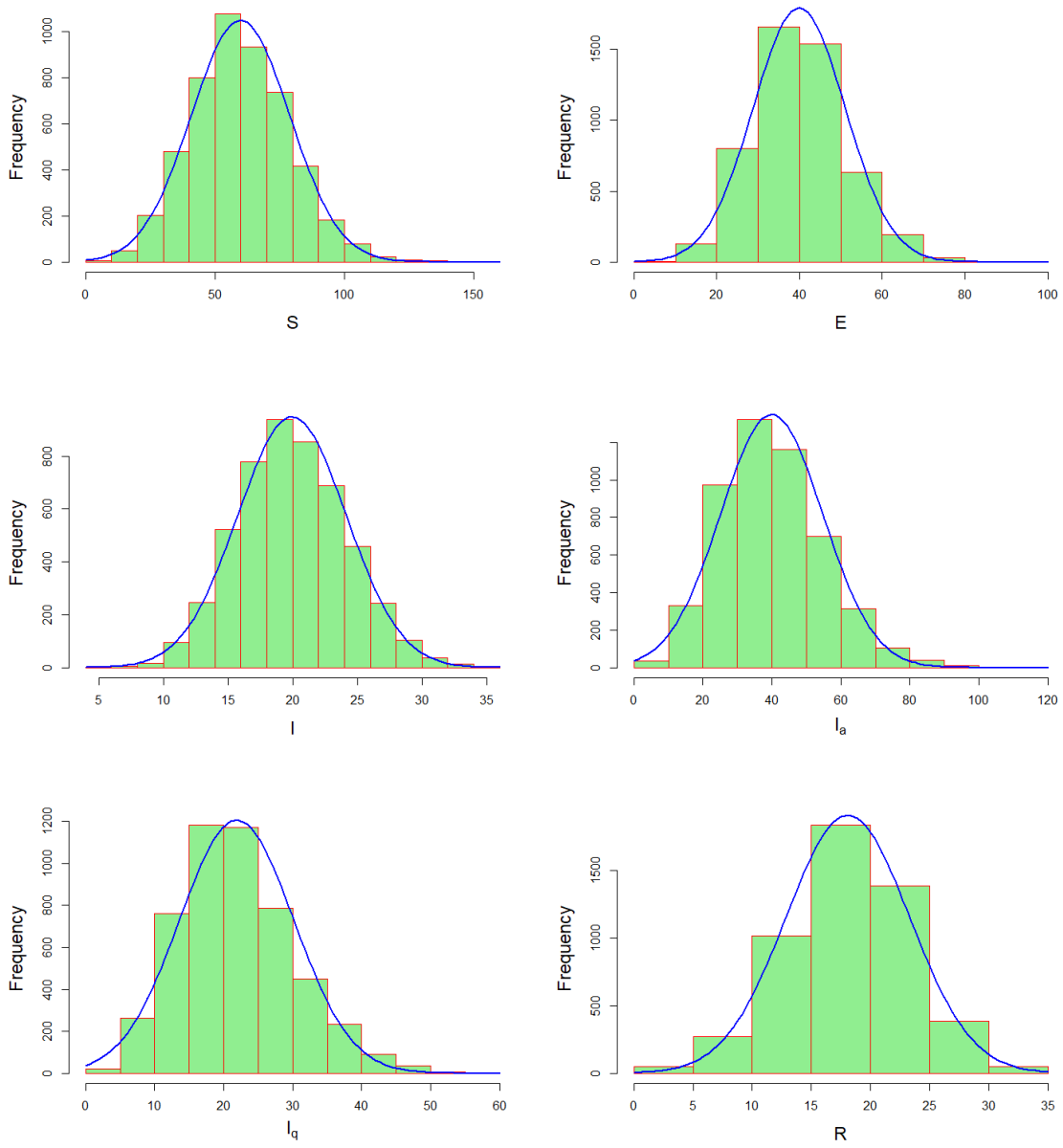


Figure 10: Histogram of the densities at the time point 150 **of the system** (5)**. The parameters are taken from Fig. 5 with** $\theta_1 = 0.3, \theta_2 = 0.2, \theta_3 = 0.1, \theta_4 = 0.3, \theta_5 = 0.2, \theta_6 = 0.2, \theta_7 = 0.1$**.**

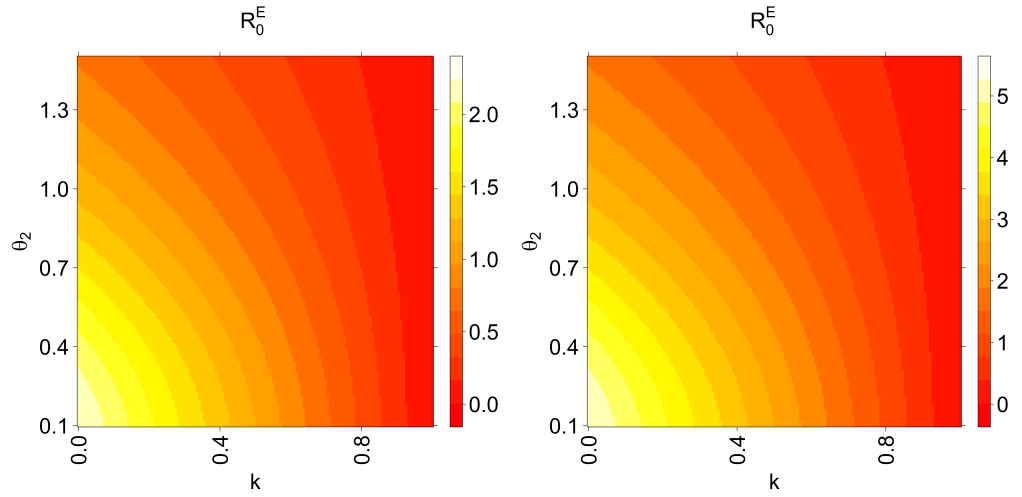


Figure 11: Heat map diagram of R_E^0 with respect to k and θ_2 for the system (5). The parameters are taken from Table 1 with $\eta_2 = 0.1$. The left figure corresponding to $\beta = 0.74$ and right figure corresponding to $\beta = 1.74$ respectively.

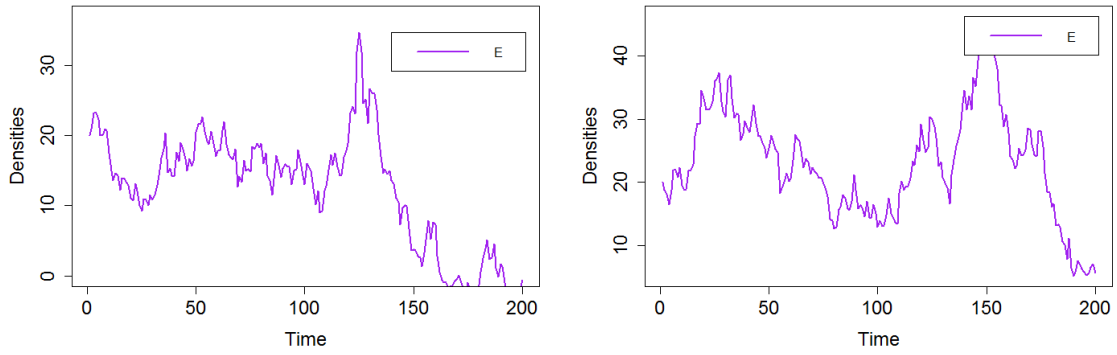


Figure 12: Two different sample paths are drawn for the parameter set same as Fig. 11(a) with $k = 0.6544$.

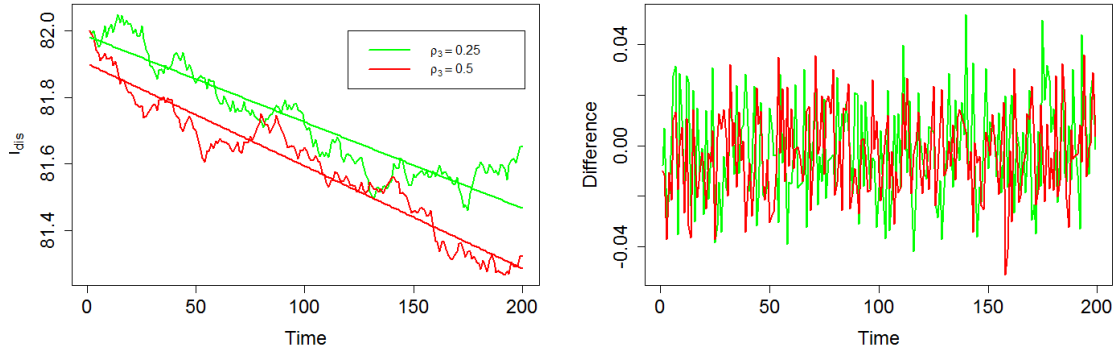


Figure 13: a) Average paths for I_{dis} are drawn for two different values of parameter ρ_3, the other parameters are same as Fig. 8. b) The difference plot corresponding to a).

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 23, No. 1, 2025 (New Series), pp 177–187
https://www.ssca.org.in/journal

Sample Size Determination for Clinical Studies when Crisp Inputs are not available

Sai Sarada Vedururu[1], R.Vishnu Vardhan[2] and K.V.S. Sarma[3]
[1]*Department of Mathematics GITAM (Deemed to be) University, Hyderabad.*
[2]*Department of Statistics, Pondicherry University, Puducherry*
[3]*Former Professor of Statistics, Sri Venkeswara University, Tirupati*

Received: 04 November 2022; Revised: 25 November 2023; Accepted: 01 May 2024

Abstract

This paper deals with the classical problem of determining the minimum sample size (n) required in clinical studies for estimating the population prevalence p' of a characteristic. The popularly used formula for n requires prior knowledge on p' which may not be known as a crisp value. It can be estimated from a pilot study or specified as a range of values reflecting some uncertainty. In the first part we characterise n as a random variable whose values depend on the uncertainty in the anticipated p' modelled by a Beta distribution and thereby determine the expected sample size and its variance. In the second part of this paper we also propose a novel method to improve the formula by considering a triplet where a,b and c denote the minimum, most likely and maximum values of p', derive a new formula and show that it is more consistent than the classical method. We demonstrate the utility of the formula with illustrations and compare them with alternative ways of presenting the inputs.

Key words: Sample size; Triangular distribution; Triplet estimates; PERT.

AMS Subject Classifications: 0000

1. Introduction

Sample size determination is a basic requirement in the design and analysis of clinical studies including community trails. If enough subjects are not included in the study, the real effect or situation prevailing in the target group cannot be estimated correctly. A large sample needs more resources to achieve the desired precision than a small sample. Further, non-sampling errors erupt while executing a large study. Statistical methods offer a scientific approach to determine the minimum sample size such that sample-related risks of incorrect decisions are minimized. We discuss some interesting issues on sample size determination with clinical objectives as background, but the arguments apply to other areas too. We can broadly divide clinical studies into two viz.,

a) *Descriptive studies* in which the chief objective is to describe a population by estimating the characteristics from sample data and

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

b) *Comparative studies* in which comparison of outcomes among two or more groups (like mean or percentage) is the main objective.

There are two broad ways of summarizing any data depending on whether the outcome is a count or a measurement. In case of dichotomous categorical data, the proportion of the outcome of interest is the summary measure, denoted by p' which is called prevalence when observed over a period of time. If the events are observed in new cases, p' is called incidence or occurrence rate. An estimate of p' is $p = \frac{k}{n}$ where k subjects out of n are found to have the outcome of interest and n denotes the sample size. The true proportion in the population is however unknown unless the entire population is studied. The behaviour of p is modelled by Binomial distribution.

On the other hand, when the outcome in the sample is a measured characteristic like blood glucose level, it is summarized as the arithmetic mean (m) (or median in some cases) of the sample values along with standard deviation(s) of the values. The behaviour of m, over different samples of the population is modelled by normal distribution.

In both situations, a $100(1-\alpha)\%$ Confidence Interval (CI) can be provided around the sample estimate such that the true mean/proportion falls in this interval with desired confidence usually 95%. According to the Central Limit Theorem in statistics, the sampling distribution of both p or m tend to be normal when n is large. Hence the 95% CI for p or m are constructed making use of the standardized normal variable (Z).

1.1. Sample size for *p* based on margin of error

We briefly outline the method of determining the sample size with desired *margin of error* (d) when the objective is to estimate (a) the proportion of dichotomous outcomes or (b) the mean of a characteristic. This method is known as precision-based method because smaller margin of error leads to higher precision.

Let p be the estimate of p'obtained from a random sample of size n drawn from the population. Then the $100(1-\alpha)\%$ CI for the population prevalence is given by

$$\left\{ p - Z_{1-\alpha/2}\sqrt{\frac{p\ (1-p)}{n}}\ ,\ p + Z_{1-\alpha/2}\sqrt{\frac{p\ (1-p)}{n}} \right\} \tag{1}$$

The quantity $d = (Z_{1-\alpha/2}\sqrt{\frac{p\ (1-p)}{n}})$denotes the margin of error and $(Z_{1-\alpha/2})$is the inverse of the cumulative standard normal distribution corresponding to the chosen α(like 0.05).

The objective is to determine n such that p is contained in (1). Since the precision of the estimate increases when d is small, one way of estimating n is to keep $d \leq d_0$ where d_0 is the desired upper limit, like 5%. Solving for n in d leads to $n \geq \frac{z^2_{1-\alpha/2}p(1-p)}{d_0^2}$. Hence for a fixed choice of d, the expression for the minimum sample size will be

$$n = \frac{z^2_{1-\alpha/2}p(1-p)}{d^2} \tag{2}$$

Formula (2) is known as Cochran's formula (Cochran, 1977) applicable for large populations. When the population size is finite, like the number of employees of a company, then $n' = \frac{n}{1+\frac{n}{N}}$

gives the sample size corrected for finite population size. The chief input for implementing (2) is p.

1.2. Sensitivity of n to changes in p

The formula in (2) requires p as input which is known only when a pilot study is carried out. When pilot study is not possible, we can get p from previous research reports or by a personal guess. For instance, if p is known as 0.9 it means that there is 90% chance that the desired condition occurs. Then for $\alpha = 0.05$ we get $(Z_{1-\alpha/2}) = 1.96$ and with $d = 0.05$ we get $n = 385$. This value quickly drops to 96 if d is taken as 0.10, keeping other parameters unchanged. Approximating $(Z_{1-\alpha/2})$ by 2, the constant appearing in (2) is approximately 1600. When $p = 0.5$ we get $n = 400$ while the actual value with 1.96 is 385. Hence the reliability of n depends on the precision with which p known. 1 shows the pattern of n against p which is concave reaching a maximum of 385 at $p = 0.50$. We see

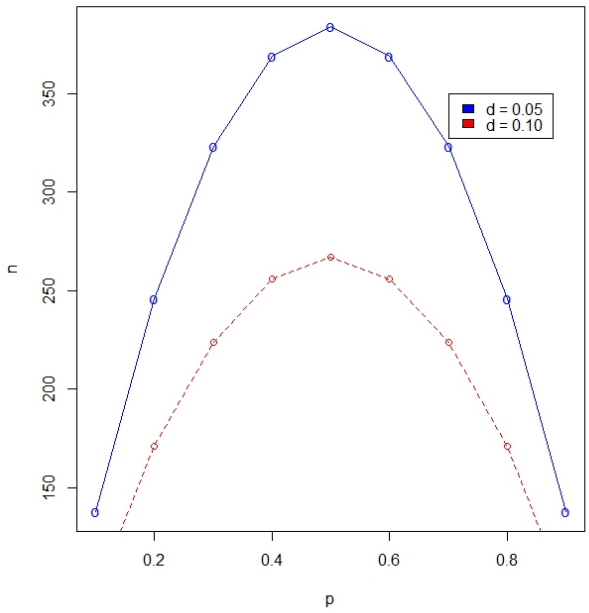


Figure 1: Sensitivity of n against p for different values of d

that n changes rapidly with d but symmetric around $p = 0.5$. The decrease in n for values of p away from 0.5 may be called the *gain due to information.* In section-2 we develop a methodology to formulate the distribution of n viewing p as a random variable using beta distribution. We also study the empirical distribution of n under the chosen model by estimating its parameters, instead of using a single p value. In section 3 we develop a new method of determining n when the input value of p is not precisely known but expressed as an *interval,* along with a middle value, which we call a *triplet.* The new estimate is proposed as a weighted average of the expected sample size at each of the three elements of the triplet. We call this *triplet estimation* and study the properties this new estimate.

2. A model for the probability distribution of n

The formula given in (2) can be stated as $n = kp(1-p)$ where $k = \frac{z^2_{1-\alpha/2}}{d^2}$ is constant for pre-determined values of α and d. We wish to identify a probability distribution for n by viewing p as a continuous random variable (Y) in $[(0,1)$ so that $n = kY(1-Y)$ and the type-1 beta distribution is natural choice for the distribution of Y specified by the density function

$$f(y) = \frac{1}{\beta(u,v)}\ y^{u-1}(1-y)^{v-1} \text{for} \quad 0<y<1,\ u,\ v>0 \tag{3}$$

The parameters (u, v) are related to E(Y) and V(Y) and given as

$$u = \ E(Y)\left\{\frac{E(Y)\,(1-E(Y))}{V(Y)} - \ 1\right\} \text{and } v = \ \{1-E(Y)\}\left\{\frac{E(Y)\,(1-E(Y))}{V(Y)} - \ 1\right\} \tag{4}$$

In fact p is the anticipated point-mass on the Bernoulli distribution which varies with the discretion of the researcher. When p is specified a fixed value there exists a single unique value of n from (3). Instead, we assume a probability distribution in the domain (0,1) with peak density at p so that we can account for the uncertainty in p and thereby determine the theoretical mean and variance of Y. The *triangular distribution* (0, p, 1) is one choice for distribution of Y which help to obtain adhoc estimates of E(Y) and V(Y), while the beta-PERT distribution on *(a,b,c)* where $a = 0$, $b = p$ and $c = 1$ is another. We use the Triangular distribution only to summarize the Bernoulli p since the truncation limits for the distribution of Y are not known at this stage. For the triangular distribution we have

$$E(Y) = \frac{\{1+p\}}{3}) \ \text{ and } \quad V(Y) = (\frac{\{p^2-p+1\}}{18} \tag{5}$$

Thus we have transformed the single anticipated p into a probability distribution and captured its mean and variance as summary. As a result, for each value of p we can uniquely identify a $\beta(u,v)$ distribution and estimate the parameters using (3).

Remark: If we use PERT (0,p,1) distribution instead of triangular distribution to estimate E(Y) and V(Y) we get $u = 1+4p$ and $v = 1+4(1-p)$ but $(u+v) = 6$ which is irrespective of u and v, which is a constraint on the parameters, not defined for the beta distribution. Hence we use triangular distribution to supply primary inputs to estimate u and v. Consider the following proposition.

Proposition-1: With Y $\sim$ Beta(u,v) the empirical distribution of n is proportional to that of Y by a constant k.

The empirical distribution of n can be obtained by simulating random deviates from $\beta(u,v)$. Table (1) gives summary of the empirical distribution of n for selected values of p, taking 95% confidence level and $d = 0.05$. This gives $k = 1536.584$ and the value of n rounded to the upper integer.

We observe the following from Table (1):

(a) The values of variance of n are much larger than the corresponding mean, due the fact the mean and variance of Y(1-Y) are multiplied by k and k^2 respectively.

Table 1: Empirical distribution of n with 1000 simulations.

p	n at p	E(Y)	V(Y)	(u,v)	E(n)	V(n)	Empirical Mean of n	Empirical variance of n
0.25	288	0.4166	0.0451	(1.826, 2.557)	373	4810.81	304	7548.13
0.35	350	0.4500	0.0429	(2.145, 2.621)	380	4349.40	314	5820.16
0.5	384	0.5000	0.0416	(2.500, 2.500)	384	4098.56	323	4991.42
0.65	350	0.5500	0.0429	(2.621, 2.145)	380	4348.08	312	6198.41
0.75	288	0.5833	0.0451	(2.557, 1.826)	373	4810.81	306	7059.36

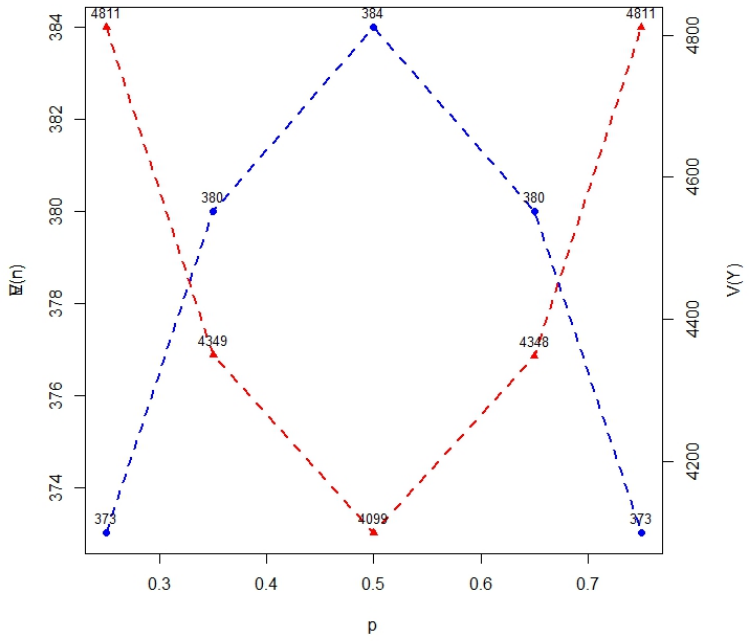


Figure 2: E(n) and V(n) as a function of p.

(b) The expected n and its variance are both symmetric around $p = 0.5$ and the empirical values also exhibit a similar pattern. When compared to the true n obtainable from (2) using the single value of p, the values of E(n) are higher and this can be because the former does not account for the impreciseness in p but E(n) takes into account a background *triangular* model to determine the mean.

(c) The empirical distribution has a shape that is similar to a beta distribution.

Figure (2) shows the pattern of E(n) and V(n) against values of p. The variance of n decreases symmetrically as p increases and reaches a minimum at $p = 0.5$ while E(n) moves in the opposite direction and reaches a maximum at the same p. In the following discussion we propose a method of summarizing the distribution of Y(1-Y) the moments of beta distribution. The empirical distribution of Y(1-Y) is shown in Figure-3.

Proposition-2: If we write T = Y(1-Y) with Y$\sim$ Beta(u,v) then the mean and variance of

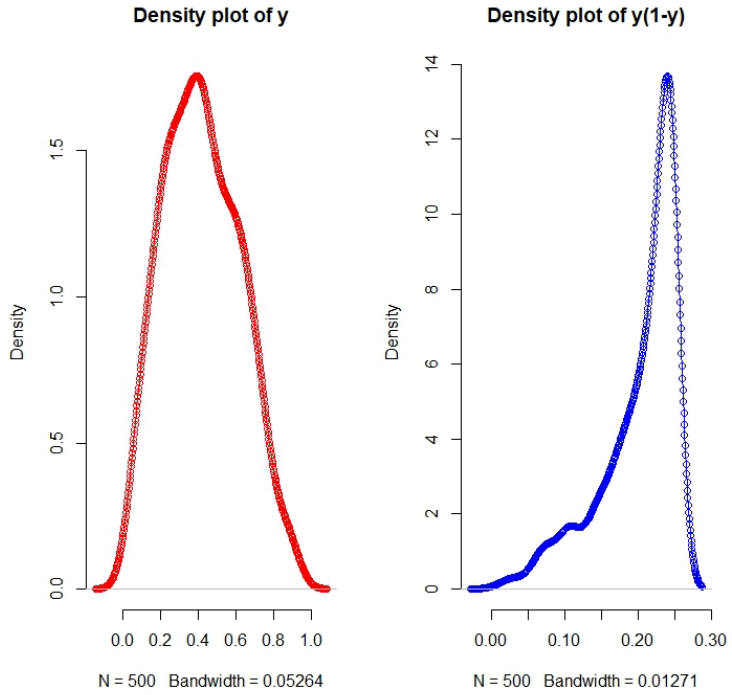


Figure 3: Empirical distribution of Y and Y(1−Y) for $p = 0.25$

T can be obtained as E(T) = E(Y){1-E(Y)} and V(T) = V(Y){1-V(Y)} which reduce to

$$E(T) = \frac{uv}{(u+v)^2} \quad \text{and} \quad V(T) = \frac{(uv)^2}{(u+v)^4(1+u+v)^2} \tag{6}$$

Further the expected sample size is

$$E(n) = k\frac{uv}{(u+v)^2} \quad \text{and} \quad V(n) = k^2\frac{(uv)^2}{(u+v)^4(1+u+v)^2} \tag{7}$$

Proof: The results follow by replacing Y and (1-Y) with their expected values and noting that $E(Y) = \frac{u}{(u+v)}$ and $E(1-Y) = \frac{v}{(u+v)}$. Similarly V(T) follows by noting that $V(Y) = \frac{uv}{(u+v)^2(1+u+v)}$ and V(1-Y) is the same as V(Y). Finally $E(n) = kE(T)$ and $V(n) = k^2V(T)$ which lead to (6) and (7). Hence the proof. With this background, we develop a new estimate of n as (i) a weighted mean of the n values obtainable at the triplet values under the beta distribution model and (ii) using PERT summary as a single input in (2).

3. The triplet estimate to handle imprecise estimates

When a single precise value of p is not available it is customary to specify the same as a triplet (p_1, p_2, p_3) where p_2 is the *most likely value* and (p_1, p_3) are the *lower* and *upper* values of p such that $p_1 < p_2 < p_3$. This approach is used in project management studies to describe the activity durations and latter summarized into mean and SD using beta distribution. Malcolm *et al.* (1959) and Clark (1962) used this approach to summarise the activity durations in project management and to estimate the time to completion the project. Books on Operations Research widely discuss this method (Taha, 2013).

Applying this logic to (p_1, p_2, p_3) we obtain $p_0 = (p_1+4p_2+p_3)/6$ as the mean prevalence. If we use this p_0 in (2) we get a single value of n denoted by n_0. Our new approach is to evaluate n at each of the three values of the triplet and summarize them as a weighted average to get a new crisp value.

We now use the method of triplet inputs to determine the sample size for estimating the population prevalence. Here is another proposition.

We now use the method of triplet inputs to determine the sample size for estimating the population prevalence. Here is another proposition.

Proposition-3: Let n_i be the sample size when the anticipated prevalence is p_i for i = 1,2,3. Then $\mathrm{E}(n_i) = k\frac{u_i v_i}{(u_i+v_i)^2}$ and $\mathrm{V}(n_i) = k^2\frac{(u_i v_i)^2}{(u_i+v_i)^4(1+u_i+v_i)^2}$ where (u_i, v_i) denote the parameters of the underlying beta distribution for i = 1,2,3 and k is the constant by design. Then the new estimate of n will be

$$n_{cap} = \sum_{i=1}^{3} w_i E(n_i\ ) \tag{8}$$

where $w_\mathrm{i} \geq 0$ and $w_1+w_2+w_3 = 1$. We call this the *triplet estimate* of n and $V\,(n_{cap}\ ) = \sum_{i=1}^{3} w_i^2 V(n_i\ )$. It also true that $V\,(n_{cap}\ ) \leq V(n_i\ )$ for i = 1,2,3.

One way of assigning weights is to take $w_2 = 0.5$ and $w_1 = w_3 = 0.25$ so that $E\,(n_2)$ receives more weight than the other two because p_2 is more likely valid than the other two values of the triplet. Another set of weights is {1/6. 4/6, 1/6} corresponding to {w_1, w_2, w_3} which are the weights used in PERT calculations.

Vardhan and Sarma (2010) have used the triplet method in the context of ROC curve analysis. Sarada *et al.* (2018); Vedururu *et al.* (2019) used this method in the context of measuring the process capability index in quality control.Venkatesu *et al.* (2019) have applied this method to redesign a control chart. In all these applications, it was found that the new estimator has lower SE than the classical point estimator.

Instead of pre-defined fixed weights, an objective way is to define weights which reflect the uncertainty in the specification of p (in terms of a triplet). We propose the following weights.

Proposition-4: The weight w_i for $\mathrm{E}(n_i)$ will be the ordinate of the $\beta(u,v)$ distribution at p_i for i = 1,2,3 and normalized to make the sum equal to unity.

This method allots weight as a function of p_i and hence accounts for the anticipated uncertainty in specifying p. We cannot determine the weights with PERT distribution, since the density of vanishes at p_2 and p_3 (truncation limits) and hence n_{cap} cannot be evaluated. Hence the full beta distribution without truncation will be used. Here is an illustration.

Illustration-3

Let us take $p_1 = 0.25$, $p_2 = 0.35$, $p_3 = 0.5$. From the intermediate results from Table-1 we see that the vector of means as (373, 380, 384) and the corresponding variance vector is (4810.76, 4348.74, 4099.11). The vector of weights from beta distribution with corresponding (u_i, v_i) becomes w = (1.5492, 1.6330, 1.6976). Dividing each weight by the sum of weights and applying (8) gives n_{cap} = 380 and $\mathrm{V}(n_{cap})$ = 1467.96 which smaller than the minimum

of the three variances.

Table (2) shows some experimental results comparing n_{cap} with the n obtainable when we use only a single value p_2 as the input in (2).

Table 2: Triplet estimate of n with arbitrary window around the middle.

p_2	Triplet (p_1, p_2, p_3)	n_2 (at p_2)	V(n_2)	n_{cap}	$V(n_{cap})$
0.25	(0.15, 0.25, 0.35)	373	4810.76	370	1682.01
0.35	(0.25, 0.35, 0. 50)	380	4348.74	380	1467.95
0.50	(0.30, 0.50, 0.65)	384	4099.11	381	1441.74
0.65	(0.45, 0.65, 0.75)	380	4348.74	380	1471.39
0.75	(0. 50, 0.70, 0.75)	377	4548.01	379	1489.95

Suppose we take fixed weights instead of deriving from beta density. We consider two types of fixed weights and compare the resulting n_{cap} and its variance.

Table 3: Triplet estimate of n under different schemes of weights.

Triplet (p_1, p_2, p_3)	$w = \{1/6, 4/6, 1/6\}$		$w = \{1/3, 1/3, 1/3\}$	
	n_{cap}	V(n_{cap})	n_{cap}	V(n_{cap})
(0.15, 0.25, 0.35)	373	2413.01	372	1634.11
(0.25, 0.35, 0. 50)	380	2180.27	379	1473.18
(0.30, 0.50, 0.65)	383	2068.96	381	1443.98
(0.45, 0.65, 0.75)	380	2181.03	379	1476.22
(0. 50, 0.70, 0.75)	378	2268.83	378	1495.32

4. Stepwise procedure

The following is a stepwise procedure to handle the calculations.

1. Obtain the anticipated prevalence as a triplet(p_1,p_2,p_3) margin of error as d and level of significance as α For each i = 1,2,3 calculate the following.

2. Transform each p_i into as a point on triangular (0,1) distribution

3. Evaluate the trial values of mean and variance as $\mu_{1\mathrm{i}}$ and ${\sigma_{1\mathrm{i}}}^2$ respectively.

4. Identify a Beta distribution on (0,1) and estimate is parameters (u_i, v_i) Using $\mu_{1\mathrm{i}}$ and ${\sigma_{1\mathrm{i}}}^2$ calculate $E(n_i) = k\dfrac{u_i v_i}{(\mu_i + \nu_i)^2}$ and $V(n_i) = k^2 \dfrac{(u_i v_i)^2}{(u_i + v_i)^4 (1 + u_i + v_i)^2}$.

5. Find $w_i = \dfrac{y_i}{\sum_{i=1}^{3} y_i}$ where y_i denotes the ordinate of the Beta distribution corresponding to p_i

6. Evaluate $n_{cap} = \sum_{i=1}^{3} w_i E(n_i)$ is the new triplet estimate of n and $V(n_{cap}) = \sum_{i=1}^{3} w_i^2 V(n_i)$

5. Alternative way of summarising the triplet

The approach used to derive n_{cap} may be called *evaluate and summarize* method because we evaluate E(n) at each component of the triplet and then summarized them as a weighted average. The variance of n_{cap} was also obtained with this logic.

Alternatively, we may summarise the triplet and then evaluate as a single value from which we can obtain E(n) and V(n) In this method we use $p_0 = (p_1+4p_2+p_3)/6$ basing on the PERT weights.

Again with given p_0 we again identify a triangular distribution with p_0 at the peak and obtain

$$E(Y) = \frac{\{1 + p_0\}}{3} \text{ and } V(Y) = \frac{\{p_0^2 - p_0 + 1\}}{18} \tag{9}$$

With these values we can identify a beta distribution with parameters say (u_0, v_0) and evaluate

$$E\left(T_0\right) = \frac{u_0 v_0}{\left(u_0 + v_0\right)^2} \text{and} V\left(T_0\right) = \frac{\left(u_0 v_0\right)^2}{\left(u_0 + v_0\right)^4 \left(1 + u_0 + v_0\right)^2} \tag{10}$$

where T_0 denotes the quantity Y(1-Y) under this method. If we call this resulting n as n_0 we get E(n_0) = $k\ E\left(T_0\right)$ and V(n_0) = $k^2\ V\left(T_0\right)$. Here is an illustration.

Illustration-4

Let us consider the triplet $(0.25, 0.35, 0.50)$. We get $p_0 = 0.675, E(Y) = 0.5583, V(Y) = 0.0433, u_0 = 2.6164, v_0 = 2.0697$. Using the k value from normal distribution with (1-α) = 0.95 and 5%margin of error (d), we get E(n_0) = 379 and V(n_0) = 4440.71. With different triplets used in Illustration-3 we get the expected sample size and variance under this method of 'summarize and evaluate' are shown in Table (4).

Table 4: Estimated sample size with a pre-summarized triplet.

Triplet (p_1, p_2, p_3)	n_0	V(n_0)
(0.15, 0.25, 0.35)	373	4810.76
(0.25, 0.35, 0. 50)	381	4321.43
(0.30, 0.50, 0.65)	384	4099.87
(0.45, 0.65, 0.75)	381	4295.74
(0. 50, 0.70, 0.75)	379	4440.71

We observe that sample size exhibits higher variance by this method when compared with the method of evaluating three n values and summarizing them with beta density as weights.

6. Conclusion

The problem of finding the minimum sample size to estimate a proportion is better explained with a statistical model instead of simply evaluating the available formula with a single anticipated value of the population proportion (p'). The triangular distribution plays a key role in transforming the single p into random variable so that its mean and variance

can be used to determine the parameters of the beta distribution, which has better shape and properties than the triangular distribution. The uncertainty about p' can be handled by a beta distribution leading to a statistically summarised estimate of n. It also helps in estimating the variance of n while the classical formula gives only single value. With this logic we have proposed a new estimate of n basing on a triplet of input values for p and summarised them as a weighted average. It is shown that the new estimate (n_{cap}) has smaller variance than the variance obtainable at each of the three p values. We have used the weights from the density of beta distribution at the triplet values, so that they reflect the baseline uncertainty in the inputs and normalized them. It is also established that this method is more objective than using other methods of fixed weights, in terms of variance of n. We conclude with the observation that sample size formula greatly depends on the accuracy of the inputs given and the often found attitude among users, *to adjust the inputs until a comfortable number is reached* should be avoided.

Acknowledgements

I am indeed grateful to the Editors for their guidance and counsel. I am very grateful to the reviewer for valuable comments and suggestions of generously listing many useful references.

Conflict of interest

The authors do not have any financial or non-financial conflict of interest to declare for the research work included in this article.

References

(2022). Pert distribution. `https://en.wikipedia.org/wiki/PERT_distribution`.

Clark, C. E. (1962). The pert model for the distribution of an activity time. *Operations Research*, **10**, 405–406.

Cochran, W. G. (1977). *Sampling Techniques*. John Wiley & Sons, Inc., New York.

Cohen, J. (2016). A power primer. *American Psychological Association*, **112**, 155–159.

Daniel, W. W. and Cross, C. L. (2018). *Biostatistics*. John Wiley & Sons, New York.

Indrayan, A. and Malhotra, R. K. (2017). *Medical Miostatistics*. CRC Press.

Malcolm, D. G., Roseboom, J. H., Clark, C. E., and Fazar, W. (1959). Application of a technique for research and development program evaluation. *Operations Research*, **7**, 646–669.

Sarada, V. S., Subbarayudu, M., and Sarma, K. (2018). Estimation of process capability index using confidence intervals of process parameters. *Research & Reviews: Journal of Statistics*, **7**, 49–57.

Taha, H. A. (2013). *Operations Research: An Introduction*. Pearson Education India.

Vardhan, R. V. and Sarma, K. (2010). Estimation of the area under the roc curve using confidence intervals of mean. *ANU Journal of Physical Sciences*, **2**, 29–39.

Vedururu, S. S., Subbarayudu, M., and Sarma, K. (2019). A new method of estimating the process capability index with exponential distribution using interval estimate of the parameter. *Stochastics and Quality Control*, **34**, 95–102.

Venkatesu, B., Abbaiah, R., and Sarada, V. S. (2018). A new method of estimating the process spread using confidence interval of sample range. *Research Review International Journal of Multidisciplinary*, **8**.

Venkatesu, B., Abbaiah, R., and Sarada, V. S. (2019). On the estimating the operating characteristic of shewart control chart for means using interval estimates of process mean and spread.

Yamane, T. (1973). *Statistics: An Introductory Analysis*. Harper & Row New York.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 23, No. 1, 2025 (New Series), pp 189–196
https://www.ssca.org.in/journal

Construction of Order-of-Addition-Orthogonal Array Designs

Muhsina A.[1], **Baidya Nath Mandal**[2], **Rajender Parsad**[3] **and Sukanta Dash**[3]
[1]*The Graduate School, ICAR-Indian Agricultural Research Institute, New Delhi-110012*
[2]*ICAR-Indian Agricultural Research Institute, Gauria Karma, Jharkhand- 825405*
[3]*ICAR-Indian Agricultural Statistics Research Institute, Pusa, New Delhi-110012*

Received: 04 January 2024; Revised: 22 May 2024; Accepted: 25 May 2024

Abstract

Experiments that account for sequential order of components are order-of-addition (OofA) experiments and a full design of such experiments requires $m!$ runs for any m components. Current literature focuses on the construction of fractional designs that are optimal and efficient under the models available to date. This paper provides a systematic construction method of order-of-addition orthogonal arrays (OofA-OA) which were proved as optimal fractional OofA designs. The number of independent, synergistic and antagonistic pairs possible for any m components is also determined. An important balance property of OofA-OA is also explained.

Key words: Order-of-addition; Orthogonal array; Pair-wise order model; Optimality.

1. Introduction

The sequence by which ingredients or components are added into a system may have some definite effect on the response or final output. Experiments that deal with such sequential order of adding components are termed as order-of-addition (OofA) experiments. In early research, designs for cross over experiments constructed by Williams (1949) in which each experimental unit will be given a set of m treatments in a sequential order, were extensively used for OofA experiments. Order-of-addition experiments have been applied in agriculture (Wagner, 1995), food science (Jourdain *et al.*, 2009), cell biology (Black *et al.*, 2001), medical biology (Ding *et al.*, 2015) and many other fields in order to explore the optimal order of components added into the system. These experiments have shown that qualitative and/or quantitative outcome may vary depending on the sequence in which ingredients were added. The foremost reference to an OofA experiment; "the lady tasting tea" wherein only two ingredients, tea and milk, for which the taste of final product was determined by the order in which the ingredients were added (Fisher, 1971). Karim *et al.* (2000) performed an OofA experiment to study the effect of cocoa flavonoids on the vasodilatory capacity of rabbits. Also in engineering, Wilson (2018) proposed an approach to compute the expected utility

Corresponding Author: Baidya Nath Mandal
Email: mandal.stat@gmail.com

when the number of tasks to perform is too large and the sequencing of these tasks has some importance on the expected utility.

An OofA experiment involving m components yield $m!$ orders among which an optimal order has to be screened out using appropriate designs. We hereby call the ingredients or materials in an OofA experiment as components. Each order can be viewed as a permutation of m components, $m \geq 2$. A full design with all the $m!$ orders may not be possible to accommodate while designing the experiment when m is too large. For example, $m = 9$ gives 362,880 orders which is impossible to be contained in a single experiment. This makes us to choose a fraction or subset of the full design so that it may be accommodated in an experiment. Randomly choosing the orders from all the possible orders is relatively inefficient (Zhao *et al.*, 2020). There are many models developed so far for the experimentation of OofA problems. See Peng *et al.* (2019), Mee (2020) and Yang *et al.* (2021) for the models and related optimality proofs therein. An early model, pair-wise ordering (PWO) of effects proposed by Van Nostrand (1995) assumes that sequential order of components affects the response through pair-wise order effects or pseudo factor effects. The readers are referred to Lin and Peng (2019), Voelkel and Gallagher (2019), Tsai (2022), Zhao *et al.* (2020), Winker *et al.* (2020) and Chen *et al.* (2020) for the construction of PWO designs which satisfy efficiency, optimality and relatively smaller run size criterion.

Many designs were constructed for the OofA experiments under the PWO model. Among them, an optimal fractional design, order-of-addition-orthogonal array (OofA-OA) was introduced by Voelkel (2019). The concept behind orthogonal arrays (OA) were used to generate OofA-OA as there is a need to keep the balance while framing a design for OofA experiments. Zhao *et al.* (2021) proposed a systematic construction method for OofA-OAs which is regarded as superior among all the fractional PWO designs. Furthermore, Zhao *et al.* (2022) investigated the existence of OofA-OA with strength 3 and stated that OofA-OAs with strength 3 excel more in terms of balance properties than OofA-OAs with strength 2. In this paper, we propose a systematic method of constructing OofA-OA for any value of m from an existing OofA-OA with $m - 1$ components.

2. Preliminaries

Even though many models including component-position model by Yang *et al.* (2021) have been developed for OofA experimentation, PWO model is the most promising and acceptable model as it is simple and easy to understand. We consider PWO model for the current research. Let us suppose that there are m components which results in $m!$ orders, each of which is a permutation of these m components and it is denoted by $\mathbf{a} = (a_1, ..., a_m)^T$. Let us write OA_f to denote the full OofA design with $m!$ rows and m columns. If we denote y_k as the response due to kth order,

$$y_k = \beta_0 + \sum_{i=1}^{m-1} \sum_{j=i+1}^{m} \beta_{ij} z_{ij} + \epsilon_k \tag{1}$$

where β_0 denotes the overall mean, β_{ij} is the PWO effects of ith and jth component, ϵ_k represents the error term with mean zero and constant variance. To better understand the PWO factors $z_{ij}(\mathbf{a})$ defined by Van Nostrand (1995), we suppose that $m = 3$ components and an order $1 \to 3 \to 2$, means 1st component followed by 3rd and 2nd components are

added in succession, denoted as $\mathbf{a} = (1, 3, 2)$. Then, the PWO factors z_{ij} become $z_{12} = 1, z_{13} = 1, z_{23} = -1$. For denoting the PWO factors z_{ij}, two components are taken at a time from m components such that $1 \leq i < j \leq m$, yielding $\binom{m}{2}$ PWO factors. As there are $m(m-1)/2$ PWO factors and one general mean effect term in the model (1), $p = \binom{m}{2} + 1$ parameters have to be estimated from the model. Model (1) can be expressed in matrix form as

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2}$$

A fractional OofA design d with run size n is said to be ϕ-optimal if its moment matrix $\mathbf{M} = \frac{1}{n}\mathbf{X}'\mathbf{X}$ (where $\mathbf{X}$ denotes the model matrix) is equal to the moment matrix of the full design. Interestingly, Peng *et al.* (2019) proved the optimality of full OofA design in terms of several popular optimality criteria. Additionally, Zhao *et al.* (2021) established that any ϕ-optimal fractional OofA design is certainly an OofA-OA.

We denote P_f as the full PWO design and P_d as the fractional PWO design where d denotes the fractional OofA design with run size n. Any pair of PWO factors (z_{ij}, z_{kl}) can be called as

Synergistic pair : if $i = k$ or $j = l$
Antagonistic pair : if $i = l$ or $j = k$
Independent pair : if $i \neq k, l$ or $j \neq k, l$ (no common component).

In a full PWO design, the frequencies of all t-tuples in any t column subarray for these different pairs are as follows. We denote n_{++} as the number of times $(+, +)$ happens in a pair of PWO factors (z_{ij}, z_{kl}). Similarly, n_{+-}, n_{--}, n_{-+} are also defined. For

Synergistic pair: $n_{++} = m!/3, n_{+-} = m!/6, n_{-+} = m!/6, n_{--} = m!/3$
Antagonistic pair: $n_{++} = m!/6, n_{+-} = m!/3, n_{-+} = m!/3, n_{--} = m!/6$
Independent pair: $n_{++} = m!/4, n_{+-} = m!/4, n_{-+} = m!/4, n_{--} = m!/4$

If the ratios among the frequencies of all t-tuples in any t column subarray of P_f equal to the ratios among the frequencies of all t-tuples in any t column subarray of P_d, then d is said to be the OofA-OA(N, m, t).

3. Construction of OofA-OA from an existing OofA-OA

In this section, a method of construction of OofA-OA with $m+1$ components from an OofA-OA with m components is described. As we know, the run size for an OofA-OA is a multiple of 12, the resulting design obtained will have a run size $12h(m+1)$ where $1 \leq h \leq (m!/12) - 1$. We denote the existing design d as OofA-OA$(12h, m, 2)$ and the resultant design d' as OofA-OA$(12h(m+1), m+1, 2)$.

Theorem 1: If there exists an OofA-OA for m components, an OofA-OA for $m+1$ components can be obtained from it by placing the $(m+1)$th component in every possible position of each run of the existing OofA-OA.

Table 1: An OofA-OA(12,5,2)

1	5	3	2	4
1	5	4	2	3
2	1	4	3	5
2	3	4	1	5
2	5	1	3	4
2	5	4	3	1
3	1	4	2	5
3	5	1	2	4
3	5	4	2	1
4	1	3	2	5
4	5	1	2	3
4	5	3	2	1

Proof: Adding $(m+1)$th component in $m+1$ positions of each run of the existing OofA-OA results in $m+1$ runs per existing run in the new design. Since we add $(m+1)$th component in every possible position of every run of the existing design, the ratio of frequencies among $n_{++}, n_{+-}, n_{--}, n_{-+}$ in any two columns of the new design d' will be,

$$\frac{n_{++}}{n_{--}} = 1$$

for any synergistic, antagonistic and independent pairs. Similarly, the ratio of

$$\frac{n_{++}}{n_{+-}} = \frac{n_{++}}{n_{-+}} = \frac{n_{--}}{n_{+-}} = \frac{n_{--}}{n_{-+}} = \begin{cases} 2, & \text{for any synergistic pair} \\ 1/2, & \text{for any antagonistic pair} \\ 1, & \text{for any independent pair} \end{cases}$$

These ratios are equal to that of full design P_f and hence are OofA-OA. This completes the proof. □

Example 1: Consider an OofA-OA(12,5,2) from which an OofA-OA for 6 components may be constructed. Table 1 displays the design of OofA-OA(12,5,2) and Table 2 shows the OofA-OA(72,6,2). Here $h = 1$ and the resulting design has run size 72. Here, the component 6 is added in every 6 positions of the OofA-OA(12,5,2) to generate OofA-OA(72,6,2).

As we know, an OofA design with m components has $\binom{m}{2}$ PWO factors and these PWO factors in an OofA-OA can be classified as synergistic pairs, antagonistic pairs and independent pairs. Theorem 2 states the number of synergistic, antagonistic and independent pairs possible for an OofA-OA with m components.

Theorem 2: An OofA-OA with m components have $\binom{\binom{m}{2}}{2} - 3\binom{m}{3}$ independent pairs, $\binom{m}{3}$ antagonistic pairs and $2\binom{m}{3}$ synergistic pairs.

Proof: For an m component OofA design, there are $\binom{m}{2}$ PWO factors under the PWO model. Total number of possible pairs of PWO factors are $\binom{\binom{m}{2}}{2}$ which include all the independent, synergistic and antagonistic pairs. Now, we determine the number of antagonistic

Table 2: An OofA-OA(72,6,2)

6 1 5 3 2 4	6 2 1 4 3 5	6 2 5 1 3 4	6 3 1 4 2 5
1 6 5 3 2 4	2 6 1 4 3 5	2 6 5 1 3 4	3 6 1 4 2 5
1 5 6 3 2 4	2 1 6 4 3 5	2 5 6 1 3 4	3 1 6 4 2 5
1 5 3 6 2 4	2 1 4 6 3 5	2 5 1 6 3 4	3 1 4 6 2 5
1 5 3 2 6 4	2 1 4 3 6 5	2 5 1 3 6 4	3 1 4 2 6 5
1 5 3 2 4 6	2 1 4 3 5 6	2 5 1 3 4 6	3 1 4 2 5 6
6 1 5 4 2 3	6 2 3 4 1 5	6 2 5 4 3 1	6 3 5 1 2 4
1 6 5 4 2 3	2 6 3 4 1 5	2 6 5 4 3 1	3 6 5 1 2 4
1 5 6 4 2 3	2 3 6 4 1 5	2 5 6 4 3 1	3 5 6 1 2 4
1 5 4 6 2 3	2 3 4 6 1 5	2 5 4 6 3 1	3 5 1 6 2 4
1 5 4 2 6 3	2 3 4 1 6 5	2 5 4 3 6 1	3 5 1 2 6 4
1 5 4 2 3 6	2 3 4 1 5 6	2 5 4 3 1 6	3 5 1 2 4 6
6 3 5 4 2 1	6 4 1 3 2 5	6 4 5 1 2 3	6 4 5 3 2 1
3 6 5 4 2 1	4 6 1 3 2 5	4 6 5 1 2 3	4 6 5 3 2 1
3 5 6 4 2 1	4 1 6 3 2 5	4 5 6 1 2 3	4 5 6 3 2 1
3 5 4 6 2 1	4 1 3 6 2 5	4 5 1 6 2 3	4 5 3 6 2 1
3 5 4 2 6 1	4 1 3 2 6 5	4 5 1 2 6 3	4 5 3 2 6 1
3 5 4 2 1 6	4 1 3 2 5 6	4 5 1 2 3 6	4 5 3 2 1 6

pairs. Let $(z_{ij}\ z_{kl})$ be an antagonistic pair for which $i = l$ or $j = k$ is possible. We generally write $(z_{ij}\ z_{kl})$ such that $i < j$ and $k < l$. Thus, only three distinct components are needed for forming an antagonistic pair. Now, 3 distinct components can be taken from m components in $\binom{m}{3}$ ways. Hence, number antagonistic pairs is $\binom{m}{3}$. For synergistic pair, $(z_{ij}\ z_{kl})$, there are two options: (i) $i = k$. If so there are only three components, *i.e.* i, j, l. (ii) $j = l$. If so there are only three components, *i.e.* i, j, k. For both these options, $\binom{m}{3}$ pairs are possible. So, $2\binom{m}{3}$ synergistic pairs are possible for m component OofA-OA. Therefore, number of independent pairs is $\binom{\binom{m}{2}}{2} - 3\binom{m}{3}$. This completes the proof. □

4. Some results on OofA-OA

An OofA-OA of run size N have the following property as specified in Theorem 3.

Theorem 3: If a fractional OofA design with run size N is an OofA-OA$(N, m, 2)$, then,
$n_{++} = n_{--} = N/3, n_{+-} = n_{-+} = N/6$ for any synergistic pair
$n_{++} = n_{--} = N/6, n_{+-} = n_{-+} = N/3$ for any antagonistic pair and
$n_{++} = n_{--} = N/4, n_{+-} = n_{-+} = N/4$ for any independent pair.

Proof: As there are four two-tuples $(++, --, +-, -+)$ in an OofA-OA of strength 2, for any independent pairs of PWO factors, the frequencies of these two-tuples will be same to satisfy the equality of ratio of frequencies of these two-tuples in an OofA-OA with respect to the full OofA design. Effortlessly, we can write, $n_{++} = n_{--} = n_{+-} = n_{-+} = N/4$ for any independent pair. Obviously, the minimum run size required for an OofA-OA of strength 2 is 12. An OofA-OA$(N, m, 2)$ will always be a multiple of 12 which means N is a multiple of 12. To satisfy the ratio mentioned in the proof of Theorem 1, again we need, $n_{++} = n_{--} = N/3, n_{+-} = n_{-+} = N/6$ for any synergistic pair and $n_{++} = n_{--} = N/6, n_{+-} = n_{-+} = N/3$

for any antagonistic pair. This confirms Theorem 3. □

Theorem 4: For any OofA-OA $(N, m, 2)$, consider any two synergistic pairs $(z_{im}\ z_{jm})$ containing the mth component, the corresponding z_{ij} has the following symbols with frequency as given below

Two-tuples $(z_{im}\ z_{jm})$	z_{ij}	Frequency
$++$	$+$	N/6
$++$	$-$	N/6
$+-$	$+$	N/6
$-+$	$-$	N/6
$--$	$+$	N/6
$--$	$-$	N/6

Proof: For an OofA-OA$(N, m, 2)$, for any synergistic pair, $n_{++} = n_{--} = N/3, n_{+-} = n_{-+} = N/6$ according to Theorem 3. We can see that n_{++} and n_{--} for the two-tuples $(z_{im}\ z_{jm})$ is $n_{++} = n_{--} = 2\frac{N}{6} = N/3$. Now, if z_{im} is $+1$ and z_{jm} is -1, z_{ij} will be $+1$ and vice versa. For example, if $1 \to 5$, $z_{15} = +1$; $5 \to 2$, $z_{25} = -1$, then, $z_{12} = +1$. This completes the proof. □

Example 2: Consider an OofA-OA (12,5,2) given in Voelkel (2019). The array is given in transpose form.

$$\begin{pmatrix} 1 & 1 & 2 & 2 & 2 & 2 & 3 & 3 & 3 & 4 & 4 & 4 \\ 5 & 5 & 1 & 3 & 5 & 5 & 1 & 5 & 5 & 1 & 5 & 5 \\ 3 & 4 & 4 & 4 & 1 & 4 & 4 & 1 & 4 & 3 & 1 & 3 \\ 2 & 2 & 3 & 1 & 3 & 3 & 2 & 2 & 2 & 2 & 2 & 2 \\ 4 & 3 & 5 & 5 & 4 & 1 & 5 & 4 & 1 & 5 & 3 & 1 \end{pmatrix}'.$$

The corresponding PWO matrix is given as

$$\mathbf{P} = \begin{array}{c} \begin{matrix} z_{12} & z_{13} & z_{14} & z_{15} & z_{23} & z_{24} & z_{25} & z_{34} & z_{35} & z_{45} \end{matrix} \\ \begin{pmatrix} + & + & + & + & - & + & - & + & - & - \\ + & + & + & + & + & - & - & - & - & - \\ - & + & + & + & + & + & + & - & + & + \\ - & - & - & + & + & + & + & + & + & + \\ - & + & + & - & + & + & + & + & - & - \\ - & - & - & - & + & + & + & - & - & - \\ + & - & + & + & - & - & + & + & + & + \\ + & - & + & - & - & + & - & + & + & - \\ - & - & - & - & - & - & - & + & + & - \\ + & + & - & + & - & - & + & - & + & + \\ + & + & - & - & + & - & - & - & - & + \\ - & - & - & - & - & - & - & - & - & + \end{pmatrix} \end{array}$$

The columns of $\mathbf{P}$ matrix are labelled as $z_{12}, z_{13}, z_{14}, z_{15}, z_{23}, z_{24}, z_{25}, z_{34}, z_{35}$ and z_{45} in the respective order. Consider the synergistic pair (z_{25}, z_{35}) and the frequencies of z_{23} along with the symbol, it is clear that some balance properties are followed as in Table 3.

Table 3: The frequencies of two-tuples of an OofA-OA(12,5,2)

Two-tuples (z_{25}, z_{35})	z_{23}	Frequency
$++$	$+$	2
$++$	$-$	2
$+-$	$+$	2
$-+$	$-$	2
$--$	$+$	2
$--$	$-$	2

It is very interesting to see that this property exists for any OofA-OA. According to Zhao *et al.* (2021), when an OofA-OA is projected onto any $s(\geq 4)$ components, all the $s!$ orders occur equal number of times. Even though, the OofA-OA given in example 2 does not obey order balance property as specified in Zhao *et al.* (2021), balancing of frequency of two-tuples given in Theorem 2 is satisfied. In other words, this property can be utilized to check if a given fractional OofA design is OofA-OA even if it does not satisfy the order balance property.

5. Concluding remarks

Being PWO model as the most promising and acceptable model for OofA problems, the fractional designs under this model which are optimal with regard to any popular optimality criteria has been of considerable interest to the researchers. The OofA-OA is such a fractional design under this model that satisfies D-, A-, M.S.- and χ^2- optimality criteria. In this scenario, we propose a systematic method of constructing OofA-OA having $m+1$ components from an existing OofA-OA with m components. As the resulting design is OofA-OA, it retains efficiency, optimality and balance property. The proposed method is easy to understand and lacks complexity for the construction. However, the run size of the proposed OofA-OA is a fixed number and is not flexible. Hence, we advise future research on systematic construction of OofA-OA with flexible run sizes for which OofA-OA exists. We further introduce a balance property which is applicable to any OofA-OA even if it does not obey the order balance property.

Acknowledgements

The authors are thankful to the Director, ICAR-Indian Agricultural Research institute, New Delhi and The Graduate School, IARI for supporting the research work. Authors are thankful to the Chair Editor and Anonymous Reviewer for their support.

Conflict of interest

The authors do not have any financial or non-financial conflict of interest to declare for the research work included in this article.

References

Black, B. E., Holaska, J. M., Lvesque, L., Ossareh-Nazari, B., Gwizdek, C., Dargemont, C., and Paschal, B. M. (2001). NXT1 is necessary for the terminal step of Crm1-mediated nuclear export. *Journal of Cell Biology*, **152**, 141–155.

Chen, J., Mukerjee, R., and Lin, D. K. J. (2020). Construction of optimal fractional order-of-addition designs via block designs. *Statistics and Probability Letters*, **161**, 1–11.

Ding, X., Matsuo, K., Xu, L., Yang, J., and Zheng, L. (2015). Optimized combinations of bortezomib, camptothecin, and doxorubicin show increased efficacy and reduced toxicity in treating oral cancer. *Anti-Cancer Drugs*, **26**, 547–554.

Fisher, R. A. (1971). *The Design of Experiments*. 9th Ed., Macmillan, London.

Jourdain, L. S., Schmitt, C., Leser, M. E, Murray, B. S., and Dickinson, E. (2009). Mixed layers of sodium caseinate + dextran sulfate: Influence of order of addition to oil -water interface. *Langmuir*, **25**, 10026–10037. doi: 10.1021/la900919w.

Karim, M., McCormick, K., and Kappagoda, C. T. (2000). Effects of cocoa extracts on endothelium-dependent relaxation. *The Journal of Nutrition*, **130**, 2105S–2108S.

Lin, D. K. J. and Peng, J. Y. (2019). Design and analysis of order of addition experiments: A review and some thoughts. *Quality Engineering*, **31**, 49–59.

Mee, R. W. (2020). Order-of-addition modeling. *Statistica Sinica*, **30**, 1543–1559.

Peng, J., Mukerjee, R., and Lin, D. K. J. (2019). Design of order-of-addition experiments. *Biometrika*, **106**, 683–694.

Tsai, S. (2022). Generating optimal order-of-addition designs with flexible run sizes. *Journal of Statistical Planning and Inference*, **218**, 147–163.

Van Nostrand, R. C. (1995). *Design of experiments where the order of addition is important*. In ASA Proceedings of the Section on Physical and Engineering Sciences, American Statistical Association, Alexandria, Virginia, 155–160.

Voelkel, J. G. (2019). The designs of order-of-addition experiments. *Journal of Quality Technology*, **51**, 230–241.

Voelkel, J. G. and Gallagher, K. P. (2019). The design and analysis of order-of-addition experiments: An introduction and case study, *Quality Engineering*, **31(4)**, 627–638.

Wagner, J. J. (1995). Sequencing of feed ingredients for ration mixing. South Dakota Beef Report. `http://openprairie.sdstate.edu/sd_beefreport_1995/15`.

Williams, E. J. (1949). Experimental designs balanced for the estimation of residual effects of treatments. *Australian Journal of Scientific Research*, Series A, Physical Sciences, **2**, 149–168.

Wilson, K. J. and Henderson, D. A. (2018). Emulation of utility functions over a set of permutations: sequencing reliability growth tasks. *Technometrics*, **60**, 273–285.

Winker, P., Chen, J., and Lin, D. K. J. (2020). *Contemporary Experimental Design, Multivariate Analysis and Data Mining*, Springer Nature Switzerland AG 2020. `https://doi.org/10.1007/978-3-030-46161-4_6`.

Yang, J. F., Sun, F., and Xu, H. (2021). A component-position model, analysis and design for order-of -addition experiments. *Technometrics*, **63**, 212–224.

Zhao, Y., Lin, D. K. J., and Liu, M. (2020). Designs for order of addition experiments. *Journal of Applied Statistics*, **48**, 1475–1495.

Zhao, Y., Lin, D. K. J., and Liu, M. (2021). Optimal designs for order-of-addition experiments. *Computational Statistics and Data Analysis*, **165**, 1–15.

Zhao, S., Dong, Z., and Zhao, Y. (2022). Order-of-addition orthogonal arrays with high strength. *Mathematics*, **10**, 1187. `https://doi.org/10.3390/math10071187`.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 23, No. 1, 2025 (New Series), pp 197–216
https://www.ssca.org.in/journal

A Hybrid ARIMA-GARCH Type Copula Approach for Agricultural Price Forecasting

B. Manjunatha[1], Ranjit Kumar Paul[2], Ramasubramanian V.[3], Amrit Kumar Paul[2], Md. Yeasin[2], Mrinmoy Ray[2], G. Avinash[4] and Chandan Kumar Deb[2]

[1]*The Graduate School, ICAR-Indian Agricultural Research Institute, New Delhi*
[2]*ICAR-Indian Agricultural Statistics Research Institute, New Delhi*
[3]*ICAR-National Academy of Agricultural Research Management, Hyderabad*
[4] *ICMR-National Institute of Occupational Health, Ahmedabad*

Received: 19 April 2024; Revised: 05 June 2024; Accepted: 07 June 2024

Abstract

Agricultural commodity prices frequently exhibit inherent noise and volatility, attributable to market dynamics. This paper undertakes a comprehensive analysis of price volatility concerning key oil seed crops (Safflower, Mustard, Groundnut) and pulses (Lentil, Chickpea, Green gram) across two markets for each commodity in the Indian agricultural sector. The present study aims to improve the accuracy of price forecasting by utilizing the Bivariate Auto Regressive Integrated Moving Average (ARIMA)-Generalized Auto Regressive Conditional Heteroskedasticity (GARCH) type-Copula model. Monthly agricultural commodity price datasets for key oil seed crops and pulse crops spanning January 2010 to December 2022 have been used to evaluate the predictive performance of this model. Comparative evaluations are carried out against conventional time series models, namely Multivariate GARCH (MGARCH)-Dynamic Conditional Correlation (DCC) model and the Univariate ARIMA-GARCH model. Empirical findings demonstrate that the Bivariate ARIMA-GARCH type-Copula model outperformed the conventional time series models considered in forecasting performance. This superiority is evidenced by evaluation metrics, including Root Mean Square Error (RMSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). Moreover, this study utilized the Diebold–Mariano test to highlight the predictive accuracy of the Bivariate ARIMA-GARCH type-Copula model for the dataset under consideration, surpassing conventional time series models. The integration of Copulas with the ARIMA-GARCH type model shows promise for enhancing price forecasting accuracy, offering valuable insights for researchers and policymakers navigating the dynamic agricultural market landscape in India.

Key words: Gaussian Copula; Simulation; Student t-Copula; Time Series; Volatility.

AMS Subject Classifications: 62K05, 05B05

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

1. Introduction

Agriculture is a crucial component of the Indian economy, supporting over 47% of the population's livelihood. As stated in the 2022-23 Indian Economic Survey, the agricultural sector has demonstrated consistent growth, averaging an annual rate of 4.6% over the past six years. Agricultural commodities price data are often characterized by inherent noise and volatility due to the nature of the market. This is largely due to the rapid response of these prices to changes in supply and demand conditions, as well as the impact of weather-related factors on farm production. Moreover, asymmetric phenomena can also arise in price series, where prices tend to behave differently during economic downturns as opposed to periods of growth. It is common for agricultural price series to exhibit periods of stability, followed by periods of high volatility. These fluctuations are a common feature of the agricultural commodity market. Monitoring volatility in agricultural commodity prices can have a significant impact on a nation's overall economic performance. As such, agricultural commodity price forecasts are essential in enabling decision-makers to formulate economic policies and strategies that are in line with anticipated changes (Bhardwaj *et al.* (2014)).

One of the predominant statistical methodologies employed in forecasting price series is the Auto Regressive Integrated Moving Average (ARIMA) model as established by Box and Jenkins (1970). Nevertheless, the inherent assumptions of linearity and homoscedastic error variance within the ARIMA framework might not adequately accommodate the forecasting challenges posed by volatile agricultural commodity prices. In recognition of this limitation, Engle (1982) introduced the Auto Regressive Conditional Heteroscedastic (ARCH) model, subsequently refined by Bollerslev (1986) into the Generalized ARCH (GARCH) model. Volatility within agricultural commodity price series often exhibit both symmetric and asymmetric patterns. Although the GARCH model effectively captures the magnitude of shocks, it may not fully capture the directional characteristics of these shocks. Consequently, alternative asymmetric GARCH-type models have been devised, such as the Exponential GARCH (EGARCH) model proposed by Nelson (1991), the GJR-GARCH model introduced by Glosten *et al.* (1993), and the Asymmetric Power ARCH (APARCH) model presented by Ding *et al.* (1993). Various studies have endeavored to apply both ARIMA and GARCH models in forecasting agricultural commodity prices. Examples of such investigations include those conducted by Paul *et al.* (2009), Bhardwaj *et al.* (2014) and Dinku (2021). Moreover, the integration of ARIMA and GARCH methodologies, known as ARIMA-GARCH models, has emerged as a viable approach for forecasting agricultural commodity prices. This fusion has been demonstrated in research conducted by (Mitra and Paul (2017) and Merabet *et al.* (2021)).

The dynamics of agricultural price volatilities exhibit interdependency across commodities and markets, prompting an increased scholarly emphasis on quantifying the interdependence within agricultural price series data. However, conventional Time Series (TS) models, such as ARIMA and GARCH models, often neglect the pivotal aspect of interdependency among different series. To address this deficiency, the Vector Auto Regressive (VAR) model was introduced, enabling the exploration of linear interrelationships among multiple TS. VAR model's efficacy in capturing the volatile nature of TS data is limited. In response, the Multivariate GARCH (MGARCH) model emerged as a potential solution to this challenge.

A variety of MGARCH models have been developed over time. Engle and Kroner (1995) introduced the BEKK (Baba, Engle, Kraft, and Kroner) model, which represents a multivariate extension of the GARCH model and offers substantial flexibility in modeling. Bollerslev (1990) proposed the Constant Conditional Correlation (CCC) model, providing a relatively flexible approach that combines univariate GARCH models while assuming constant correlation among series over time. Additionally, Engle (2002) introduced the Dynamic Conditional Correlation (DCC) model, a novel class of Multivariate GARCH (MGARCH) model that combines the flexibility of univariate GARCH models with a parsimonious parametric framework for modeling correlations. Several studies have demonstrated the superiority of MGARCH models compared to univariate GARCH models in forecasting agricultural commodity prices (Wang and Wu (2012); Aziz and Iqbal (2016)). The application of MGARCH models for modeling the degree of interactions among various volatile agricultural commodities and markets is widely documented in the literature (Musunuru (2014); Sanjuán-López and Dawson (2017)).

MGARCH models often rely on assumptions of Multivariate Normal (MVN) distribution or Multivariate t (MV-t) distributions for the innovations. MVN distributions assume that each variable follows a univariate normal distribution, which may not hold true in many real world situations where variables exhibit non-normal distributions or complex relationships. Additionally, the Pearson correlation coefficient used in MVN distribution assumes linearity in the relationships between variables, limiting its ability to capture non-linear relationships that are often present. This limitation extends to MV-t distributions as well. To address these shortcomings, Copula-GARCH models have been introduced, where GARCH model combined with Copula model. The Copula is employed to capture dependency between related TS by focusing on their joint distribution and offering flexibility in modeling complex nonlinear dependencies. Sklar (1959)'s theorem is central to the theory of Copulas which states that "any multivariate distribution function can be represented as a composition of its univariate marginal distributions and a Copula", where the Copula captures all the dependencies in the joint distribution. In other words, Copula-based modelling provides the capacity to isolate the dependence structure from marginal distributions of the related TS. Various applications of Copula-GARCH model for portfolio risk estimation on financial TS data can be found in Weiß (2013); Lu *et al.* (2014); Karmakar (2017).

Previous studies utilized the ARIMA-GARCH copula model, initially fitting individual TS data and then employing residuals to model Copulas for joint distributions. These copula models used to analyze correlations among different TS, exploring various statistical measures such as skewness, kurtosis, and fat-tails (Li *et al.* (2020); Shahriari *et al.* (2023)). However, an evident research gap exists as copula models have not been utilized for forecasting future data points. Understanding the price dynamics is crucial, especially in agriculture. The present study pioneers using the Bivariate Copula-GARCH type model for forecasting agricultural prices. After fitting individual TS data to the ARIMA-GARCH type model, residuals are used to fit copula models for joint distributions. Future data points are forecasted by simulating observations from the estimated bivariate distribution function. This advanced modeling technique enhances forecasting accuracy for the data under consideration.

The rest of the manuscript is organized as follows: Section 2 presents a description of the models utilized, Section 3 discusses empirical findings, and Section 4 offers concluding remarks.

2. Material and methods

2.1. ARIMA model

The Box Jenkins ARIMA model, represented in Eq. (1), stands as the predominant technique for forecasting TS data:

$$\phi_p(B)(1-B)^d y_t = c + \theta_q(B)\varepsilon_t \tag{1}$$

where,

$$\phi_p(B) = 1 - \phi_1 B - \phi_2 B^2 - \cdots - \phi_p B^p$$

$$\theta_q(B) = 1 - \theta_1 B - \theta_2 B^2 - \cdots - \theta_q B^q$$

Here, y_t represents the value of current time; c is the constant term; B denotes backward shift operator; ε_t represents the error term; $\phi_1, \phi_2, \ldots, \phi_p$ denote Auto-Regressive (AR) coefficients of order p; $\theta_1, \theta_2, \ldots, \theta_q$ represent Moving Average (MA) coefficients of order q; d is the order of differencing.

2.2. ARCH and GARCH models

ARIMA models are limited in their ability to capture the volatility inherent in TS data and cannot adequately describe changes in conditional variances observed in real-world datasets. To address the inadequacies of ARIMA model, Engle (1982) proposed Auto Regressive Conditional Heteroscedastic (ARCH) model represented in Eq. (2). The ARCH model for the series $\{\varepsilon_t\}$ is characterized by defining the conditional distribution of ε_t given the information available up to time $t-1$, denoted as Ψ_{t-1}. The ARCH model for the series ε_t can be expressed as:

$$\varepsilon_t | \Psi_{t-1} \sim N(0, h_t) \quad \text{and} \quad \varepsilon_t = \sqrt{h_t}\nu_t$$

where h_t is conditional variance, ν_t is identically and independently distributed (iid) innovations with zero mean and unit variance. The conditional variance h_t is defined as

$$h_t = \alpha_0 + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 \tag{2}$$

The conditions of $\alpha_0 > 0, \alpha_i \geq 0 \forall i$ and $\sum_{i=1}^{q} \alpha_i < 1$ are necessary and sufficient to guarantee non-negativity and a finite conditional variance for h_t. Here, α_i denotes the coefficients indicating the impact of past shocks on the current volatility.

In response to certain shortcomings of the ARCH model, such as the rapid decay of the unconditional autocorrelation function of squared residuals, non-parsimony *etc.*, Bollerslev (1986) introduced the Generalized ARCH (GARCH) model. The variance equation of GARCH model is represented in Eq. 3 as:

$$h_t = \alpha_0 + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + \sum_{j=1}^{p} \beta_j h_{t-j} \tag{3}$$

where $\alpha_0 > 0, \alpha_i \geq 0 \forall i, \beta_j \geq 0 \forall j$. Here, β_j denotes the coefficients indicating the impact of past volatilities on the current volatility. While the GARCH model excels at capturing overall volatility in TS, it falls short when it comes to asymmetric impacts of positive and negative events. To address this limitation, various asymmetric GARCH-type of models have evolved namely EGARCH, GJR-GARCH and APARCH model stated subsequently.

2.3. Asymmetric GARCH-type models

2.3.1. EGARCH model

EGARCH model addresses asymmetric volatility without parameter constraints. It models the conditional variance, h_t, as an asymmetric function of lagged disturbances, defined by Eq. (4).

$$\ln(h_t) = \alpha_0 + \sum_{j=1}^{p} \beta_j \ln(h_{t-j}) + \sum_{i=1}^{q} \left(\alpha_i \left| \frac{\varepsilon_{t-i}}{\sqrt{h_{t-i}}} \right| + \lambda_i \frac{\varepsilon_{t-i}}{\sqrt{h_{t-i}}} \right) \tag{4}$$

where λ_i represents the asymmetric parameter, capturing asymmetric effects due to external shocks.

2.3.2. GJR-GARCH model

GJR-GARCH model considers the impact of ε_{t-1}^2 on the conditional variance, depending on the sign of ε_{t-1}. They introduced an indicator variable to capture this sign dependence. The GJR-GARCH model is represented in Eq. (5).

$$h_t = \alpha_0 + \sum_{j=1}^{p} \beta_j h_{t-j} + \sum_{i=1}^{q} \alpha_i \varepsilon_{t-i}^2 + \gamma \varepsilon_{t-1}^2 I_{t-1} \tag{5}$$

where γ $(-1 < \gamma < 1)$ denote the asymmetric parameter, and I_{t-1} is the indicator variable, such that

$$I_{t-1} = \begin{cases} 1 & \text{if } \varepsilon_{t-1} < 0 \\ 0 & \text{if } \varepsilon_{t-1} \geq 0 \end{cases}$$

2.3.3. APARCH model

The APARCH model incorporates asymmetric power into the conditional variance, specified as represented in Eq. (6).

$$h_t^{\delta/2} = \alpha_0 + \sum_{j=1}^{p} \beta_j h_{t-j}^{\delta/2} + \sum_{i=1}^{q} \alpha_i \left(|\varepsilon_{t-i}| - \gamma \varepsilon_{t-i}\right)^{\delta} \tag{6}$$

where γ $(-1 < \gamma < 1)$ denotes the asymmetric parameter, and $\delta(> 0)$ denotes the power term parameter. An application of different asymmetric GARCH type models can be found in Rakshit *et al.* (2021).

2.4. ARIMA-GARCH type models

ARIMA-GARCH type models integrate ARIMA for capturing linear dynamics and various GARCH models (e.g., GARCH, EGARCH, GJR-GARCH, and APARCH) to address volatility clustering. The ARIMA component accounts for linear behaviour in the first stage, thereby leaving nonlinear components in residuals. Paul *et al.* (2014) developed formulae for out-of-sample forecast using ARIMA-GARCH model. Paul (2015) applied ARIMA-GARCH model for forecasting volatility in agricultural crop yield.

The presence of serial autocorrelation in residuals from the ARIMA model is typically assessed using the Ljung-Box test, a statistical test proposed by Ljung and Box (1978). Meanwhile, the existence of heteroscedasticity in these residuals is evaluated through the ARCH Lagrange Multiplier (LM) test, introduced by Engle (1982). If serial correlation and heteroscedasticity are detected in the residuals based on the results of the Ljung-Box test and ARCH-LM tests, respectively, the residuals are then subjected to a GARCH model. GARCH is employed to model these residual patterns comprehensively, thereby capturing both mean and volatility dynamics effectively. This approach ensures a thorough analysis of both linear and nonlinear components in the data, enhancing the overall modelling accuracy and robustness. The schematic representation of ARIMA-GARCH type model is illustrated in Figure 1.

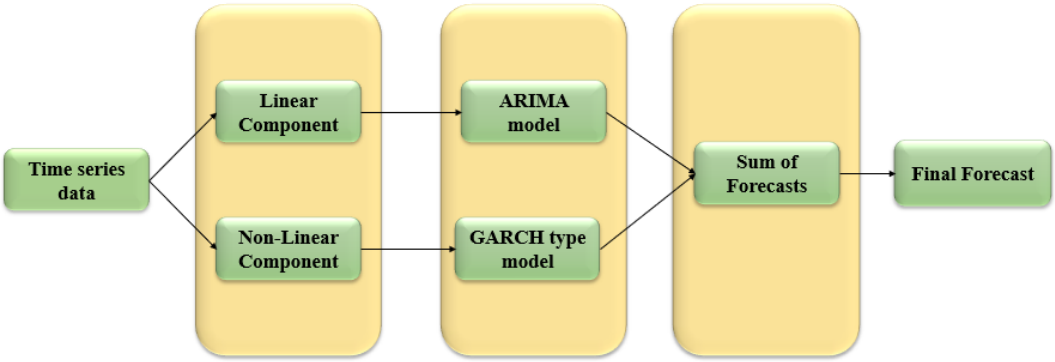


Figure 1: ARIMA-GARCH type model

2.5. Copula

Copulas have been introduced by applied mathematician, Sklar (1959). Copula comes from the Latin word "copulature" which means "to join together". Copulas are handled by

utilizing Probability Integral Transformation (PIT) and Inverse Probability Integral Transformation (Inverse PIT) which are described subsequently.

PIT: Suppose that a random variable X has a continuous distribution for which the Cumulative Distribution Function (CDF) is F_X. Then the random variable U defined by PIT as $U = F_X(X)$ has a standard uniform distribution.

Inverse PIT: Given a continuous standard uniform variable U and an invertible CDF G_X^{-1}, the random variable X defined by Inverse PIT $X = G_X^{-1}(U)$ has distribution function G_X.

Accordingly, the formal definition of Copula is as follows: Let $X = (X_1, X_2, \ldots, X_d)$ be a vector of random variables with their marginal CDFs $F_1, F_2, \ldots, F_d$ as continuous functions. By applying the PIT to each component, obtain the U vector containing $U_1, U_2, \ldots, U_d$ random variables; here each variable will follow standard uniform distribution as

$$U = (U_1, U_2, \ldots, U_d) = [F_1(X_1), F_2(X_2), \ldots, F_d(X_d)]$$

Then, Copula C is a joint cumulative distribution function of d random variables given by

$$C(U_1, U_2, \ldots, U_d) = H[G_1^{-1}(U_1), G_2^{-1}(U_2), \ldots, G_d^{-1}(U_d)]$$

To overcome the limitation of Pearson correlation coefficient, Copula modeling utilizes Spearman's rank correlation coefficient, a nonparametric measure of correlation. It avoids distributional assumptions and linear relationships. Nonparametric correlation measures allow flexible analysis, accommodating non-linear patterns and non-normal data.

2.5.1. Bivariate gaussian copula

Let Φ_{xy} be the distribution function of a standardised bivariate normal CDF and Φ^{-1} be the inverse of standard normal CDF, and ρ is the Spearman rank correlation coefficient (i.e. dependence parameter) between the components. Then the bivariate Gaussian Copula CDF is expressed as shown in Eq. (7).

$$C_\rho(u_1, u_2) = \Phi_{xy}[\Phi^{-1}(u_1), \Phi^{-1}(u_2); \rho] \tag{7}$$

Let $s = \Phi^{-1}(u_1)$ and $t = \Phi^{-1}(u_2)$, then the Gaussian Copula density is given by Eq. (8).

$$c_\rho(u_1, u_2) = \frac{1}{\sqrt{1-\rho^2}} \exp\left\{\frac{-(\rho^2 s^2 + \rho^2 t^2 - 2\rho st)}{2(1-\rho^2)}\right\} \tag{8}$$

2.5.2. Bivariate Student-t copula

When the interest focuses on modelling data which exhibits heavy-tailed behaviour, the Student-t Copula may be used instead of the Gaussian Copula.

Let t_{xy} be the distribution function of a standardised bivariate Student-t CDF and t^{-1} be the inverse of standard Student's t CDF with η degrees of freedom and ρ dependence

parameter is the Spearman rank correlation coefficient between the components, then the bivariate Student-t Copula CDF is expressed as shown in Eq. (9).

$$C_{\rho\eta}(u_1, u_2) = t_{xy}[t_\eta^{-1}(u_1), t_\eta^{-1}(u_2); \rho] \tag{9}$$

Let $s = t_\eta^{-1}(u_1)$ and $r = t_\eta^{-1}(u_2)$, then the Student-t Copula density is given by Eq. (10) as follows:

$$c_{\eta\rho}(u_1, u_2) = \frac{\Gamma\left(\frac{\eta+2}{2}\right)\Gamma\left(\frac{\eta}{2}\right)}{\sqrt{1-\rho^2}\Gamma^2\left(\frac{\eta+1}{2}\right)} \left\{\left(1+\frac{s^2}{\eta}\right)\left(1+\frac{r^2}{\eta}\right)\right\}^{(\eta+1)/2} \left(1+\frac{s^2+r^2-2\rho sr}{\eta(1-\rho^2)}\right)^{-(\eta+2)/2} \tag{10}$$

2.6. Bivariate ARIMA-GARCH type-Copula model

2.6.1. ARIMA-GARCH type model selection

ARIMA-GARCH type models, *viz.*, ARIMA-GARCH, ARIMA-EGARCH, ARIMA-GJR-GARCH and ARIMA-APARCH, are fitted to the two TS data independently. The optimal ARIMA-GARCH type model is selected based on minimum value of Akaike Information Criterion (AIC) and Bayesian Information Criterion (BIC) for the two TS. From the optimal ARIMA-GARCH type model, mean and volatility forecasts are obtained for both TS. While these models capture temporal dependency within the individual TS, a Copula model is employed to capture dependency between two TS.

2.6.2. Copula modeling

The residuals of the fitted ARIMA-GARCH type models for the two TS are employed for Copula modeling. The Spearman rank correlation coefficient ρ is utilized to assess the relationship between the residuals of the two TS. If the residuals of the two fitted TS are not significantly correlated, then the forecasts from the optimal ARIMA-GARCH type models for each TS are considered the final predictions. However, if the residuals are significantly correlated, Copula modeling is employed. In Copula modeling, the residuals of both TS are transformed using the PIT. The transformed values and estimated dependence parameter ρ of copula are then utilized to fit both Gaussian Copula and Student-t Copula functions. The optimal Copula function is selected based on the AIC and BIC criteria. The schematic representation of the methodology of Bivariate ARIMA-GARCH type-Copula model is illustrated in Figure 2.

2.6.3. One day ahead forecast through simulation

The optimal Copula function used to obtain the bivariate distribution (joint distribution) of residuals from an ARIMA-GARCH-type model is applied to two TS. By simulating a large number of observations from the estimated bivariate distribution function through random sampling, multiple potential future scenarios are generated. These scenarios incorporate uncertainty, variability, and the complex relationships between the residuals of the two TS, helping capture the range of possible future outcomes more comprehensively. The step

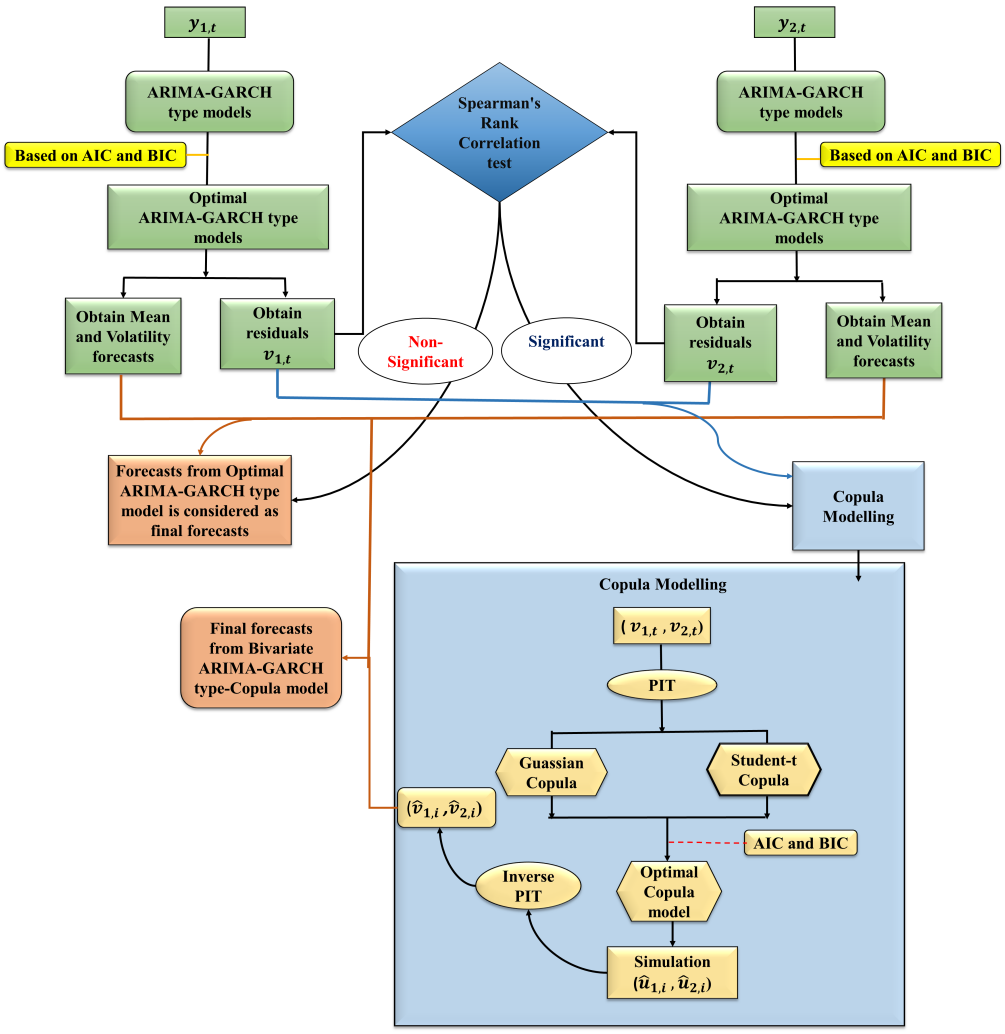


Figure 2: Flow chart of Bivariate ARIMA-GARCH type-Copula

by step algorithm to obtain one day ahead forecast through simulation can be summarized as follows:

1. Simulate n pairs of random samples $(\hat{u}_{1,i}, \hat{u}_{2,i})$ from the estimated Optimal Copula function. Here $\hat{u}_{1,i}$ and $\hat{u}_{2,i}$ denote the simulated values for the residuals of the optimal ARIMA-GARCH type models of the first and second TS, respectively, where $1 \leq i \leq n$.

2. To ensure that the simulated values of residuals are in their respective original scales, inverse PIT is applied to obtain transformed values $(\hat{v}_{1,i}, \hat{v}_{2,i})$.

3. Multiply $\hat{v}_{1,i}$ by the respective predicted one-day ahead volatility $\sqrt{h_{1,t}}$ from the optimal GARCH type model for the first TS, and multiply $\hat{v}_{2,i}$ by the respective predicted one-day ahead volatility $\sqrt{h_{2,t}}$ from the optimal GARCH type model for the second TS.

$$\hat{\varepsilon}_{1,i} = \hat{v}_{1,i}\sqrt{h_{1,t}} \quad \text{and} \quad \hat{\varepsilon}_{2,i} = \hat{v}_{2,i}\sqrt{h_{2,t}}$$

4. Obtain $(\hat{\mu}_{1,i}, \hat{\mu}_{2,i})$ by adding the mean forecast from the ARIMA model for the first and second TS to the $\hat{\varepsilon}_{1,i}$ and $\hat{\varepsilon}_{2,i}$ respectively.

$$\hat{\mu}_{1,i} = \hat{\mu}_{1,t} + \hat{\varepsilon}_{1,i} \quad \text{and} \quad \hat{\mu}_{2,i} = \hat{\mu}_{2,t} + \hat{\varepsilon}_{2,i}$$

5. Take average to obtain one day ahead forecasts $(\hat{k}_{1,t}, \hat{k}_{2,t})$ of both TS

$$\hat{k}_{1,t} = \frac{1}{n}\sum_{i=1}^{n}\hat{\mu}_{1,i} \quad \text{and} \quad \hat{k}_{2,t} = \frac{1}{n}\sum_{i=1}^{n}\hat{\mu}_{2,i}$$

$\hat{k}_{1,t}$ and $\hat{k}_{2,t}$ are considered as one day ahead forecasts from Bivariate ARIMA-GARCH type-Copula model for first and second TS, respectively.

3. Data and empirical findings

3.1. Data description

In this study, we collected monthly agricultural commodity price data for three oilseed crops and three pulse crops from two primary markets for each commodity. The data was obtained from the AGMARKNET portal of the Ministry of Agriculture and Farmers Welfare, Government of India (`https://agmarknet.gov.in/`), covering the period from January 2010 to December 2022. The selection of major markets was based on their significant arrival quantities. The chosen markets are detailed below:

Oilseeds:

- Safflower: Latur (Maharashtra) and Kalaburagi (Karnataka)
- Mustard: Sri Ganganagar (Rajasthan) and Satna (Madhya Pradesh)
- Groundnut: Gondal (Gujarat) and Bikaner (Rajasthan)

Pulses:

- Lentil: Banda (Uttar Pradesh) and Narsinghpur (Madhya Pradesh)
- Chickpea: Hinganghat (Maharashtra) and Dewas (Madhya Pradesh)
- Green gram: Bhagat Ki Kothi (Rajasthan) and Kalaburagi (Karnataka)

Each agricultural commodity price dataset contained 156 observations, the series was divided into training and testing sets. The training set consisted of 144 months of observations, which were used for model building. The last 12 months of observations were

Table 1: Descriptive statistics of monthly agricultural commodity price data

Commodity	Markets	Mean	S.D.	C.V (%)	Skew	Kurt	Minimum	Maximum
Safflower	Latur	3301.18	844.03	25.57	0.81	0.08	1985.71	5755.00
	Kalaburagi	3236.95	940.20	29.05	0.71	-0.12	1795.95	5944.72
Mustard	Sri Ganganagar	3922.62	1268.29	32.33	1.16	0.72	2153.15	7679.16
	Satna	3690.72	1246.44	33.77	1.18	0.72	1900.00	7397.23
Groundnut	Gondal	4518.99	1065.89	23.59	0.28	-0.90	2796.08	6933.18
	Bikaner	4055.43	951.53	23.46	0.49	-0.41	2456.59	6540.67
Lentil	Banda	4328.20	1220.39	28.20	0.45	-0.76	2279.00	7277.69
	Narsinghpur	4255.75	1183.62	27.81	0.45	-0.93	2466.35	7022.16
Chickpea	Hinganghat	3627.42	1090.69	30.07	0.89	1.81	1835.11	7629.75
	Dewas	3859.89	1260.67	32.66	1.13	2.84	1835.09	8871.43
Green gram	Bhagat Ki Kothi	5413.40	1267.25	23.41	-0.18	-0.75	2025.00	8270.67
	Kalaburagi	5254.95	1125.54	21.42	-0.21	-0.68	2612.50	8132.74

S.D.: Standard Deviation, C.V.: Coefficient of Variation, Skew: Skewness, Kurt: Kurtosis

kept for validating the model. The Table 1 presents key statistics for various commodities across different markets.

Green gram in Bhagat Ki Kothi market stands out with the highest mean price, while safflower in Kalaburagi market records the lowest. Mustard in Satna market exhibits the highest coefficient of variation (C.V.), indicating considerable price variability, while green gram in Kalaburagi shows the lowest. Dewas for Chickpea reports the highest maximum price, and safflower in Kalaburagi reflects the lowest minimum. Positively skewed distributions are observed in most of the agricultural commodity markets except the green gram market in Bhagat Ki Kothi and Kalaburagi, which display negative skewness. Leptokurtic distributions are evident in Mustard in Sri Ganganagar, Satna, Chickpea in Hinganghat and Dewas, while Safflower in Latur exhibit approximately mesokurtic distributions. The remaining commodities markets demonstrate platykurtic distributions.

3.2. Test for normality

To evaluate normality of agricultural commodity price data, most widely used statistical tests, *viz.*, Jarque-Bera test (Jarque and Bera (1987)) and Shapiro-Wilk test (Shapiro and Wilk (1965)) were employed. The results of these normality tests are presented in Table 2, indicating that the majority of agricultural commodity markets show significant deviations from normality at 1% level as evidenced by low p-values (<0.01), the Green gram prices in the Bhagat Ki Kothi and Kalaburagi markets are significant at the 5% level with p-value below 0.05 from the Jarque-Bera test and Shapiro-Wilk tests. Hence all the agricultural commodity markets price data considered were non-normal.

3.3. Test for stationarity

The stationarity of data is a crucial property of TS analysis. A series is considered stationary if it maintains a constant mean and variance over time. To assess stationarity, several statistical tests namely the Augmented Dickey-Fuller (ADF) test (Dickey and Fuller (1979)), the Phillips-Perron (PP) test (Phillips and Perron (1988)) and the Kwiatkowski-Phillips-Schmidt-Shin (KPSS) test (Kwiatkowski *et al.* (1992)) were employed. The null hypothesis for the ADF and PP tests states that the series is non-stationary, while for the KPSS test, it suggests that the series is stationary. Table 3 presents stationarity test results

Table 2: Normality test results of agricultural commodity price data of different markets

Commodity	Markets	Jarque-Bera test		Shapiro-Wilk test	
		Statistic	p-value	Statistic	p-value
Safflower	Latur	17.322	0.0002	0.936	< 0.0001
	Kalaburagi	13.489	0.0012	0.944	< 0.0001
Mustard	Sri Ganganagar	39.849	< 0.0001	0.871	< 0.0001
	Satna	40.732	< 0.0001	0.861	< 0.0001
Groundnut	Gondal	17.056	0.0029	0.966	0.0008
	Bikaner	17.397	0.0025	0.967	0.0009
Lentil	Banda	8.978	0.0113	0.957	0.0002
	Narsinghpur	10.663	0.0048	0.944	< 0.0001
Chickpea	Hinganghat	44.062	< 0.0001	0.932	< 0.0001
	Dewas	89.206	< 0.0001	0.915	< 0.0001
Green gram	Bhagat Ki Kothi	8.185	0.0124	0.978	0.0154
	Kalaburagi	8.948	0.0138	0.978	0.0168

for agricultural commodity price data across various markets. All agricultural commodity markets price series are deemed non-stationary, as indicated by the p-values.

Table 3: Stationarity test results of agricultural commodity price data of different markets

Commodity	Markets	ADF test		PP test		KPSS test	
		Statistic	p-value	Statistic	p-value	Statistic	p-value
Safflower	Latur	-2.042	0.559	-4.202	0.874	1.788	< 0.01
	Kalaburagi	-2.012	0.572	-8.827	0.608	2.251	< 0.01
Mustard	Sri Ganganagar	-2.283	0.458	-8.068	0.652	2.179	< 0.01
	Satna	-2.357	0.427	-6.532	0.741	2.163	< 0.01
Groundnut	Gondal	-2.224	0.483	-17.24	0.125	1.566	< 0.01
	Bikaner	-2.232	0.479	-9.945	0.495	1.941	< 0.01
Lentil	Banda	-1.820	0.651	-9.350	0.578	1.586	< 0.01
	Narsinghpur	-1.789	0.663	-8.301	0.638	1.474	< 0.01
Chickpea	Hinganghat	-3.109	0.114	-16.948	0.142	1.201	< 0.01
	Dewas	-2.731	0.272	-14.232	0.298	1.111	< 0.01
Green gram	Bhagat Ki Kothi	-1.967	0.596	-16.988	0.195	0.985	< 0.01
	Kalaburagi	-2.918	0.193	-12.751	0.383	0.868	< 0.01

3.4. Residual analysis

Suitable ARIMA model is selected based on minimum AIC and BIC criteria and also observing the significance of autocorrelation and partial autocorrelation functions. Subsequently, the residuals from the ARIMA model undergo diagnostics measures.

3.4.1. Ljung-Box test for serial autocorrelation

The Ljung-Box test is utilized to assess the presence of serial autocorrelation in residuals from the ARIMA model. The null hypothesis suggests that the residuals exhibit no

autocorrelation for a fixed number of lags. A rejection of this hypothesis indicates the presence of serial autocorrelation. Table 4 presents Ljung-Box test results, revealing significant autocorrelation in agricultural commodity price data from all specified markets.

3.4.2. ARCH lagrange multiplier (LM) test for heteroscedasticity

The ARCH LM test is employed to evaluate the existence of heteroscedasticity in residuals. It examines whether the variance of the residuals is constant over time or not. The null hypothesis states that the residuals are homoscedastic, while a rejection of null hypothesis suggests the presence of heteroscedasticity. In addition Table 4 presents ARCH-LM test results, indicating heteroscedasticity in agricultural commodity price data across all specified markets.

Table 4: Ljung-Box and ARCH-LM test statistic results of agricultural commodity price data of different markets

Commodity	Markets	Ljung-Box	ARCH-LM
Safflower	Latur	10.23	44.19
	Kalaburagi	14.51	37.82
Mustard	Sri Ganganagar	6.88	58.44
	Satna	6.95	78.67
Groundnut	Gondal	7.81	28.57
	Bikaner	12.91	41.36
Lentil	Banda	10.45	27.47
	Narsinghpur	10.94	46.33
Chickpea	Hinganghat	22.14	40.34
	Dewas	12.16	62.04
Green gram	Bhagat Ki Kothi	7.28	23.98
	Kalaburagi	14.76	20.36

Note: The test statistics provided in the table lead to p-values of less than 0.01 for all cases.

3.4.3. Broock-Dechert-Scheinkman (BDS) test for nonlinearity

The nonparametric Broock-Dechert-Scheinkman (BDS) test (Broock *et al.* (1996)) is utilized to test the nonlinearity of the residual series. This test assesses whether the residuals exhibit nonlinear dependence. The null hypothesis assumes that the residuals are independently and identically distributed (iid). A rejection of this hypothesis indicates nonlinearity in the residuals. The results of the BDS test, presented in Table 5, indicate the possible presence of nonlinear patterns in the residuals of the ARIMA model at 1% significance level in all the agricultural markets price series.

It is evident that autocorrelation, heteroscedasticity, and nonlinearity is detected in the residuals based on the results of the aforementioned tests, hence residuals are then subjected to a GARCH type models such as standard GARCH, EGARCH, GJR-GARCH, and APARCH. Through rigorous evaluation, the optimal ARIMA-GARCH model is selected based on criteria such as the AIC and the BIC, as outlined in Table 6. Subsequently, the estimated parameters of the best-fitted model are detailed in Table 7.

Table 5: BDS test results of agricultural commodity price data of different markets

Commodity	Markets	Dimension (m)			
		0.5σ	1.0σ	1.5σ	2.0σ
Safflower	Latur	10.73	8.40	7.50	5.48
	Kalaburagi	7.21	7.05	4.14	1.11
Safflower	Latur	16.45	10.01	8.34	6.58
	Kalaburagi	13.13	11.22	7.34	2.47
Mustard	Sri Ganganagar	5.18	7.14	7.46	5.18
	Satna	8.16	9.42	9.32	6.95
Mustard	Sri Ganganagar	9.25	10.86	10.57	7.89
	Satna	9.03	11.31	11.77	9.31
Groundnut	Gondal	10.56	5.81	3.95	2.54
	Bikaner	10.96	9.81	5.85	4.69
Groundnut	Gondal	21.56	11.31	8.45	6.83
	Bikaner	17.26	13.04	7.95	7.24
Lentil	Banda	9.94	7.43	5.70	3.59
	Narsinghpur	4.62	4.09	3.21	2.09
Lentil	Banda	13.45	9.84	7.60	5.14
	Narsinghpur	11.01	8.82	7.02	4.27
Chickpea	Hinganghat	20.41	10.74	7.21	6.58
	Dewas	14.29	9.36	7.43	6.70
Chickpea	Hinganghat	25.08	12.06	8.28	7.94
	Dewas	23.37	13.91	10.15	8.94
Green gram	Bhagat Ki Kothi	21.21	14.81	10.87	7.55
	Kalaburagi	5.71	5.38	5.21	5.64
Green gram	Bhagat Ki Kothi	28.27	18.66	12.74	8.44
	Kalaburagi	11.21	9.69	7.81	6.84

Note: The test statistics provided in the table lead to p-values less than 0.01 for all cases.

The correlation between the residuals of ARIMA-GARCH type models for two markets of the same agricultural commodity is examined through Spearman's rank correlation, and the results are shown in Table 8, indicating a significant correlation between the residuals of ARIMA-GARCH type models for two markets of all agricultural commodities at the one percent level. Subsequently, the residuals of the two markets were transformed via the PIT. The transformed values are then utilized to fit both Gaussian Copula and Student-t Copula models, and their AIC and BIC values are presented in Table 9. The results indicate that in all cases, the Student-t Copula model is the optimal Copula, with the lowest AIC and BIC values. This suggests that the Student-t Copula model provides a better goodness-of-fit compared to the Gaussian Copula.

After fitting the Student-t Copula model to the residuals of the optimal ARIMA-GARCH models for the two considered markets, proceed to simulate $n = 1000$ pairs of random samples from the estimated Student-t Copula function. Next, obtain one-day-ahead forecasts from the Bivariate ARIMA-GARCH-Copula model using algorithm 2.6.3. Repeat this one-day-ahead forecast procedure for each day in the test dataset, employing algorithm 2.6.3.

Table 6: Optimal ARIMA-GARCH type model for different commodity markets

Commodity	Markets	Optimal ARIMA-GARCH type model	AIC	BIC
Safflower	Latur	ARIMA (2,1,1) - GARCH (1,1)	12.107	12.211
	Kalaburagi	ARIMA (2,1,0) - APARCH (1,0)	13.121	13.268
Mustard	Sri Ganganagar	ARIMA (1,1,0) - GARCH (1,1)	13.797	13.902
	Satna	ARIMA (2,1,1) - APARCH (1,0)	13.103	13.270
Groundnut	Gondal	ARIMA (2,1,0) - APARCH (1,0)	13.231	13.398
	Bikaner	ARIMA (1,1,0) - APARCH (1,0)	14.461	14.628
Lentil	Banda	ARIMA (2,1,1) - APARCH (1,1)	14.165	14.290
	Narsinghpur	ARIMA (2,1,0) - APARCH (1,1)	14.139	14.293
Chickpea	Hinganghat	ARIMA (2,1,0) - GARCH (1,1)	14.539	14.664
	Dewas	ARIMA (2,1,0) - GARCH (1,1)	14.838	14.922
Green gram	Bhagat Ki Kothi	ARIMA (2,1,1) - APARCH (1,1)	15.088	15.213
	Kalaburagi	ARIMA (2,1,1) - APARCH (1,1)	14.779	14.904

Table 7: Parameter estimates of ARIMA-GARCH type models

Commodity	Markets	ϕ_1	ϕ_2	θ_1	α_1	β_1	γ	δ
Safflower	Latur	0.364	0.638	0.747	0.686	0.206	-	-
		(<0.001)	(<0.001)	(<0.001)	(<0.001)	(0.009)	-	-
	Kalaburagi	1.207	-0.239	-	0.418	-	0.178	3.218
		(<0.001)	(0.003)	-	(<0.001)	-	(<0.001)	(<0.001)
Mustard	Sri Ganganagar	0.913	-	-	0.362	0.589	-	-
		(<0.001)	-	-	(<0.001)	(<0.001)	-	-
	Satna	1.327	-0.438	0.104	0.633	-	0.057	3.499
		(<0.001)	(<0.001)	(0.031)	(<0.001)	-	(<0.001)	(<0.001)
Groundnut	Gondal	1.284	-0.406	-	0.503	-	0.242	3.072
		(<0.001)	(<0.001)	-	(<0.001)	-	(<0.001)	(<0.001)
	Bikaner	0.887	-	-	0.087	-	0.952	3.500
		(<0.001)	-	-	(<0.001)	-	(<0.001)	(<0.001)
Lentil	Banda	0.885	-0.148	0.510	0.099	0.567	0.480	3.127
		(<0.001)	(0.018)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)
	Narsinghpur	0.978	-0.189	-	0.049	0.493	0.898	3.358
		(<0.001)	(0.014)	-	(<0.001)	(<0.001)	(<0.001)	(<0.001)
Chickpea	Hinganghat	1.198	-0.294	-	0.384	0.581	-	-
		(<0.001)	(0.003)	-	(<0.001)	(<0.001)	-	-
	Dewas	0.758	0.249	0.911	0.272	0.702	-	-
		(<0.001)	(0.002)	(<0.001)	(<0.001)	(<0.001)	-	-
Green gram	Bhagat Ki Kothi	0.433	0.570	0.674	0.340	0.058	0.254	3.445
		(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)	(<0.001)
	Kalaburagi	1.054	-0.548	-0.523	0.395	0.108	0.104	3.268
		(<0.001)	(<0.001)	(0.046)	(<0.001)	(<0.001)	(<0.001)	(<0.001)

Note: The values in the parenthesis indicates the p-value

In evaluating the forecasting performance of the ARIMA-GARCH type-Copula, a comparative analysis is conducted against established traditional methodologies. Specifically, the efficacy of the ARIMA-GARCH-Copula model is compared with that of the Univariate ARIMA-GARCH type and MGARCH-DCC models. The assessment of accuracy utilizes key evaluation metrics, namely Root Mean Square Error (RMSE) (Eq.11), Mean Absolute Error (MAE) (Eq.12), and Mean Absolute Percentage Error (MAPE) (Eq.13), applied to the test dataset.

Table 8: Correlation analysis of ARIMA-GARCH type models

Commodity	Markets	Spearman's Rank Correlation Coefficient
Safflower	Latur and Kalaburagi	0.644
Mustard	Sri Ganganagar and Satna	0.554
Groundnut	Gondal and Bikaner	0.469
Lentil	Banda and Narsinghpur	0.674
Chickpea	Hinganghat and Dewas	0.577
Green gram	Bhagat Ki Kothi and Kalaburagi	0.527

Note: p-values of correlation coefficient are less than 0.01 for all cases.

Table 9: Comparison of Copula models

Commodity	Markets	Gaussian Copula		Student-t Copula	
		AIC	BIC	AIC	BIC
Safflower	Latur and Kalaburagi	-352.43	-346.33	**-354.36**	**-351.31**
Mustard	Sri Ganganagar and Satna	-476.47	-473.42	**-481.54**	**-475.44**
Groundnut	Gondal and Bikaner	-272.15	-269.10	**-278.62**	**-272.52**
Lentil	Banda and Narsinghpur	-444.05	-441.00	**-495.60**	**-489.50**
Chickpea	Hinganghat and Dewas	-351.78	-348.73	**-393.26**	**-387.16**
Green gram	Bhagat Ki Kothi and Kalaburagi	-224.19	-221.14	**-250.90**	**-244.80**

$$\text{RMSE} = \sqrt{\frac{1}{m}\sum_{t=1}^{m}(y_t - \hat{y}_t)^2} \tag{11}$$

$$\text{MAE} = \frac{1}{m}\sum_{t=1}^{m}|y_t - \hat{y}_t| \tag{12}$$

$$\text{MAPE} = \frac{1}{m}\sum_{t=1}^{m}\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100 \tag{13}$$

where y_i and $\hat{y}_i$ represent the actual and predicted values, respectively, and m is the number of observations in test dataset.

Table 10 provides a comparison of model forecasting performance considering RMSE, MAE and MAPE. The findings consistently reveal that the Bivariate ARIMA-GARCH type-Copula model outperforms both the MGARCH-DCC model and the Univariate ARIMA-GARCH type model across all agricultural commodity market price series. This superiority is underscored by the model's ability to achieve the lowest RMSE, MAE and MAPE values.

In addition to traditional accuracy metrics, the Diebold-Mariano (DM) test proposed by Diebold and Mariano (2002) is used to compare the forecasting performance of two competing models. The fundamental premise of the DM test lies in its null hypothesis, which posits that both forecasting models exhibit the same level of accuracy. By comparing the forecast errors of the Bivariate ARIMA-GARCH type–Copula model and benchmark models

Table 10: Comparison of forecasting performance of different models

Commodity	Markets	BAGC model			MGARCH-DCC model			UAGC model		
		RMSE	MAE	MAPE (%)	RMSE	MAE	MAPE (%)	RMSE	MAE	MAPE (%)
Safflower	Latur	147.43	102.34	1.96	208.72	159.94	3.07	415.67	354.97	6.78
	Kalaburagi	367.02	293.90	5.49	422.53	343.90	6.54	453.45	392.31	7.58
Mustard	Sri Ganganagar	279.00	212.94	3.36	474.58	418.43	7.43	1065.31	978.72	15.55
	Satna	257.06	195.56	3.19	355.44	310.18	4.98	447.56	358.06	6.01
Groundnut	Gondal	329.16	224.28	3.58	522.77	510.90	7.94	933.32	766.58	11.82
	Bikaner	462.97	368.28	6.72	642.34	479.93	8.93	1078.72	982.64	16.43
Lentil	Banda	179.07	147.05	2.29	331.77	281.12	4.46	538.12	431.15	6.53
	Narsinghpur	260.87	227.61	3.68	527.40	490.60	8.01	751.30	655.51	10.59
Chickpea	Hinganghat	221.58	167.26	3.95	300.89	263.71	6.29	608.56	463.61	10.94
	Dewas	206.01	167.42	3.76	422.29	345.91	8.01	589.72	475.63	10.26
Green gram	Bhagat Ki Kothi	283.54	239.46	3.86	388.31	352.34	5.51	761.36	632.49	9.89
	Kalaburagi	234.50	202.74	3.27	502.84	378.53	5.87	848.16	704.32	11.35

Note: BAGCM:Bivariate ARIMA-GARCH type -Copula model; MGARCH-DCC: Multivariate GARCH DCC model; UAGCM: Univariate ARIMA-GARCH type-Copula

Table 11: Diebold-Mariano test for pairwise comparison of Copula based model with benchmark models

Commodity	Markets	Benchmark Models	
		MGARCH-DCC model	Univariate ARIMA-GARCH type model
Safflower	Latur	-3.3075 (0.0052)	-8.5562 (<0.0001)
	Kalaburagi	-2.2874 (0.0385)	-7.2749 (<0.0001)
Mustard	Sri Ganganagar	-4.7645 (0.0003)	-6.5871 (<0.0001)
	Satna	-4.7203 (0.0004)	-8.467 (<0.0001)
Groundnut	Gondal	-4.862 (0.0003)	-10.504 (<0.0001)
	Bikaner	-5.5891 (<0.0001)	-11.713 (<0.0001)
Lentil	Banda	-3.1998 (0.0064)	-4.9542 (0.0002)
	Narsinghpur	-3.8288 (0.0018)	-6.0619 (<0.0001)
Chickpea	Hinganghat	-3.6141 (0.0028)	-8.5482 (<0.0001)
	Dewas	-2.1141 (0.0428)	-12.388 (<0.0001)
Green gram	Bhagat Ki Kothi	-3.6864 (0.0025)	-5.8317 (<0.0001)
	Kalaburagi	-2.1669 (0.0479)	-4.7079 (0.0004)

(MGARCH-DCC model and Univariate ARIMA-GARCH type model), the DM test evaluates whether there exists a statistically significant difference in their predictive capabilities. Table 11 presents the statistic values and their corresponding p-values (in parentheses) of the DM test, comparing the predictive accuracy of the Bivariate ARIMA-GARCH type–Copula model with benchmark models on the test datasets. The results suggest that the forecasting performance of the Bivariate ARIMA-GARCH type–Copula model significantly outperforms both the MGARCH-DCC model and the Univariate ARIMA-GARCH type model.

4. Conclusions

This study focused on analyzing the price volatility of oilseed crops *viz.*, safflower, mustard, and groundnut, as well as pulses *viz.*, lentil, chickpea, and green gram across two markets for each commodity. By employing the Bivariate ARIMA-GARCH type-Copula model, the accuracy of price forecasting in the agricultural sector was studied. This study highlights the importance of incorporating Copulas into advanced modeling techniques to capture the complex interdependencies and joint distributions of agricultural commodity

prices. The research findings demonstrate that the Bivariate ARIMA-GARCH type-Copula model surpassed both the MGARCH-DCC model and the Univariate ARIMA-GARCH type model in terms of forecasting performance. The evaluation metrics *viz.*, RMSE, MAPE, and MAE, consistently indicated the superior predictive ability of the Bivariate ARIMA-GARCH type-Copula model across all agricultural commodity market price series. Furthermore, the Diebold-Mariano (DM) test results provided additional validation of the Bivariate ARIMA-GARCH type-Copula model's outperformance compared to the alternative models. This signifies the robustness and reliability of Bivariate ARIMA-GARCH type-Copula model in capturing the joint distribution of commodity prices and improving forecasting accuracy. In the dynamic realm of agriculture, understanding price dynamics and volatility drivers is paramount. Combining Copulas with the ARIMA-GARCH model holds promise for better price predictions. By using advanced modeling, researchers and policymakers can improve forecasting accuracy. This study underscores the importance of continuous monitoring and analysis of agricultural commodity prices to mitigate risks and optimize market strategies in the ever-evolving landscape of the Indian economy.

Acknowledgements

The facilities provided by Indian Council of Agricultural Research - Indian Agricultural Statistics Research Institute (ICAR- IASRI), New Delhi and the funding granted to the first author by ICAR in the form of an ICAR-SRF fellowship are duly acknowledged for carrying out this study, which is a part of his doctoral research being pursued at ICAR-IASRI. In addition, thanks are due to the Graduate School, ICAR-Indian Agricultural Research Institute (ICAR-IARI), New Delhi for their support provided. The authors also thank the Chair Editor and reviewer for helpful comments which led to considerable improvement in the paper.

Conflict of interest

The authors do not have any financial or non-financial conflict of interest to declare for the research work included in this article.

References

Aziz, Z. and Iqbal, J. (2016). Comparing volatility forecasts of univariate and multivariate GARCH models: Evidence from the asian stock markets. *Journal of International and Global Economic Studies*, **9**, 67–78.

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.

Bollerslev, T. (1990). Modelling the coherence in short-run nominal exchange rates: a multivariate generalized ARCH model. *The Review of Economics and Statistics*, **72**, 498–505.

Box, G. E. P. and Jenkins, G. (1970). *Time Series Analysis: Forecasting and Control*. Holden-Day, San Francisco, CA, USA.

Broock, W. A., Scheinkman, J. A., Dechert, W. D., and LeBaron, B. (1996). A test for independence based on the correlation dimension. *Econometric Reviews*, **15**, 197–235.

Dickey, D. A. and Fuller, W. A. (1979). Distribution of the estimators for autoregressive time series with a unit root. *Journal of the American Statistical Association*, **74**, 427–431.

Diebold, F. X. and Mariano, R. S. (2002). Comparing predictive accuracy. *Journal of Business and Economic Statistics*, **20**, 134–144.

Ding, Z., Granger, C. W. J., and Engle, R. F. (1993). A long memory property of stock market returns and a new model. *Journal of Empirical Finance*, **1**, 83–106.

Dinku, T. (2021). Modeling price volatility for selected agricultural commodities in ethiopia: The application of garch models. *International Journal of Environmental Science*, **6**, 264–277.

Engle, R. F. (1982). Autoregressive conditional heteroscedasticity with estimates of the variance of united kingdom inflation. *Econometrica: Journal of the Econometric Society*, **50**, 987–1007.

Engle, R. F. (2002). Dynamic conditional correlation: A simple class of multivariate generalized autoregressive conditional heteroskedasticity models. *Journal of Business and Economic Statistics*, **20**, 339–350.

Engle, R. F. and Kroner, K. F. (1995). Multivariate simultaneous generalized ARCH. *Econometric Theory*, **11**, 122–150.

Glosten, L. R., Jagannathan, R., and Runkle, D. E. (1993). On the relation between the expected value and the volatility of the nominal excess return on stocks. *The Journal of Finance*, **48**, 1779–1801.

Jarque, C. M. and Bera, A. K. (1987). A test for normality of observations and regression residuals. *International Statistical Review*, **55**, 163–172.

Karmakar, M. (2017). Dependence structure and portfolio risk in indian foreign exchange market: A GARCH-EVT-Copula approach. *The Quarterly Review of Economics and Finance*, **64**, 275–291.

Kwiatkowski, D., Phillips, P. C. B., Schmidt, P., and Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root? *Journal of Econometrics*, **54**, 159–178.

Li, T., Zhong, J., and Huang, Z. (2020). Potential dependence of financial cycles between emerging and developed countries: Based on ARIMA-GARCH Copula model. *Emerging Markets Finance and Trade*, **56**, 1237–1250.

Ljung, G. M. and Box, G. E. P. (1978). On a measure of lack of fit in time series models. *Biometrika*, **65**, 297–303.

Lu, X. F., Lai, K. K., and Liang, L. (2014). Portfolio value-at-risk estimation in energy futures markets with time-varying copula-GARCH model. *Annals of Operations Research*, **219**, 333–357.

Merabet, F., Zeghdoudi, H., Yahia, R. H., and Saba, I. (2021). Modelling of oil price volatility using ARIMA-GARCH models. *Advances in Mathematics*, **10**, 2361–2380.

Mitra, D. and Paul, R. K. (2017). Hybrid time-series models for forecasting agricultural commodity prices. *Model Assisted Statistics and Applications*, **12**, 255–264.

Musunuru, N. (2014). Modeling price volatility linkages between corn and wheat: a multivariate GARCH estimation. *International Advances in Economic Research*, **20**, 269–280.

Nelson, D. B. (1991). Conditional heteroscedasticity in asset returns: A new approach. *Econometrica*, **59**, 347–370.

Paul, R., Prajneshu, and Ghosh, H. (2009). Garch nonlinear time series analysis for modelling and forecasting of india's volatile spices export data. *Journal of the Indian Society of Agricultural Statistics*, **63**, 123–131.

Paul, R. K. (2015). ARIMAX-GARCH-WAVELET model for forecasting volatile data. *Model Assisted Statistics and Applications*, **10**, 243–252.

Paul, R. K., Ghosh, H., and Prajneshu (2014). Development of out-of-sample forecast formulae for ARIMAX-GARCH model and their application. *Journal of the Indian Society of Agricultural Statistics*, **68**, 85–92.

Phillips, P. C. B. and Perron, P. (1988). Testing for a unit root in time series regression. *Biometrika*, **75**, 335–346.

Rakshit, D., Paul, R. K., and Panwar, S. (2021). Asymmetric price volatility of onion in india. *Indian Journal of Agricultural Economics*, **76**, 245–260.

Sanjuán-López, A. I. and Dawson, P. J. (2017). Volatility effects of index trading and spillovers on us agricultural futures markets: A multivariate GARCH approach. *Journal of Agricultural Economics*, **68**, 822–838.

Shahriari, S., Sisson, S. A., and Rashidi, T. (2023). Copula ARMA-GARCH modelling of spatially and temporally correlated time series data for transportation planning use. *Transportation Research Part C: Emerging Technologies*, **146**, 103969.

Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, **52**, 591–611.

Sklar, A. M. (1959). Fonctions de repartition an dimensions et leurs marges. *Publications of the Institute of Statistics of the University of Paris*, **8**, 229–231.

Wang, Y. and Wu, C. (2012). Forecasting energy market volatility using GARCH models: Can multivariate models beat univariate models? *Energy Economics*, **34**, 2167–2181.

Weiß, G. N. (2013). Copula-GARCH versus dynamic conditional correlation: an empirical study on VaR and ES forecasting accuracy. *Review of Quantitative Finance and Accounting*, **41**, 179–202.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 23, No. 1, 2025 (New Series), pp 217–223
https://www.ssca.org.in/journal

Environmentally Responsible Index Tracking: Maintaining Performance while Reducing Carbon Footprint of the Portfolio

Lakshmi M. V.[1], **Soudeep Deb**[2] **and Rituparna Sen**[3]
[1]*Indian Institute of Science Education and Research, Tirupati AP 517507*
[2]*Indian Institute of Management Bangalore KA 560076*
[3]*Indian Statistical Institute Bangalore KA 560059*

Received: 07 January 2024; Revised: 06 June 2024; Accepted: 09 June 2024

Abstract

Amid the global crisis of climate change, urgent action is imperative. In this study, we develop two types of decarbonized indices, which render a dynamic hedging approach for passive investors. Focusing on long-term returns with minimal active trading and risk exposure, we create the decarbonized indices for NIFTY-50, a benchmark index for the Indian market. Proposed methodology relies on suitable optimization techniques to choose the portfolio weights that minimize the tracking error while significantly reducing carbon footprints. These indices are shown to perform better than existing benchmarks, especially during major climate events. They are likely to offer investors a buffer to adapt to climate policies and carbon pricing. Since these indices align with the net-zero objective and foster climate-resilient advancements, they also offer actionable pathways to address climate challenges while maintaining financial objectives.

Key words: Climate change; Decarbonized index; Market index; NIFTY-50; Tracking error.

1. Introduction

Climate change, a significant challenge in recent times, not only impacts health, environment and the ecosystem, but also poses a large aggregate risk to the financial systems. This necessitates the development of analytical tools that can offer enhanced indexation of financial markets by considering Environmental, Social and Governance (ESG) factors. Such techniques are critical to solve the inefficiency of fundamental financial markets, especially in developing countries like India.

In this paper, we present methods to create two decarbonized indices from established benchmarks, and demonstrate their efficacy for the Indian economy. Specifically, we show that the resulting index significantly lowers total carbon impact, acting as a hedge against climate risks. Our method relies on tracking error (TE), a metric representing the variation

Corresponding Author: Rituparna Sen
Email: rsen@isibang.ac.in

of the difference in composition between a portfolio and its benchmark index. The relationship of TE to ESG has remained largely unexplored. The mimicking portfolio approach of Lamont (2001) is theoretically appealing but challenging to implement. In a more relevant study, Andersson *et al.* (2016) introduced decarbornized indices from the benchmark by minimizing TE subject to suitable constraints based on carbon footprints of the constituent companies. Mezali and Beasley (2013) earlier used quantile regression with a mixed-integer linear programming formulation. Li *et al.* (2022) constructed a robust model that maximizes ESG score, while minimizing the risk and maximizing the return simultaneously.

It is further important to note that the existing sustainability-themed indices in the Indian stock market, namely S&P BSE GREENEX, BSE Carbonex, and NIFTY100 Enhanced ESG Index (Patel and Kumari, 2020; C and Nishad, 2021) prioritize tracking the performance of companies based on their carbon emissions, ESG score and efforts to mitigate climate risk, without focusing on the parent index's performance. They use market capitalization for weighting, without any effort to replicate the performance of the dropped stocks. To circumvent this, we develop an optimized index by minimizing tracking error. It is more effective in capturing lost contributions from dropped stocks by compensating from other highly correlated stocks that remain in the portfolio.

We describe this methodology in Section 2. The application of the methods on Indian market is illustrated in Section 3. By utilizing real-time data for in-sample and out-of-sample calculations, we show how the index attempts to bridge the divide between theory and practice. The paper ends with a succinct summary and scopes of future work in Section 4.

2. Methodology

Throughout this article, we work with the Indian stock market index NIFTY-50, that tracks 50 largest Indian companies listed in the National Stock Exchange. To explain the method, let these $N = 50$ stocks be sorted by their carbon footprints in decreasing order. For the i^{th} stock, r_i, m_i, q_i denote the return, market capitalization and carbon footprint, respectively. Bold-faced letters $\boldsymbol{r}, \boldsymbol{m}, \boldsymbol{q}$ denote the corresponding vectors for all stocks. Following extant literature, the portfolio return of the benchmark is indicated by $R^b = (\boldsymbol{w}^b)^T \boldsymbol{r}$, where $\boldsymbol{w}^b = (w_i^b)_{1 \le i \le N}$ is the vector of portfolio weights taken to be proportional to the market capitalization,

$$w_i^b = \frac{m_i}{\sum_{i=1}^N m_i}. \tag{1}$$

Let $\boldsymbol{w^d}$ be the vector of weights for the proposed decarbonized index, R^d being the corresponding return. Our objective is to minimize the tracking error and find (sd indicates standard deviation)

$$\boldsymbol{w^d} = \underset{\boldsymbol{w}=(w_i)_{1 \le i \le N}}{\arg\min} \ (\mathrm{TE}) = \underset{\boldsymbol{w}=(w_i)_{1 \le i \le N}}{\arg\min} \left[\mathrm{sd} \left(\sum_{i=1}^N (w_i - w_i^b) r_i \right) \right]. \tag{2}$$

To avoid computing the large dispersion matrix of returns in (2), we use the Fama and French (2012) factor model. It allows us to decompose the return into weighted sum of common factor returns and specific returns. If r_{it} and r_{ft} denote the return of the i^{th} stock

and the risk-free rate at time t, then the model is

$$r_{it} - r_{ft} = \beta_{i0} + \beta_{i1}\mathrm{SMB}_t + \beta_{i1}\mathrm{HML}_t + \beta_{i3}\mathrm{WML}_t + \beta_{i4}\mathrm{MF}_t + e_{it}, \tag{3}$$

where e_{it} is the error, β_{ij} denotes the factor loading; SMB, HML, WML and MF indicate the size effect (small-minus-big), value effect (high-minus-low), momentum factor (winners-minus-losers), and market factor. Let F_j denote these factors, with dispersion matrix Ω. Also, let $\boldsymbol{\beta}$ be the matrix of loadings and Δ be the diagonal matrix of specific risk variances. Then, the dispersion of the excess returns is $\boldsymbol{\beta}\Omega\boldsymbol{\beta}^T + \Delta$. Consequently, the volatility of any portfolio with returns $\boldsymbol{r}$ and weights $\boldsymbol{w}$ is $\sqrt{\boldsymbol{w}^T(\boldsymbol{\beta}\Omega\boldsymbol{\beta}^T + \Delta)\boldsymbol{w}}$. This, in (2), implies

$$\boldsymbol{w^d} = \underset{\boldsymbol{w}=(w_i)_{1\le i\le N}}{\arg\min} \sqrt{(\boldsymbol{w} - \boldsymbol{w^b})^T (\boldsymbol{\beta}\Omega\boldsymbol{\beta}^T + \Delta)(\boldsymbol{w} - \boldsymbol{w^b})}. \tag{4}$$

To strike a balance between reducing carbon footprints and preserving diversity in the composition, we employ two distinct methodologies to construct decarbonized indices (DCI). Each methodology has its own advantages and disadvantages, as we explicate below.

In the first approach, we exclude k worst performers in carbon intensity, and the remaining stocks are re-weighted to minimize TE. Here, the DCI is constructed using weights w_i^d, obtained by solving (4) subject to the constraints

$$\begin{aligned} &\sum_{i=1}^{N} w_i^d = 1, \text{ with} \\ &w_i^d = 0, \text{ for } i = 1, 2, \ldots, k, \text{ and } 0 \le w_i^d \le 1, \text{ for } i = k+1, \ldots, N. \end{aligned} \tag{5}$$

We solve this minimization problem using the Trust-Region Constrained Algorithm (TRCA), which is useful to deal with the following problem:

$$\text{minimize } f(x), \text{ subject to } c^{lb} \le c(x) \le c^{ub}, \; x^{lb} \le x \le x^{ub}. \tag{6}$$

It can take multiple linear and non-linear constraints as inputs (Conn *et al.*, 2000). The objective function is approximated by a quadratic model restricted to the trust-region centered at the initial guess or the current point. The algorithm works by iteratively improving the initial guess (Kimiaei, 2022). We omit technicalities of the algorithm, and refer to Byrd *et al.* (1987) for further details.

Our second methodology includes all stocks without specifically targeting those with high carbon footprints. In this case, the minimization problem (4) is solved by setting a threshold C for the total footprint of the index. This approach ensures a largely unchanged composition, maintaining its diversity, yet reducing the footprint. Mathematically, we find the weights in (4) considering

$$\begin{aligned} &\sum_{i=1}^{N} w_i^d = 1, \text{ with} \\ &\sum_{i=1}^{N} q_i w_i^d \le C \text{ and } 0 \le w_i^d \le 1, \text{ for } i = 1, \ldots, N. \end{aligned} \tag{7}$$

For implementation, we again rely on TCRA discussed before. Hereafter, this index is denoted as DCI_2 while the first index is coded as DCI_1.

A brief comparison of the ideology behind the construction of the indices is critical here. A potential drawback of the first approach is that it may lead to a less diverse index composition. Lower diversity leads to higher volatility and risk. On the positive side, possibility of inclusion in the index can serve as an incentive for the high-emission companies to proactively reduce their emissions. Contrastingly, the overall carbon footprint reduction with the second approach is significant but limited when compared to the first approach.

3. Application

We consider NIFTY-50 data for 5 years, 2017-18 until 2022-23. To quantify the carbon footprint of the stocks, we consider greenhouse-gas intensity per sale and total carbon-dioxide emissions (abbreviated as GHG and CO2 hereafter) as proxies. Then, four decarbonized indices are created from each benchmark, using the two methods and the two proxies. We rely on Bloomberg and Yahoo!Finance for obtaining these data. The factors data for (3) are obtained from IIM-A Data Library (Agarwalla *et al.*, 2013). Comprehensive information about stocks used for our calculations are detailed in Table 1.

Table 1: Number of stocks included (St.Incl), omitted (St.Omit) and corresponding omission percentage of market capitalization (MktCap.Omit) in the construction of DCI.

	GHG			CO2		
Period	St.Incl	St.Omit	MktCap.Omit	St.Incl	St.Omit	MktCap.Omit
2017-18	30	20	35.3%	32	18	33.7%
2018-19	33	17	28.5%	35	15	26.9%
2019-20	34	16	25.3%	36	14	23.7%
2020-21	35	15	23.6%	38	12	21.1%
2021-22	35	15	23.6%	38	12	21.0%

Our analysis broadly consists of three parts – determining optimal values of k and C for calculating the two DCI, generating optimal portfolio weights using a window of one year for five years (in-sample calculations), calculating the monthly performance of DCI and comparing their performances with the benchmark (out-of-sample calculations). It is useful to present a brief summary of our main findings first. As expected, DCI_2 maintains the composition yet provides a lower carbon footprint than benchmark index, whereas DCI_1 renders an even lower carbon footprint because of excluding several stocks. In-sample calculations illustrate that the second index offers a very low active risk as compared to the benchmark. On the other hand, out-of-sample results demonstrate that both indices outperform the benchmark during major climate events throughout the five years.

Delving deeper into our analysis, recall that the optimal values of k and C in our methods (refer to (5) and (7)) are determined through an assessment of TE using 5 years of data. Here, a series of optimizations are executed for a range of k (5%-50% of N) and C (50%-95%). In GHG, optimum k is 6 and C is 80%, whereas the numbers are 5 and 70% for CO2. These values are employed in the subsequent steps.

Next, in Figure 1, we compare the carbon footprints of the decarbonized indices with the considered benchmark in each case. A substantial reduction in the carbon footprint of the index is evident, achieving more than 50% reduction in method-1. This methodology can be expanded to consider sector compositions and optimize while maintaining fixed sector representations. With method-2, reductions of around 20-30% were achieved in different cases, which should be perceived as a significant accomplishment without alterations to sector representations.

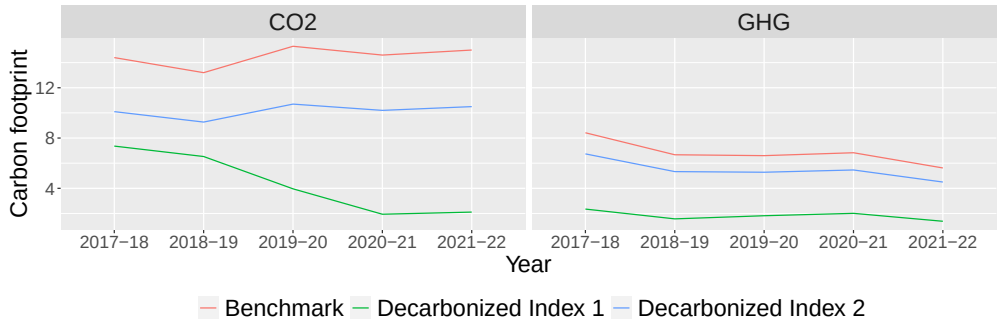


Figure 1: Comparison of carbon footprints of the considered benchmark index and the decarbonized indices

Turn attention to in-sample estimation of TE for the four DCIs constructed using a moving window of one-year and optimal values of k and C. We provide a summary in Table 2. Please refer to the supplement for additional figures and discussions on this. The risk on the benchmark portfolio is measured by $\text{sd}(R^b)$, whereas the TE of DCI relative to the benchmark can be calculated by $\text{sd}(R^d - R^b)/\text{sd}(R^b)$. These in-sample estimations reveal significant carbon footprint reductions in both methods, with low TE in most cases. Interestingly, DCI_1 for CO2 exhibits high TE due to the exclusion of valuable stocks. DCI_2, meanwhile, demonstrate low TE everywhere because it avoids dropping valuable stocks.

Table 2: Risk on the Benchmark portfolio (BM) and tracking error of the decarbonized indices relative to the benchmark index in each Method.

Period	GHG			CO2		
	BM	TE(DCI_1)	TE(DCI_2) (in 10^{-3})	BM	TE(DCI_1)	TE(DCI_2) (in 10^{-3})
2017-18	27.25	2.72	1.13	26.14	2.54	3.07
2018-19	53.62	1.40	6.31	51.55	1.28	1.37
2019-20	159.7	0.52	1.49	153.8	4.09	0.74
2020-21	159.8	0.79	2.96	150.4	9.0	0.45
2021-22	176.9	1.07	9.22	166.3	6.36	0.52

Our last point of discussion is the out-of-sample performance, where monthly returns are computed for 2018-19 to 2022-23 using weights generated from in-sample calculations conducted in the previous year. A comparison is made between the monthly performance of the decarbonized indices, the actual benchmark, and the considered benchmark. We observe that the constructed indices track the considered benchmark very closely and on the average outperform the benchmark index. We then explore whether during climate events,

the decarbonized indices exhibit superior performance compared to their parent benchmark indices. To investigate this effect, we identify and highlight significant climate events from the past few years in the out-of-sample results of our indices. Figure 2 illustrate these findings. We observe that both indices outperform the benchmark in terms of out-of-sample returns in at least seven of the twelve such events. Particularly, DCI_1 of GHG outperforms the benchmark in 75% of the events.

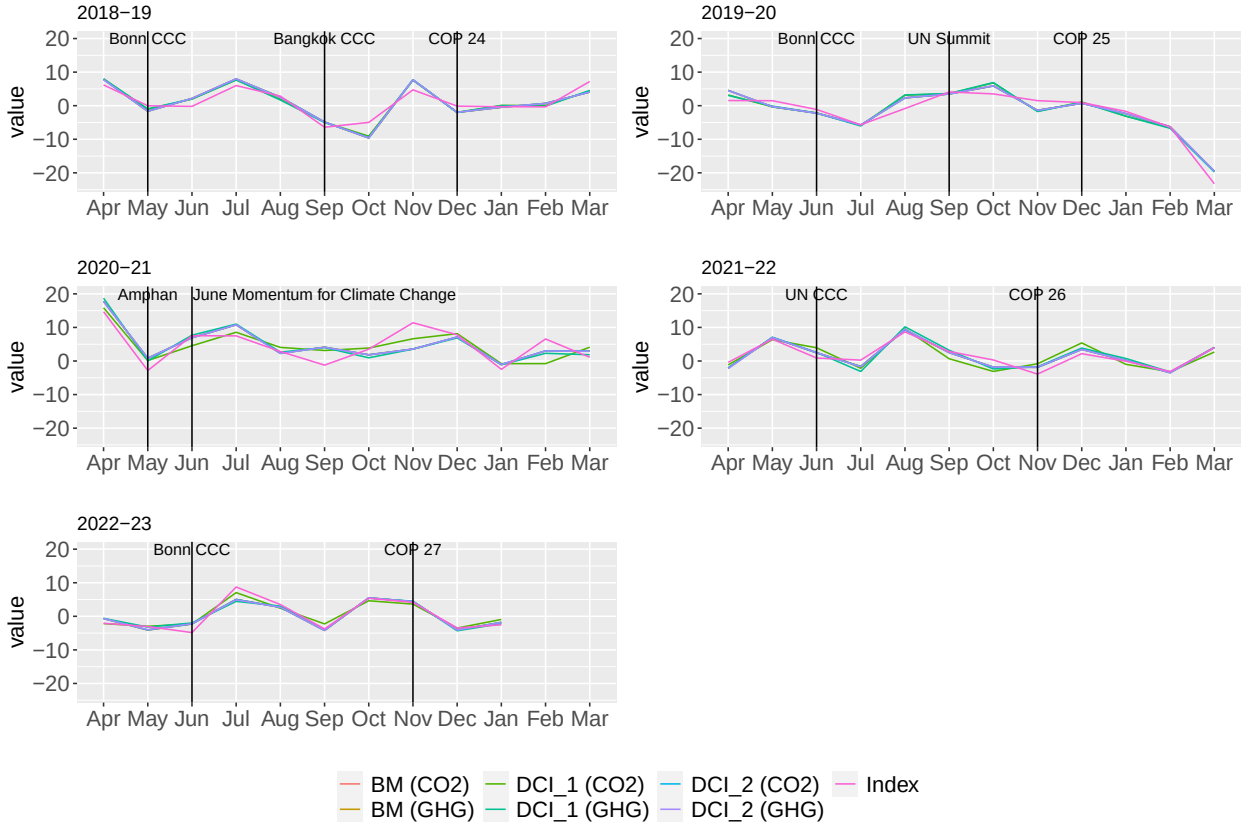


Figure 2: Out-of-sample performance of difference indices during important climate events across five years. BM stands for considered benchmark, DCI is proposed decarbonized index following the two methods. CCC stands for climate change conference.

4. Conclusion

With the World Resources Institute (Friedrich *et al.*, 2020) identifying China, USA and India as top GHG emitters, there arises a compelling need for decarbonized indices in India. We devised two novel optimization methods for creating practical decarbonized indices which complement existing green indices and foster investment awareness. These indices offer real-world utility, granting investors time to acclimate to economic shifts and financial uncertainties. Leveraging real-time data, they mitigate risks tied to climate policy execution. For long-term passive investors, these indices hold promise over clean energy options. They exhibit comparable returns to benchmark indices, gaining an edge once carbon pricing and stringent emissions policies take effect, potentially outperforming benchmarks.

Measurement of company-wise GHG emissions is crucial for constructing the indices.

We faced the challenge of missing data due to poor reporting. This in turn impacted benchmark composition, potentially excluding stocks sensitive to climate change and policies. Regression results showed limited explanatory power of common factors for stock returns. In future, we plan to extend the current method to deal with missingness. We also believe that consideration of sector compositions with suitable data can enhance future results. Moreover, we have laid the theoretical framework for integrating ideas like Value-at-Risk in optimization. These quantities might capture the extreme movements in prices during climate events in a better fashion.

Conflict of interest

The authors do not have any financial or non-financial conflict of interest to declare for the research work included in this article.

References

Agarwalla, S. K., Jacob, J., and Varma, J. R. (2013). Four factor model in Indian equities market. Working paper, Indian Institute of Management, Ahmedabad.

Andersson, M., Bolton, P., and Samama, F. (2016). Hedging climate risk. *Financial Analysts Journal*, **72**, 13–32.

Byrd, R. H., Schnabel, R. B., and Shultz, G. A. (1987). A trust region algorithm for nonlinearly constrained optimization. *SIAM Journal on Numerical Analysis*, **24**, 1152–1170.

C, N. and Nishad, T. M. (2021). Carbon reduction and sustainable investment: a way to sustainable development. *Energy Economics Letters*, **8**, 134–144.

Conn, A. R., Gould, N. I., and Toint, P. L. (2000). *Trust Region Methods*. SIAM.

Fama, E. F. and French, K. R. (2012). Size, value, and momentum in international stock returns. *Journal of Financial Economics*, **105**, 457–472.

Friedrich, J., Ge, M., Pickens, A., and Vigna, L. (2020). This interactive chart shows changes in the world's top 10 emitters. *World Resources Institute*, **10**.

Kimiaei, M. (2022). An active set trust-region method for bound-constrained optimization. *Bulletin of the Iranian Mathematical Society*, **48**, 1–25.

Lamont, O. A. (2001). Economic tracking portfolios. *Journal of Econometrics*, **105**, 161–184.

Li, X., Xu, F., and Jing, K. (2022). Robust enhanced indexation with esg: An empirical study in the chinese stock market. *Economic Modelling*, **107**, 105711.

Mezali, H. and Beasley, J. E. (2013). Quantile regression for index tracking and enhanced indexation. *Journal of the Operational Research Society*, **64**, 1676–1692.

Patel, S. K. and Kumari, P. (2020). Indian stock market movements and responsiveness of sustainability indices: a risk adjusted analysis. *International Management Review*, **16**, 55–64.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 23, No. 1, 2025 (New Series), pp 225–247
https://www.ssca.org.in/journal

A Comprehensive Review of Data Science, Artificial Intelligence, and Big Data Analytics in Indian Official Statistics

Prasily P. and Manoharan M.
Department of Statistics, University of Calicut, Kerala, 673635

Received: 30 May 2023; Revised: 01 May 2024; Accepted: 09 June 2024

Abstract

This article examines how Data Science (DS), Artificial Intelligence (AI), and Big Data Analytics (BDA) are used in the Indian digital government to produce official statistics. Official statistics play a crucial role in shaping policy decisions, and scientific advances have made it possible to extract insights and patterns from vast amounts of data. The article examines the current state of official statistics in India and explores how the digital government is applying DS, AI, and BDA to upgrade statistical analysis. The article also discusses the challenges of executing these advanced technologies, including data quality and privacy concerns. Furthermore, the review highlights some recent developments and schemes for the benefit of DS, AI, and BDA in Indian official statistics, including the application of Machine Learning (ML) and predictive modeling. The article concludes with recommendations for future research and policy in this area, highlighting the need for equality between technological innovation and ethical considerations to ensure the precise and responsible use of official statistics in the Indian digital government.

Key words: Official statistics; Data Science; Artificial Intelligence; Big Data Analytics.

AMS Subject Classifications: 62A01, 97K80.

1. Introduction

Indian policy started to take shape from the days of colonial rule, with official statistics playing a key role. The British government started to collect and publish data on the Indian economy during the mid-nineteenth century. Data was collected on agriculture, demography, and trade from official records. The government continued to formulate and monitor economic and social policies based on official statistics after independence. The Central Statistical Office (CSO), established in 1951, became the nodal agency for collecting, compiling, and disseminating official statistics. (`https://unstats.un.org/unsd/wsd/docs/India_wsd_history.pdf`). Rao (2010) argues that India's rapid economic growth and ambitious development agenda make official statistics even more important.

Corresponding Author: Prasily P.
Email: prasilyprasi29@gmail.com

1.1. Background and significance of official statistics in India

India, which has a population of 1.42 billion, heavily depends on official statistics for policy-making decisions.(https://worldpopulationreview.com/countries, https://www.statista.com/statistics/263766/total-population-of-india/). Developing economic policies and programs in India is driven by official statistics. The Indian constitution mandates that the central government collect and publish official statistics associated with the country's economy, society, and population.

Official statistics are significant to Indian government officials and policymakers, and numerous steps have done to grow their availability and accuracy. The National Statistical System (NSS) gathers information on several socioeconomic factors via sample surveys. The NSS contains various organizations such as the CSO, the National Sample Survey Office (NSSO), and the Registrar General and Census Commissioner of India (RGCCI) (https://mospi.gov.in/142-present-indian-statistical-system-organisation).

The government uses official statistics for some functions, including planning, analyzing, and monitoring policies and programs. For instance, official statistics used to assess the progress of the Sustainable Development Goals (SDGs) and various flagship schemes of the government such as Sashakt Bharat - Sabal Bharat (Empowered and Resilient India), Swachh Bharat - Swasth Bharat (Clean and Healthy India), Samagra Bharat - Saksham Bharat (Inclusive and Entrepreneurial India), Satat Bharat – Sanatan Bharat (Sustainable India) and Sampanna Bharat- Samriddh Bharat (Prosperous and Vibrant India) (https://sustainabledevelopment.un.org/content/documents/26162Main_Messages_India.pdf). Non-governmental organizations, academics, and researchers are also using official statistics to study the economic facets.

The National Policy on Official Statistics (NPOS) (https://mospi.gov.in/sites/default/files/announcements/draft_policy_17may18.pdf) outlines the draft policy by the Government of India's Ministry of Statistics and Programme Implementation (MOSPI). It covers fundamental principles of official statistics, objectives, government policy initiatives, and mechanisms for regulating core statistics. This policy underlines the professional independence, confidentiality of data, and maintaining statistical standards to provide relevant and accurate empirical data to inform economic and social policies. It further addresses the decentralization of the statistical system in India and the involvement of various government bodies, including the MOSPI, Directorates of Economics & Statistics, and the National Statistical Commission. It also prioritizes ensuring quality, promoting data sharing, developing capacity, and cooperating internationally in official statistics. The revised draft (https://mospi.gov.in/sites/default/files/announcements/Draft_National_Policy_on_Official_Statistics.pdf) brings into perspective the transformative power of data and statistics in achieving sustainable development and inclusive growth. The initiatives embraced to reform and empower the institutional framework of the official statistical system in India are discussed, such as involving better coordination, international data standards adoption, SDG monitoring via the National Indicator Framework, and the Collection of Statistics Act amendment. It also indicates India's election to the United Nations Statistical Commission (UNSC) for 2024-2027 highlights its responsibility for maintaining global statistical efficiency and integrity.

Official data are significant, but gathering and analyzing them in India is fraught with

difficulties. The absence of timely and reliable data, especially in socio-economic indicators such as poverty and employment, represents a significant issue. The inconsistent definition and assessment of numerous indicators throughout the nation's states and regions is another problem. India's development planning and policymaking process both heavily rely on official statistics. The Indian government has taken measures to improve official statistics, but some issues still need to be resolved for reliable and timely data dissemination.

1.2. Overview of the role of digital government in official statistics

In recent years, the importance of using digital technologies and data analytics in official statistics has increased, and India is no exception. Digital government initiatives can enhance official statistics through better collection, analysis, and dissemination, leading to higher-quality, more timely, and comprehensive data.

The government of India has launched several initiatives to improve the availability and accessibility of official statistics using digital technologies. For example, the MOSPI launched the National Data Sharing and Accessibility Policy (NDSAP) (`https://dst.gov.in/national-data-sharing-and-accessibility-policy-0`) and the National Data and Analytics Platform (NDAP) (`https://ndap.niti.gov.in/`) to facilitate data sharing between different government departments and to improve the availability of official statistics.

NDSAPs main objective is to promote data sharing and reuse by defining standards for sharing and guidelines for data management. National governments have made their data available in public open formats for researchers, policymakers, and the general public.

NDAP is a web-based platform that opens up official statistics and datasets collected by different departments in the government. The platform aims to promote data-driven decision-making by making it easier for users to access and analyze official statistics. The most notable definiteness omitted from the platform is the appliance of data visualization and analysis.

Additionally, the Indian government has launched several other initiatives to improve the collection and analysis of official statistics through digital government. Mobile apps and cloud computing are examples of digital technologies used to improve data collection and analysis. The government is also exploring the application of data analytics and ML to automate data analysis and improve the accuracy and timeliness of official statistics.

Digital technology can transform the process of collecting, processing, and publishing official data, making it more accurate and accessible for policymakers. Utilizing digital technologies requires carefully evaluating data quality, privacy, and security concerns.

A particular focus will be placed on data science, AI, and BDA in India's official statistics as part of this review. The objectives of this review are to:

- Explore the historical and contemporary significance of official statistics in India.
- Identify the current challenges facing official statistics in India.
- Evaluate the potential of digital government in addressing these challenges.

- Analyze the implications of digital government for the reliability and quality of official statistics in India.
- Provide recommendations for future research and policy development in this area.

2. Overview of official statistics in India

Government agencies in India collect official statistics to inform policy and decision-making. The National Statistics Office (NSO) is responsible for providing official statistics covering population, economy, social welfare, natural resources, environment, and management. The data security, quality, privacy, accessibility, and coverage of data remain challenges despite progress. India's government is using digital technologies like NDAP and NDSAP to improve official statistics.

2.1. Brief history of official statistics in India

Official statistics in India have a long and intriguing history dating back to the colonial era. The CSO, founded in 1861 by the British colonial government, was the country of India's first official statistics office. The main objective of this agency was to provide statistical information for the British government's economic and administrative policies in India (Rao, 2010; Ghosh *et al.*, 1999)

After India gained independence in 1947, the Central Statistical Organization was renamed the Central Statistical Office and became part of the Ministry of Planning. The CSO's role has evolved to include gathering, compiling, and disseminating official statistics.

According to Sarma (1958), the government of India has taken pivotal measures to streamline the statistical system as part of its development of a system for official statistics. It supports economic aspects in planning implemented during the First Five Year Plan (1951-1956). In the consecutive five-year plan (1956-1961), National Sample Survey Offices (NSSOs) introduced the system to conduct surveys on different topics.

In 2005, the discussion of setting up the National Statistical Commission (NSC) was to provide guidance and direction to the statistical system in India (`https://mospi.gov.in/national-statistical-commission-0`). NSC was formed in 2006 to provide quality and integrity in statistics to the society. Rao (2013) discusses the NSC and its functions. It outlines the historical background of statistical data collection in the country and the role of the NSC in coordinating statistical activities. NSC is responsible for identifying core national statistics, formulating national policies related to the statistical system, and improving public confidence in official statistics. It has a pivotal role in shaping the official statistical system and meeting the statistical requirements of the nation.

The Indian government has prioritized the use of data analytics and digital technology in official statistics during the last few years. The government has started some programs, such as the NDSAP and NDAP, which were covered in the preceding section, to increase the availability and accessibility of official data. In general, the statistical system of India has gradually developed with an increasing focus on data for planning and policymaking. The government's focus on digital technology and data analytics expects change in future official statistics.

2.2. Current status and challenges of official statistics in India

The official statistics system in India has come a long way since its establishment. Despite efforts to improve the statistical system, several challenges remain.

The lack of readily available and high-quality data is one of the significant challenges facing official statistics in India. While India has made main progress in data collection and dissemination, there are still gaps in the data availability, particularly for specific social and economic indicators. It can also be problematic to determine the quality of the data collected; some data sources are not updated regularly or do not accurately capture the whole scope of the indicator being measured.

Another challenge is the better coordination and harmonization between different agencies involved in data collection and dissemination. Multiple agencies collect and publish data in India, and there is a lack of coordination between them, which can lead to inconsistencies and discrepancies in the data reported.

The dissemination of official statistics is also a challenge in India. Despite recent advances, official statistics are not always easily accessible to the general public, limiting their use and impact. Moreover, the statistical system must be more transparent to build trust in its accuracy and reliability.

The COVID-19 pandemic has highlighted some of these challenges, with the need for timely and accurate data becoming even more critical. The authors Hantrais *et al.* (2021) in the period of pandemic has highlighted the need for better investment in the statistical system to ensure its resilience and effectiveness in times of crisis.

To address these challenges, the Indian government has taken several steps to improve the statistical system, initiatives like the NDSAP and the NDAP aimed at increasing the availability and accessibility of official statistics. The government has also established a new NSC to provide guidance and direction to the statistical system and enhance the quality of official statistics.

Overall, while there have been significant improvements in the official statistics system in India, there is still a long way to go to ensure that official statistics are of high quality, accessible to all, and able to inform policy and decision-making effectively. For more details, refer to (`https://www.thehinducentre.com/publications/policy-watch/credible-data-for-public-good-constraints-challenges-and-the-way-ahead/article65971093.ece`).

2.3. Role of digital government in addressing these challenges

Official statistics in India face many challenges that the digital government can help address. The use of digital technologies and platforms can help improve data collection, quality, and dissemination, as well as increase transparency and accountability in the statistical system (Rana *et al.*, 2020; Chatterjee, 2020; Vijai, 2019).

One way in which digital government can improve official statistics is through the use of technology in data collection. Surveys conducted on mobile devices and online questionnaires can help increase the efficiency and accuracy of data collection, specifically in

hard-to-reach places. To reduce data entry errors and improve data collection efficiency, the Indian government has launched the NSSO mobile application. The use of digital technology can also help in real-time data collection, thereby ensuring the availability of timely data.

Another way in which digital government can improve official statistics is through the use of advanced analytics techniques such as BDA and ML. Using these techniques can uncover data insights missed by traditional statistics. India's NDAP integrates and analyzes administrative and survey data using BDA and AI.

Digital platforms, online portals, dashboards, and mobile applications can improve the accessibility and user-friendliness of official statistics. India's NDSAP promotes open data sharing and easy access to official statistics.

Furthermore, the digital government can play a role in increasing transparency and accountability in the statistical system. The use of digital platforms can help in the monitoring and reporting of statistical data and make it easier for stakeholders to identify any issues. Integrity and authenticity of official statistics can be ensured through blockchain, building trust.

To sum up, the digital government can improve official statistics in India. By utilizing digital technologies and platforms, the statistical system can become more efficient, accurate, and accessible, leading to better quality and impact of official statistics on policy and decision-making.

3. Digital government and official statistics in India

In India, the concept of digital government aims to use technology to enhance the effectiveness and efficiency of public services, including gathering and distributing official statistics. A digital government may transform official statistics collection, processing, and analysis with the help of big data and AI. Initiatives such as the NDAP and the NDSAP are examples of how the government of India is using digital technologies to improve the quality and accessibility of official statistics. Data security, privacy concerns, and the need to maintain data quality remain crucial factors for digital governance in India's official statistics (Alvarenga *et al.*, 2020; Tripathi and Dungarwal, 2020).

3.1. Role of data science in official statistics in India

Data science is playing an increasingly important role in official statistics in India. A large and complex dataset is analyzed using statistical and computational methods, including machine learning and predictive modeling. In India, the NSO is using data science to improve the accuracy and timeliness of official statistics. For more information, see Ashofteh and Bravo (2021).

NSO is developing a BDA platform to analyze various datasets, including agriculture (Sinha and Dhanalakshmi, 2022; Guntukula, 2020), health (Subrahmanya *et al.* (2022)), energy, and environment, etc. The platform aims to use advanced data science techniques to analyze large datasets to provide insights and inform policy decisions.

The use of data science in official statistics in India has the potential to enhance the

accuracy and efficiency of data collection and analysis. Data quality and privacy remain challenges, including the need for skilled professionals. The NSO has established guidelines and protocols for data privacy and confidentiality while also providing training programs to develop skills in data science.

3.2. Use of AI in official statistics in India

AI is also playing an increasingly important role in official statistics in India. AI involves the process and analysis of the data using computational models and algorithms, and it may automate several steps in data collecting and analysis (Chatterjee *et al.*, 2022; Vijai and Wisetsri, 2021).

In India, the NSO has been researching the use of AI in official statistics to upgrade the efficiency and accuracy of data collection and processing. For example, using ML algorithms to analyze satellite photos; for instance, may be used to estimate agricultural yields using AI. Ilyas *et al.* (2023).

The use of AI in official statistics in India has the potential to improve the accuracy and efficiency of data collection and analysis. The development and implementation of AI solutions still face many challenges, including data quality and privacy concerns Sharma *et al.* (2022). According to Ashofteh and Bravo (2021), the NSO has started training programs to advance expertise in AI and ML and has set rules and processes to guarantee data quality and privacy.

The National Institution for Transforming India (NITI) Ayog provides a report `https://www.niti.gov.in/sites/default/files/2021-08/Part2-Responsible-AI-12082021.pdf` focusing on the accelerated adoption of AI technology in India. AI plays a significant role in the national strategy, emphasizing the diversity, digital divide, scale, and lack of awareness in the Nation as factors that can amplify the risks associated with AI. In February 2021, an approach paper titled "Principles of Responsible AI" is scheduled, drawing on consultations and the Indian Constitution. AI can improve healthcare, agriculture, education, and entertainment, especially during the COVID-19 epidemic. It highlights how essential it is to use technology sensibly, reflecting the Prime Minister's remarks at the Davos Summit. It also discusses the necessity for a multidisciplinary approach to solve issues and foster confidence in AI systems, as well as the operationalization of responsible AI principles and the roles of the public, corporate, and research sectors.

3.3. Use of BDA in official statistics in India

The use of BDA in official statistics has become increasingly important in India Dubey *et al.* (2019a). BDA is a method for studying huge, complicated information to uncover insights and patterns that might help policymaking. In India, the NSO has been exploring the application of BDA to improve the accuracy and timeliness of data collection and analysis. The NSO has initiated a project to develop a comprehensive BDA platform for the research of various datasets, including those related to agriculture (Tantalaki *et al.*, 2019; Misra *et al.*, 2020), health (Chinnaswamy *et al.*, 2019; Li *et al.*, 2021), and the environment Dubey *et al.* (2019b, 2020). Nonetheless, issues like the lack of qualified personnel and the necessity to guarantee the privacy and quality of data continue to exist.

Despite facing challenges, BDA in official statistics has displayed promising outcomes in India. The BDA program has been used in India to evaluate and monitor various government programs, such as the Pradhan Mantri Jan Dhan Yojana (PMJDY) financial inclusion program (`https://pib.gov.in/Pressreleaseshare.aspx?PRID=1649091`) and the PMFBY crop insurance program (`https://transformingindia.mygov.in/scheme/pradhan-mantri-fasal-bima-yojana/`). Using BDA, the government accurately identified scheme beneficiaries and improved resource targeting. BDA is used to assess the effectiveness of environmental policies and programs, such as the Swachh Bharat Abhiyan (Clean India Mission) campaign (`https://www.pmindia.gov.in/en/major_initiatives/swachh-bharat-abhiyan/`). Consequently, BDA can facilitate data-driven policy decisions by the government by transforming Indian official statistics.

The goal of good governance in democratic countries is to ensure the provision of public services through effective participation to ensure accountability, responsiveness, and transparency. Meeting SDGs is one way of accomplishing this. Converged governance efforts at the grassroots level are required to achieve sustained development, which generates continuous baseline data. The amalgam of structured and unstructured data through BDA and emerging information and communication technologies (ICTs) can revolutionize governance processes and support data-backed decision-making Malhotra *et al.* (2018).

In addition, BDA can also help identify trends and patterns in official statistics that might otherwise go unnoticed. Analyzing social, economic, and environmental issues can be improved by combining data from social media, geospatial, and survey sources. BDA can be used for disease tracking and predicting crop yields based on weather patterns. As a result, the government can take proactive measures to address and prevent potential problems. For example, Mamatha *et al.* (2023) used BDA to track the spread of diseases, while Jaber *et al.* (2022) predicted crop yields using BDA. Policymakers can also use BDA insights to monitor and evaluate the effectiveness of government policies and programs, providing valuable data-driven feedback.

3.4. Exploring integrated data systems (IDS)

IDS are indispensable tools in National Statistics, ensuring data quality and efficient data collection for digital surveys. These systems aim to maximize information accessibility while minimizing user effort. The integration of data from various sources is a key aspect of the transformation of national statistics systems in the digital age Gokhberg *et al.* (2020). This integration should be supported by well-defined data governance frameworks Križman and Tissot (2022) and should consider the specific features of the digital economy Kasianova *et al.* (2021).

The integration process can be challenging, as emphasized by Sakshaug and Steorts (2023). Their discussion on merging surveys and administrative data underscores the complexities involved, including ethical considerations and computational burdens. Obtaining consent for data linkage and improving accuracy through computational techniques emerge as critical focal points.

Another innovative approach, as explored by Haim *et al.* (2023) involves a user-centered paradigm for data collection, where participants contribute digital trace data for academic research. This novel method combines survey data with donated data to unlock

deeper insights. Challenges such as methodological and ethical considerations are addressed alongside software solutions aimed at enhancing usability and reducing drop-outs.

The necessity for data and statistics in monitoring SDGs is addressed by Abbas *et al.* (2023). The authors highlight challenges in data dissemination and suggest AI as a potential solution. Furthermore, it proposes capacity development projects and a comprehensive indicator utilizing AI for processing data and producing official statistics.

The digital era brings challenges and opportunities, as discussed by Hassani and MacFeely (2023) Their comprehensive framework for digital data governance emphasizes the evolving landscape due to emerging technologies, underlining the importance of ethics and trustworthiness. Vavilova and Ketova (2023) developed an analytical system for regional socio-economic processes using official data and dynamic models to forecast and examine time, territory, and age-related indicators.

Daraio *et al.* (2022) proposes a completeness-aware integration approach to enhance data quality. Gootzen *et al.* (2023) introduces a quality framework for combining survey, administrative, and big data, showcasing its application in case studies involving mobility and virus detection data.

These studies elucidate the significance of IDS in shaping the future of National Statistics, offering valuable insights and innovative solutions to meet the challenges of the digital age.

3.5. Benefits and challenges of digital government in official statistics

Digital government has brought about many benefits in official statistics. Data science, AI, and BDA extract insights and patterns from vast data. As a result, statistical analysis is now more accurate and effective, which helps policymakers make better choices. But, there are also some difficulties with the help of the digital government in official statistics.

One of the benefits of digital government in official statistics is improved data quality. The utilization of digital systems guarantees the standardization and consistency of data collection, processing, and storage. As a result, data collection and analysis are more accurate, and statistical outputs are more reliable. Furthermore, digital systems enable real-time data collection, allowing for the latest statistical information essential in today's rapidly changing world.

Another benefit of digital government in official statistics is increased efficiency. Utilizing cutting-edge tools like ML and BDA speeds up data processing and analysis, requiring less time and effort than traditional statistical methods. Consequently, this enables policymakers to make more timely and informed decisions, leading to improved governance.

However, ensuring data privacy and security is one of the biggest challenges in adopting digital government in official statistics. Digital data storage requires data protection regulations and privacy laws to prevent breaches and misuse.

In today's society, a digital divide is growing between those with and without access to technology. India's digital government use in official statistics may leave behind marginal-

ized communities due to the significant digital divide, leading to unequal representation in statistical outputs. The need to overcome the digital divide and ensure that all communities have access to technology is evident from this. It is imperative to close the digital divide and ensure all communities have access to digital technology in light of these findings.

In conclusion, the benefits of digital government in official statistics are numerous, including improved data quality and increased efficiency. However, the approval of digital government also comes with several challenges, including data privacy and security and the digital divide. Balancing advanced technologies with ethical considerations is crucial for accessible digital government benefits in official statistics.

4. Case studies

4.1. Case study 1: Digital India and official statistics

Through the Digital India program, India aims to become a knowledge economy and a society empowered by digital technology. Improvements in official statistics will be a key focus of this initiative since they shape decisions regarding policy. This case study examines the implementation of digital India in official statistics and identifies its benefits and challenges.

4.1.1. Literature review

A knowledge-based economy is the objective of the Digital India initiative. Enhancing digital literacy and infrastructure and promoting digital services are essential to the program's success. Furthermore, digital technology will be essential to raising the standard of official data.

The use of technology in official statistics has been a topic of interest in the literature for several years. The study by Saxena (2018) explores the impact of demographic variables on the perception of corruption in e-government services in India. Hierarchical regression analysis shows that only gender influences the perception of corruption, with men perceiving a decrease and women perceiving an increase post-launch of the Digital India initiative. The study fills a gap in the literature by highlighting the importance of considering demographic variables in understanding citizens' perceptions of corruption in developing countries. Its small sample size and narrow focus on demographic variables limit the study.

The article by Rao (2019) examines the processes of identity creation in digital India through the use of Aadhaar. It challenges the distinction between identification and identity and shows how Aadhaar procedures create or deny conditions for belonging. It involves stitching together a digital signature, documentary proof of identity, and personal recognition to become a rights-bearing individual. Aadhaar adds a new layer of procedures on top of older methods of recognition, insisting on unique individual recognition while also recognizing a specific status.

Aadhaar is India's biometric program, which captures iris scans, fingerprints, facial photos, and demographic data from over 90% of the population. Nair (2021) argues that Aadhaar prompts a re-evaluation and contestation of individualism in postcolonial India because it dataficates the body. Additionally, it suggests that the program facilitates belonging

in the emerging technocratic imagination of a digital India.

The authors Gautam *et al.* (2022) conducted a study to examine the impact of financial technology on digital literacy in India, using the poverty score as a moderating variable. They found that Kisan Credit Cards (KCCs) had a positive association with literacy rate, while ATMs had a negative one. However, both KCCs and ATMs had a beneficial effect on literacy when interacting with poverty scores. The study's findings have implications for policymakers to understand the situation at the ground level while forming new policies for society's betterment. The authors suggest that ordinary people should take advantage of financial technology and get motivated toward digital literacy. The study by Gautam and Kanoujiya (2022) examined the impact of regional rural banks on digital literacy and rural development in India, using data from 29 Indian states and two union territories over three fiscal years. The study concluded that regional rural banks support digital literacy and rural development, and it advised banks and the government to concentrate on these issues to advance financial inclusion and rural development.

The article by Al Dahdah and Mishra (2022) examines India's transition to digital healthcare via the Rashtriya Swasthya Bima Yojana (RSBY) program and its use of smart cards. The authors discuss the politics of digitized public-private welfare policy and question the value of a program that aims to deliver affordable, high-quality healthcare to the private health market. The authors analyze digital access to healthcare in RSBY, questioning the role of digital technologies in transforming healthcare access in India. The study by Kameswaran *et al.* (2023) examines the challenges faced by people with visual impairments in India when accessing digital banking technology. The authors argue that there is a gap in research on the challenges faced by people with disabilities in obtaining accessible technology in the first place. Through qualitative research, the authors find that participants faced social and technical difficulties and engaged in advocacy work to secure and maintain access to digital banking. They expand on the view of advocacy as a form of access work performed by people with visual impairments.

The Department of Science and Technology (DST) has launched several pioneering initiatives in the realms of Data Science, Big Data, and the Internet of Things (IoT). These programs underscore the potential of data science in official statistics while also highlighting pertinent challenges (`https://dst.gov.in/data-science-research-initiative`, `https://dst.gov.in/big-data-initiative-1`, `https://dst.gov.in/internet-things-iot-research-initiative`). To enhance innovation policy delivery and monitoring in their respective sectors, DST also introduced the Automotive Sectoral System of Innovation (IASSI) and the Indian ICT Sectoral System of Innovation (IICTSSI) in 2023. (`https://dst.gov.in/sites/default/files/Indian%20Automotive%20Sectorial%20System%20of%20Innovation%20%28IASSI%29%20Report_0.pdf`, `https://dst.gov.in/sites/default/files/Indian%20ICT%20Sectorial%20System%20of%20Innovation%20%28IISSI%29%20Report_0.pdf`). Despite encountering challenges, both initiatives offer evidence-based development priorities and policy options, emphasizing the importance of effective management and connectivity for driving innovation and economic value.

4.1.2. Case study analysis

- Overview of official statistics in India:
  The MOSPI is in charge of the nation of Indian official statistics system. The system is responsible for gathering, putting together, and disseminating official statistics on different socio-economic indicators. The system consists of organizations, including NSSO, CSO, and RGCCI.

- Background on digital India program:
  Using technology to empower Indian society and economy, Digital India is a government initiative. Digital infrastructure, digital literacy, and digital services are intended to be created through the program. `https://pib.gov.in/PressReleaseIframePage.aspx?PRID=1885962`). Through this program, citizens can access government services and information digitally. It focuses on infrastructure, governance, and services.

  The program includes initiatives such as the creation of digital infrastructure, such as the National Optical Fibre Network (NOFN) (`https://ddd.gov.in/scheme/bharat-net/`); the origination of digital literacy programs, the development of e-governance platforms and the promotion of digital financial services. Through this program, the government aims to empower citizens by providing them with digital tools and services that enhance their participation in the country's economic, social and political spheres.

- Evaluation of the implementation of digital India in official statistics:
  The implementation of digital India in official statistics has been ongoing since the launch of the initiative in 2015 (`https://csc.gov.in/digitalIndia`). Some key initiatives undertaken by the Ministry of Electronics and Information Technology (MeitY) under the Digital India program are Aadhaar, DigiLocker, Open Government Data Platform, etc. (`https://pib.gov.in/PressReleaseIframePage.aspx?PRID=1885962`). The initiative has focused on several areas; including the following:

  – Digitization of data collection: The initiative has aimed to digitize data collection processes to improve the accuracy and timeliness of official statistics.
  – Development of digital platforms: The initiative has aimed to develop digital platforms for the dissemination of official statistics, such as the MOSPI website and mobile applications.
  – Use of data analytics: The initiative has aimed to leverage the power of data analytics to extract insights and patterns from official statistics.

Overall, the implementation of digital India in official statistics has led to several benefits; including the following:

- Improved accuracy and timeliness of official statistics: Digitalization has facilitated the speedy publication of official statistics and the reduction of errors.

- Increased accessibility of official statistics: The development of digital platforms has made official statistics more accessible to the general public, researchers, and policymakers.

- Increased efficiency of official statistics: The use of data analytics has improved the efficiency of official statistics analysis, allowing policymakers to make more informed decisions based on them.

However, the implementation of digital India in official statistics has also faced several challenges, including the following:

- Quality of data: The quality of data collected through digital platforms may be affected by problems such as incomplete or inaccurate data or bias in the sampling process.
- Privacy concerns: The digitization of data collection processes raises concerns about the privacy and confidentiality of individual data.
- Infrastructure challenges: The implementation of digital India in official statistics requires significant investment in digital infrastructure, which may be a challenge for some regions.

India's government uses advanced tools for data cleaning, standardization, and validation to improve data quality. For instance, the MOSPI has established the National Data Quality Forum (NDQF) to improve data quality across government agencies (`https://ndqf.in/`). The NDQF has implemented data quality scorecards and audits to ensure the accuracy and reliability of data. By spotting mistakes and abnormalities in data sets, the application of ML algorithms for predictive modeling has also improved the quality of data (Ngiam and Khor, 2019; Gruson *et al.*, 2019; Sharma *et al.*, 2020).

The MOSPI unveiled the digital India Mobile Van (`https://pib.gov.in/PressReleaseIframePage.aspx?PRID=1895957`). Mobile Vans are unique initiatives under the program that provide digital literacy and allow remote and inaccessible areas in the country to access digital services. The vans are equipped with computers and other digital accessories, such as printers and scanners, which are used to provide several digital services. It includes digital services for connecting to the Internet, imparting digital literacy, and rendering government services electronically, such as enabling individuals to use the Internet for registering their birth and death certificates, etc. In addition to reaching out to women, seniors, and people with disabilities with limited access to digital infrastructure, the initiative ensures that digital services reach out to the most vulnerable communities. This initiative offsets the digital divide and ensures that marginalized groups access digital services from their residences, making it more inclusive and empowering citizens in different parts of the country. It has a significant contribution to ensuring the success of the Digital India program since all citizens have access to digital services irrespective of their geographical location or economic status.

On the other hand, the implementation of digital technologies in official statistics comes with various challenges, like data privacy and security issues. To solve this, the government of India has undertaken several measures. Firstly, the Data Protection Bills have been set (`https://www.meity.gov.in/writereaddata/files/The%20Digital%20Personal%20Data%20Protection%20Bill%2C%202022.pdf`). This bill primarily regulates how people's data can be collected, stored, and processed in India. Secondly, the National

Cyber Security Coordinator (NCSC) has been set to coordinate and oversight cybersecurity activities in government (https://pib.gov.in/PressReleaseIframePage.aspx?PRID=1556474).

In conclusion, the execution of digital technologies in official statistics through initiatives such as Digital India has significantly improved the efficiency and effectiveness of government decision-making. It is still necessary to address quality issues and privacy concerns to ensure the correct and responsible use of official statistics. The Indian government has taken various measures to address these challenges, and continued efforts in this direction will be crucial for the success of digital government and official statistics in India.

4.2. Case study 2: Use of data science and BDA in the Indian census

The Indian census is one of the world's major administrative tasks, with over 1.39 billion people residing in India (https://statisticstimes.com/demographics/country/india-population.php). To make informed decisions, particularly in the healthcare, education, and infrastructure sectors, policymakers and government officials need the census. The application of data science and BDA has become more and more necessary for the analysis of the massive volumes of data collected during the census. This case study will explore the use of data science and BDA in the Indian census, focusing on their benefits and challenges.

4.2.1. Background

India has a census system that gathers socioeconomic and demographic data from each home every ten years. Businesses, researchers, and policymakers benefit from the census's valuable data. The introduction of digital technology has made the census more accurate and efficient than it was under the paper-based system. Under the British Raj, India conducted its first census in 1872. (https://censusindia.gov.in/nada/index.php/catalog/40444/download/44078/DROP_IN_ARTICLE-05.pdf). The Office of the Registrar General and Census Commissioner of India (ORGI), which is in charge of compiling and disseminating census data, conducts the census. (https://censusindia.gov.in/census.website/node/378).

4.2.2. Use of data science

With the tools it has provided for data collection, processing, analysis, and dissemination, data science has been crucial in the census. The census of India has been using data science and machine learning algorithms to improve the accuracy and efficiency of its operations. An example is the use of ML algorithms to improve the quality of data collection. The census uses Paper Data Capture Operation (PDC), which includes an automated data collection system that uses optical character recognition (OCR) to read the data collected from paper forms (https://www2.census.gov/programs-surveys/decennial/2020/program-management/planning-docs/PDC_detailed_operational_plan.pdf). Additionally, the system alerts any data error such as missing or inaccurate entries for inspection by using ML methods. The gathered information becomes more accurate and trustworthy as a result.

4.2.3. Use of BDA

The article by Chatfield *et al.* (2018) focuses on the state of big data and BDA use in the National census context of four countries: Australia, Ireland, Mexico, and the U.S.A. The study found that the census agencies in these countries are at varying stages in digitally transforming their census process, products, and services through assimilating and using big data and BDA. However, the cross-case analysis of government websites and documents revealed emerging challenges in creating public value in the national census context, including BDA capability development, cross-agency data access & integration, and data security, privacy, and trust. Based on the insights gained, the article proposes a research model to explore the possible links among these challenges, BDA use, and public value creation.

The study by Marathe *et al.* (2020) presents a data science pipeline that integrates techniques such as ML, Statistics, Data Visualization, and Geographic Information System (GIS) for open big data in sustainable development. Using this pipeline, the Pune Municipal Corporation applied the geo-enabled tree census dataset to its tree census data. The study focuses on the visualization of big data, ward-wise analysis, and identification of marginalized species that require urgent attention from the authorities. A new biodiversity index is introduced in this study to address the limitations of existing indices when applied to cities in the Indian subcontinent. Overall, this study highlights the potential of data science techniques in analyzing big data and providing insights into sustainable development.

There are no studies utilizing BDA for the Indian census currently. BDA can be used in the Indian census to analyze large volumes of data and extract insights and patterns. By collaborating with technology companies, such as IBM and Microsoft, ORGI can develop BDA tools for the census. In addition to helping identify population characteristics like age, gender, education, and occupation, these tools will help analyze census data. As a result, policymakers and planners can gain a better understanding of demographic trends and patterns.

4.2.4. Challenges

Despite the advantages of data science and BDA in the Indian census, some issues still need to be resolved. One of the main challenges is data privacy and security. The census collects sensitive personal information, and there is a risk of this data being misused/breached. The ORGI has enforced strict protocols for regulating data handling and storage.

A significant number of Indians still lack access to digital technology due to the digital divide. Census data for these populations may be inaccurate or underrepresented as a result. Using offline methods and training field enumerators to collect data using non-digital means has been one of the strategies used by the ORGI to reach out to these populations.

4.2.5. Conclusion

Data science and BDA can improve data quality, accuracy, and efficiency in the Indian census. However, several challenges are addressed, including data privacy and security and the digital divide. To overcome these obstacles and guarantee that the census data is accurate, dependable, and secure, the ORGI should develop strategies and protocols. It is essential to balance technological innovation and ethical considerations when using census

data to ensure the responsible use of advanced technologies.

4.3. Case study 3: Artificial intelligence and official statistics in India

The National Sample Survey (NSS) data is used as an example in this case study to examine the application of AI in Indian government statistics. AI has the power to completely change the methods used to gather, handle, and evaluate official statistics. The study offers insights into the application of AI in the NSS data and looks at the advantages and difficulties of utilizing it in official statistics.

4.3.1. Background

Official data have always been gathered and published by the Indian government, going back to the colonial era. On the other hand, there is increasing interest in investigating the application of AI in official statistics due to the quick evolution of technology. AI can make data collection and analysis more quick, accurate, and efficient. This case study especially looks at the NSS data as an example of how AI is being used in official statistics in India. With a broad scope of social and economic variables covered, the NSS is the biggest household survey carried out in India. The NSS is India's largest household survey, covering socioeconomic indicators used by policymakers, researchers, and companies to understand its socioeconomic situation. Interviews and self-completed questionnaires are used for data collection. Traditional data collection and analysis are time-consuming.

AI is increasingly used in official statistics in India to improve data analysis, prediction, and decision-making. The study by Chawla *et al.* (2022) examines the role of AI and Information Management (IM) in India's energy transition, which has been strained due to rapid urbanization and modernization. Despite India's status as the global IT heart and having above-average research output in AI, it has not fully leveraged its benefits in the energy sector. The study analyzes proposed strategies, current policies, and available literature to highlight the challenges and barriers to developing and using AI and IM in India's energy sector. The study suggests that policymakers in India must take adaptive and swift actions toward developing comprehensive AI and IM policies to extract maximum benefits from the ongoing transition of the energy sector.

The article by Chatterjee *et al.* (2022) explores the public value generated by AI-enabled services from the perspective of Indian citizens. An analysis of 315 interviews is conducted using the Partial Least Square-Structural Equation Modeling (PLS-SEM) technique based on IT assimilation theory and public value theory. The study finds that the assimilation of AI-enabled services positively impacts citizens' satisfaction and generates public value. It also identifies risk factors that may influence the uptake of such services. The paper contributes to understanding the benefits and challenges of AI-enabled services in the public sector. For instance, the government has started several initiatives that use AI and ML algorithms to improve the timeliness and accuracy of official statistics. One such example is the use of chatbots for data collection and analysis. India introduces a WhatsApp chatbot to spread knowledge about the coronavirus and request social media platforms to stop the spread of false information. (`https://techcrunch.com/2020/03/21/india-whatsapp-mygov-corona-helpdesk-bot/`) The chatbot responds immediately to user inquiries, speeding up response times and increasing data accuracy.

Another example is the application of predictive modeling to estimate population growth and migration patterns. The study by Devi *et al.* (2022) analyzes the Land Use and Land Cover (LULC) change rate of Cochin, an urbanized coastal city in India. A contrast of the observed and simulated LULCs of 2020 validated the model's simulation. The model demonstrated acceptable LULC dynamics, with an overall accuracy of 87.5%. The future scenarios of LULC, projected till 2100, show an increase in built-up lands and a shrinkage of natural land covers, such as forests and water bodies. The urban growth indicator confirms the extreme transformation of the area in terms of urbanization. The study suggests establishing appropriate urban planning and management policies for sustainable environmental conservation.

In addition to the census, the government has also launched several AI-based initiatives to improve the collection and analysis of data in various sectors, including health, agriculture, and education. For instance, the National Health Stack (NHS) is a government initiative that aims to digitize health records and use AI to analyze the data to improve healthcare services. (`https://abdm.gov.in:8081/uploads/NHS_Strategy_and_Approach_1_89e2dd8f87.pdf`).

However, some challenges are involved. The study by Sharma *et al.* (2022) explores the interrelationships and challenges of implementing AI in India's Public Manufacturing Sector (PMS). AI integration with PMSs is challenging due to low data quality, inadequate understanding of cognitive technologies, privacy concerns, and the high cost of implementing cognitive projects. The study proposes a model for decision-makers and managers to develop intelligent AI-enabled systems for manufacturing organizations in emerging economies. The study highlights the need to address these challenges to enhance the scope of AI implementation in the PMS sector.

4.3.2. Methodology

This case study employs a qualitative research methodology, focusing on the study of secondary data sources. Based on literature, official reports, and interviews with field experts, the study examines the use of AI in NSS data. The analysis is guided by the following research questions:

- What are the benefits of using AI in official statistics, specifically in the NSS data?
- What are the challenges associated with implementing AI in official statistics?
- How has AI been implemented in the NSS data, and what are the implications of this implementation?

4.3.3. Findings

According to the study, AI has the potential to greatly increase the timeliness, accuracy, and efficiency of data gathering and processing in official statistics. AI can automate complex tasks like data imputation, validation, and cleaning. It can enhance data quality and lower mistake rates. Large data sets may be rapidly analyzed by AI, which enables researchers to identify patterns and insights that would be difficult to discover manually.

Moreover, AI can assist in lowering the expenses related to data collection and processing by replacing human labor.

Even with these benefits, there are still difficulties in integrating AI into official data. The quality of the data is one of the main obstacles. Unless AI algorithms are trained on accurate or neutral data, they will provide biased or inaccurate results. Data security and privacy are other issues since AI needs access to a lot of personal information. In low-resource environments, it might be difficult to find qualified workers to develop, deploy, and maintain AI systems.

In the case of the NSS data, AI has been implemented in several ways, such as using Natural Language Processing (NLP) techniques to extract data from open-ended survey questions and using ML algorithms to impute missing data. However, there are still challenges associated with the implementation of AI in the NSS data, such as the need for more training data to improve the accuracy of the algorithms.

4.3.4. Conclusion

In conclusion, the use of AI in official statistics has the potential to completely transform India's data collection, processing, and analysis methods. Incorporating AI into the statistics system is a positive step for the government. Addressing AI's issues and worries is essential, especially those about data security and privacy. AI has the potential to be a helpful tool in official statistics and advance national development with the correct policies and approaches.

5. Future directions and concluding remarks

There are several potentials for the future of official statistics in India as the benefits of data science, AI, and BDA continue to develop. When these cutting-edge technologies are combined, data gathering, processing, and distribution may become more accurate and efficient. Predictive modeling and ML may also help anticipate future trends and patterns, which can give policymakers and decision-makers important information. There are advantages and drawbacks, such as concerns about data privacy and quality, the digital divide, and ethics.

Digital government has a critical role in shaping the future of official statistics in India. It is possible to improve the accuracy, efficiency, and accessibility of official statistics by using advanced technologies. The digital government may also help different government departments and stakeholders engaged in official statistics collaborate and coordinate. It can also ensure that the benefits of official statistics are shared equitably among all sections of society, including marginalized communities. However, the digital government must also ensure that the ethical considerations associated with the advantages of advanced technologies are addressed and that the benefits of official statistics are shared equitably.

Collaboration among policymakers, researchers, and practitioners is essential for maximizing the benefits and minimizing the challenges of advanced technologies. Some recommendations for each group include:

- Policymakers: Policymakers must prioritize investment in technology and infrastruc-

ture to support the implementation of advanced technologies in official statistics. Additionally, these technologies must address ethical considerations, data quality, and privacy concerns. Furthermore, they must prioritize capacity building and training to ensure that government officials have the necessary skills and knowledge to implement these technologies effectively.

- Researchers: Researchers need to investigate the possible advantages and challenges associated with the implementation of advanced technologies in official statistics. In addition, they must develop and share best practices for the responsible use of these technologies. Furthermore, they must collaborate with government agencies to ensure that research findings are translated into policy and practice.
- Practitioners: Practitioners involved in official statistics must prioritize the development of data quality and management frameworks to ensure that data is accurate, reliable, and timely. The benefits of official statistics must also be shared equitably across society, including marginalized groups. Furthermore, they must engage in ongoing professional development to ensure they have the necessary skills and knowledge to implement advanced technologies effectively.

To sum up, the use of advanced technologies, such as data science, AI, and BDA, has the potential to transform the field of official statistics in India. Cooperation between researchers, policymakers, and practitioners will maximize benefits while minimizing obstacles. Digital governments must address the ethical issues associated with these technologies and ensure that all sections of society can share the benefits of official statistics.

In conclusion, this review article explored the use of data science, AI, and BDA in official statistics in India's digital government. This study found that the implementation of digital government programs has greatly enhanced the gathering, processing, and distribution of official statistics in India. While AI has aided in the development of predictive modeling and pattern recognition, data science and BDA have made it possible to collect and analyze data more thoroughly and accurately. However, there are several difficulties with using these technologies in official statistics, such as bias, privacy, and data quality concerns. To guarantee that the use of cutting-edge technology in official statistics is morally acceptable, responsible, and accurate, policymakers and practitioners need to address these issues.

The implications of this study are significant for both practice and policy. It emphasizes the importance of implementing cutting-edge technologies in official statistics to improve data quality and facilitate data analysis. Also, it emphasizes the importance of thoroughly considering ethical and privacy concerns before introducing new technologies. Digital government can promote the application of cutting-edge technology in official statistics, and government support is essential for advancing pertinent infrastructure and expertise. While using these technologies, policymakers should ensure that official statistics remain accurate and unbiased.

The future of official statistics in India depends on the development of digital government and the use of advanced technologies, such as data science, AI, and BDA. Future research in this area should explore the ethical and privacy concerns surrounding the benefit of these technologies in official statistics. Additionally, research can explore how these technologies can improve the quality and accuracy of official statistics in India's rural and

regional regions. Furthermore, research should examine how advanced technologies can facilitate the accessibility and dissemination of official statistics to decision-makers, researchers, and the general public.

Acknowledgements

The authors wish to express sincere gratitude to the Editors for their valuable guidance and counsel. We are deeply thankful to the anonymous referees for their constructive comments and insightful suggestions on the earlier version of this manuscript.

Conflict of interest

The authors do not have any financial or non-financial conflict of interest to declare for the research work included in this article.

References

Abbas, S. W., Hamid, M., Alkanhel, R., and Abdallah, H. A. (2023). Official statistics and big data processing with artificial intelligence: Capacity indicators for public sector organizations. *Systems*, **11**, 424.

Al Dahdah, M. and Mishra, R. K. (2022). Digital health for all: The turn to digitized healthcare in India. *Social Science & Medicine*, **319**, 114968.

Alvarenga, A., Matos, F., Godina, R., and CO Matias, J. (2020). Digital transformation and knowledge management in the public sector. *Sustainability*, **12**, 5824.

Ashofteh, A. and Bravo, J. M. (2021). Data science training for official statistics: A new scientific paradigm of information and knowledge development in national statistical systems. *Statistical Journal of the IAOS*, **37**, 771–789.

Chatfield, A. T., Ojo, A., Puron-Cid, G., and Reddick, C. G. (2018). Census big data analytics use: International cross case analysis. In *Proceedings of the 19th Annual International Conference on Digital Government Research: Governance in the Data Age*, pages 1–10.

Chatterjee, S. (2020). AI strategy of India: policy framework, adoption challenges and actions for government. *Transforming Government: People, Process and Policy*, **14**, 757–775.

Chatterjee, S., Khorana, S., and Kizgin, H. (2022). Harnessing the potential of artificial intelligence to foster citizens' satisfaction: An empirical study on India. *Government Information Quarterly*, **39**, 101621.

Chawla, Y., Shimpo, F., and Sokołowski, M. M. (2022). Artificial intelligence and information management in the energy transition of India: lessons from the global IT heart. *Digital Policy, Regulation and Governance*, **24**, 17–29.

Chinnaswamy, A., Papa, A., Dezi, L., and Mattiacci, A. (2019). Big data visualisation, geographic information systems and decision making in healthcare management. *Management Decision*, **57**, 1937–1959.

Daraio, C., Leo, S. D., and Scannapieco, M. (2022). Accounting for quality in data integration systems: a completeness-aware integration approach. *Scientometrics*, **127**, 1465–1490.

Devi, A. B., Deka, D., Aneesh, T. D., Srinivas, R., and Nair, A. M. (2022). Predictive modelling of land use land cover dynamics for a tropical coastal urban city in Kerala, India. *Arabian Journal of Geosciences*, **15**, 399.

Dubey, R., Gunasekaran, A., Childe, S. J., Blome, C., and Papadopoulos, T. (2019a). Big data and predictive analytics and manufacturing performance: integrating institutional theory, resource-based view and big data culture. *British Journal of Management*, **30**, 341–361.

Dubey, R., Gunasekaran, A., Childe, S. J., Bryde, D. J., Giannakis, M., Foropon, C., Roubaud, D., and Hazen, B. T. (2020). Big data analytics and artificial intelligence pathway to operational performance under the effects of entrepreneurial orientation and environmental dynamism: A study of manufacturing organisations. *International Journal of Production Economics*, **226**, 107599.

Dubey, R., Gunasekaran, A., Childe, S. J., Papadopoulos, T., Luo, Z., Wamba, S. F., and Roubaud, D. (2019b). Can big data and predictive analytics improve social and environmental sustainability? *Technological Forecasting and Social Change*, **144**, 534–545.

Gautam, R. S. and Kanoujiya, J. (2022). Role of regional rural banks in rural development and its influences on digital literacy in India. *Iconic Research and Engineering Journals*, **5**, 92–101.

Gautam, R. S., Rastogi, S., Rawal, A., Bhimavarapu, V. M., Kanoujiya, J., and Rastogi, S. (2022). Financial technology and its impact on digital literacy in India: Using poverty as a moderating variable. *Journal of Risk and Financial Management*, **15**, 311.

Ghosh, J. K., Maiti, P., Rao, T. J., and Sinha, B. K. (1999). Evolution of statistics in India. *International Statistical Review/Revue Internationale de Statistique*, **67**, 13–34.

Gokhberg, L., Kuznetsova, T., Abdrakhmanova, G., Fursov, K., Nechaeva, E., Shahsnov, S., and Suslov, A. (2020). Prospective model of official statistics for the digital age. *Higher School of Economics Research Paper No. WP BRP*, **111**.

Gootzen, Y., Daas, P. J. H., and van Delden, A. (2023). Quality framework for combining survey, administrative and big data for official statistics. *Statistical Journal of the IAOS*, **39**, 439–446.

Gruson, D., Helleputte, T., Rousseau, P., and Gruson, D. (2019). Data science, artificial intelligence, and machine learning: opportunities for laboratory medicine and the value of positive regulation. *Clinical Biochemistry*, **69**, 1–7.

Guntukula, R. (2020). Assessing the impact of climate change on Indian agriculture: Evidence from major crop yields. *Journal of Public Affairs*, **20**, e2040.

Haim, M., Leiner, D., and Hase, V. (2023). Integrating data donations in online surveys. *Medien & Kommunikationswissenschaft*, **71**, 130–137.

Hantrais, L., Allin, P., Kritikos, M., Sogomonjan, M., Anand, P. B., Livingstone, S., Williams, M., and Innes, M. (2021). Covid-19 and the digital revolution. *Contemporary Social Science*, **16**, 256–270.

Hassani, H. and MacFeely, S. (2023). Driving excellence in official statistics: Unleashing the potential of comprehensive digital data governance. *Big Data and Cognitive Computing*, **7**, 134.

Ilyas, Q. M., Ahmad, M., and Mehmood, A. (2023). Automated estimation of crop yield using artificial intelligence and remote sensing technologies. *Bioengineering*, **10**, 125.

Jaber, M. M., Ali, M. H., Abd, S. K., Jassim, M. M., Alkhayyat, A., Aziz, H. W., and Alkhuwaylidee, A. R. (2022). Predicting climate factors based on big data analytics based agricultural disaster management. *Physics and Chemistry of the Earth, Parts A/B/C*, **128**, 103243.

Kameswaran, V., Y, V., and Marathe, M. (2023). Advocacy as access work: How people with visual impairments gain access to digital banking in India. *Proceedings of the ACM on Human-Computer Interaction*, **7**, 1–23.

Kasianova, N., Kendiuknov, O., and Pishenina, T. (2021). Features of the assessment of state and a prospective for the development of the digital economy. In *1st International Scientific Conference" Legal Regulation of the Digital Economy and Digital Relations: Problems and Prospects of Development"(LARDER 2020)*, pages 233–239. Atlantis Press.

Križman, I. and Tissot, B. (2022). Data governance frameworks for official statistics and the integration of alternative sources. *Statistical Journal of the IAOS*, **38**, 947–955.

Li, W., Chai, Y., Khan, F., Jan, S. R. U., Verma, S., Menon, V. G., and Li, X. (2021). A comprehensive survey on machine learning-based big data analytics for IoT-enabled smart healthcare system. *Mobile Networks and Applications*, **26**, 234–252.

Malhotra, C., Anand, R., and Singh, S. (2018). Applying big data analytics in governance to achieve sustainable development goals (SDGs) in India. *Data Science Landscape: Towards Research Standards and Protocols*, **38**, 273–291.

Mamatha, K., Samantha, S., and Prasad, K. K. (2023). Crop yield prediction using deep learning. In *ICDSMLA 2021: Proceedings of the 3rd International Conference on Data Science, Machine Learning and Applications*, pages 93–102. Springer.

Marathe, A., Mirchandani, K., Chordiya, K., and Stephen, K. (2020). Big data analytics for sustainable cities: Pune tree census data exploratory analysis. In *2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*, pages 1–7. IEEE.

Misra, N., Dixit, Y., Al-Mallahi, A., Bhullar, M. S., Upadhyay, R., and Martynenko, A. (2020). IoT, big data, and artificial intelligence in agriculture and food industry. *IEEE Internet of Things Journal*, **9**, 6305–6324.

Nair, V. (2021). Becoming data: biometric IDs and the individual in 'Digital India'. *Journal of the Royal Anthropological Institute*, **27**, 26–42.

Ngiam, K. Y. and Khor, W. (2019). Big data and machine learning algorithms for health-care delivery. *The Lancet Oncology*, **20**, e262–e273.

Rana, N. P., Luthra, S., and Rao, H. R. (2020). Key challenges to digital financial services in emerging economies: the Indian context. *Information Technology & People*, **33**, 198–229.

Rao, T. J. (2010). Official statistics in India: The past and the present. *Journal of Official Statistics*, **26**, 215.

Rao, T. J. (2013). National statistical commission and Indian official statistics. *Resonance*, **18**, 1062–1072.

Rao, U. (2019). Population meets database: Aligning personal, documentary and digital identity in Aadhaar-enabled India. *South Asia: Journal of South Asian Studies*, **42**, 537–553.

Sakshaug, J. W. and Steorts, R. C. (2023). Recent advances in data integration. *Journal of Survey Statistics and Methodology*, **11**, 513–517.

Sarma, N. (1958). Economic development in India: The first and second five year plans. *Staff Papers-International Monetary Fund*, **6**, 180–238.

Saxena, S. (2018). Perception of corruption in e-government services post-launch of "Digital India": Role of demographic variables. *Digital Policy, Regulation and Governance*, **20**, 163–177.

Sharma, A., Jain, A., Gupta, P., and Chowdary, V. (2020). Machine learning applications for precision agriculture: A comprehensive review. *IEEE Access*, **9**, 4843–4873.

Sharma, M., Luthra, S., Joshi, S., and Kumar, A. (2022). Implementing challenges of artificial intelligence: Evidence from public manufacturing sector of an emerging economy. *Government Information Quarterly*, **39**, 101624.

Sinha, B. B. and Dhanalakshmi, R. (2022). Recent advancements and challenges of Internet of Things in smart agriculture: A survey. *Future Generation Computer Systems*, **126**, 169–184.

Subrahmanya, S. V. G., Shetty, D. K., Patil, V., Hameed, B. Z., Paul, R., Smriti, K., Naik, N., and Somani, B. K. (2022). The role of data science in healthcare advancements: applications, benefits, and future prospects. *Irish Journal of Medical Science (1971-)*, **191**, 1473–1483.

Tantalaki, N., Souravlas, S., and Roumeliotis, M. (2019). Data-driven decision making in precision agriculture: The rise of big data in agricultural systems. *Journal of Agricultural & Food Information*, **20**, 344–380.

Tripathi, M. and Dungarwal, M. (2020). Digital India: Role in development. *International Journal of Home Science*, **6**, 388–92.

Vavilova, D. D. and Ketova, K. V. (2023). Information and analytical system for the analysis of regional socio-economic processes based on the integrated use of dynamic models of various types. *Journal Of Applied Informatics*, **18**, 97–110.

Vijai, C. (2019). FinTech in India–opportunities and challenges. *SAARJ Journal on Banking & Insurance Research (SJBIR)*, **8**.

Vijai, C. and Wisetsri, W. (2021). Rise of artificial intelligence in healthcare startups in India. *Advances in Management*, **14**, 48–52.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 23, No. 1, 2025 (New Series), pp 249–272
https://www.ssca.org.in/journal

Three-State Markov Probability Distributions for the Stock Price Prediction

Tirupathi Rao Padi[1], Sarode Rekha[2] and Gulbadin Farooq Dar[3]
[1,2]*Department of Statistics Pondicherry University, Puducherry-605014*
[3]*Department of Statistics, St. Joseph's University, Bangalore-560027*

Received: 05 December 2023; Revised: 09 April 2024; Accepted: 10 June 2024

Abstract

This paper focuses on predicting the movement of State Bank of India (SBI) stock prices using the Markov model, a challenging task in financial markets. It comprises two main sections: Firstly, it formulates probability distributions for various states using Markov model parameters, deriving Pearson's coefficients like average, variance, skewness, and kurtosis. Secondly, real-time SBI data is gathered and divided into five datasets representing each business day. Numerical calculations are performed using R software, computing parameters such as transition probability matrix (TPM) and initial probability vector (IPV) for each dataset. Expected returns and closing price predictions are determined, validated through the Chi-square test for goodness of fit, and assessed for robustness using Akaike information criterion (AIC) and Bayesian information criterion (BIC). The model is designed to facilitate optimal investment strategies and could benefit from user-friendly digital interfaces for traders. It explores indicators such as timing for buying/selling, probability of price movements, expected gains/losses, and estimated closing prices to enhance understanding of SBI's market behaviour in the Indian context.

Key words: Markov model; Share price; Probability distribution; Transition probability; Initial probability.

AMS Subject Classifications: 62K05, 05B05.

1. Introduction

Mathematical and stochastic modeling are pivotal tools in unraveling the intricacies of the stock market and projecting its future trends. This paper is dedicated to crafting a three-state Markov probability distribution model to analyse stocks and anticipate their forthcoming price fluctuations.

Certainly, at the core of every nation's economic structure lies an indispensable link with its stock market. This intricate connection serves as the lifeblood of the country's financial stability, embodying a sophisticated network where a diverse spectrum of individuals and entities converge. Participants engage in a multifaceted interplay of buying and

Corresponding Author: Sarode Rekha
Email: statisticsrekha@gmail.com

selling an extensive array of financial instruments within this bustling marketplace. This dynamic interaction propels economic activities and nurtures an environment teeming with opportunities for trading and investment.

Amidst this whirlwind of financial transactions, traders emerge as the linchpin, fueled by the relentless pursuit of optimal outcomes. Their endeavours are marked by a meticulous analysis of market trends, a process that involves an exhaustive examination of historical data, intricate technical analyses, and a keen understanding of global economic influences. Armed with this knowledge, traders navigate the intricate pathways of the market, making calculated and strategic decisions.

Crucially, these traders are not guided by mere intuition but rather by a commitment to informed decision-making. They implement sophisticated risk management strategies, diligently assessing potential risks and rewards. Their objective is crystal clear: to optimise profits and minimise losses. Every move within this dynamic financial landscape is a result of careful consideration, a balance between seizing opportunities and mitigating risks.

In the grand tapestry of the stock market, the significance of this strategic decision-making cannot be overstated. It not only influences individual financial destinies but also ripples through the larger economic fabric of the nation. The stock market becomes a barometer, reflecting the collective confidence and sentiment of investors, thereby shaping the economic trajectory of the entire country.

In essence, the stock market embodies more than just financial transactions; it symbolizes the aspirations, strategies, and challenges of a nation's economic journey. Armed with their expertise and insights, traders play a pivotal role in shaping this intricate landscape, where every decision made resonates far beyond individual portfolios, weaving into the intricate tapestry of a nation's economic prosperity.

A Markov regime ARCH model used to investigate and analyse the volatility within market behaviour (Cai (1994)). There is sufficient evidence on the usage of Wiener–Hopf results for solving the option pricing problems with the Markov processes (Jobert and Rogers (2006)). When applied to forecast data from the stock market, the HMM with fuzzy model innovation produced results that were more accurate than those from forecasting models like ARIMA, ANN, *etc.* (Hassan (2009)). A flexible Mixed HMM approach that considers temporal and spatial variability. This method is adaptable because it can handle the distinctive features of financial time series data, such as asymmetry, kurtosis, and unobserved heterogeneity (Dias *et al.* (2010)). HMM and support vector machines were used to predict the movement of the stock price (Rao and Hong (2010)). The stock price dynamics were examined through a semi-Markov return model (D'Amico and Petroni (2012)). A finite state Markov chain model was used to evaluate share price movements in the share market (Choji *et al.* (2013)). The utilization of a Markov-switching using GARCH approach has provided a method for predicting the volatility in the Tehran Stock Exchange-TSE (Abounoori *et al.* (2016)). The Nigerian Stock Exchange market has utilized the Markov chain model for analysing its behaviour (Adesokan *et al.* (2017)). The Markov chain model was used to forecast the stock price movement of the Taiwanese company High Tech Computer (Huang *et al.* (2017)). The Markov chain is used in forecasting the behaviour of the Nepal Stock Exchange Index (Bhusal (2017)). The Markov chain model is used to predict the stock market trend in the context of the Indian stock market (Padi *et al.* (2022)). The HMM was utilised to

properly comprehend the financial factors in the stock market, and the results were more helpful for portfolio managers in making the best choices (Dar *et al.* (2022)). The impact of international trade on the share prices of the Industrial Bank of Korea was assessed through the utilization of stochastic prediction modelling (Dar *et al.* (2023)).

Numerous studies have predominantly concentrated on classical methodologies for either developing new models or applying existing Markov models to forecast market behaviour. However, there exists a dearth of research on deriving probability distributions for sequences of states and estimating parameters through predictive modeling, specifically tailored to Markov processes. Delving into the probability distributions of transitional states can furnish more precise information inputs. The parametric estimation within the Markov model and its extension into probability distributions have been largely overlooked by probability researchers.

In response to this research gap, our study underscores the importance of Markov modeling in formulating probability distributions by constructing the Markov model based on parameters such as TPM and IPV. We have mathematically derived explicit relationships for various statistical measures using these formulated probability distributions. Focusing on three states - *Rise State*, *Stable State*, and *Fall State* - of SBI shares, our General Markov model entails two key parameters: TPM, governing transitions among states, and IPV, describing the likelihood of each state's initial occurrence. Our primary objective is to establish probability distributions separately for *Rise State*, *Stable State*, and *Fall State* across all segregated data sets for different business days. We have derived explicit mathematical relationships for diverse statistical measures and Pearson's coefficients. Sensitivity analysis has been conducted by determining Markov model parameters, obtaining probability distributions, and analysing statistical measures to gain a comprehensive understanding of SBI share price behaviour. Additionally, our model encompasses an additional study where expected returns and closing prices of SBI are computed using the formulas outlined in Section 2.7.

2. Stochastic model

The Markov model is a type of mathematical model that focuses on predicting the next event based on the event that happened just before it, without considering events from a long time ago. This means, it doesn't have a memory of past events beyond the most recent one. The schematic diagram for the model is placed below.

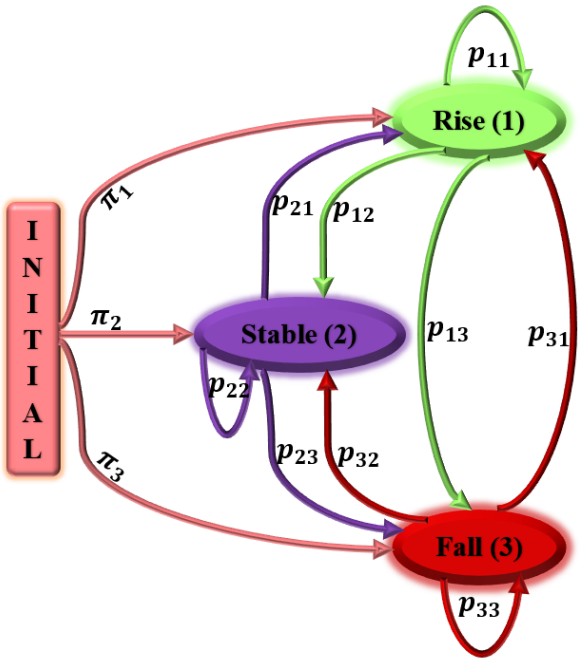


Figure 1: Schematic Diagram of Three-State Markov Model

In this study, the main aim is to figure out the likelihood of different states happening. These states are divided into three categories: *Rise State*, *Stable State*, and *Fall State*. The Markov model consists of two key parameters namely TPM and IPV.

2.1. Transition probability matrix (TPM)

A Transition Probability Matrix (TPM) is often called a Stochastic Matrix. It is defined as

$$P = Y_{n-1} \overset{Y_n}{\left(P_{jk} \right)} \qquad \forall j,k = 1,2,3$$

$P\{Y_n = k/Y_0 = 1, Y_1 = 2, ..., Y_{n-1} = j\} = P[Y_n = k/Y_{n-1} = j] = P_{jk}$ be the transition probability from jth state to kth state. Every TPM must satisfy the following conditions like,

- The matrix must possess equal numbers of rows and columns; *i.e.*, TPM is a squared matrix.
- Each element within the matrix must represent a probability; *i.e.*, $P_{jk} \geq 0$.
- The sum of each row must be equivalent to one; *i.e.*, $\sum_{k=1}^{3} P_{jk} = 1, \forall j,k = 1,2,3.$

It earns the label "Doubly Stochastic Matrix" when the sums of both its each row and each column are equal to one.

2.2. Initial probability vector (IPV)

The initial probability vector determines the chance of happening in a particular state. It is denoted by π.

$$\pi = (\pi_1, \pi_2, \pi_3)$$

2.3. Notations and terminology

π_k : Initial probability for the k^{th} state, $\pi_k \geq 0$; for all k=1,2,3; $\sum_{k=1}^{3} \pi_k = 1$; $\pi_k = \frac{n_k}{n}$; $n = \sum_{k=1}^{3} n_k$, *i.e.*, Total number of observations considered for the study in the specific business day

p_{jk}: The transition probability between states j and k represents the likelihood of moving from state j to state k in a given system or process.
i.e., $P\{Y_n = k/Y_{n-1} = j\} \geq 0$; $0 \leq p_{jk} \leq 1$ and $\sum_{k=1}^{3} p_{jk} = 1 \forall j = 1, 2, 3.$

j : Origin state

k : Destination state

y_t: Share price of the SBI on t^{th} day

Δy_t: $y_t - y_{t-1}$; The difference between the current day (t) share price and previous day (t-1) share price in SBI

dy_t: Derivative of the share price's return at time 't'; $dy_t = \frac{\Delta y_t}{y_{t-1}}$

R : *Rise State* occurs in SBI; $R = \left(dy_t \geq \mu + \frac{3\sigma}{\sqrt{n}}\right)$

S : *Stable State* occurs in SBI; $S = \left(\mu - \frac{3\sigma}{\sqrt{n}} < dy_t < \mu + \frac{3\sigma}{\sqrt{n}}\right)$

F : *Fall State* occurs in SBI; $F = \left(dy_t \leq \mu - \frac{3\sigma}{\sqrt{n}}\right)$

μ: Mean of dy_t

σ: standard deviation of dy_t

n : Total number of observations in the business day

m : Number of estimated values for testing the goodness of fit

O_i: Observed share value on i^{th} day; i=1,2, ..., n

E_i: Estimated share value on i^{th} day; i=1,2, ..., n

v : Number of parameters in the study of specific business day

$Y(\omega_1)$: Number of times *Rise State* occurs, $[Y(\omega_1) = y] = 0, 1$

$Y(\omega_2)$: Number of times *Stable State* occurs, $[Y(\omega_2) = y] = 0, 1$

$Y(\omega_3)$: Number of times *Fall State* occurs, $[Y(\omega_3) = y] = 0, 1$

2.4. Probability distribution and some statistical measures for *Rise State*

2.4.1. The probability distribution for *Rise State*

Let us consider a random variable denoted by $Y(\omega_1) = y$ which represents the happening of the *Rise State*. This variable can assume values 0 and 1, where '0' signifies its absence of the *Rise State* and '1' signifies its presence of the *Rise State*.

$$P[Y(\omega_1) = y] = \begin{cases} \sum_{k=1}^{3}\sum_{j=2}^{3} \pi_k p_{kj} & ;\text{for } y = 0 \\ \sum_{k=1}^{3} \pi_k p_{k1} & ;\text{for } y = 1 \\ 0 & ;\text{otherwise}(y \geq 2) \end{cases} \tag{1}$$

2.4.2. Statistical measures for *Rise State*

The Average Occurrence of *Rise State*

$$\mu_R = \sum_{k=1}^{3} \pi_k p_{k1} \tag{2}$$

The Variance of a *Rise State*

$$\sigma_R^2 = \mu_R^2\Big(\sum_{k=1}^{3}\sum_{j=2}^{3} \pi_k p_{kj}\Big) + (1-\mu_R)^2\Big(\sum_{k=1}^{3} \pi_k p_{k1}\Big) \tag{3}$$

The Third Central Moment for *Rise State*

$$\mu_{3R} = -\mu_R^3\Big(\sum_{k=1}^{3}\sum_{j=2}^{3} \pi_k p_{kj}\Big) + (1-\mu_R)^3\Big(\sum_{k=1}^{3} \pi_k p_{k1}\Big) \tag{4}$$

The Coefficient of skewness for *Rise State*

$$\begin{aligned} \beta_{1R} = &\Big[-\mu_R^3\Big(\sum_{k=1}^{3}\sum_{j=2}^{3} \pi_k p_{kj}\Big) + (1-\mu_R)^3\Big(\sum_{k=1}^{3} \pi_k p_{k1}\Big)\Big]^2 \times \\ &\Big[\mu_R^2\Big(\sum_{k=1}^{3}\sum_{j=2}^{3} \pi_k p_{kj}\Big) + (1-\mu_R)^2\Big(\sum_{k=1}^{3} \pi_k p_{k1}\Big)\Big]^{-3} \end{aligned} \tag{5}$$

Coefficient of Kurtosis for *Rise State*

$$\beta_{2R} = \Big[\mu_R^4\Big(\sum_{k=1}^{3}\sum_{j=2}^{3} \pi_k p_{kj}\Big) + (1-\mu_R)^4\Big(\sum_{k=1}^{3} \pi_k p_{k1}\Big)\Big]\Big[\mu_R^2\Big(\sum_{k=1}^{3}\sum_{j=2}^{3} \pi_k p_{kj}\Big) + (1-\mu_R)^2\Big(\sum_{k=1}^{3} \pi_k p_{k1}\Big)\Big]^{-2} \tag{6}$$

The Coefficient of Variation for *Rise State*

$$C.V_R = \Big[\mu_R^2\Big(\sum_{k=1}^{3}\sum_{j=2}^{3} \pi_k p_{kj}\Big) + (1-\mu_R)^2\Big(\sum_{k=1}^{3} \pi_k p_{k1}\Big)\Big]^{1/2}\Big[\sum_{k=1}^{3} \pi_k p_{k1}\Big]^{-1}\% \tag{7}$$

2.4.3. Moment generating function for *Rise State*

$$M_{YR}(t) = \Big(\sum_{k=1}^{3}\sum_{j=2}^{3}\pi_k p_{kj}\Big) + e^{t}\Big(\sum_{k=1}^{3}\pi_k p_{k1}\Big) \tag{8}$$

2.4.4. Characteristic function for *Rise State*

$$\phi_{YR}(t) = \Big(\sum_{k=1}^{3}\sum_{j=2}^{3}\pi_k p_{kj}\Big) + e^{it}\Big(\sum_{k=1}^{3}\pi_k p_{k1}\Big) \tag{9}$$

2.4.5. Probability generating function for *Rise State*

$$P_{SR}(t) = \Big(\sum_{k=1}^{3}\sum_{j=2}^{3}\pi_k p_{kj}\Big) + S\Big(\sum_{k=1}^{3}\pi_k p_{k1}\Big) \tag{10}$$

2.5. Probability distribution and some statistical measures for *Stable State*

2.5.1. The probability distribution for *Stable State*

Let us consider a random variable denoted by $Y(\omega_2) = y$ which represents the happening of the *Stable State*. This variable can assume values 0 and 1, where '0' signifies its absence of the *Stable State* and '1' signifies its presence of the *Stable State*.

$$P[Y(\omega_2) = y] = \begin{cases} \sum\limits_{k=1}^{3}\sum\limits_{j=1,j\neq 2}^{3}\pi_k p_{kj} & ;\text{for } y = 0 \\ \sum\limits_{k=1}^{3}\pi_k p_{k2} & ;\text{for } y = 1 \\ 0 & ;\text{otherwise}(y \geq 2) \end{cases} \tag{11}$$

2.5.2. Statistical measures for *Stable State*

The Average Occurrence of *Stable State*

$$\mu_S = \sum_{k=1}^{3}\pi_k p_{k2} \tag{12}$$

The Variance of a *Stable State*

$$\sigma_S^2 = \mu_S^2\Big(\sum_{k=1}^{3}\sum_{j=1,j\neq 2}^{3}\pi_k p_{kj}\Big) + (1-\mu_S)^2\Big(\sum_{k=1}^{3}\pi_k p_{k2}\Big) \tag{13}$$

The Third Central Moment for *Stable State*

$$\mu_{S3} = -\mu_S^3\Big(\sum_{k=1}^{3}\sum_{j=1,j\neq 2}^{3}\pi_k p_{kj}\Big) + (1-\mu_S)^3\Big(\sum_{k=1}^{3}\pi_k p_{k2}\Big) \tag{14}$$

The Coefficient of Skewness for *Stable State*

$$\beta_{1S} = \Big[-\mu_S^3 \Big(\sum_{k=1}^{3} \sum_{j=1, j\neq 2}^{3} \pi_k p_{kj} \Big) + (1-\mu_S)^3 \Big(\sum_{k=1}^{3} \pi_k p_{k2} \Big) \Big]^2 \Big[\mu_S^2 \Big(\sum_{k=1}^{3} \sum_{j=1, j\neq 2}^{3} \pi_k p_{kj} \Big) (1-\mu_S)^2 \Big(\sum_{k=1}^{3} \pi_k p_{k2} \Big) \Big]^{-3} \tag{15}$$

Coefficient of Kurtosis for *Stable State*

$$\beta_{2S} = \Big[\mu_S^4 \Big(\sum_{k=1}^{3} \sum_{j=1, j\neq 2}^{3} \pi_k p_{kj} \Big) + (1-\mu_S)^4 \Big(\sum_{k=1}^{3} \pi_k p_{k2} \Big) \Big] \Big[\mu_S^2 \Big(\sum_{k=1}^{3} \sum_{j=1, j\neq 2}^{3} \pi_k p_{kj} \Big) (1-\mu_S)^2 \Big(\sum_{k=1}^{3} \pi_k p_{k2} \Big) \Big]^{-2} \tag{16}$$

Coefficient of variation for *Stable State*

$$C.V_S = \Big[\mu_S^2 \Big(\sum_{k=1}^{3} \sum_{j=1, j\neq 2}^{3} \pi_k p_{kj} \Big) + (1-\mu_S)^2 \Big(\sum_{k=1}^{3} \pi_k p_{k2} \Big) \Big]^{1/2} \Big(\sum_{k=1}^{3} \pi_k p_{k2} \Big)^{-1} \% \tag{17}$$

2.5.3. Moment generating function for *Stable State*

$$M_{YS}(t) = \Big(\sum_{k=1}^{3} \sum_{j=1, j\neq 2}^{3} \pi_k p_{kj} \Big) + e^t \Big(\sum_{k=1}^{3} \pi_k p_{k2} \Big) \tag{18}$$

2.5.4. Characteristic function for *Stable State*

$$\phi_{YS}(t) = \Big(\sum_{k=1}^{3} \sum_{j=1, j\neq 2}^{3} \pi_k p_{kj} \Big) + e^{it} \Big(\sum_{k=1}^{3} \pi_k p_{k2} \Big) \tag{19}$$

2.5.5. Probability generating function for *Stable State*

$$P_{SS}(t) = \Big(\sum_{k=1}^{3} \sum_{j=1, j\neq 2}^{3} \pi_k p_{kj} \Big) + S \Big(\sum_{k=1}^{3} \pi_k p_{k2} \Big) \tag{20}$$

2.6. Probability distribution and some statistical measures for *Fall State*

2.6.1. The Probability distribution for *Fall State*

Let us consider a random variable denoted by $Y(\omega_3) = y$ which represents the happening of the *Fall State*. This variable can assume values 0 and 1, where '0' signifies its

absence of the *Fall State* and '1' signifies its presence of the *Fall State*.

$$P[Y(\omega_3)=y]=\begin{cases}\sum\limits_{k=1}^{3}\sum\limits_{j=1}^{2}\pi_k p_{kj} & ;\text{for } y=0\\ \sum\limits_{k=1}^{3}\pi_k p_{k3} & ;\text{for } y=1\\ 0 & ;\text{otherwise}(y\geq 2)\end{cases} \tag{21}$$

2.6.2. Statistical measures for *Fall State*

The Average Occurrence of *Fall State*

$$\mu_F=\sum_{k=1}^{3}\pi_k p_{k3} \tag{22}$$

The Variance of a *Fall State*

$$\sigma_F^2=\mu_F^2\Big(\sum_{k=1}^{3}\sum_{j=1}^{2}\pi_k p_{kj}\Big)+(1-\mu_F)^2\Big(\sum_{k=1}^{3}\pi_k p_{k3}\Big) \tag{23}$$

The Third Central Moment for *Fall State*

$$\mu_F^3=-\mu_F^3\Big(\sum_{k=1}^{3}\sum_{j=1}^{2}\pi_k p_{kj}\Big)+(1-\mu_F)^3\Big(\sum_{k=1}^{3}\pi_k p_{k3}\Big) \tag{24}$$

The Coefficient of Skewness for *Fall State*

$$\begin{aligned}\beta_{1F}=&\Big[-\mu_F^3\Big(\sum_{k=1}^{3}\sum_{j=1}^{2}\pi_k p_{kj}\Big)+(1-\mu_F)^3\Big(\sum_{k=1}^{3}\pi_k p_{k3}\Big)\Big]^2\times\\ &\Big[\mu_F^2\Big(\sum_{k=1}^{3}\sum_{j=1}^{2}\pi_k p_{kj}\Big)+(1-\mu_F)^2\Big(\sum_{k=1}^{3}\pi_k p_{k3}\Big)\Big]^{-3}\end{aligned} \tag{25}$$

Coefficient of Kurtosis for *Fall State*

$$\beta_{2F}=\Big[\mu_F^4\Big(\sum_{k=1}^{3}\sum_{j=1}^{2}\pi_k p_{kj}\Big)+(1-\mu_F)^4\Big(\sum_{k=1}^{3}\pi_k p_{k3}\Big)\Big]\Big[\mu_F^2\Big(\sum_{k=1}^{3}\sum_{j=1}^{2}\pi_k p_{kj}\Big)+(1-\mu_F)^2\Big(\sum_{k=1}^{3}\pi_k p_{k3}\Big)\Big]^{-2} \tag{26}$$

The Coefficient of Variation for *Fall State*

$$C.V_F=\Big[\mu_F^2\Big(\sum_{k=1}^{3}\sum_{j=1}^{2}\pi_k p_{kj}\Big)+(1-\mu_F)^2\Big(\sum_{k=1}^{3}\pi_k p_{k3}\Big)\Big]^{1/2}\Big(\sum_{k=1}^{3}\pi_k p_{k3}\Big)^{-1}\% \tag{27}$$

2.6.3. Moment generating function for *Fall State*

$$M_{YF}(t)=\Big(\sum_{k=1}^{3}\sum_{j=1}^{2}\pi_k p_{kj}\Big)+e^t\Big(\sum_{k=1}^{3}\pi_k p_{k3}\Big) \tag{28}$$

2.6.4. Characteristic function for *Fall State*

$$\phi_{YF}(t) = \Big(\sum_{k=1}^{3}\sum_{j=1}^{2}\pi_k p_{kj}\Big) + e^{it}\Big(\sum_{k=1}^{3}\pi_k p_{k3}\Big) \tag{29}$$

2.6.5. Probability generating function for *Fall State*

$$P_{SF}(t) = \Big(\sum_{k=1}^{3}\sum_{j=1}^{2}\pi_k p_{kj}\Big) + S\Big(\sum_{k=1}^{3}\pi_k p_{k3}\Big) \tag{30}$$

2.7. Predictions of returns on income

2.7.1. Expected returns on SBI shares

The explicit mathematical relation for computing expected share returns

$$[E.S.R]_{3\times 1} = [p_{jk}]^n_{3\times 3}[M.S]_{3\times 1}; \forall\, n = 1, 2, ... \tag{31}$$

E.S.R =Expected share price returns
P^n = Limiting Probability Matrix (Computed using TPM)
M.S =Mean state

2.7.2. Prediction of closing prices of SBIs shares

The explicit mathematical relation for predicted Closing prices of SBI shares

$$P.S.P = (Y_{Rt} \times \mu_R) + (Y_{St} \times \mu_S) + (Y_{Ft} \times \mu_F) \tag{32}$$

where,

Y_{Rt} = Expected closing price of the SBIs share on the current day for the *Rise State*
Y_{St}= Expected closing price of the SBIs share on the current day for the *Stable State*
Y_{Ft}= Expected closing price of the SBIs share on the current day for the *Fall State*
μ_R= Average chance for occurrence of the *Rise State*
μ_S = Average chance for occurrence of the *Stable State*
μ_F = Average occurrence for occurrence of the *Fall State*

2.8. Validation of the model

2.8.1. Testing for model's goodness of fit

The Chi-Square test statistic, denoted as χ^2 , is utilized to assess the goodness of fit between observed and expected categorical data. In the context of comparing observed (original) and expected (predicted) share prices, the formula for χ^2 is:

$$\chi^2 = \sum_{i=1}^{m}\frac{[O_i - E_i]^2}{E_i} \sim \chi^2_{m-1} \tag{33}$$

2.8.2. Computation of AIC and BIC

The formulas for calculating AIC (Akaike Information Criterion) and BIC (Bayesian Information Criterion) are as follows:

$$AIC = -2\, loglikelihood + 2v \tag{34}$$

$$BIC = -2\, loglikelihood + v\ 2logn \tag{35}$$

3. Data description for the developed model

Figure 2 provides a clear depiction of the data and methodology employed in the current study. It delineates the detailed procedures utilized to assess the results with precision and thoroughness.

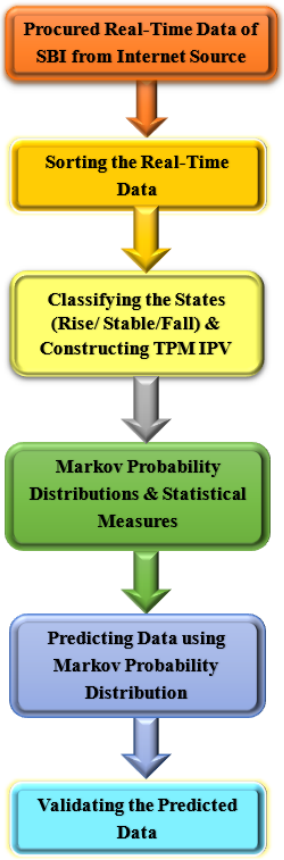


Figure 2: Three-State Markov model data description & methodology flow chart

3.1. Data source and organization of the data

The detailed description of Figure 2; in order to utilise the Markov model that was developed, real-time data on the closing prices of SBI (State Bank of India) stocks was considered. This real-time data, crucial for evaluating the model's effectiveness, consisted of 251 observations collected over a period spanning from 2^{nd} May 2022 to 5^{th} May 2023.

These observations were meticulously sourced from the renowned financial platform, Yahoo Finance, accessible via the internet link (https://in.finance.yahoo.com).

The dataset, which served as the foundation for this analysis, was specifically focused on the closing prices of SBI stocks. Closing prices, in the context of stock market analysis, represent the final prices at which a stock trades during a regular trading session. These prices are often used to assess the overall performance of a particular stock.

This dataset, constituting 251 data points, is of paramount importance for evaluating the Markov model's predictive capabilities in real-world scenarios. It forms the basis upon which the model's predictions and effectiveness in forecasting SBI stock prices are tested and validated. The historical closing prices, meticulously organized and structured, were compiled into a comprehensive sample data template, as detailed in Table 1. This template serves as the primary reference for the subsequent analysis and assessment of the Markov model's accuracy and reliability in predicting the closing prices of SBI stocks during the specified period.

Table 1: SBI's sample data matrix

S.No.	Date	Closing Price
1	02-05-2022	491
2	04-05-2022	479.649994
3	05-05-2022	480
⋮	⋮	⋮
249	03-05-2023	570.5
250	04-05-2023	580
251	05-05-2023	576.5

3.2. Data formulation

In light of the observed influence of market seasonality on closing prices concerning specific weekdays, the 251 collected observations were categorized based on business days (Monday, Tuesday, Wednesday, Thursday, and Friday). The Sample data matrix of Mondays data placed in Table 2. Remaining business days also done like Mondays data.

Table 2: SBI's Monday data matrix

S.No.	Date	Closing Price	Returns	State	Transition
1	02-05-2022	491	-	-	-
2	09-05-2022	475.899994	-0.03075	F	-
3	16-05-2022	455	-0.04392	F	FF
⋮	⋮	⋮	⋮	⋮	⋮
49	10-04-2023	526.299988	-0.00085	S RS	
50	17-04-2023	544	0.033631	R	SR
51	24-04-2023	554.599976	0.019485	R	RR

To delve deeper into this segmentation and its impact, individual sensitivity studies were undertaken for each business day. Prior to these studies, the data from all five datasets

were pooled together, facilitating comprehensive analysis. Within each dataset, a meticulous classification was performed, focusing on the transient state of returns. This systematic approach allowed for a detailed exploration of how market dynamics and price fluctuations varied across different weekdays, shedding light on the intricate relationship between market behaviour and specific business days.

3.3. Data disclosure

The states are determined according to the values of dY_t and are categorized into three distinct types: *Rise*(State-1) when the condition $dY_t \geq \mu + \frac{3\sigma}{\sqrt{n}}$ is met, *Stable* (State-2) when the condition $\mu - \frac{3\sigma}{\sqrt{n}} < dY_t < \mu + \frac{3\sigma}{\sqrt{n}}$ is satisfied, and *Fall* (State-3) when the condition $dY_t \leq \mu - \frac{3\sigma}{\sqrt{n}}$ holds true. In these definitions, μ represents the mean, σ represents the standard deviation of dY_t, and n signifies the number of observations within the segregated dataset.

Classification of states for Monday, Tuesday, Wednesday, Thursday, and Friday are placed in the Figures 3, 4, 5, 6, and 7 respectively.

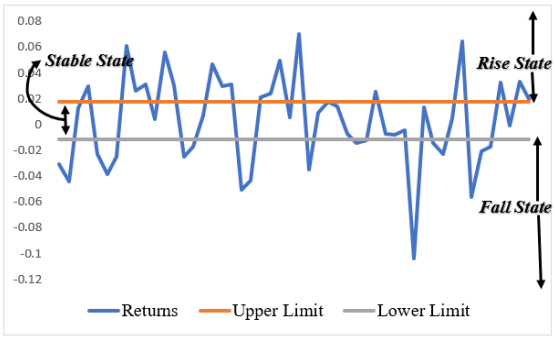


Figure 3: Classification of states in Monday data

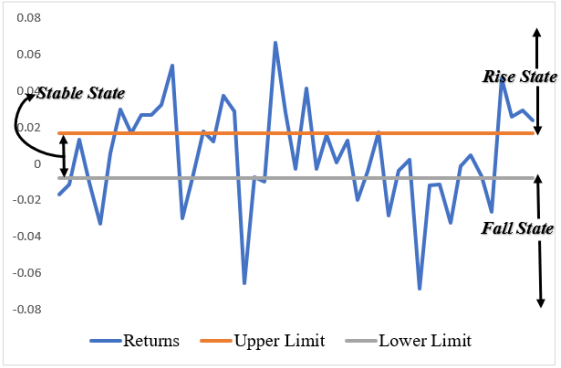


Figure 4: Classification of states in Tuesday data

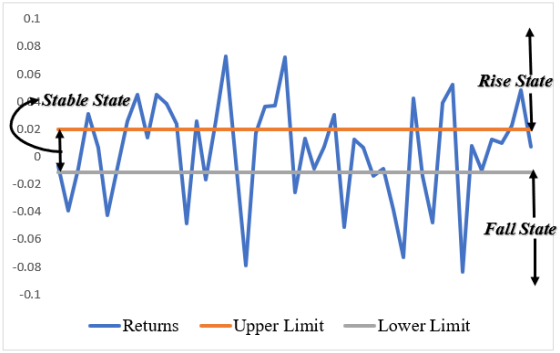


Figure 5: Classification of states in Wednessday data

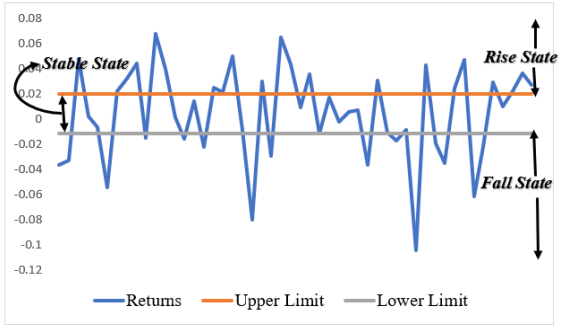


Figure 6: Classification of states in Thursday data

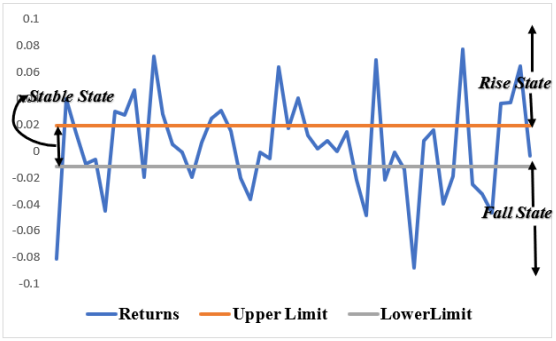


Figure 7: Classification of states in Friday data

Markov model is a composition of TPM, and IPV, which are computed with real-time data through R programming. Separate probability distributions, and statistical characteristics like average, variance, third central moments, skewness, kurtosis *etc.* are obtained for all segregated data sets. However, we have considered the averages for computing the predicted closing prices. The expected returns for SBI of all data sets are calculated by a formula as in section 2.7.1. We have obtained the predicted values (about 10 observations) of expected returns using the notion of sections 2.7.1 and 2.7.2. The developed Markov model is validated with the Chi-Square test for all data sets individually. AIC and BIC are also computed for each data set separately for the model's goodness of fit.

4. Results and discussion

The below parameters are placed in sections 4.1 and 4.2 which are computed by the above methodology.

4.1. Transition probability matrix (TPM) for SBI share closing prices

The explored TPM for Monday, Tuesday, Wednesday, Thursday, and Friday sets are as follows.

4.1.1. Transition behaviour of the market from monday to friday

The explored TPMs from Monday to Friday data placed in below Table 3.

Table 3: Transition Probabilities for all Business Days in a Week

Day	Transition Probabilities								
	RR	RS	RF	SR	SS	SF	FR	FS	FF
Monday	0.4706	0.2353	0.2941	0.4286	0.3571	0.2143	0.2222	0.2778	0.5
Tuesday	0.625	0.1875	0.1875	0.3125	0.375	0.3125	0.1429	0.5	0.3571
Wednesday	0.4445	0.2222	0.3333	0.3529	0.3529	0.2941	0.3077	0.5385	0.1538
Thursday	0.45	0.3	0.25	0.1538	0.3077	0.5385	0.625	0.1875	0.1875
Friday	0.4	0.4	0.2	0.1667	0.5	0.3333	0.375	0.25	0.375

The graphical representation of the above table is placed in Figure 8. It gives a clear interpretation of the transition behaviour of all business days in a week.

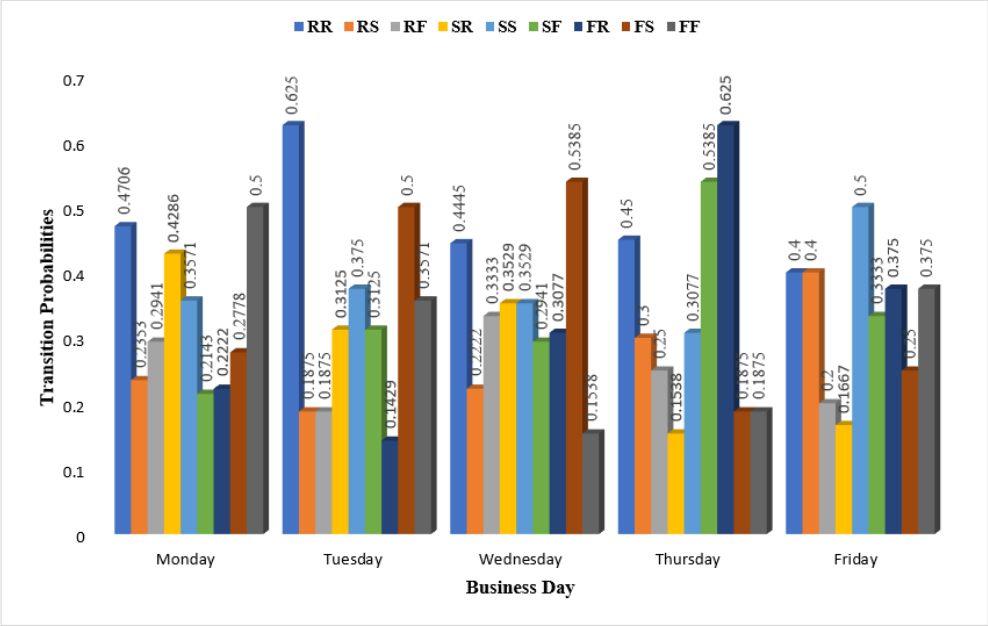


Figure 8: Transition probabilities for all business days in a week

From the above Table 3 and Figure 8 transition probabilities on Monday data set, it is observed that *Fall State* in the current day given that *Fall State* in the previous day is having

the highest likelihood (50%); similarly *Rise State* in the current day given that *Rise State* in the previous day is having second highest likelihood (47.06%); *Rise State* in the current day given that *Stable State* in the previous day is having third highest likelihood (42.86%); and *Fall State* in the current day given that *Stable State* in the previous day having least likelihood (21.43%).

Based on the transition probabilities gleaned from Table 3 and Figure 8 of the Tuesday data set, it is observed that *Rise State* in the current day given that *Rise State* in the previous day is having the highest likelihood (62.5%); similarly *Stable State* in the current day given that *Fall State* in the previous day is having second highest likelihood (50%); *Stable State* in the current day given that *Stable State* in the previous day is having third highest likelihood (37.5%); and *Rise State* in the current day given that *Fall State* in the previous day having least likelihood (14.29%).

In analysing the transition probabilities extracted from Table 3 and Figure 8 of the Wednesday data set, it is observed that *Stable State* on the current day given that *Fall State* in the previous day is having the highest likelihood (53.85%); similarly *Rise State* in the current day given that *Rise State* in the previous day is having second highest likelihood (44.45%); *Stable State* in the current day given that *Stable State* in the previous day and *Rise State* in the current day and *Stable State* in the previous day both are having third highest likelihood (35.29%); and *Fall State* in the current day given that *Fall State* in the previous day having least likelihood (15.38%).

Examining the transition probabilities sourced from Table 3 and Figure 8 of the Thursday data set, it is observed that *Rise State* in the current day given that *Fall State* in the previous day is having the highest likelihood (62.5%); similarly, *Fall State* in the current day given that *Stable State* in the previous day is having the second highest likelihood (53.85%); *Rise State* in the current day given that *Rise State* in the previous day is having third highest likelihood (45%); and *Rise State* in the current day given that *Stable State* in the previous day having least likelihood (15.38%).

From the above Table 3 and Figure 8 transition probabilities of Friday data set, it is observed that *Stable State* in the current day given that *Stable State* in the previous day is having the highest likelihood (50%); similarly *Rise State* in the current day given that *Rise State* previous day and *Stable State* in the current day and *Rise State* in the previous day are having second highest likelihood (40%); *Fall State* in the current day given that *Fall State* in the previous day and *Rise State* in the current day and *Fall State* in the previous day are having third highest likelihood (37.5%); and *Fall State* in the current day given that *Rise State* in the previous day having least likelihood (20%).

These findings highlight distinct patterns in SBI's share prices throughout the week, indicating varying transient behaviours. This information can be invaluable for portfolio managers, enabling them to assess how SBI's share prices transition between *Rise*, *Stable*, and *Fall* states each day. These indicators provide crucial insights, allowing managers to strategize effectively, capitalize on profit opportunities, and implement corrective measures to mitigate losses.

4.2. Initial probability vector (IPV) for SBI share prices

After a thorough process of the real-time data, the IPVs of the *Rise*, *Stable*, and *Fall* states are obtained.

4.2.1. Indicators of *Rise*, *Stable*, and *Fall* States on Monday to Friday

The indicators on the chances of *Rise*, *Stable*, and *Fall* states the data under study are as below.

Table 4: Initial Probabilities for all Business Days in a Week

Day	Initial Probabilities		
	Rise	*Stable*	*Fall*
Monday	0.36	0.28	0.36
Tuesday	0.3617	0.3404	0.2979
Wednesday	0.3673	0.3673	0.2654
Thursday	0.42	0.26	0.32
Friday	0.30	0.38	0.32

The graphical representation of the above Table 4 is placed in Figure 9.

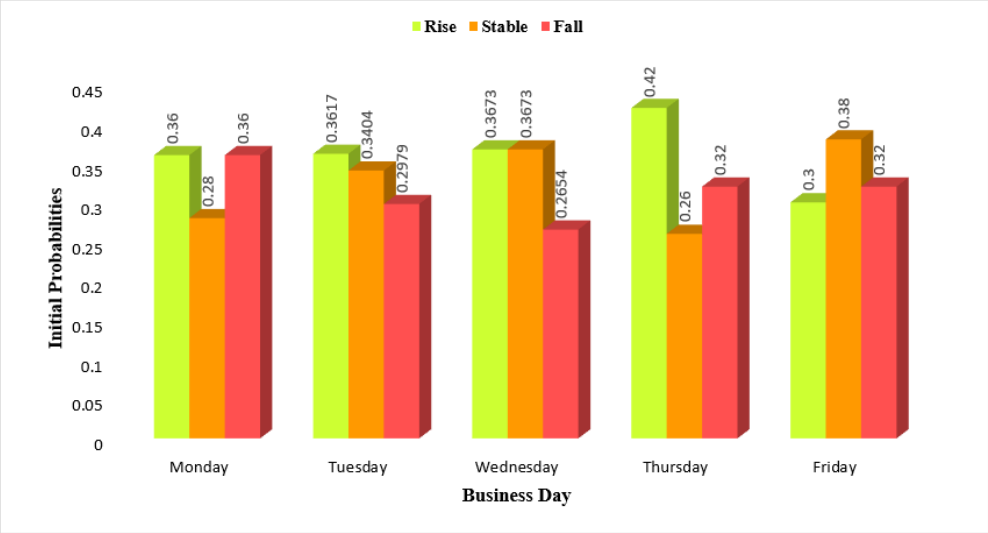


Figure 9: Initial probabilities for all business days in a week

From the above Table 4 and Figure 9, it is observed that on Monday, the likelihood of both *Rise State* and *Fall State* is equal at 36%. From Tuesday to Thursday, *Rise State* consistently has a higher likelihood than the other states, with Thursday having the highest probability at 42%, followed by Wednesday at 36.73%. Conversely, on Friday, the likelihood of the *Rise State* drops to its lowest at 30% compared to the other states. This suggests a strategy for short-term traders: consider selling shares during the middle of the week when the probability of a price increase is notably higher.

4.3. Probability distributions for *Rise*, *Stable*, and *Fall* states

The probability distributions for *Rise*, *Stable*, and *Fall* states of all business days in a week (Monday, Tuesday, Wednesday, Thursday, and Friday) are as in Table 5.

Table 5: Probability distributions of all states

Day	Chance of happening of the state		
	Rise	*Stable*	*Fall*
Monday	0.3694	0.2847	0.3459
Tuesday	0.375	0.3444	0.2806
Wednesday	0.3745	0.3541	0.2713
Thursday	0.429	0.266	0.305
Friday	0.3033	0.39	0.3067

Figure 10 illustrates the occurrence of *Rise*, *Stable*, and *Fall* states graphically. It provides a visual representation of the frequency of each state over the observed period.

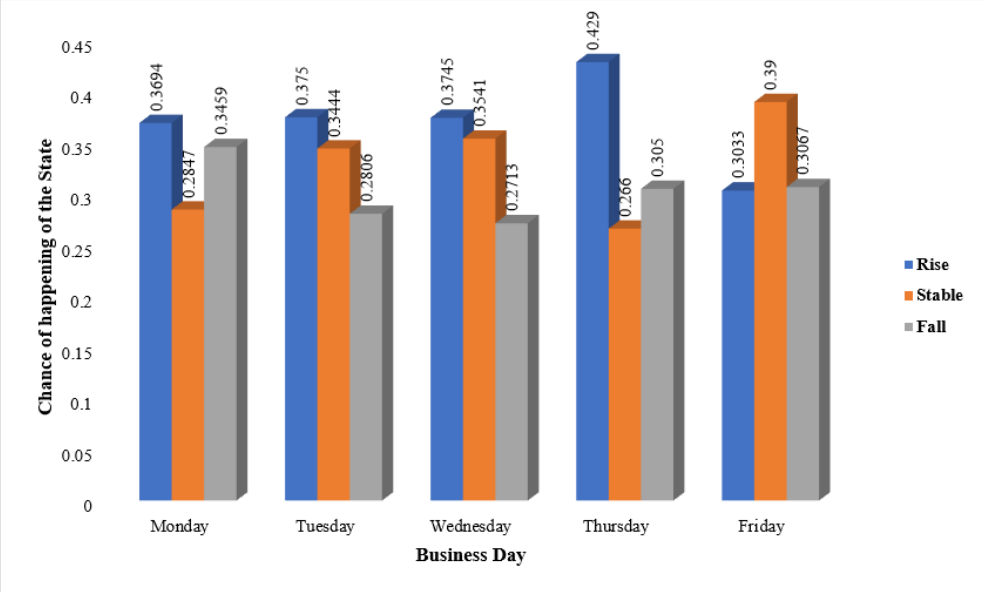


Figure 10: Chance of happening of *Rise*, *Stable*, and *Fall* states

According to the data presented in Table 5 and Figure 10, there is a noticeable trend in the occurrence of *Rise* and *Fall* states across different days of the week. Specifically, the likelihood of the *Rise State* is highest on Thursdays, closely followed by Tuesdays. Similarly, Fridays exhibit the highest probability of the *Rise State*, with Wednesdays following closely behind. In contrast, the *Fall State* is more likely to occur on Mondays, with Fridays showing the next highest probability.

This information suggests certain patterns or tendencies in market behaviour throughout the week. Traders may find it useful to be aware of these tendencies when making decisions about trading strategies, timing of trades, and risk management. For instance, understanding that Thursdays often have a higher chance of experiencing the *Rise State* could

influence traders to adjust their positions accordingly or anticipate potential market movements. Similarly, knowledge of increased *Fall State* occurrences on Mondays might prompt traders to exercise caution or implement specific risk mitigation measures at the beginning of the trading week. Overall, awareness of these patterns can help traders make more informed decisions and navigate market dynamics more effectively.

Statistical measures/characteristics are useful in understanding the behaviour of the probability distributions.

4.4. Discussion on statistical measures

In order to have a better understanding of the model behaviour and the probability distributions, the statistical measures are computed and placed in Table 6.

Table 6: Statistical measures for *Rise*, *Stable* and *Fall* states

State	Statistical Measure	Monday	Tuesday	Wednesday	Thursday	Friday
Rise State	Average	0.3694	0.375	0.3745	0.429	0.3033
	Variance	0.2329	0.2344	0.2343	0.245	0.2113
	3rd central moment	0.0608	0.0586	0.0588	0.0348	0.0831
	Beta -1	0.2928	0.2667	0.2687	0.0823	0.7321
	Beta -2	1.2928	1.2667	1.2687	1.0823	1.7321
	C.V.	130.652	129.099	129.223	115.369	150.953
Stable State	Average	0.2847	0.3444	0.3541	0.266	0.39
	Variance	0.2036	0.2258	0.2287	0.1952	0.2379
	3rd central moment	0.0877	0.0703	0.0667	0.0914	0.0523
	Beta -1	0.9104	0.4288	0.3721	1.1218	0.2034
	Beta -2	1.9104	1.4288	1.3721	2.1218	1.2035
	C.V.	158.505	137.966	135.045	166.114	125.064
Fall State	Average	0.3459	0.2806	0.2713	0.305	0.3067
	Variance	0.2262	0.2019	0.1977	0.212	0.2126
	3rd central moment	0.0697	0.0886	0.0904	0.0827	0.0822
	Beta -1	0.4199	0.954	1.0582	0.7175	0.7032
	Beta -2	1.4199	1.954	2.0582	1.7175	1.7032
	C.V.	137.519	160.124	163.885	150.953	150.361

4.4.1. Discussion on the results

The results presented in Table 6 indicates that the *Rise State* is more frequently observed from Monday to Thursday compared to the *Stable* and *Fall* states. Specifically, there is a higher probability of the *Rise State* occurring during these days. Furthermore, Thursday stands out as the day with the highest likelihood for the *Rise State* in comparison to other business days.

Conversely, the *Stable State* exhibits a higher probability of occurrence on Fridays, suggesting a distinct pattern at the end of the week. This observation implies that different states (*Rise*, *Stable*, and *Fall*) exhibit varying likelihoods on different days, providing valuable

insights into the underlying patterns of the data. The below Figure 11 shows the graphical representation of this content.

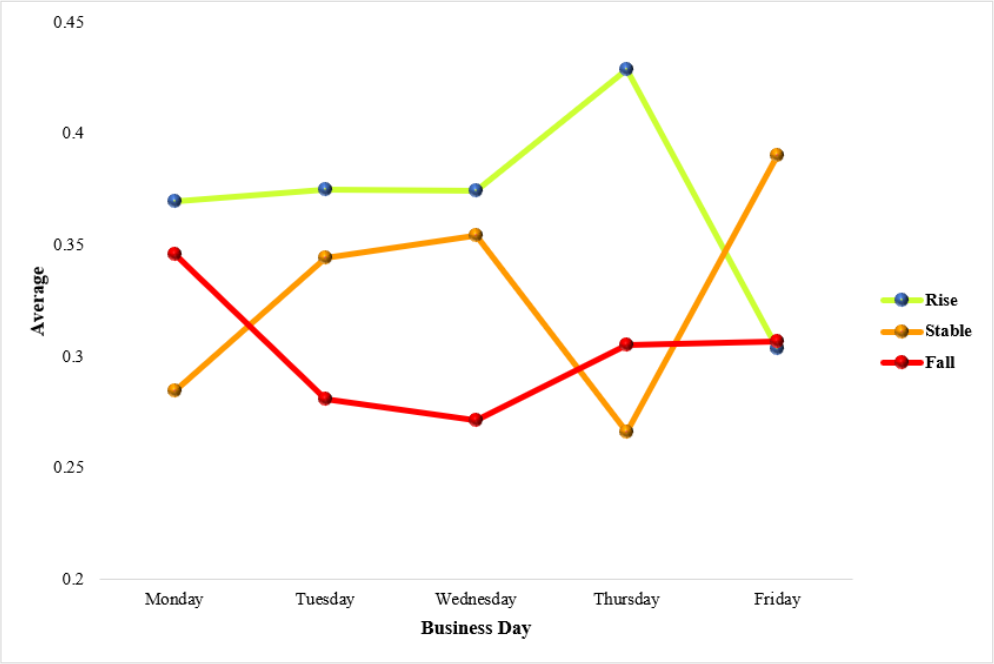


Figure 11: Averages of *Rise*, *Stable*, and *Fall* states in different days in a week

The analysis of the provided Table 6 reveals interesting patterns in the variability (variance) of different states (*Rise*, *Stable*, and *Fall*) across the weekdays. In the *Rise State*, similar variances are observed from Monday to Thursday, indicating consistent behaviour during these days. However, on Friday, there is a notable decrease in variance, suggesting a more stable trend compared to the preceding days.

In the *Stable State*, the highest variance is observed on Friday, signifying fluctuations, and unpredictability in the stock market towards the end of the week. Conversely, Thursday stands out with the least variance in this state, indicating a more stable and predictable market behaviour on that day.

For the *Fall State*, high variance is noted on Monday, suggesting significant fluctuations at the beginning of the week. In contrast, Wednesday exhibits the least variance in this state, indicating a relatively calmer and more predictable market environment.

Interestingly, the data emphasizes that Thursday is characterized by the least variance across all states (*Rise*, *Stable*, and *Fall*). This suggests that Thursdays tend to have a more stable market behaviour, making them potentially favourable for certain investment strategies.

These observations provide valuable insights for investors, indicating specific days of the week when the stock market is either more stable or prone to fluctuations. Investors could potentially use this information to inform their trading decisions, adapting their strategies based on the observed patterns of variance in different market states across weekdays.

The positive skewness indicated by the non-negative third central moment across all states (*Rise*, *Stable*, and *Fall*) implies that in the stock market, there are more frequent occurrences of small or moderate gains. These modest gains are a common feature, suggesting relative stability in stock prices. However, the presence of occasional significant upward shifts in stock prices, although infrequent, contributes to the overall positive skewness.

For investors, this pattern highlights the regularity of stable or moderately positive market movements, punctuated by occasional notable upticks. Recognizing these infrequent but substantial positive shifts is vital for investors seeking opportunities for significant profits. However, it also underscores the need for prudent risk management, as these occasional large movements can result in substantial losses if not carefully navigated. Understanding this skewed distribution is essential for making informed investment decisions in the stock market.

The kurtosis values being less than three for all states (*Rise*, *Stable*, and *Fall*) on every business day indicate a platykurtic distribution in the stock market.

The observation of the lowest coefficient of variation in the *Rise State* on Thursday (115.369) implies that this particular day showcases a remarkable consistency and stability in stock market performance, graphically it is presented in Figure 12.

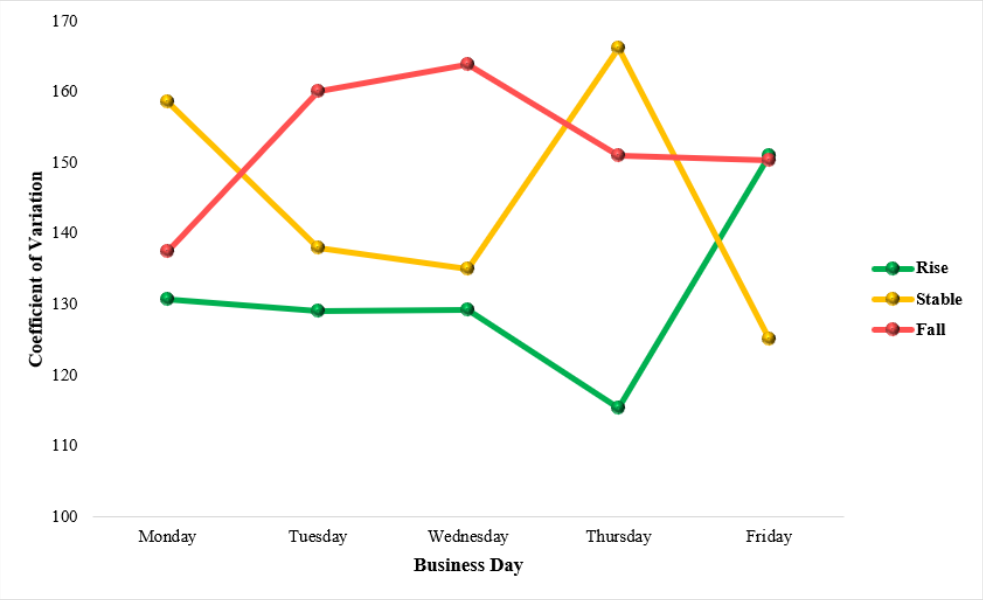


Figure 12: Coefficient of variation for *Rise*, *Stable*, and *Fall* states in different days in a week

Hence, Figure 12 may advise to short-term traders that Thursday might be an opportune day to consider selling stocks to maximise returns. The lower coefficient of variation indicates reduced volatility and fluctuations, indicating a more predictable market environment. This stability can provide short-term traders with confidence in making strategic decisions, potentially leading to optimal returns on their investments. Understanding these patterns in the coefficient of variation aids traders in identifying favourable moments for executing trades and capitalizing on market stability.

4.5. Expected (predicted) returns for SBI's shares

The expected returns computed for all days in a week are separately computed using the formula mentioned in Section 2.7.1.

4.5.1. Expected returns for SBIs all business days data of all states

The given below are the expected SBI share price returns in 10 business days due to *Rise*, *Stable*, and *Fall* states.

Table 7: Expected returns for *Rise*, *Stable*, and *Fall* states

State	Day	Monday	Tuesday	Wednesday	Thursday	Friday
Rise	t=1	0.009207387	0.015662265	0.003799441	0.008274521	0.013172060
	t=2	0.004862423	0.009527704	0.004649036	0.003889552	0.005524245
Stable	t=1	0.010790754	0.002533674	0.002325153	0.013513428	0.001853549
	t=2	0.006335824	0.004617461	0.004317823	0.006204354	0.002929969
Fall	t=1	-0.006832183	-0.003926799	0.007331084	0.016880182	0.004984202
	t=2	0.001627426	0.002101875	0.003548923	0.005802842	0.006345211

Analysing the resulted Table 7, it is observed there is expected returns of *Fall State* on Monday and Tuesday are negative, it may indicate to the traders there is a risk factor involved in share market, so these results may advise to the short-term traders to adopt risk tolerance and portfolio strategy to overcome the loss on investment.

4.5.2. Estimated (Predicted) closing prices of SBI shares

The SBI's closing prices are predicted using the linearity formula which is placed in Section 2.7.2 The predicted share prices are

Table 8: Predicted closing prices

Week	Monday	Tuesday	Wednesday	Thursday	Friday
First	556.8796	578.5471	572.9164	582.9601	579.2679
Second	559.2062	570.899	575.3413	585.924	582.0315

Figure 13 depicts the observed and predicted closing prices of SBI shares. These forecasts are valuable for short-term traders, enabling them to discern patterns in SBI share prices and make informed decisions for trading in the upcoming week.

4.6. Validation of the model

4.6.1. Chi-square test

The Markov model developed was assessed using the Chi-Square test for goodness of fit, considering both expected and observed values over two weeks (business days only). The test's null hypothesis (H_0) posits that the developed model fits the data well, meaning there is no significant difference between the observed and expected closing prices of SBI shares.

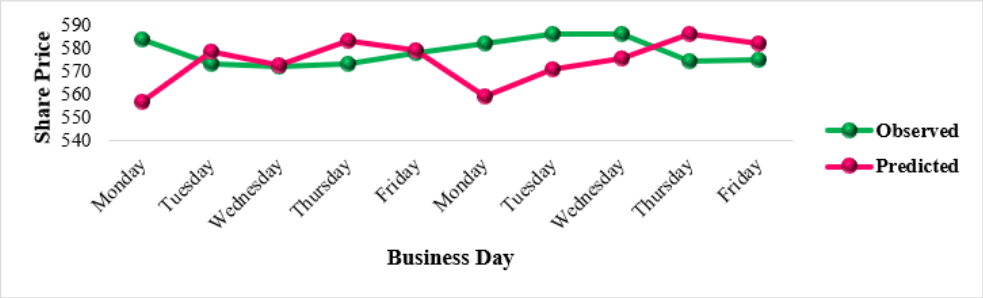


Figure 13: Predicted Closing Prices of First and Second Week

The calculated probability value (p-value) for SBI is 0.9719 with 9 degrees of freedom. The result indicates that the stated hypothesis is not rejected, confirming that the developed model aligns with the data.

4.6.2. AIC and BIC

Additionally, the model's robustness was evaluated using the Akaike information criterion (AIC) and Bayesian information criterion (BIC).

The AIC and BIC were calculated using the formulae mentioned in the Section 2.8.2. For Monday, Tuesday, Wednesday, Thursday, and Friday, the AIC values are 70.80734, 57.5237, 63.0746, 69.68, and 64.1920, while the corresponding BIC values are 91.6174, 75.8102, 81.7867, 90.49, and 83.1102. These findings indicate that the AIC and BIC values are lowest for Tuesday data, followed by Wednesday's data. Consequently, the results obtained from the developed model affirm that the upward trend in share value is notably more consistent during the middle of the week.

Acknowledgements

We are thankful to the DST INSPIRE. Ms. Sarode Rekha is supported by INSPIRE fellowship (IF190741) of the Department of Science and Technology (DST), Government of India.

Conflict of interest

The authors do not have any financial or non-financial conflict of interest to declare for the research work included in this article.

References

Abounoori, E., Elmi, Z. M., and Nademi, Y. (2016). Forecasting Tehran stock exchange volatility; markov switching garch approach. *Physica A: Statistical Mechanics and its Applications*, **445**, 264–282.

Adesokan, I., Ngare, P., and Kilishi, A. (2017). Analyzing expected returns of a stock using the markov chain model and the capital asset pricing model. *Applied Mathematical Sciences*, **11**, 2777–2788.

Bhusal, M. K. (2017). Application of markov chain model in the stock market trend analysis of Nepal. *International Journal of Scientific & Engineering Research*, **8**, 1733–1745.

Cai, J. (1994). A markov model of switching-regime arch. *Journal of Business & Economic Statistics*, **12**, 309–316.

Choji, D. N., Eduno, S. N., and Kassem, G. T. (2013). Markov chain model application on share price movement in stock market. *Computer Engineering and Intelligent Systems*, **4**, 84–95.

Dar, G. F., Ahn, Y.-H., Dar, Q. F., and Ma, J.-H. (2023). Impact of international trade on the share prices of the industrial bank of Korea using stochastic prediction modeling. *Journal of Global Business and Trade*, **19**, 71–90.

Dar, G. F., Padi, T. R., Rekha, S., and Dar, Q. F. (2022). Stochastic modeling for the analysis and forecasting of stock market trend using hidden markov model. *Asian Journal of Probability and Statistics*, **18**, 43–56.

Dias, J. G., Vermunt, J. K., and Ramos, S. (2010). Mixture hidden markov models in finance research. In *Advances in Data Analysis, Data Handling and Business Intelligence: Proceedings of the 32nd Annual Conference of the Gesellschaft für Klassifikation eV, Joint Conference with the British Classification Society (BCS) and the Dutch/Flemish Classification Society (VOC), Helmut-Schmidt-University, Hamburg, July 16-18, 2008*, pages 451–459. Springer.

D'Amico, G. and Petroni, F. (2012). A semi-markov model for price returns. *Physica A: Statistical Mechanics and its applications*, **391**, 4867–4876.

Hassan, M. R. (2009). A combination of hidden markov model and fuzzy model for stock market forecasting. *Neurocomputing*, **72**, 3439–3446.

Huang, J.-C., Huang, W.-T., Chu, P.-T., Lee, W.-Y., Pai, H.-P., Chuang, C.-C., and Wu, Y.-W. (2017). Applying a markov chain for the stock pricing of a novel forecasting model. *Communications in Statistics-theory and Methods*, **46**, 4388–4402.

Jobert, A. and Rogers, L. C. (2006). Option pricing with markov-modulated dynamics. *SIAM Journal on Control and Optimization*, **44**, 2063–2078.

Padi, T. R., Dar, G. F., and Rekha, S. (2022). Stock market trend analysis and prediction using markov chain approach in the context of Indian stock market. *IOSR Journal of Mathematics*, **18**, 40–45.

Rao, S. and Hong, J. (2010). Analysis of hidden markov models and support vector machines in financial applications. *University of California at Berkeley: Berkeley, CA, USA*.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 23, No. 1, 2025 (New Series), pp 273–281
https://www.ssca.org.in/journal

An Economic Analysis of Two-Node Tandem Queue with Feedback

Ankita Roy Chowdhury[1] **and Indra**[2]
[1]*D.A.V College for Girls, Yamunanagar, Haryana-135001, India*
[2]*Department of Statistics and Operational Research, Kurukshetra University, Kurukshetra, Haryana-136119, India*

Received: 30 November 2023; Revised: 15 June 2024; Accepted: 19 June 2024

Abstract

Queueing theory is basically a mathematical descriptive theory, as compared to optimisation theory, which focuses on maximising or minimising an objective function under restrictions. It attempts to define, visualise, and anticipate queues in order to obtain a better understanding of them and to provide solutions. In this paper, we obtain steady state solution for a two-node tandem queueing model with feedback. Because of its expanding usefulness in - simulating manufacturing facilities, computer/communication networks, production and assembly lines, hospitals, transportation systems, banks, and so on - queueing networks with feedback are now an area of major study and application interest. Various performance measures along with cost and profit analysis for the system have been presented in the paper.

Key words: Tandem queues; Feedback; Steady-state; Matrix geometric technique; Cost analysis.

AMS Subject Classifications: 62K05, 05B05

1. Introduction

Tandem queues are gaining popularity in recent years because to their broad practicality in simulating and analysing real-world problems such as communication networks, hospital administration, maintenance and repair facilities, and many others. In a tandem queueing system, service facilities are linked in series, and the customer passes through all of the service stations before exiting the system. The earliest work on sequence of queues in series was credited to Taylor and Jackson (1954). They obtained steady-state solution for aircraft availability subtjected to maintenance rates and spare engines supply. Burke (1956) proved that the distribution of the output of a queue with Poisson arrivals, exponential service times and infinite capacity was also Poisson with same mean value as the arrival rate, thus each queue could be treated independently. This theorem forms the basis for queues in series. Some other notable contributions are due to Reich (1957), Reich

Corresponding Author: Ankita Roy Chowdhury
Email: ankitaroy2122@gmail.com

(1963), Niu (1977), Morse (1958), Hillier and Boling (1972). Song and Ali (2009) presented a discrete-time tandem queue model and determined mean and variance of the queue length in closed form expression. Yarmand and Down (2013) proposed an algorithm for assigning servers to stations for a tandem queue with no buffer to maximize throughput. He and Chao (2014) applied matrix-analytic method to solve a queue model with K-servers in series and no waiting space in between. They showed with the help of numerical results that allocating servers to different stations in decreasing order of their service speeds didn't optimize the whole system. Van Do (2015) obtained a closed-form solution for two-stage markovian tandem queue with heterogeneous servers and showed that Eigen values could be found explicitly. Tandem queue has been analysed by several researchers using different techniques such as product form solution, Runge-Kutta's method and many more. But not much appears to have been done using analytical approach. Present study considers Tandem queue with feedback, which has been solved using Matrix-Geometric Approach.

Queues with feedback typically reflect scenarios in which a customer returns to the server having received the service (due to incomplete or unsatisfactory service). Such queues with feedback are prominent in the health care, telecommunications, and manufacturing industries where the likelihood of rework is significant. Finch (1959) introduced the concept of feedback in cyclic queues. Takacs (1963) analysed a single server queue with feedback. van der Mei *et al.* (2002) evaluated response time in a two-node queueing network with feedback. Tang and Zhao (2008) analysed GI/G/1 at each node for two-node tandem model with feedback using matrix analytic method. Chowdhury and Indra (2020) presented prediction for two-node tandem queue with feedback having state and time dependent service rates using probability generating function

In dealing with Queueing Systems with complicated topologies, the Matrix Geometric Approach outperforms the conventional Probability Generating Function Approach. Raj and Chandrasekar (2016) used the matrix geometric approach to evaluate a queueing system with device failure, standby server, and phase type service and repair. Indra and Rajan (2017) considered a Markovian queue with two heterogeneous and intermittently available servers subject to catastrophes and obtained the solutions using matrix geometric approach. Shoukry *et al.* (2018) compared the M/M/1 model with and without breakdown using a matrix geometric approach. Chaudhry *et al.* (2018) used a matrix geometric approach to obtain solutions to finite and infinite discrete queues which involved heavy-tailed distributions for service times.

The remaining work is structured as: - firstly we present the model description and assumptions, followed by its governing equations, solutions and performance measures. We also present cost and profit analysis for the model.

2. Model assumptions and descriptions

We have considered two-node tandem queue with feedback. Customers arrive at first node to receive services by server 1. If server 1 is idle, it provides service immediately, otherwise customers join the queue. On service completion at first node, the customer joins the second queue or proceeds to avail service at second node if server 2 is idle. The inter-arrival times and service times are independent and follow exponential distribution with parameters $\boldsymbol{\lambda}$, $\boldsymbol{\mu_1}$ at first node and $\boldsymbol{\mu_2}$ at second node respectively. If the customer

is satisfied with the service when it is completed, it exits the system with probability '$\boldsymbol{q}$' (disperse probability). If it finds the service unsatisfactory or wants the service again, it re-joins the queue with probability '$\boldsymbol{p}$' (feedback probability).

3. Solution of the model

The infinitesimal generator matrix Q of the system is given by:

$$Q = \begin{pmatrix} B_{00} & B_{01} & 0 & 0 & 0 & 0 & .. & .. \\ B_{10} & A_1 & A_2 & 0 & 0 & 0 & .. & .. \\ 0 & A_0 & A_1 & A_2 & 0 & 0 & .. & .. \\ 0 & 0 & A_0 & A_1 & A_2 & 0 & .. & .. \\ : & : & \ddots & \ddots & \ddots & \ddots & : & .. \\ : & : & : & : & : & .. & : & .. \\ : & : & : & : & : & .. & : & .. \end{pmatrix}$$

where,

$$B_{00} = \begin{pmatrix} -\lambda & 0 & 0 & 0 & 0 & .. & .. & .. & .. & .. \\ q\mu_1 & -(q\mu_2+\lambda) & 0 & 0 & 0 & .. & .. & .. & .. & .. \\ 0 & q\mu_2 & -(q\mu_2+\lambda) & 0 & 0 & .. & .. & .. & .. & .. \\ 0 & 0 & q\mu_2 & -(q\mu_2+\lambda) & 0 & .. & .. & .. & .. & .. \\ : & : & : & : & \ddots & \ddots & \ddots & \ddots & : & .. \\ : & : & : & : & \ddots & \ddots & \ddots & \ddots & : & .. \\ 0 & 0 & 0 & 0 & 0 & .. & q\mu_2 & -(q\mu_2+\lambda) & 0 & .. \\ : & : & : & : & : & .. & : & : & : & .. \end{pmatrix}$$

$$B_{01} = \begin{pmatrix} \lambda & 0 & 0 & 0 & 0 & .. & .. & .. & .. & .. \\ 0 & \lambda & 0 & 0 & 0 & .. & .. & .. & .. & .. \\ 0 & 0 & \lambda & 0 & 0 & .. & .. & .. & .. & .. \\ 0 & 0 & 0 & \lambda & 0 & .. & .. & .. & .. & .. \\ : & : & : & \ddots & \ddots & .. & \ddots & \ddots & : & .. \\ : & : & : & \ddots & \ddots & .. & \ddots & \ddots & : & .. \\ 0 & 0 & 0 & 0 & 0 & .. & 0 & \lambda & 0 & .. \\ : & : & : & : & : & .. & : & : & : & .. \end{pmatrix}$$

$$B_{10} = \begin{pmatrix} 0 & \mu_1 & 0 & 0 & 0 & .. & .. & .. & .. & .. \\ 0 & 0 & \mu_1 & 0 & 0 & .. & .. & .. & .. & .. \\ 0 & 0 & 0 & \mu_1 & 0 & .. & .. & .. & .. & .. \\ 0 & 0 & 0 & 0 & \mu_1 & .. & .. & .. & .. & .. \\ : & : & : & \ddots & \ddots & .. & \ddots & \ddots & : & .. \\ : & : & : & \ddots & \ddots & .. & \ddots & \ddots & : & .. \\ 0 & 0 & 0 & 0 & 0 & .. & 0 & 0 & 0 & .. \\ : & : & : & : & : & .. & : & : & : & .. \end{pmatrix}$$

$$A_0 = \begin{pmatrix} 0 & \mu_1 & 0 & 0 & 0 & .. & .. & .. & .. & .. \\ 0 & 0 & \mu_1 & 0 & 0 & .. & .. & .. & .. & .. \\ 0 & 0 & 0 & \mu_1 & 0 & .. & .. & .. & .. & .. \\ 0 & 0 & 0 & 0 & \mu_1 & .. & .. & .. & .. & .. \\ : & : & : & \ddots & \ddots & .. & \ddots & \ddots & : & .. \\ : & : & : & \ddots & \ddots & .. & \ddots & \ddots & : & .. \\ 0 & 0 & 0 & 0 & 0 & .. & 0 & 0 & 0 & .. \\ : & : & : & : & : & .. & : & : & : & .. \end{pmatrix}$$

$$A_1 = \begin{pmatrix} -(\lambda+\mu_1) & 0 & 0 & .. & .. & .. & .. & .. \\ q\mu_2 & -(q\mu_2+\mu_1+\lambda) & 0 & .. & .. & .. & .. & .. \\ 0 & q\mu_2 & \ddots & .. & .. & .. & .. & .. \\ 0 & 0 & \ddots & \ddots & .. & .. & .. & .. \\ : & : & : & \ddots & \ddots & \ddots & : & .. \\ : & : & : & \ddots & \ddots & \ddots & : & .. \\ 0 & 0 & 0 & .. & q\mu_2 & -(q\mu_2+\mu_1+\lambda) & 0 & .. \\ : & : & : & .. & : & : & : & .. \end{pmatrix}$$

$$A_2 = \begin{pmatrix} \lambda & 0 & 0 & 0 & .. & .. & .. & .. & .. \\ 0 & \lambda & 0 & 0 & .. & .. & .. & .. & .. \\ 0 & 0 & \lambda & 0 & .. & .. & .. & .. & .. \\ 0 & 0 & 0 & \lambda & .. & .. & .. & .. & .. \\ : & : & : & \ddots & .. & \ddots & \ddots & : & .. \\ : & : & : & \ddots & .. & \ddots & \ddots & : & .. \\ 0 & 0 & 0 & 0 & .. & 0 & 0 & \lambda & .. \\ : & : & : & : & .. & : & : & : & .. \end{pmatrix}$$

The state of this queueing network can be described by the vector [n_1,n_2], where n_1 indicates the number of customers at the first node and n_2 indicates the number of customers at the second node. We have $n_1 = 0, 1, 2, 3...$ and $n_2 = 0, 1, 2, 3...$
We define $\pi_{i,j} = P\{n_1{=}i,\ n_2{=}j\} = \lim_{t\to\infty} P\{n_1(t){=}i,\ n_2\ (t){=}j\}$ where (i,j) represents the state space. The steady-state probability vector $\boldsymbol{\pi}$ is given by,

$$\boldsymbol{\pi} = (\boldsymbol{\pi_0}, \boldsymbol{\pi_1}, \boldsymbol{\pi_2}, ... \quad ... \quad ... \quad ... \quad ..., \boldsymbol{\pi_n}, \boldsymbol{\pi_{n+1}}, ... \quad ... \quad ...) \tag{1}$$

where $\boldsymbol{\pi_k}$=[$\pi(k,0)$,$\pi(k,1)$, $\pi(k,2)$,... ...],k=0,1,2,3 ...
The Ergodicity condition is checked for the given model by

$$\pi_A A_2 e < \pi_A A_0 e \tag{2}$$

where π_A is the solution for π_AA=0 and $A = A_0 + A_1 + A_2$.
The steady-state probabilities π_k are related geometrically to each other as $\pi_k{=}\pi_1\ R^{k-1}$,

$\forall \quad k \geq 2$. Here, R is called the rate matrix and is the minimal non-negative solution of equation,

$$A_2 + RA_1 + R^2A_0 = 0 \tag{3}$$

The steady-state probabilities are obtained by solving the following equations:

$$\boldsymbol{\pi Q = 0} \tag{4}$$

$$\boldsymbol{\pi e = 1} \tag{5}$$

where '$\boldsymbol{e}$' is a column vector with each component equal to one.
The normalizing equation is given by:

$$\theta = \pi_0 e + \pi_1 (I - R)^{-1} e \tag{6}$$

First of all the given process is checked for ergodicity condition. If the condition holds, then we proceed to obtain the rate matrix using equation (3). π_0 and π_1 are obtained using equations (4) and (5). Finally, the normalizing constant θ is computed to normalize π_0 and π_1.

4. Performance measures

We calculate some performance measures using the steady-state probabilities, obtained by employing equation (4) & equation (5), for the system as follows:
i) "Mean number of customers in the system ($\boldsymbol{MNS}$)"

$$\boldsymbol{MNS} = \sum_{n=1}^{\infty} \boldsymbol{n\pi_n} \tag{7}$$

ii) "Mean number of customers in the queue ($\boldsymbol{MNQ}$)"

$$\boldsymbol{MNQ} = \sum_{\mathbf{n}=1}^{\infty} \boldsymbol{(n-1)\pi_n} \tag{8}$$

iii) "Probability that the two servers are busy ($\boldsymbol{P_B}$)"

$$\boldsymbol{P_B = (1 - \pi_0)} \tag{9}$$

iv) "Probability that the two servers are idle ($\boldsymbol{P_I}$)"

$$\boldsymbol{P_I = \pi_0} \tag{10}$$

v) "Mean waiting time in the system ($\boldsymbol{MWS}$)"

$$\boldsymbol{MWS} = \frac{\boldsymbol{MNS}}{\boldsymbol{\lambda}} \tag{11}$$

vi) "Mean waiting time in the queue $\boldsymbol{(MWQ)}$"

$$\boldsymbol{MWQ} = \frac{\boldsymbol{MNQ}}{\boldsymbol{\lambda}} \tag{12}$$

5. Cost and profit model

Let C_1 be the cost associated with a customer present in the queue, C_2 be the cost when server is busy, and C_3 be the cost when server is idle. So, we have the expected cost function as,

$$Total \quad Expected \quad Cost \quad (\mathbf{TEC}) = \boldsymbol{C_1 * MNQ + C_2 * P_B + C_3 * P_I} \tag{13}$$

Similarly, for an expected profit function, we have

$$Total \quad Expected \quad Profit \quad (\mathbf{TEP}) = \boldsymbol{\rho} * \mathbf{MNS} - \mathbf{TEC} \tag{14}$$

where $\boldsymbol{\rho}$ is the revenue and ρ * MNS is the total revenue (**TR**).

6. Sensitivity analysis

We have performed sensitivity analysis by changing the values of the parameters, in order to arrive at a decision. For calculation, let C_1=10, C_2=15, C_3=5 and ρ =150. The performance measurements are determined in conjunction with the overall estimated cost and profit. By altering the values of the parameters under consideration, several cost and profit graphs have been plotted.

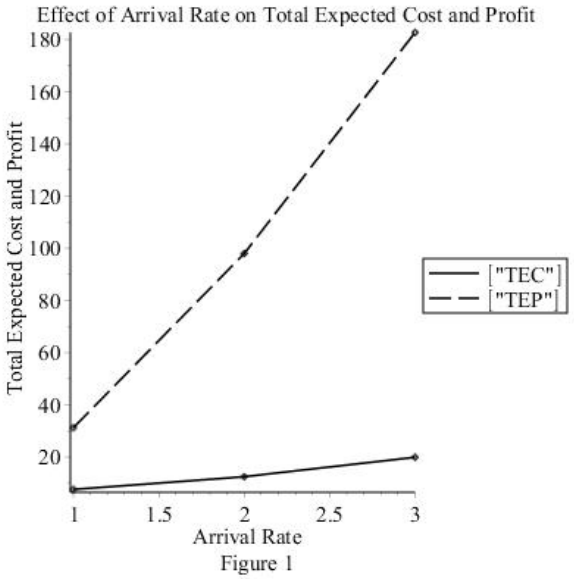


Figure 1: Effect of arrival rate on "total expected cost and total expected profit" for fixed values of ($\boldsymbol{\mu_1}$ = 5, $\boldsymbol{\mu_2}$ = 9, $\boldsymbol{q}$=0.8)

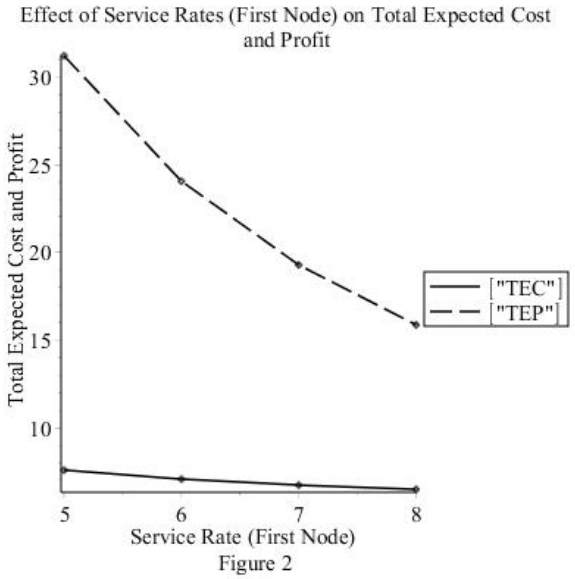


Figure 2: Effect of service rate (First Node) on "total expected cost and total expected profit" for fixed values of ($\boldsymbol{\lambda}$ = 1, $\boldsymbol{\mu_2}$ = 9, $\boldsymbol{q}$=0.8)

7. Discussion

The model analyses two-node tandem queue with feedback using Matrix Geometric Approach. Using system size probabilities, some performance measures of the system have been derived. Based on these performance measures, we have developed expected cost and profit functions. Further, we have performed sensitivity analysis to obtain numerical results

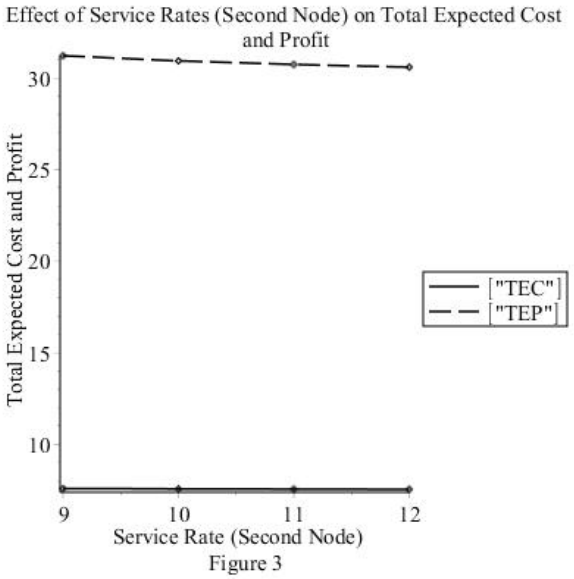


Figure 3: Effect of service rate (Second Node) on "total expected cost and total expected profit" for fixed values of ($\lambda = 1$, $\mu_1 = 5$, q=0.8)

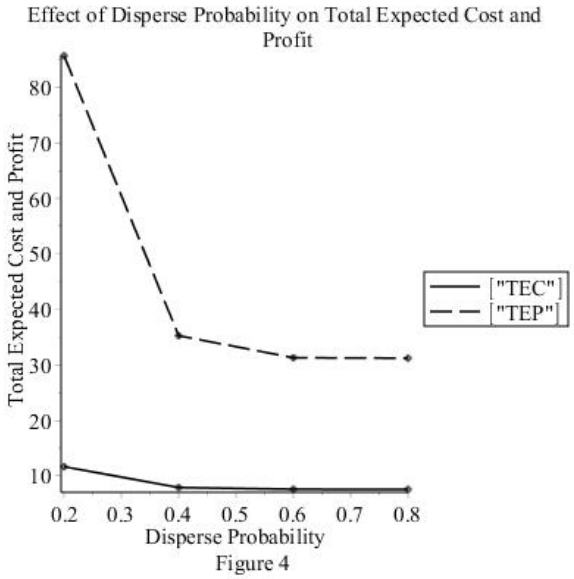


Figure 4: Effect of disperse probability on "total expected cost and total expected profit" for fixed values of ($\lambda = 1$, $\mu_1 = 5$, $\mu_2 = 9$)

for various performance measures, total expected cost and total expected profit for the parameters considered in the model. There is a slight increment in the total expected cost but on the other hand there is a handsome increment in the total expected profit as arrival rate increases. Further, total expected cost decreases as service rate (first node) increases. Also, total expected profit decreases slightly as service rate (first node) increases due to the cost associated with probability of server being idle. In addition to, total expected cost and total expected profit decreases slightly as service rate (second node) increases. The total expected cost and total expected profit decreases as disperse probability increases *i.e.* if customer decides not take a feedback.

Future considerations

Many real life congestion problems which have special structural properties can be easily solved using matrix-geometric technique even if the dimensions are of higher order. The work can be further extended for markovian and non-markovian queueing networks by considering different parameters along with their transient solutions.

Declarations

Funding- Not Applicable

Acknowledgements

We thank the Editors and the reviewers for their constructive comments that helped improve this paper

Conflict of interest

The authors do not have any financial or non-financial conflict of interest to declare for the research work included in this article.

References

Burke, P. J. (1956). The output of a queuing system. *Operations Research*, **4**, 699–704.

Chaudhry, M. L., Indra, and Rajan, V. (2018). Analytically simple and computationally efficient solution to Geo/G/1 and Geo/G/1/N queues involving heavy-tailed distributions for service times. *Calcutta Statistical Association Bulletin*, **70**, 74–85.

Chowdhury and Indra (2020). Prediction of two-node tandem queue with feedback having state and time dependent service rates. In *Journal of Physics: Conference Series*, volume 1531, page 012063. IOP Publishing.

Finch, P. (1959). Cyclic queues with feedback. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **21**, 153–157.

He, Q.-M. and Chao, X. (2014). A tollbooth tandem queue with heterogeneous servers. *European Journal of Operational Research*, **236**, 177–189.

Hillier, F. and Boling, R. (1972). *Optimal Allocation of Work in Production Line Systems with Variable Operations Times*. Technical report, Department of Operations Research, Stanford University.

Indra and Rajan (2017). Queuing analysis of markovian queue having two heterogeneous servers with catastrophes using matrix geometric technique. *International Journal of Statistics and Systems*, **12**, 205–212.

Morse, P. M. (1958). *Queues, Inventories and Maintenance*. John Willey & Sons.

Niu, S. (1977). *Bounds and Comparisons for Some Queueing Systems*. ORC 77-32. Technical report, Operations Research Center, University of California, Berkeley.

Raj, M. and Chandrasekar, B. (2016). Matrix geometric method for queueing model with state-dependent arrival of an unreliable server and ph service. *Mathematica Aeterna*, **6**, 107–116.

Reich, E. (1957). Waiting times when queues are in tandem. *The Annals of Mathematical Statistics*, **28**, 768–773.

Reich, E. (1963). Note on queues in tandem. *The Annals of Mathematical Statistics*, **34**, 338–341.

Shoukry, E., Salwa, M., and Boshra, A. (2018). Matrix geometric method for M/M/1 queueing model with and without breakdown atm machines. *American Journal of Engineering Research (AJER)*, **7**, 246–252.

Song, X. and Ali, M. M. (2009). A performance analysis of discrete-time tandem queues with markovian sources. *Performance Evaluation*, **66**, 524–543.

Takacs, L. (1963). A single-server queue with feedback. *Bell System Technical Journal*, **42**, 505–519.

Tang, J. and Zhao, Y. Q. (2008). Stationary tail asymptotics of a tandem queue with feedback. *Annals of Operations Research*, **160**, 173–189.

Taylor, J. and Jackson, R. (1954). An application of the birth and death process to the provision of spare machines. *Journal of the Operational Research Society*, **5**, 95–108.

van der Mei, R. D., Gijsen, B., van den Berg, J., et al. (2002). Response times in a two-node queueing network with feedback. *Performance Evaluation*, **49**, 99–110.

Van Do, T. (2015). A closed-form solution for a tollbooth tandem queue with two heterogeneous servers and exponential service times. *European Journal of Operational Research*, **247**, 672–675.

Yarmand, M. H. and Down, D. G. (2013). Server allocation for zero buffer tandem queues. *European Journal of Operational Research*, **230**, 596–603.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 23, No. 1, 2025 (New Series), pp 283–301
https://www.ssca.org.in/journal

Bayesian Estimation of Scale Parameter of Inverted Kumaraswamy Distribution under Various Combinations of Different Priors and Loss Functions

Ableen Kaur, Parmil Kumar and Hemani Sharma
Department of Statistics, University of Jammu, Jammu, J &K - 180006

Received: 28 July 2023; Revised: 17 April 2024; Accepted: 19 June 2024

Abstract

Bayesian estimation, a non-classical method of estimation has emerged as one of the most accepted method in statistical inference. In this paper, the Bayesian estimators of the parameters of Inverted Kumaraswamy distribution under two priors, namely Gamma and Uniform have been obtained considering three different cases: (i) when α is known and β is unknown, (ii) when α is unknown and β is known, and (iii) when α and β both are unknown. The symmetric and asymmetric loss functions *viz.*, Linear exponential (LINEX), Squared error (SE) and Entropy loss (EL) functions have been used for the Bayesian estimation. Lindley's approximation (L-approximation) has been used to obtain approximate Bayes estimators. Their performance was compared using simulated risks. An intensive simulation study is carried out with the help of Matlab and R software to examine the behavior of estimators based on their relative mean square errors.

Key words: Bayesian estimation; Inverted Kumaraswamy distribution; Lindley's approximation; Relative mean square error; Symmetric and asymmetric loss function.

AMS Subject Classifications: 62F15

1. Introduction

In recent literature, several novel distributions have been proposed for describing various real life situations in many applied sciences. Kumaraswamy (1980) obtained a distribution, which is derived from beta distribution after fixing some parameters in beta distribution. But it has a closed-form cumulative distribution function which is invertible and for which the moments do exist. If X follows Kum(α,β), then the Cumulative Distribution Function CDF is given by

$$F(x) = (1 - (1 - x)^{\alpha})^{\beta}; 0 < x < 1, \alpha, \beta > 0$$

The distribution is appropriate to natural phenomena whose outcomes are bounded from both sides, such as the individuals' heights, test scores, temperatures and hydrological daily data of rain fall.

Corresponding Author: Ableen Kaur
Email: ableenkaur23@gmail.com

Abd Al-Fattah *et al.* (2017) derived the Inverted Kumaraswamy (IKum) distribution from Kumaraswamy (Kum) distribution using the transformation $T = \frac{1}{X} - 1$ so if X follows Kum(α,β) where α and β are shape parameters, then the T has a IKum distribution with CDF

$$F(t) = (1 - (1 + t)^{-\alpha})^{\beta}; t > 0, \alpha, \beta > 0 \tag{1}$$

and probability density function (pdf)

$$f(t) = \alpha\beta(1 + t)^{-(\alpha+1)}(1 - (1 + t)^{-\alpha})^{\beta-1}; t > 0, \alpha, \beta > 0 \tag{2}$$

Also the Reliability and Hazard rate functions are given by

$$R(t) = P(T > t) = 1 - F(t) = 1 - (1 - (1 + t)^{-\alpha})^{\beta} \tag{3}$$

$$H(t) = \frac{f(t)}{R(t)} = \frac{\alpha\beta(1 + t)^{-(\alpha+1)}(1 - (1 + t)^{-\alpha})^{\beta-1}}{1 - (1 - (1 + t)^{-\alpha})^{\beta}} \tag{4}$$

Abd Al-Fattah *et al.* (2017) found IKum distribution to be a right skewed distribution, which according to Moustafa and Mahmoud (2018), will affect long term reliability predictions, producing optimistic predictions of rare events occurring in the right tail of the distribution compared with other distributions. Also the IKum distribution provides good fit to several data in literature.

The inverse distributions, also known as inverted or reciprocal distributions, have been widely applied to a wide variety of scenarios in this context. Lately, many researchers have considered and studied the properties of inverted distributions. For example, Tiao and Cuttman (1965) obtained Inverted Dirichlet distribution and its application to a problem in bayesian inference. Prakash (2012) studied the inverted exponential model and Flaih *et al.* (2012) presented exponentiated inverted Weibull distribution. Iqbal *et al.* (2017) developed a general form of IKum distribution. Fan and Gui (2022) studied the statistical inference of inverted exponentiated Rayleigh distribution based on joint progressively type-II censored data. Aldahlan *et al.* (2022) estimated the parameters of the Beta Inverted Exponential Distribution under Type-II Censored Samples. Sana *et al.* (2023) considered the problem of estimation of unknown parameters based on lower record values for Inverted Kumaraswamy distribution using Lindley's approximation.

To our best knowledge no such Bayesian analysis for Inverted Kumaraswamy distribution under these combinations of priors and loss functions has been done.

The paper is carried out as follows: In Section 2 the likelihood function is obtained, followed by the derivation of posterior distribution of the unknown parameter in all three cases under the considered priors in Section 3. In Section 4 different loss functions are used to compute the estimates of the parameters. Section 5 depicts the simulation study conducted for performance evaluation along with the results in tabular form. The study is concluded in Section 6, followed by references used for literature review.

2. Likelihood function for the inverted Kumaraswamy distribution

Let $X_1, X_2, ..., X_n$ be a random sample of size n taken from the Inverted Kumaraswamy distribution. Then the likelihood function for the given sample observations is

$$L(x;\alpha,\beta) = \alpha^n\beta^n \prod_{i=1}^{n} (1+x_i)^{-(\alpha+1)}(1-(1+x_i)^{-\alpha})^{\beta-1}$$

$$L(x;\alpha,\beta) = \alpha^n\beta^n \prod_{i=1}^{n} \frac{(1+x_i)^{-(\alpha+1)}}{(1-(1+x_i)^{-\alpha})} e^{\beta\sum_{i=1}^{n}\ln\left(1-(1+x_i)^{-\alpha}\right)} \tag{5}$$

3. Priors and posterior distributions for the unknown parameters of inverted Kumaraswamy distrubution

In Bayesian estimation selection of appropriate prior for the parameters is a crucial step. In this paper, we consider one informative and one non-informative prior. The corresponding posterior distributions were derived for each case.

3.1. CASE I: When β is unknown and α is known

3.1.1. Posterior distribution under gamma prior

$$\pi_1(\beta|a,b) = \frac{e^{-b\beta}\beta^{a-1}b^a}{\Gamma(a)}; \beta,a,b>0 \tag{6}$$

Using the likelihood function (5) and the prior (6), the posterior distribution for the parameter β becomes

$$\begin{aligned}\pi_1(\beta|x) &= \frac{L(x;\alpha,\beta)*\pi_1(\beta|a,b)}{\int_0^\infty L(x;\alpha,\beta)*\pi_1(\beta|a,b)d\beta} \\ &= \frac{\alpha^n\beta^n\prod_{i=1}^{n}\frac{(1+x_i)^{-(\alpha+1)}}{((1-(1+x_i)^{-\alpha})}e^{\beta\sum_{i=1}^{n}\ln\left(1-(1+x_i)^{-\alpha}\right)}\frac{e^{-b\beta}\beta^{a-1}b^a}{\Gamma(a)}}{\int_0^\infty \alpha^n\beta^n\prod_{i=1}^{n}\frac{(1+x_i)^{-(\alpha+1)}}{((1-(1+x_i)^{-\alpha})}e^{\beta\sum_{i=1}^{n}\ln(1-(1+x_i)^{-\alpha})}\frac{e^{-b\beta}\beta^{a-1}b^a}{\Gamma(a)}d\beta}\end{aligned}$$

$$\pi_1(\beta|x) = \frac{\beta^{(n+a)-1}\exp(-\beta R)R^{n+a}}{\Gamma(n+a)}, \tag{7}$$

where $R(x,\alpha) = b - \sum_{i=1}^{n}\ln(1-(1+x_i)^{-\alpha})$

3.1.2. Posterior distribution under uniform prior

$$\pi_2(\beta|k) = k; \beta,k>0 \tag{8}$$

Using the likelihood function (5) and the prior (8), the posterior distribution for the parameter β becomes

$$\begin{aligned}\pi_2(\beta|x) &= \frac{L(x;\alpha,\beta)*\pi_2(\beta|k)}{\int_0^\infty L(x;\alpha,\beta)*\pi_2(\beta|k)d\beta} \\ &= \frac{\alpha^n\beta^n\prod_{i=1}^n \frac{(1+x_i)^{-(\alpha+1)}}{(1-(1+x_i)^{-\alpha})}e^{\beta\sum_{i=1}^n \ln\left(1-(1+x_i)^{-\alpha}\right)}k}{\int_0^\infty \alpha^n\beta^n\prod_{i=1}^n \frac{(1+x_i)^{-(\alpha+1)}}{(1-(1+x_i)^{-\alpha})}e^{\beta\sum_{i=1}^n \ln\left(1-(1+x_i)^{-\alpha}\right)}kd\beta}\end{aligned}$$

$$\pi_2(\beta|x) = \frac{\beta^n \exp(-\beta T)T^{n+1}}{\Gamma(n+1)}, \tag{9}$$

where $T(x,\alpha) = -\sum_{i=1}^n \ln\left(1-(1+x_i)^{-\alpha}\right)$

3.2. CASE II: When β is known and α is unknown

3.2.1. Posterior distribution under gamma prior

$$\pi_1(\alpha|a,b) = \frac{e^{-b\beta}\beta^{a-1}b^a}{\Gamma(a)}; \alpha,a,b>0 \tag{10}$$

Using the likelihood function (5) and the prior (10), the posterior distribution for the parameter β becomes

$$\begin{aligned}\pi_1(\alpha|x) &= \frac{L(x;\alpha,\beta)*\pi_1(\alpha|a)}{\int_0^\infty L(x;\alpha,\beta)*\pi_1(\alpha|a)d\alpha} \\ &= \frac{\alpha^n\beta^n\prod_{i=1}^n(1+x_i)^{-(\alpha+1)}(1-(1+x_i)^{-\alpha})^{\beta-1}\frac{e^{-b\beta}\beta^{a-1}b^a}{\Gamma(a)}}{\int_0^\infty \alpha^n\beta^n\prod_{i=1}^n(1+x_i)^{-(\alpha+1)}(1-(1+x_i)^{-\alpha})^{\beta-1}\frac{e^{-b\beta}\beta^{a-1}b^a}{\Gamma(a)}d\alpha} \\ &= \frac{\alpha^{n+a-1}e^{-b\alpha}\prod_{i=1}^n(1+x_i)^{-(\alpha+1)}(1-(1+x_i)^{-\alpha})^{\beta-1}}{\int_0^\infty \alpha^{n+a-1}e^{-b\alpha}\prod_{i=1}^n(1+x_i)^{-(\alpha+1)}(1-(1+x_i)^{-\alpha})^{\beta-1}d\alpha} \\ &= K_1^{-1}\alpha^{n+a-1}e^{-b\alpha}\prod_{i=1}^n(1+x_i)^{-(\alpha+1)}\left(1-(1+x_i)^{-\alpha}\right)^{\beta-1}\end{aligned}$$

where,

$$K_1 = \int_0^\infty \alpha^{n+a-1}e^{-b\alpha}\prod_{i=1}^n(1+x_i)^{-(\alpha+1)}\left(1-(1+x_i)^{-\alpha}\right)^{\beta-1}d\alpha \tag{11}$$

3.2.2. Posterior distribution under uniform prior

$$\pi_2(\alpha|a) = k; k>0 \tag{12}$$

$$\begin{aligned}\pi_2(\alpha|x) &= \frac{L(x;\alpha,\beta) * \pi_2(\alpha|a)}{\int_0^\infty L(x;\alpha,\beta) * \pi_2(\alpha|a)d\alpha} \\ &= \frac{\alpha^n\beta^n \prod_{i=1}^n (1+x_i)^{-(\alpha+1)}(1-(1+x_i)^{-\alpha})^{\beta-1}k}{\int_0^\infty \alpha^n\beta^n \prod_{i=1}^n (1+x_i)^{-(\alpha+1)}(1-(1+x_i)^{-\alpha})^{\beta-1}kd\alpha} \\ &= \frac{\alpha^n \prod_{i=1}^n (1+x_i)^{-(\alpha+1)}(1-(1+x_i)^{-\alpha})^{\beta-1}}{\int_0^\infty \alpha^n \prod_{i=1}^n (1+x_i)^{-(\alpha+1)}(1-(1+x_i)^{-\alpha})^{\beta-1}d\alpha} \\ &= K_2^{-1}\alpha^n \prod_{i=1}^n (1+x_i)^{-(\alpha+1)}\Big(1-(1+x_i)^{-\alpha}\Big)^{\beta-1}\end{aligned}$$

where,

$$K_2 = \int_0^\infty \alpha^n \prod_{i=1}^n (1+x_i)^{-(\alpha+1)}\Big(1-(1+x_i)^{-\alpha}\Big)^{\beta-1}d\alpha \tag{13}$$

3.3. CASE III: When α and β both are unknown

3.3.1. Posterior distribution under gamma prior

Suppose the parameters are independent and follow Gamma distribution,

$$\pi(\alpha|a_1,b_1) \propto \alpha^{a_1-1}e^{-b_1\alpha}; \alpha>0, a_1, b_1>0$$
$$\pi(\beta|a_2,b_2) \propto \beta^{a_2-1}e^{-b_2\beta}; \beta>0, a_2, b_2>0$$

where, $a_1\&b_1$ and $a_2\&b_2$, are non-negative hyperparameters and are known. The joint prior distribution for α and β is given by

$$\pi_{11}(\alpha,\beta|a_1,b_1,a_2,b_2) \propto \alpha^{a_1-1}\beta^{a_2-1}e^{-b_1\alpha_1-b_2\beta}$$

The joint posterior density function of parameters α and β is obtained as

$$\pi_{11}(\alpha,\beta|x) = \frac{L(x;\alpha,\beta) * \pi_{11}(\alpha,\beta|a_1,b_1,a_2,b_2)}{\int_0^\infty\int_0^\infty L(x;\alpha,\beta) * \pi_{11}(\alpha,\beta|a_1,b_1,a_2,b_2)d\alpha d\beta}$$

the above equation cannot be obtained in closed form so in order to find the Bayes estimator of the parameters we have used Lindley approximation method. The joint posterior density function can be written as

$$\pi_{11}(\alpha,\beta|x) \propto \alpha^{n+a_1-1}\beta^{n+a_2-1} \prod_{i=1}^n (1+x_i)^{-(\alpha+1)}\Big(1-(1+x_i)^{-\alpha}\Big)^{\beta-1}e^{-b_1\alpha_1-b_2\beta} \tag{14}$$

3.3.2. Posterior distribution under uniform prior

Suppose the parameters are independent and follow Uniform distribution,

$$\pi(\alpha|k_1) = k_1; \alpha>0, k_1>0$$
$$\pi(\beta|k_2) = k_2; \beta>0, k_2>0$$

The joint prior distribution for α and β is given by

$$\pi_{12}(\alpha, \beta|k_1, k_2) = k_1 k_2$$

The joint posterior density function of parameters α and β is obtained as

$$\pi_{12}(\alpha, \beta|x) = \frac{L(x; \alpha, \beta) * \pi_{12}(\alpha, \beta|k_1, k_2)}{\int_0^\infty \int_0^\infty L(x; \alpha, \beta) * \pi_{12}(\alpha, \beta|k_1, k_2) d\alpha d\beta}$$

the above equation cannot be obtained in closed form so in order to find the Bayes estimator of the parameters we have used Lindley approximation method. The joint posterior density function can be written as

$$\pi_{11}(\alpha, \beta|x) \propto \alpha^n \beta^n \prod_{i=1}^{n} (1 + x_i)^{-(\alpha+1)} \Big(1 - (1 + x_i)^{-\alpha}\Big)^{\beta-1} \tag{15}$$

4. Bayesian estimation under different loss functions

This section presents the Bayes estimates of the unknown parameter obtained under three loss functions *viz.*, Linear exponential, Squared error and Entropy loss functions.

4.1. Case I: When β is unknown and α is known

4.1.1. Bayesian estimation by using gamma prior under different loss functions

• Bayes estimator under LINEX loss function

The LINEX loss function is given by

$$L(\hat{\beta}, \beta) = \exp(q_1(\hat{\beta} - \beta)) - h(\hat{\beta} - \beta) - 1; q_1, h \neq 0 \tag{16}$$

By using LINEX loss function as given in (16), the risk function is given by

$$R(\hat{\beta}, \beta) = E[L(\hat{\beta}, \beta)] = \int_0^\infty \exp(q_1(\hat{\beta} - \beta) - h(\hat{\beta} - \beta) - 1).\pi_1(\beta|x) d\beta$$

$$= \int_0^\infty \Big[\exp(q_1\hat{\beta}).\exp(-q_1\beta) - h\hat{\beta} + h\beta - 1\Big] . \frac{\beta^{(n+a)-1} \exp(-\beta R) R^{n+a}}{\Gamma(n+a)} d\beta$$

$$= \exp(q_1\hat{\beta}) \frac{R^{n+a}}{[R + q_1]^{n+a}} + \frac{h}{R}(n + a) - (h\hat{\beta} + 1)$$

Now solving $\frac{\partial R(\hat{\beta}, \beta)}{\partial \hat{\beta}} = 0$, we get

$$\exp(q_1\hat{\beta}) = \frac{h}{q_1} \left(\frac{R + q_1}{R}\right)^{n+a}$$

Taking log on both sides, we obtain the Bayes estimator as

$$\hat{\beta}_{LINEX} = \frac{1}{q_1} \left[\ln\left(\frac{h}{q_1}\right) + (n + a) \ln\left(\frac{R + q_1}{R}\right)\right] \tag{17}$$

• **Bayes estimator under squared error loss function**

The squared error loss function is given by

$$L(\hat{\beta},\beta) = (\hat{\beta}-\beta)^2 \tag{18}$$

By using Squared error loss function as given in (18), the risk function is given by

$$R(\hat{\beta},\beta) = E[L(\hat{\beta},\beta)] = \int_0^\infty \left[(\hat{\beta}-\beta)^2\right].\pi_1(\beta|x)d\beta$$

$$= \int_0^\infty \left[(\hat{\beta}-\beta)^2\right].\frac{\beta^{(n+a)-1}\exp(-\beta R)R^{n+a}}{\Gamma(n+a)}d\beta$$

$$= \hat{\beta}^2 + \frac{(n+a+1)(n+a)}{R^2} - 2\hat{\beta}\frac{(n+a)}{R}$$

Now solving $\frac{\partial R(\hat{\beta},\beta)}{\partial\hat{\beta}} = 0$, we get

$$\implies 2\hat{\beta} - 2\frac{(n+a)}{R} = 0$$

$$\implies \hat{\beta}_{SELF} = \frac{(n+a)}{R} \tag{19}$$

• **Bayes estimator under entropy loss function**

The entropy loss function is given by

$$L(\hat{\beta},\beta) = b[\Delta - \ln(\Delta) - 1]; b > 0 \tag{20}$$

Assuming $b = 1$,$\Delta = \frac{\hat{\beta}}{\beta}$, we have

$$L(\hat{\beta},\beta) = \left[\left(\frac{\hat{\beta}}{\beta}\right) - \ln\left(\frac{\hat{\beta}}{\beta}\right) - 1\right] \tag{21}$$

By using Entropy loss function as given in (21), the risk function is given by

$$R(\hat{\beta},\beta) = E[L(\hat{\beta},\beta)] = \int_0^\infty \left[\left(\frac{\hat{\beta}}{\beta}\right) - \ln\left(\frac{\hat{\beta}}{\beta}\right) - 1\right].\pi_1(\beta|x)d\beta$$

$$= \int_0^\infty \left[\left(\frac{\hat{\beta}}{\beta}\right) - \ln\left(\frac{\hat{\beta}}{\beta}\right) - 1\right].\frac{\beta^{(n+a)-1}\exp(-\beta R)R^{n+a}}{\Gamma(n+a)}d\beta$$

$$= \hat{\beta}.\frac{R}{n+a-1} - \ln(\hat{\beta}) + \frac{\psi(n+a)}{\Gamma n+a} - 1$$

Now solving $\frac{\partial R(\hat{\beta},\beta)}{\partial\hat{\beta}} = 0$, we get

$$\implies \frac{R}{n+a-1} - \frac{1}{\hat{\beta}} = 0$$

$$\implies \hat{\beta}_{ELF} = \frac{n+a-1}{R} \tag{22}$$

4.1.2. Bayesian estimation by using uniform prior under various loss functions

• Bayes estimator under LINEX loss function

The LINEX loss function is given by

$$L(\hat{\beta},\beta) = \exp(q_1(\hat{\beta}-\beta)) - h(\hat{\beta}-\beta) - 1, q_1, h \neq 0 \tag{23}$$

By using LINEX loss function as given in (23), the risk function is given by

$$R(\hat{\beta},\beta) = E[L(\hat{\beta},\beta)] = \int_0^\infty \exp(q_1(\hat{\beta}-\beta) - h(\hat{\beta}-\beta) - 1).\pi_2(\beta|x)d\beta$$

$$= \int_0^\infty \left[\exp(q_1\hat{\beta}).\exp(-q_1\beta) - h\hat{\beta} + h\beta - 1\right].\frac{\beta^n \exp(-\beta T)^{n+1}}{\Gamma(n+1)}d\beta$$

$$= \exp(q_1\hat{\beta})[\left(\frac{T}{(T+q_1)}\right)]^{n+1} - h\frac{(n+2)}{T} - (h\hat{\beta}+1)$$

Now solving $\frac{\partial R(\hat{\beta},\beta)}{\partial \hat{\beta}} = 0$, we get

$$\exp(q_1\hat{\beta}) = \frac{h}{q_1}\left(\frac{T+q_1}{T}\right)^{n+1}$$

Taking log on both sides, we obtain the Bayes estimator as

$$\hat{\beta}_{LINEX} = \frac{1}{q_1}\left[ln\left(\frac{h}{q_1}\right) + (n+1)log\left(\frac{T+q_1}{T}\right)\right] \tag{24}$$

• Bayes estimator under squared error loss function

The squared error loss function is given by

$$L(\hat{\beta},\beta) = (\hat{\beta}-\beta)^2 \tag{25}$$

By using Squared error loss function as given in (25), the risk function is given by

$$R(\hat{\beta},\beta) = E[L(\hat{\beta},\beta)] = \int_0^\infty \left[(\hat{\beta}-\beta)^2\right].\pi_2(\beta|x)d\beta$$

$$= \int_0^\infty \left[(\hat{\beta}-\beta)^2\right].\frac{\beta^n \exp(-\beta T)^{n+1}}{\Gamma(n+1)}d\beta$$

$$= \hat{\beta}^2 + \frac{(n+2)(n+1)}{T^2} - 2\hat{\beta}\frac{(n+1)}{T}$$

Now solving $\frac{\partial R(\hat{\beta},\beta)}{\partial \hat{\beta}} = 0$, we get

$$\implies 2\hat{\beta} - 2\frac{(n+1)}{T} = 0$$

$$\implies \hat{\beta}_{SELF} = \frac{(n+1)}{T} \tag{26}$$

• **Bayes estimator under entropy loss function**

The entropy loss function is given by

$$L(\hat{\beta}, \beta) = b[\Delta - \ln(\Delta) - 1]; b > 0 \tag{27}$$

Assuming $b = 1$,$\Delta = \frac{\hat{\alpha}}{\alpha}$, we have

$$L(\hat{\beta}, \beta) = \left[\left(\frac{\hat{\alpha}}{\alpha}\right) - \ln\left(\frac{\hat{\alpha}}{\alpha}\right) - 1\right] \tag{28}$$

By using Entropy loss function as given in (28), the risk function is given by

$$R(\hat{\beta}, \beta) = E[L(\hat{\beta}, \beta)] = \int_0^\infty \left[\left(\frac{\hat{\beta}}{\beta}\right) - \ln\left(\frac{\hat{\beta}}{\beta}\right) - 1\right] .\pi_2(\beta|x)d\beta$$

$$= \int_0^\infty \left[\left(\frac{\hat{\beta}}{\beta}\right) - \ln\left(\frac{\hat{\beta}}{\beta}\right) - 1\right] .\frac{\beta^n \exp(-\beta T)^{n+1}}{\Gamma(n+1)} d\beta$$

$$= \hat{\beta}.\frac{T}{n} - \ln(\hat{\beta}) + \frac{\psi(n+1)}{\Gamma n + 1} - 1$$

Now solving $\frac{\partial R(\hat{\beta},\beta)}{\partial \hat{\beta}} = 0$, we get

$$\implies \frac{T}{n} - \frac{1}{\hat{\beta}} = 0$$

$$\implies \hat{\beta}_{ELF} = \frac{n}{T} \tag{29}$$

4.2. Case II: When β is known and α is unknown

4.2.1. Bayesian estimation by using gamma prior under different loss functions

• **Bayes estimator under LINEX loss function**

The bayes estimator of α under LINEX loss function is given by

$$\hat{\alpha}_{LINEX} = -\frac{1}{h} \ln E[e^{-h\alpha}|x] \tag{30}$$

where,

$$E[e^{-h\alpha}|x] = \frac{\int_0^\infty \alpha^{n+a-1} e^{-\alpha(b+h)} \prod_{i=1}^n (1+x_i)^{-(\alpha+1)} (1-(1+x_i)^{-\alpha})^{\beta-1} d\alpha}{\int_0^\infty \alpha^{n+a-1} e^{-b\alpha} \prod_{i=1}^n (1+x_i)^{-(\alpha+1)} (1-(1+x_i)^{-\alpha})^{\beta-1} d\alpha}$$

• **Bayes estimator under squared error loss function**

The bayes estimator of α under SELF is given by

$$\hat{\alpha}_{SELF} = E[\alpha|x] \tag{31}$$

where,

$$E[\alpha|x] = \frac{\int_0^\infty \alpha^{n+a} e^{-b\alpha} \prod_{i=1}^n (1+x_i)^{-(\alpha+1)} (1-(1+x_i)^{-\alpha})^{\beta-1} d\alpha}{\int_0^\infty \alpha^{n+a-1} e^{-b\alpha} \prod_{i=1}^n (1+x_i)^{-(\alpha+1)} (1-(1+x_i)^{-\alpha})^{\beta-1} d\alpha}$$

• **Bayes estimator under Entropy loss function**

The bayes estimator of α under ELF is given by

$$\hat{\alpha}_{ELF} = (E[\alpha^{-1}|x])^{-1} \tag{32}$$

where,

$$E[\alpha^{-1}|x] = \frac{\int_0^\infty \alpha^{n+a-1} e^{-b\alpha+\alpha^{-1}} \prod_{i=1}^n (1+x_i)^{-(\alpha+1)} (1-(1+x_i)^{-\alpha})^{\beta-1} d\alpha}{\int_0^\infty \alpha^{n+a-1} e^{-b\alpha} \prod_{i=1}^n (1+x_i)^{-(\alpha+1)} (1-(1+x_i)^{-\alpha})^{\beta-1} d\alpha}$$

4.2.2. Bayesian estimation by using Uniform prior under different loss functions

• **Bayes estimator under LINEX loss function**

The bayes estimator of α under LINEX loss function is given by

$$\hat{\alpha}_{LINEX} = -\frac{1}{h} \ln E[e^{-h\alpha}|x] \tag{33}$$

where,

$$E[e^{-h\alpha}|x] = \frac{\int_0^\infty \alpha^n e^{-h\alpha} \prod_{i=1}^n (1+x_i)^{-(\alpha+1)} (1-(1+x_i)^{-\alpha})^{\beta-1} a\alpha}{\int_0^\infty \alpha^n \prod_{i=1}^n (1+x_i)^{-(\alpha+1)} (1-(1+x_i)^{-\alpha})^{\beta-1} d\alpha}$$

• **Bayes estimator under squared error loss function**

The bayes estimator of α under SELF is given by

$$\hat{\alpha}_{SELF} = E[\alpha|x] \tag{34}$$

where,

$$E[\alpha|x] = \frac{\int_0^\infty \alpha^{n+1} \prod_{i=1}^n (1+x_i)^{-(\alpha+1)} (1-(1+x_i)^{-\alpha})^{\beta-1} a\alpha}{\int_0^\infty \alpha^n \prod_{i=1}^n (1+x_i)^{-(\alpha+1)} (1-(1+x_i)^{-\alpha})^{\beta-1} d\alpha}$$

• **Bayes estimator under entropy loss function**

The bayes estimator of α under ELF is given by

$$\hat{\alpha}_{ELF} = (E[\alpha^{-1}|x])^{-1} \tag{35}$$

where,

$$E[\alpha^{-1}|x] = \frac{\int_0^\infty \alpha^{n-1} \prod_{i=1}^n (1+x_i)^{-(\alpha+1)} (1-(1+x_i)^{-\alpha})^{\beta-1} a\alpha}{\int_0^\infty \alpha^n \prod_{i=1}^n (1+x_i)^{-(\alpha+1)} (1-(1+x_i)^{-\alpha})^{\beta-1} d\alpha}$$

4.3. Case III: When α and β both are unknown

In previous section, we obtained the mathematical expression for the Bayes estimates of the parameters. We notice that these estimators are in the form of ratio of two integrals. Thus, Lindley's approximation method is a good alternative to solve such types of problems see Lindley (1980). Therefore, we briefly discuss about this approximation technique and apply it to evaluate the Bayesian estimates by considering the function I(x), defined as follows;

$$I(x) = E[\alpha, \beta|x] = \frac{\int u(\alpha,\beta)e^{L(\alpha,\beta)+G(\alpha,\beta)}d(\alpha,\beta)}{\int e^{L(\alpha,\beta)+G(\alpha,\beta)}d(\alpha,\beta)} \tag{36}$$

where,

$u(\alpha, \beta)$ is the function of α and β only; $L(\alpha, \beta)$ is the log likelihood function; $G(\alpha, \beta)$ is the log of joint prior density.

According to Lindley (1980), if ML estimates of the parameters are available and n is sufficiently large then the above ratio of the integral can be approximated as:

$$I(x) = u(\hat{\alpha},\hat{\beta}) + \tfrac{1}{2}[(\hat{u}_{\beta\beta} + 2\hat{u}_\beta\hat{p}_\beta)\hat{\sigma}_{\beta\beta} + (\hat{u}_{\alpha\beta} + 2\hat{u}_\alpha\hat{p}_\beta)\hat{\sigma}_{\alpha\beta} + (\hat{u}_{\beta\alpha} + 2\hat{u}_\beta\hat{p}_\alpha)\hat{\sigma}_{\beta\alpha} + (\hat{u}_{\alpha\alpha} + 2\hat{u}_\alpha\hat{p}_\alpha)\hat{\sigma}_{\alpha\alpha}] + \tfrac{1}{2}[(\hat{u}_\beta\hat{\sigma}_{\beta\beta} + \hat{u}_\alpha\hat{\sigma}_{\beta\alpha})(\hat{L}_{\beta\beta\beta}\hat{\sigma}_{\beta\beta} + \hat{L}_{\beta\alpha\beta}\hat{\sigma}_{\beta\alpha} + \hat{L}_{\alpha\beta\beta}\hat{\sigma}_{\alpha\beta} + \hat{L}_{\alpha\alpha\beta}\hat{\sigma}_{\alpha\alpha}) + (\hat{u}_\beta\hat{\sigma}_{\alpha\beta} + \hat{u}_\alpha\hat{\sigma}_{\alpha\alpha}) \times (\hat{L}_{\alpha\beta\beta}\hat{\sigma}_{\beta\beta} + \hat{L}_{\beta\alpha\alpha}\hat{\sigma}_{\beta\alpha} + \hat{L}_{\alpha\beta\alpha}\hat{\sigma}_{\alpha\beta} + \hat{L}_{\alpha\alpha\alpha}\hat{\sigma}_{\alpha\alpha})]$$

where, $\hat{\alpha}$ and $\hat{\beta}$ are the MLE of α and β respectively. The expressions for the MLE of the parameters of Inverted Kumaraswamy distribution have been derived by Al-Fattah et. al. (2017) and Sana et.al. (2023)

$$\hat{u}_\alpha = \frac{\partial u(\hat{\alpha},\hat{\beta})}{\partial\hat{\alpha}}, \hat{u}_\beta = \frac{\partial u(\hat{\alpha},\hat{\beta})}{\partial\hat{\beta}}, \hat{u}_{\alpha\beta} = \frac{\partial^2 u(\hat{\alpha},\hat{\beta})}{\partial\hat{\alpha}\partial\hat{\beta}}, \hat{u}_{\beta\alpha} = \frac{\partial^2 u(\hat{\alpha},\hat{\beta})}{\partial\hat{\beta}\partial\hat{\alpha}},$$

$$\hat{u}_{\alpha\alpha} = \frac{\partial^2 u(\hat{\alpha},\hat{\beta})}{\partial\hat{\alpha}^2}, \hat{u}_{\beta\beta} = \frac{\partial^2 u(\hat{\alpha},\hat{\beta})}{\partial\hat{\beta}^2}, \hat{p}_\alpha = \frac{\partial G(\hat{\alpha},\hat{\beta})}{\partial\hat{\alpha}}, \hat{p}_\beta = \frac{\partial G(\hat{\alpha},\hat{\beta})}{\partial\hat{\beta}},$$

$$\hat{L}_{\alpha\alpha} = \frac{\partial^2 L(\hat{\alpha},\hat{\beta})}{\partial\hat{\alpha}^2}, \hat{L}_{\beta\beta} = \frac{\partial^2 L(\hat{\alpha},\hat{\beta})}{\partial\hat{\beta}^2}, \hat{L}_{\alpha\alpha\alpha} = \frac{\partial^3 L(\hat{\alpha},\hat{\beta})}{\partial\hat{\alpha}^3}, \hat{L}_{\alpha\alpha\beta} = \frac{\partial^3 L(\hat{\alpha},\hat{\beta})}{\partial\hat{\alpha}\partial\hat{\alpha}\partial\hat{\beta}},$$

$$\hat{L}_{\beta\beta\alpha} = \frac{\partial^3 L(\hat{\alpha},\hat{\beta})}{\partial\hat{\beta}\partial\hat{\beta}\partial\hat{\alpha}}, \hat{L}_{\beta\alpha\beta} = \frac{\partial^3 L(\hat{\alpha},\hat{\beta})}{\partial\hat{\beta}\partial\hat{\alpha}\partial\hat{\beta}}, \hat{L}_{\alpha\beta\beta} = \frac{\partial^3 L(\hat{\alpha},\hat{\beta})}{\partial\hat{\alpha}\partial\hat{\beta}\partial\hat{\beta}}, \hat{L}_{\beta\alpha\alpha} = \frac{\partial^3 L(\hat{\alpha},\hat{\beta})}{\partial\hat{\beta}\partial\hat{\alpha}\partial\hat{\alpha}}$$

4.3.1. Bayesian estimation by using gamma prior under different loss functions

• Bayes estimator under squared error loss function

After substitution, the equation (14) reduces like Lindleys integral, therefore, for the

Bayes estimates of the parameter α under squared error loss function are,

$$u(\alpha,\beta)=\alpha$$
$$L(\alpha,\beta)=n\ln\alpha+n\ln\beta-(\alpha+1)\sum\ln(1+x_i)+(\beta-1)\sum\ln(1-(1+x_i)^{-\alpha})$$
$$G(\alpha,\beta)=(a_1-1)\ln\alpha+(a_2-1)\ln\beta-b_1\alpha-b_2\beta$$

It may be verified that,

$$u_\alpha=1, u_{\alpha\alpha}=u_{\alpha\beta}=u_{\beta\alpha}=u_{\beta\beta}=0$$
$$p_\alpha=\frac{a_1-1}{\alpha}-b_1, p_\beta=\frac{a_2-1}{\beta}-b_2$$
$$L_\alpha=\frac{n}{\alpha}-\sum\ln(1+x_i)+(\beta-1)\sum\frac{(1+x_i)^{-\alpha}}{1-(1+x_i)^{-\alpha}}\ln(1+x_i)$$
$$L_{\alpha\alpha}=\frac{-n}{\alpha^2}-(\beta-1)\sum\frac{(1+x_i)^{-\alpha}(\ln(1+x_i))^2}{[1-(1+x_i)^{-\alpha}]^2}$$
$$L_{\alpha\alpha\alpha}=\frac{2n}{\alpha^3}-(\beta-1)\sum\frac{\left((1+x_i)^{3\alpha}-(1+x_i)^{-\alpha}\right)}{[1-(1+x_i)^{-\alpha}]^4}(\ln(1+x_i))^3$$
$$L_{\alpha\beta}=\sum\frac{(1+x_i)^{-\alpha}\ln(1+x_i)}{[1-(1+x_i)^{-\alpha}]}=L_{\beta\alpha}$$
$$L_{\alpha\alpha\beta}=-\sum\frac{(1+x_i)^{-\alpha}(\ln(1+x_i))^2}{[1-(1+x_i)^{-\alpha}]^2}=L_{\beta\alpha\alpha}=L_{\alpha\beta\alpha}$$
$$L_\beta=\frac{n}{\beta}+\sum\ln(1-(1+x_i)^{-\alpha})$$
$$L_{\beta\beta}=\frac{-n}{\beta^2}$$
$$L_{\beta\beta\beta}=\frac{2n}{\beta^3}$$
$$L_{\beta\beta\alpha}=L_{\beta\alpha\beta}=L_{\alpha\beta\beta}=0$$

If α and β are orthogonal then $\sigma_{ij}=0$ for $i\neq j$ and $\sigma_{ij}=-\frac{1}{L_{ij}}$ for $i=j$.

After evaluation of all U-terms, L-terms, and p-terms at the point $(\hat{\alpha},\hat{\beta})$ and using the above expression, the approximate Bayes estimator of α under SELF is,

$$\hat{\alpha}^L_{SELF}=\hat{\alpha}+\hat{u}_\alpha\hat{p}_\alpha\hat{\sigma}_{\alpha\alpha}+0.5(\hat{u}_\alpha\hat{\sigma}_{\alpha\alpha}\hat{\sigma}_{\beta\beta}\hat{L}_{\alpha\beta\beta}+\hat{u}_\alpha\hat{\sigma}^2_{\alpha\alpha}\hat{L}_{\alpha\beta\beta}) \tag{37}$$

and similarly the Bayes estimate fir β under SELF is, $u_\beta=1, u_{\alpha\alpha}=u_{\alpha\beta}=u_{\beta\alpha}=u_{\beta\beta}=0$ and remaining L-terms and p-terms will be same as above. Thus we have

$$\hat{\beta}^L_{SELF}=\hat{\beta}+\hat{u}_\beta\hat{p}_\beta\hat{\sigma}_{\beta\beta}+0.5(\hat{u}_\beta\hat{\sigma}_{\alpha\alpha}\hat{\sigma}_{\beta\beta}\hat{L}_{\alpha\alpha\beta}+\hat{u}_\beta\hat{\sigma}^2_{\beta\beta}\hat{L}_{\beta\beta\beta}) \tag{38}$$

• Bayes estimator under LINEX loss function

The approximate Bayes estimator of α under LINEX is evaluated by taking $u(\alpha,\beta) = e^{-h\alpha}, h > 0$, $u_\alpha = -he^{-h\alpha}, u_{\alpha\alpha} = h^2e^{-h\alpha}, u_{\alpha\beta} = u_{\beta\alpha} = u_{\beta\beta} = 0$ and remaining L-terms and p-terms will be same as above. Thus we have

$$\hat{\alpha}^L_{LINEX} = \hat{\alpha} + \hat{u}_\alpha\hat{p}_\alpha\hat{\sigma}_{\alpha\alpha} + 0.5(\hat{u}_\alpha\hat{\sigma}_{\alpha\alpha}\hat{\sigma}_{\beta\beta}\hat{L}_{\alpha\beta\beta} + \hat{u}_\alpha\hat{\sigma}^2_{\alpha\alpha}\hat{L}_{\alpha\beta\beta}) \tag{39}$$

and similarly the Bayes estimate fir β under LINEX is evaluated by taking $u(\alpha,\beta) = e^{-h\beta}, h > 0$, $u_\beta = -he^{-h\beta}, u_{\beta\beta} = h^2e^{-h\beta}, u_{\alpha\beta} = u_{\beta\alpha} = u_{\alpha\alpha} = 0$ and remaining L-terms and p-terms will be same as above. Thus we have

$$\hat{\beta}^L_{LINEX} = \hat{\beta} + \hat{u}_\beta\hat{p}_\beta\hat{\sigma}_{\beta\beta} + 0.5(\hat{u}_\beta\hat{\sigma}_{\alpha\alpha}\hat{\sigma}_{\beta\beta}\hat{L}_{\alpha\alpha\beta} + \hat{u}_\beta\hat{\sigma}^2_{\beta\beta}\hat{L}_{\beta\beta\beta}) \tag{40}$$

• Bayes estimator under entropy loss function

The approximate Bayes estimator of α under ELF is evaluated by taking $u(\alpha,\beta) = e^{\alpha^{-1}}$, $u_\alpha = -\frac{e^{\alpha^{-1}}}{\alpha^2}, u_{\alpha\alpha} = \frac{e^{\alpha^{-1}}}{\alpha^3}\left[\frac{1}{\alpha} + 2\right], u_{\alpha\beta} = u_{\beta\alpha} = u_{\beta\beta} = 0$ and remaining L-terms and p-terms will be same as above. Thus we have

$$\hat{\alpha}^L_{ELF} = \hat{\alpha} + \hat{u}_\alpha\hat{p}_\alpha\hat{\sigma}_{\alpha\alpha} + 0.5(\hat{u}_\alpha\hat{\sigma}_{\alpha\alpha}\hat{\sigma}_{\beta\beta}\hat{L}_{\alpha\beta\beta} + \hat{u}_\alpha\hat{\sigma}^2_{\alpha\alpha}\hat{L}_{\alpha\beta\beta}) \tag{41}$$

and similarly the Bayes estimate for β under LINEX is evaluated by taking $u(\alpha,\beta) = e^{\beta^{-1}}$, $u_\beta = -\frac{e^{\beta^{-1}}}{\beta^2}, u_{\beta\beta} = \frac{e^{\beta^{-1}}}{\beta^3}\left[\frac{1}{\beta} + 2\right], u_{\alpha\beta} = u_{\beta\alpha} = u_{\alpha\alpha} = 0$ and remaining L-terms and p-terms will be same as above. Thus we have

$$\hat{\beta}^L_{ELF} = \hat{\beta} + \hat{u}_\beta\hat{p}_\beta\hat{\sigma}_{\beta\beta} + 0.5(\hat{u}_\beta\hat{\sigma}_{\alpha\alpha}\hat{\sigma}_{\beta\beta}\hat{L}_{\alpha\alpha\beta} + \hat{u}_\beta\hat{\sigma}^2_{\beta\beta}\hat{L}_{\beta\beta\beta}) \tag{42}$$

4.3.2. Bayesian estimation by using uniform prior under different loss functions

• Bayes estimator under squared error loss function

After substitution, the equation (15) reduces like Lindleys integral, therefore, for the Bayes estimates of the parameter α under squared error loss function are,

$$u(\alpha,\beta) = \alpha$$
$$L(\alpha,\beta) = n\ln\alpha + n\ln\beta - (\alpha+1)\sum\ln(1+x_i) + (\beta-1)\sum\ln(1-(1+x_i)^{-\alpha})$$
$$G(\alpha,\beta) = \ln k_1 + \ln k_2$$

It may be verified that,

$$u_\alpha = 1, u_{\alpha\alpha} = u_{\alpha\beta} = u_{\beta\alpha} = u_{\beta\beta} = 0$$

$$p_\alpha = p_\beta = 0$$

$$L_\alpha = \frac{n}{\alpha} - \sum \ln(1+x_i) + (\beta-1)\sum \frac{(1+x_i)^{-\alpha}}{1-(1+x_i)^{-\alpha}} \ln(1+x_i)$$

$$L_{\alpha\alpha} = \frac{-n}{\alpha^2} - (\beta-1)\sum \frac{(1+x_i)^{-\alpha}(\ln(1+x_i))^2}{[1-(1+x_i)^{-\alpha}]^2}$$

$$L_{\alpha\alpha\alpha} = \frac{2n}{\alpha^3} - (\beta-1)\sum \frac{\left((1+x_i)^{3\alpha} - (1+x_i)^{-\alpha}\right)}{[1-(1+x_i)^{-\alpha}]^4}(\ln(1+x_i))^3$$

$$L_{\alpha\beta} = \sum \frac{(1+x_i)^{-\alpha}\ln(1+x_i)}{[1-(1+x_i)^{-\alpha}]} = L_{\beta\alpha}$$

$$L_{\alpha\alpha\beta} = -\sum \frac{(1+x_i)^{-\alpha}(\ln(1+x_i))^2}{[1-(1+x_i)^{-\alpha}]^2} = L_{\beta\alpha\alpha} = L_{\alpha\beta\alpha}$$

$$L_\beta = \frac{n}{\beta} + \sum \ln(1-(1+x_i)^{-\alpha})$$

$$L_{\beta\beta} = \frac{-n}{\beta^2}$$

$$L_{\beta\beta\beta} = \frac{2n}{\beta^3}$$

$$L_{\beta\beta\alpha} = L_{\beta\alpha\beta} = L_{\alpha\beta\beta} = 0$$

If α and β are orthogonal then $\sigma_{ij} = 0$ for $i \neq j$ and $\sigma_{ij} = -\frac{1}{L_{ij}}$ for $i = j$.

After evaluation of all U-terms, L-terms, and p-terms at the point $(\hat{\alpha}, \hat{\beta})$ and using the above expression, the approximate Bayes estimator of α under SELF is,

$$\hat{\alpha}^L_{SELF} = \hat{\alpha} + \hat{u}_\alpha \hat{p}_\alpha \hat{\sigma}_{\alpha\alpha} + 0.5(\hat{u}_\alpha \hat{\sigma}_{\alpha\alpha}\hat{\sigma}_{\beta\beta}\hat{L}_{\alpha\beta\beta} + \hat{u}_\alpha \hat{\sigma}^2_{\alpha\alpha}\hat{L}_{\alpha\beta\beta}) \tag{43}$$

and similarly the Bayes estimate for β under SELF is, $u_\beta = 1, u_{\alpha\alpha} = u_{\alpha\beta} = u_{\beta\alpha} = u_{\beta\beta} = 0$ and remaining L-terms and p-terms will be same as above. Thus we have

$$\hat{\beta}^L_{SELF} = \hat{\beta} + \hat{u}_\beta \hat{p}_\beta \hat{\sigma}_{\beta\beta} + 0.5(\hat{u}_\beta \hat{\sigma}_{\alpha\alpha}\hat{\sigma}_{\beta\beta}\hat{L}_{\alpha\alpha\beta} + \hat{u}_\beta \hat{\sigma}^2_{\beta\beta}\hat{L}_{\beta\beta\beta}) \tag{44}$$

• Bayes estimator under LINEX loss function

The approximate Bayes estimator of α under LINEX is evaluated by taking $u(\alpha, \beta) = e^{-h\alpha}, h > 0$, $u_\alpha = -he^{-h\alpha}, u_{\alpha\alpha} = h^2e^{-h\alpha}, u_{\alpha\beta} = u_{\beta\alpha} = u_{\beta\beta} = 0$ and remaining L-terms and p-terms will be same as above. Thus we have

$$\hat{\alpha}^L_{LINEX} = \hat{\alpha} + \hat{u}_\alpha \hat{p}_\alpha \hat{\sigma}_{\alpha\alpha} + 0.5(\hat{u}_\alpha \hat{\sigma}_{\alpha\alpha}\hat{\sigma}_{\beta\beta}\hat{L}_{\alpha\beta\beta} + \hat{u}_\alpha \hat{\sigma}^2_{\alpha\alpha}\hat{L}_{\alpha\beta\beta}) \tag{45}$$

and similarly the Bayes estimate fir β under LINEX is evaluated by taking $u(\alpha, \beta) = e^{-h\beta}, h > 0$, $u_\beta = -he^{-h\beta}, u_{\beta\beta} = h^2e^{-h\beta}, u_{\alpha\beta} = u_{\beta\alpha} = u_{\alpha\alpha} = 0$ and remaining L-terms and p-terms will be same as above. Thus we have

$$\hat{\beta}^L_{LINEX} = \hat{\beta} + \hat{u}_\beta \hat{p}_\beta \hat{\sigma}_{\beta\beta} + 0.5(\hat{u}_\beta \hat{\sigma}_{\alpha\alpha}\hat{\sigma}_{\beta\beta}\hat{L}_{\alpha\alpha\beta} + \hat{u}_\beta \hat{\sigma}^2_{\beta\beta}\hat{L}_{\beta\beta\beta}) \tag{46}$$

• Bayes estimator under entropy loss function

The approximate Bayes estimator of α under ELF is evaluated by taking $u(\alpha,\beta) = e^{\alpha^{-1}}$, $u_\alpha = -\frac{e^{\alpha^{-1}}}{\alpha^2}, u_{\alpha\alpha} = \frac{e^{\alpha^{-1}}}{\alpha^3}\left[\frac{1}{\alpha}+2\right], u_{\alpha\beta} = u_{\beta\alpha} = u_{\beta\beta} = 0$ and remaining L-terms and p-terms will be same as above. Thus we have

$$\hat{\alpha}^L_{ELF} = \hat{\alpha} + \hat{u}_\alpha\hat{p}_\alpha\hat{\sigma}_{\alpha\alpha} + 0.5(\hat{u}_\alpha\hat{\sigma}_{\alpha\alpha}\hat{\sigma}_{\beta\beta}\hat{L}_{\alpha\beta\beta} + \hat{u}_\alpha\hat{\sigma}^2_{\alpha\alpha}\hat{L}_{\alpha\beta\beta}) \tag{47}$$

and similarly the Bayes estimate for β under LINEX is evaluated by taking $u(\alpha,\beta) = e^{\beta^{-1}}$, $u_\beta = -\frac{e^{\beta^{-1}}}{\beta^2}, u_{\beta\beta} = \frac{e^{\beta^{-1}}}{\beta^3}\left[\frac{1}{\beta}+2\right], u_{\alpha\beta} = u_{\beta\alpha} = u_{\alpha\alpha} = 0$ and remaining L-terms and p-terms will be same as above. Thus we have

$$\hat{\beta}^L_{ELF} = \hat{\beta} + \hat{u}_\beta\hat{p}_\beta\hat{\sigma}_{\beta\beta} + 0.5(\hat{u}_\beta\hat{\sigma}_{\alpha\alpha}\hat{\sigma}_{\beta\beta}\hat{L}_{\alpha\alpha\beta} + \hat{u}_\beta\hat{\sigma}^2_{\beta\beta}\hat{L}_{\beta\beta\beta}) \tag{48}$$

5. Simulation study

Next, a simulation study was conducted to investigate the performance of Bayes estimators of the unknown parameter for case I, i,e, when α is known and β is unknown under two priors discussed in this paper. The study was executed for different sample sizes specifically for n= 20,50,70,100,150,200. The observations were generated from Inverted Kumaraswamy distribution using the quantile function. For the expression of quantile function refer Al-Fattah *et al.* (2017). The Bayes estimates were obtained using LINEX, SELF and Entropy loss function. For Gamma prior the values of hyperparameters considered are (a=0.01, b=0.01). In our study 6000 samples were generated. The Bayes estimates were compared in terms of relative mean square errors.

6. Conclusion

In this paper, we estimated the unknown parameter of IKum distribution considering three different cases: (i)when α is known and β is unknown, (ii)when α is unknown and β is known, and (iii)when α and β both are unknown, using two prior distributions under three different loss functions, though simulations were carried out for caseI only. Relative MSE were also derived using the following formula:

$$MSE = \frac{\sum_{i=1}^{N}(Estimator - Truevalue)^2}{N}$$

$$RelativeMSE = \frac{MSE}{Truevalue}$$

where $N = 6000$.

Concluding Remarks

$*$ From Table 1, the Relative MSE of estimator under Entropy loss function is minimum for $\alpha = 0.4$, $\beta = 0.8$ and the Relative MSE of estimator under Squared error loss funtion is minimum for $\alpha = 0.5$, $\beta = 1$.

$*$ From Table 2 also, estimator under Entropy loss function stands most efficient for $\alpha = 0.4$, $\beta = 0.8$, whereas for $\alpha = 0.5$, $\beta = 1$, estimator under Squared error loss function holds minimum relative error.

∗ From the graphs it is observed that Relative MSE decreases as sample size increases.

As a future research work, this paper can be extended in many ways. The simulation results for the other two cases(when α is unknown and β is known, and when α and β both are unknown) can also be computed and the obtained results can be represented in the form of graphs. The obtained estimators can also be applied to real life data for illustrative purposes. Additionally, the estimation of entropy in this setup may also be considered.

Acknowledgements

I am indeed grateful to the Editors for their guidance and counsel. I am very grateful to the reviewer for valuable comments and suggestions of generously listing many useful references.

Conflict of interest

The authors do not have any financial or non-financial conflict of interest to declare for the research work included in this article.

References

Afaq, S. P., Ahmad, and Ahmed, A. (2015). Preference of priors for the exponentiated exponential distribution under different loss functions. *International Journal of Modern Mathematical Sciences*, **13**, 307–321.

Ahmad, S. P. and Fatima K. (2017). Preference of priors for the generalized inverse Rayleigh distribution under different loss functions. *Journal of Statistics Applications and Probability Letters*, **4**, 73–90.

Aldahlan, M. A., Bakoban, R. A., and Alzahrani, L. S. (2022). On estimating the parameters of the Beta inverted exponential distribution under type-II censored samples. *Mathematics*, **10**, 506–542.

Al-Fattah, A. A., E-Helbawy, A. A., and AL-Dayian G. R. (2017). Inverted Kumaraswamy distribution: Properties and estimation. *Pakistan Journal of Statistics*, **33**, 37–61.

Dey, S. (2012). Bayesian estimation of the parameter and reliability function of an inverse Rayleigh distribution. *Malaysian Journal of Mathematical Sciences*, **6**, 113–124.

El-Din, M. M. and Abu-Moussa, M. (2018). On estimation and prediction for the inverted Kumaraswamy distribution based on general progressive censored samples. *Pakistan Journal of Statistics and Operation Research*, **14**, 717–736.

Fan, J. and Gui, W. (2022). Statistical inference of inverted exponentiated Rayleigh distribution under joint progressively type-II censoring. *Entropy*, **24**, 171.

Flaih, A., Elsalloukh, H., Mendi E., and Milanova, M. (2012). The exponentiated inverted Weibull distribution.*Applied Mathematics and Information Sciences*, **6**, 167–171.

Hasan, M. R. and Baizid, A. R. (2016). Bayesian estimation under different loss functions using Gamma prior for the case of exponential distribution. *Jouranl of Scientific Research*, **9**, 67–78.

Iqbal, Z., Tahir, M. M., Riaz, N., Ali, S. A., and Ahmad, M. (2017). Generalized inverted Kumaraswamy distribution: Properties and application. *Open Journal of Statistics*, **7**, 645-662.

Kazmi, A. M. S., Aslam, M., and Ali. S. (2012). Preference of prior for the class of lifetime distributions under different loss functions. *Pakistan Journal of Statistics*, **28**, 467–487.

Kumaraswamy, P. (1980). A generalized probability density function for double bounded random processes. *Journal of Hydrology*, **46**, 79–88.

Prakash, G. (2012). Inverted exponential distribution under a Bayesian viewpoint. *Journal of Modern Applied Statistical Methods*, **11**, 190–202.

Reshi, J. A., Ahmed, A., and Ahmad, S. P. (2014). Bayesian analysis of scale parameter of the generalized inverse Rayleigh model using different loss functions. *International Journal of Modern Mathematical Sciences*, **10**, 151–162.

Rastogi, M. K. and Oguntunde, P. E. (2019). Classical and Bayes estimation of reliability characteristics of the Kumaraswamy-inverse exponential distribution. *International Journal of System Assurance Engineering and Management*, **10**, 190–200.

Sana, Faizan, M., and Khan, A. A. (2023). Bayesian estimation using Lindley's approximation of inverted Kumaraswamy distribution based on lower record values. *TWMS Journal of Applied and Engineering Mathematics*, **13**, 65–73.

Tiao, G. G. and Cuttman, I. (1965). The inverted Dirichlet distribution with applications, *Journal of the American Statistical Association*, **60**, 793–805.

Tummala, V. M. and Sathe, P. T. (1978). Minimum expected loss estimators of reliability and parameters of certain life time distributions. *IEEE Transactions on Reliability*, **27**, 283–285.

ANNEXURE

Table 1: Bayes estimate and Relative mean square error under Gamma prior when hyperparameters (a,b)=(0.01,0.01)

Case I: α known, β unknown							
n		$\alpha = 0.4, \beta = 0.8$			$\alpha = 0.5, \beta = 1$		
		LINEX	SELF	ELF	LINEX	SELF	ELF
20	Estimate	0.83308	0.84414	0.80328	1.03981	1.0529	0.99961
	RelMSE	0.04870	0.05210	0.04490	0.05910	0.06280	0.05390
50	Estimate	0.81030	0.81922	0.80116	1.01620	1.02001	1.00010
	RelMSE	0.01730	0.01750	0.01660	0.02080	0.02140	0.02060
70	Estimate	0.80934	0.81081	0.80090	1.0109	1.0125	0.99918
	RelMSE	0.01230	0.01260	0.01210	0.01510	0.01470	0.01460
100	Estimate	0.80651	0.80865	0.80010	1.0205	0.99102	1.00040
	RelMSE	0.00824	0.00855	0.00802	0.01100	0.01010	0.01020
150	Estimate	0.80458	0.80526	0.79951	1.0141	0.99454	1.00070
	RelMSE	0.00538	0.00550	0.00528	0.00715	0.00665	0.00696
200	Estimate	0.80325	0.80363	0.80090	1.0081	0.99723	0.99930
	RelMSE	0.00414	0.00417	0.00400	0.00530	0.00498	0.00506

Table 2: Bayes estimate and Relative mean square error under Uniform prior

Case I: α known, β unknown							
n		$\alpha = 0.4, \beta = 0.8$			$\alpha = 0.5, \beta = 1$		
		LINEX	SELF	ELF	LINEX	SELF	ELF
20	Estimate	0.87870	0.88261	0.84432	1.1579	1.00160	1.0559
	RelMSE	0.06090	0.06200	0.05420	0.10100	0.05380	0.06670
50	Estimate	0.83057	0.82968	0.81540	1.0619	0.99932	1.0178
	RelMSE	0.01880	0.01900	0.01780	0.02740	0.02050	0.02200
70	Estimate	0.82186	0.82366	0.81270	1.0430	0.99687	1.0147
	RelMSE	0.01300	0.01340	0.01250	0.01780	0.01500	0.01570
100	Estimate	0.81395	0.81584	0.80961	1.0315	1.00010	1.0089
	RelMSE	0.00858	0.00860	0.00811	0.00907	0.01190	0.01020
150	Estimate	0.80922	0.81043	0.80556	1.0205	1.00060	1.0073
	RelMSE	0.00570	0.00555	0.00544	0.00738	0.00668	0.00690
200	Estimate	0.80754	0.80732	0.80328	1.0154	0.99883	1.0057
	RelMSE	0.00418	0.00435	0.00416	0.00541	0.00484	0.00519

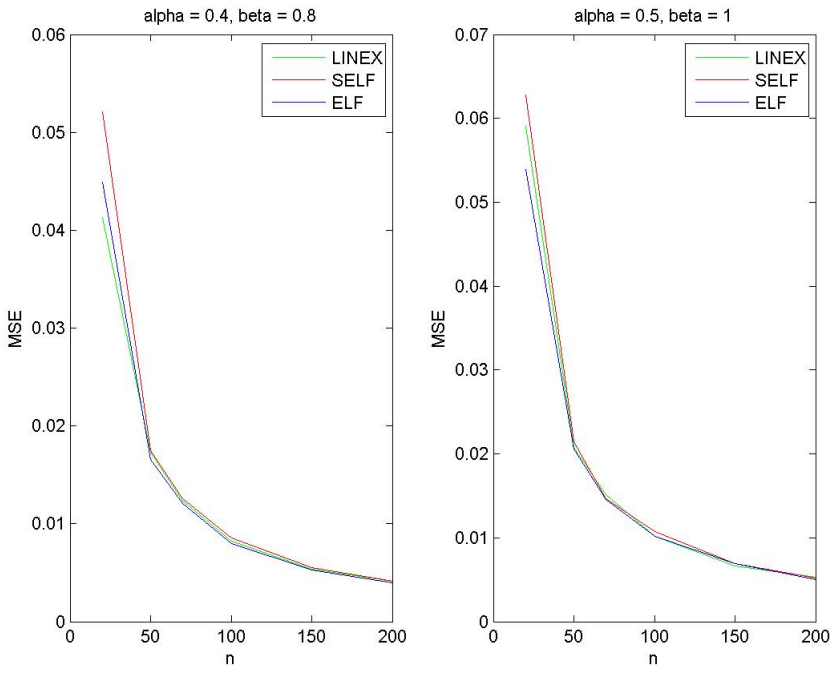


Figure 1: Relative Mean Square Error of β under Gamma prior

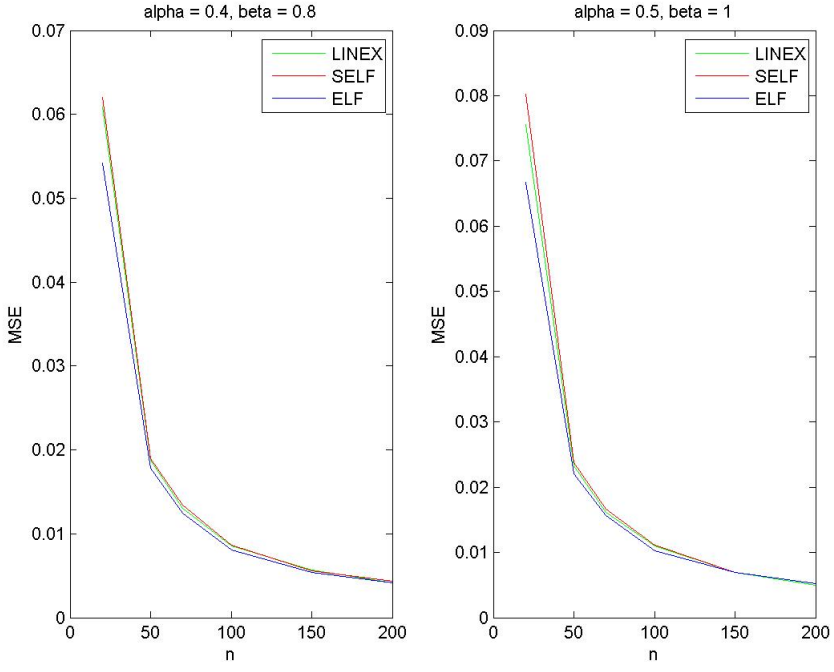


Figure 2: Relative Mean Square Error of β under Uniform prior

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 23, No. 1, 2025 (New Series), pp 303–320
https://www.ssca.org.in/journal

Determining Optimal Threshold and Some Inferential Procedures for a Skewed ROC Model in the Binary Classification Framework

Sandhya Singh[1], Saebugari Balaswamy[1] and R.Vishnu Vardhan[2]
[1]*Department of Statistics*
Indira Gandhi National Tribal University, Amarkantak, Madhya Pradesh, India
[2]*Department of Statistics, Pondicherry University, Puducherry*

Received: 31 May 2024; Revised: 20 June 2024; Accepted: 25 June 2024

Abstract

ROC curve is a useful tool in the assessment of the performance of a diagnostic test over the range of possible values of a predictor variable and the sensitivity, specificity, optimal threshold and Area under the curve (AUC) are its intrinsic measures to know the accuracy of the diagnostic test. The area under the curve is a measure of accuracy which provides the extent of correct classification of the test and also it is a measure of discrimination to compare the performance of two or more diagnostic tests. Further, the optimal threshold is a cut-off, which discriminates the populations into one of the two groups with a maximum of accurate accuracy. The Youden's Index method is the usual approach to identify the optimal threshold. The alternate approaches to compute the optimal threshold have been provided in this paper when the data is of skewed nature in the ROC context. For this purpose, ROC model is considered to show how the discriminatory ability of a test changes on changing the location and scale parameters by using a generalized half normal distribution. Further, the simulation studies are conducted to study the proposed methodology and also compared with the existing ROC models using both simulations and real datasets.

Key words: ROC Curve; Sensitivity; Specificity; AUC; Index of union; Concordance of probability.

AMS Subject Classifications: 62P10.

1. Introduction

The ROC curve was first developed by radar engineers during World War II for truly detecting enemy objects in battle fields starting in 1941 which led to its name "Receiver Operating Characteristic" (ROC) curve. Now-a-days, this technique is being extensively used in diverse areas of research such as banking, Finance, Engineering, Machine learning and Medical Sciences, *etc.* ROC curve was introduced in medicine for analysis of radiographic images (Lusted, 1971). This is an important tool applied in classification problems mostly

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

associated with evaluating the performance of the diagnostic test(s) by means of the accuracy or sensitivity measures, also to provide accuracy of the classifier/diagnostic test and helps in determining the optimal cut-off of a diagnostic test or classifier. To define ROC curve, there is a need of two intrinsic measures, such as, Sensitivity (True Positive Rate, TPR), which is the probability of a positive test result conditioned on the individual truly being positive and Specificity (True Negative Rate, TNR), which is the probability of a negative test result conditioned on the individual truly being negative. Graphically, the ROC curve can be achieved by using 1-TNR on x-axis and TPR on y-axis, resulting a smooth curve. This smooth curve is embedded with various threshold points; we tend to choose such threshold that attains minimum distance from the chance line. Though this approach is heuristic, there are other established indices that helps in determining the optimal threshold, one such index is the Youden's Index. The portion under the ROC curve is termed as the area under the curve (AUC), theoretically lies between 0 and 1. In a practical point of view, it is interpreted as that higher value of AUC indicates that the performance of marker/diagnostic test is better. Further, a test's AUC should not lie below or close to 0.5, this result in random classification and test is not considered for classification. Even though this technique's role is to classify the subjects into one of the predefined groups, it also allows allocating the new subjects into one of those groups with a proper status label. Further, much theoretical work has been done in the ROC context using different distributional assumptions and the formal statistical definition of ROC curve in terms of cumulative distribution function (CDF) is

$$ROC = 1 - G\left(F^{-1}(1-t)\right), 0 < t < 1$$

Here, F and G are the CDFs of two independent populations and the ROC model so generated is referred to as bi-distributional ROC model. The test score derived from a marker or diagnostic test do have some pattern and follows a particular distribution, then the ROC curve be developed based on that particular distribution, by which one can gets the proper fit of the data, and appropriate results with interpretation. In ROC literature, many models have been proposed based upon bi-distributional assumption such as Bi-lognormal (Dorfman and Alf, 1968, 1969), Bi-normal (Egan, 1975), Bi-gamma (Hussain, 2012), Bi-beta (Zou *et al.*, 1997), Bi-exponential (Tang and Balakrishnan, 2011), Hybrid ROC models (Balaswamy *et al.*, 2015) and many more. In the recent past, estimation of area under the ROC, for non-normal data (Balaswamy and Vardhan, 2022), Bi-Generalised Exponential ROC curve (Balaswamy and Vardhan, 2023), area under the ROC Curve in the framework of gamma mixtures (Arunima and Vishnu Vardhan, 2022), area under the multi-class ROC statistics and applications for non-normal data (Arunima and Vishnu Vardhan, 2023) are few to cite in the ROC framework.

This paper focuses on different procedure to obtain an optimal threshold other than the method of Youden's index. This provides the better and easiest way of calculating the optimal threshold. In order to demonstrate this methodology, a new ROC model is considered based upon a skewed distribution. To illustrate this skewed nature, let us consider a practical illustration. In assessing the subject's life status (alive or dead), a marker by name *Acute Physiology and Chronic Health Evaluation (APACHE II)* will be used. Mostly, the APACHE II score do not satisfy the normality assumption and possesses a skewed pattern. In such case, the conventional bi-normal ROC model may not suitable to assess the performance and threshold of APACHE II. So, there is a need to find a suitable statistical distribution that can meet the requirements of ROC model. Another marker that has similar kind of

non-normality is the *Simplified Acute Physiology Score (SAPS III)*. Hence, there is a need to look into the influence of measures of location, scale and shape to model a newer version of ROC. The present work addresses the above practical situations using the data of APACHE II and SAPS III. Along with these, simulations are also carried out to demonstrate the worst, moderate and better classification scenarios from the proposed ROC model. It is understood that these two datasets follow Generalized Half Normal distribution and for comparison purpose, two other distributions namely, the Normal and the Half-Normal are also considered.

2. Methodology

Let $(x_1, x_2) \in S$ be the test scores which are observed in healthy(0) and diseased (1) populations respectively. It is assumed that '0' and '1' population follow Generalized Half Normal Distribution with $\alpha > 0, \sigma > 0$ as shape and scale parameters, respectively. The probability density function and cumulative distribution function of Generalized Half Normal Distribution are given as follows:

$$g(x, \alpha, \sigma) = \sqrt{\frac{2}{\pi}} \left(\frac{\alpha}{x}\right) \left(\frac{x}{\sigma}\right)^{\alpha} exp\left(-\frac{1}{2}\left(\frac{x}{\sigma}\right)^{2\alpha}\right) \; ; x \geq 0$$

$$G(x, \alpha, \sigma) = 1 - 2\Phi\left[-\left(\frac{x}{\sigma}\right)^{\alpha}\right]$$

where $\Phi(\cdot)$ is the c.d.f. of the standard normal distribution. As the ROC curve is a trade-off between False Positive Rate (FPR) and True Positive Rate (TPR). Therefore, the FPR is derived by using probabilistic definition as follows

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

on further simplification, the expression for t can be obtained as

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

where $\Phi^{-1}(\cdot)$ is the inverse cumulative standard normal distribution function. Similarly, TPR expression is derived by using its probabilistic definition as follows

$$TPR = y(t) = P(S > t|1) = 2\left[1 - \Phi\left(\frac{t}{\sigma_1}\right)^{\alpha_1}\right] \tag{3}$$

substituting (2) in (3),

$$y(t) = 2\left[1 - \Phi\left(\left(\frac{\sigma_0}{\sigma_1}\right)^{\alpha_1} \left[\Phi^{-1}\left(1 - \frac{x(t)}{2}\right)\right]^{\frac{\alpha_1}{\alpha_0}}\right)\right]$$

Let, $\Phi^{-1}\left(1 - \frac{x(t)}{2}\right) = Z_x$ and on further simplification,

$$y(t) = 2\left[1 - \Phi\left(\left(\frac{\sigma_0}{\sigma_1}\right)^{\alpha_1} [Z_x]^{\frac{\alpha_1}{\alpha_0}}\right)\right]$$

Let, $\beta = \frac{\sigma_0}{\sigma_1}$ and $\alpha = \frac{\alpha_1}{\alpha_0}$. Then

$$y(t) = 2\left[1 - \Phi\left(\beta^{\alpha_1}\left[Z_x\right]^{\frac{\alpha_1}{\alpha_0}}\right)\right] \tag{4}$$

on further simplification, the expression for ROC curve is

$$y(t) = 1 - erf\left(\frac{\beta^{\alpha_1}\left[Z_x\right]^{\alpha}}{\sqrt{2}}\right) \tag{5}$$

The expression in (5) can be referred to as Generalized Half Normal ROC curve. In ROC methodology AUC measures the entire two dimensional area underneath the ROC curve.

$$AUC = \int_0^1 y(t)\mathrm{d}t$$

$$AUC = \int_0^1 1 - erf\left(\frac{\beta^{\alpha_1}\left[\Phi^{-1}\left(1 - \frac{x(t)}{2}\right)\right]^{\alpha}}{\sqrt{2}}\right)\mathrm{d}x(t) \tag{6}$$

The above expression has no closed form solution; therefore it needs to be evaluated numerically. The numerical evaluations have been carried out using Simpson's method in the results section. Let $\alpha = 1, i.e. \alpha_1 = \alpha_0 = 1$ in equation (5) and on further simplification,

$$AUC = 2 - 2\left[\Phi\left(\left(\frac{\sigma_0}{\sigma_1}\right)\left[\Phi^{-1}\left(1 - \frac{x(t)}{2}\right)\right]\right)\right] \tag{7}$$

The equation (7) is known as ROC curve for Half Normal Distribution (HN ROC curve) and the AUC for the HN ROC is given by

$$AUC = 1 - \frac{2}{\pi}\left(\frac{\sigma_0}{\sigma_1}\right) \tag{8}$$

3. Optimal threshold

The optimal threshold is very important in classification to obtain the good accuracy and to minimize the misclassification rate. Therefore, the four different methods to determine the optimal threshold that are in this paper are as follows.

Youden's index (J): This Index is a single statistic that captures the performance of a dichotomous diagnostic test. J is a function of sensitivity and specificity, such that

$$J(c) = \{Sensitivity(c) + Specificity(c) - 1\}$$

Over all cut point c; "optimal t" denotes the cut-point corresponding to J. When the value of J is maximum, optimal t is the optimum cut point value.

The closest to (0,1) criterion (ER): In this criteria, the optimal cut point is defined as the point closest to the point (0, 1) on the ROC curve.

$$ER(c) = \sqrt{(1 - TPR(c))^2 + (FPR(c))^2}$$

Mathematically, the point C_{ER} minimising the ER(c) function is called the optimal cut point value.

Concordance probability method (CZ): The concordance probability method defines the optimal cut point as the point maximizing the product of sensitivity and specificity.

$$CZ(c) = TPR(c) \times TNR(c)$$

The product gets value between 0 and 1. The concordance probability of dichotomized measure at cut point c can be expressed as the area of a rectangle associated with the ROC curve. Cut point $\hat{c_z}$ maximizing CZ(c) actually maximizes the area of the rectangle.

Index of union (IU): The optimal cut point should be chosen as the point which classifies most of individuals correctly and thus least of them incorrectly. From this point of view, Ilker (2017) proposed the index of union (IU) method to obtain the optimal threshold. This method provides an "optimal" cut point which has maximum sensitivity and specificity values at the same time. In order to find the highest sensitivity and specificity values at the same time, the AUC value is taken as the starting value of them. The above criteria correspond to the following equation,

$$IU(c) = (|TPR(c) - AUC| + |TNR(c) - AUC|)$$

The cut-point optimal t which minimizes the IU(C) function and the $|TPR(c) - TNR(c)|$ difference will be "optimal" cut point value.

Among these four methods of optimal threshold identification, choosing a one optimal threshold with good accuracy is a question. In order to answer this, Ilker (2017) compared these four methods with the mathematical optimal threshold (equating both density curves of healthy/ normal and diseased/abnormal populations and solve for the threshold). But this is not possible in all the cases of distributions, just like the case of proposed GHN ROC curve, here the closed form solution for the threshold is not possible. Therefore, keeping this in mind, we have used TPR value and their corresponding specificity values are considered to be higher. Wherever, these values are higher, that particular threshold will be of good choice with greater accuracy. Further, these four methods are tested at various sample sizes and different classification scenarios. In the next subsection, the inferential aspects of proposed ROC curve are discussed. For which, the variance of AUC is estimated through bootstrapping method as follows.

3.1. Bootstrap estimate of AUC

Since there is no closed form for AUC, its variance can be obtained using bootstrap technique. Let 'B' be the number of bootstraps obtained from the data with the sample sizes n_0 and n_1 respectively from normal and abnormal populations. Then the bootstrapped AUC estimate and its variance are given as

$$\widehat{AUC_B} = \frac{1}{B}\sum_{b=1}^{B} AUC_b \tag{9}$$

$$Var\left(\widehat{AUC_B}\right) = \frac{1}{B-1}\sum_{b=1}^{B}\left(AUC_b - \widehat{AUC}\right)^2 \tag{10}$$

3.2. Confidence intervals for AUC

Let $\widehat{AUC}$ denote the sample AUC value. For large samples, the distribution of AUC is approximately normal. Hence, a $100(1-\alpha)\%$ confidence interval for AUC may be computed using the standard normal distribution as follows

$$\widehat{AUC_B} \pm Z_{\frac{\alpha}{2}} \sqrt{Var\left(\widehat{AUC_B}\right)} \tag{11}$$

where $Z_{\frac{\alpha}{2}}$ is the $\frac{\alpha}{2}$ standard normal percentile.

3.3. Test statistic

A test with $AUC_0 = 0.5$ is considered useless as it classifies only 50% of individuals correctly. For this test, the ROC curve coincides with the chance line and $TPR = FPR$. Hence, the null and alternative hypothesis are defined as $H_0 : AUC = AUC_0 and H_1 : AUC > AUC_0$. Then the test statistic is defined as

$$Z = \frac{\widehat{AUC_B} - AUC_0}{\sqrt{Var\left(\widehat{AUC_B}\right)}} \tag{12}$$

The next subsection deals with the construction of confidence intervals for the proposed ROC Curve to explain the variability of the curve at each and every threshold value.

3.4. Confidence intervals for FPR and TPR

The $100(1-\alpha)\%$ confidence intervals for FPR and TPR, which in turn help in producing the confidence interval for GHN ROC curve. Therefore, the $100(1-\alpha)\%$ confidence intervals for FPR and TPR are as follows,

$$\widehat{FPR} \pm Z_{\frac{\alpha}{2}} \sqrt{Var(\widehat{FPR})}; \qquad \widehat{TPR} \pm Z_{\frac{\alpha}{2}} \sqrt{Var(\widehat{TPR})}$$

where variance of false positive rate and true positive rate are estimated through Delta method. The expression for $Var(\widehat{FPR}$ and $Var(\widehat{TPR}$ are

$$Var(\widehat{FPR}) = \left(\frac{\partial FPR}{\partial \sigma_0}\right)^2 Var(\hat{\sigma}_0) + \left(\frac{\partial FPR}{\partial \alpha_0}\right)^2 Var(\hat{\alpha}_0) \tag{13a}$$

$$Var(\widehat{TPR}) = \left(\frac{\partial TPR}{\partial \sigma_1}\right)^2 Var(\hat{\sigma}_1) + \left(\frac{\partial TPR}{\partial \alpha_1}\right)^2 Var(\hat{\alpha}_1) \tag{13b}$$

Further, the partial differentiations of FPR and TPR with respect to their parameters are as follows,

$$\begin{aligned}
\frac{\partial FPR}{\partial \sigma_0} &= \frac{\partial}{\partial \sigma_0}\left\{2\left[1-\Phi\left[\left(\frac{t}{\sigma_0}\right)^{\alpha_0}\right]\right]\right\}\\
\frac{\partial FPR}{\partial \sigma_0} &= \frac{2\alpha_0 t^{\alpha_0}}{\sigma_0^{\alpha_0+1}}\phi\left(\frac{t}{\sigma_0}\right)^{\alpha_0}\\
\frac{\partial FPR}{\partial \alpha_0} &= \frac{\partial}{\partial \alpha_0}\left\{2\left[1-\Phi\left[\left(\frac{t}{\sigma_0}\right)^{\alpha_0}\right]\right]\right\}\\
\frac{\partial FPR}{\partial \alpha_0} &= -2\phi\left(\frac{t}{\sigma_0}\right)^{\alpha_0}\left(\frac{t}{\sigma_0}\right)^{\alpha_0}\log\left(\frac{t}{\sigma_0}\right)\\
\frac{\partial TPR}{\partial \sigma_1} &= \frac{\partial}{\partial \sigma_1}\left\{2\left[1-\Phi\left[\left(\frac{t}{\sigma_1}\right)^{\alpha_1}\right]\right]\right\}\\
\frac{\partial TPR}{\partial \sigma_1} &= \frac{2\alpha_1 t^{\alpha_1}}{\sigma_1^{\alpha_1+1}}\phi\left(\frac{t}{\sigma_1}\right)^{\alpha_1}\\
\frac{\partial TPR}{\partial \alpha_1} &= \frac{\partial}{\partial \alpha_1}\left\{2\left[1-\Phi\left[\left(\frac{t}{\sigma_1}\right)^{\alpha_1}\right]\right]\right\}\\
\frac{\partial TPR}{\partial \alpha_1} &= -2\phi\left(\frac{t}{\sigma_1}\right)^{\alpha_1}\left(\frac{t}{\sigma_1}\right)^{\alpha_1}\log\left(\frac{t}{\sigma_1}\right)
\end{aligned}$$

Now, by substituting the above expressions in equations (13a) and (13b), we obtain the variances of FPR and TPR as,

$$\begin{aligned}
Var\left(\widehat{FPR}\right) &= \left[\frac{2\alpha_0 t^{\alpha_0}}{\sigma_0^{\alpha_0+1}}\phi\left(\frac{t}{\sigma_0}\right)^{\alpha_0}\right]^2 Var(\hat{\sigma}_0)\\
&\quad + \left[-2\phi\left(\frac{t}{\sigma_0}\right)^{\alpha_0}\left(\frac{t}{\sigma_0}\right)^{\alpha_0}\log\left(\frac{t}{\sigma_0}\right)\right]^2 Var(\hat{\alpha}_0) \qquad (14a)\\
Var\left(\widehat{TPR}\right) &= \left[\frac{2\alpha_1 t^{\alpha_1}}{\sigma_1^{\alpha_1+1}}\phi\left(\frac{t}{\sigma_1}\right)^{\alpha_1}\right]^2 Var(\hat{\sigma}_1)\\
&\quad + \left[-2\phi\left(\frac{t}{\sigma_1}\right)^{\alpha_1}\left(\frac{t}{\sigma_1}\right)^{\alpha_1}\log\left(\frac{t}{\sigma_1}\right)\right]^2 Var(\hat{\alpha}_1) \qquad (14b)
\end{aligned}$$

The variances of the parameters can be estimated through their asymptotic distributions, but in the present context the maximum likelihood estimators of the Generalized Half Normal distribution do not have closed form expressions. Therefore, the maximum likelihood parameters of these distributions can be obtained by direct maximization of log-likelihood function using the Newton-Raphson method in R. The asymptotic variances of the parameters are estimated using the Bootstrap method. Hence, the bootstrapped estimates of σ_0&α_0

and their variance are

$$\hat{\sigma}_0 = \frac{1}{B}\sum_{b=1}^{B}\sigma_{0b}$$
$$Var\left(\hat{\sigma}_0\right) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\sigma_{0b} - \hat{\sigma}_0\right)^2$$
$$\hat{\alpha}_0 = \frac{1}{B}\sum_{b=1}^{B}\alpha_{0b}$$
$$Var\left(\hat{\alpha}_0\right) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\alpha_{0b} - \hat{\alpha}_0\right)^2$$

In a similar manner, we can obtain the bootstrap estimate of $\sigma_1 \& \alpha_1$ as follows,

$$\hat{\sigma}_1 = \frac{1}{B}\sum_{b=1}^{B}\sigma_{1b}$$
$$Var\left(\hat{\sigma}_1\right) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\sigma_{1b} - \hat{\sigma}_1\right)^2$$
$$\hat{\alpha}_1 = \frac{1}{B}\sum_{b=1}^{B}\alpha_{1b}$$
$$Var\left(\hat{\alpha}_1\right) = \frac{1}{B-1}\sum_{b=1}^{B}\left(\alpha_{1b} - \hat{\alpha}_1\right)^2$$

Now, using the above variances for the parameters of Generalized Half Normal distribution along with equations (14a) and (14b), the confidence intervals for FPR and TPR are obtained. By using these confidence intervals, the confidence interval lines can be plotted along with the GHN ROC curve to show the variability of the proposed ROC Curve at each and every point on the ROC space.

In the next section, the results are carried out using simulation studies and real datasets to explain the proposed methodology and the confidence intervals are also evaluated for the summary measure AUC and the proposed ROC Curve.

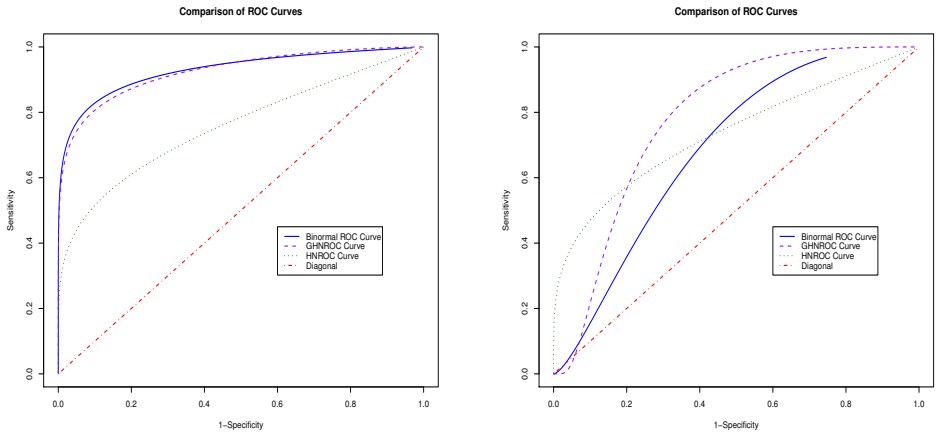
4. Results and discussions

Different simulation studies have been carried out to study the behaviour of the proposed ROC curve and also compared with the existing ROC models in literature. In this results and discussions sections, there are different subsections which will explain the necessity and importance of the proposed ROC curve in detail. The results reported in the tables are given in the appendix.

4.1. Comparison of ROC Curves - simulated datasets

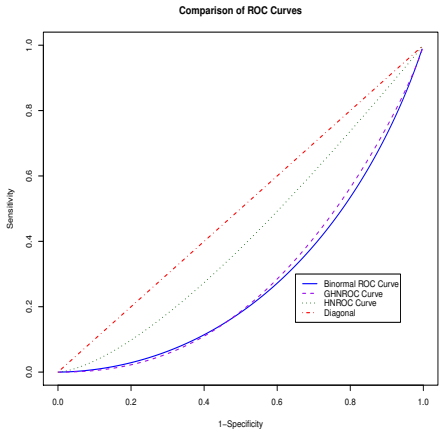
In this section different situations (Better, Moderate and Worst cases) of simulation studies in classification are considered and the results are given in Table 1, which consists of

optimal threshold, AUC, J and One sample KS test for testing the reliability of the simulated data (from GHN distribution) with GHN, Half Normal and Normal distributions. The GHN ROC model is compared with the existing ROC models like HN ROC and Binormal ROC model in all the three different situations of classification.



(a) Better case of Classification

(b) Moderate case of Classification



(c) Worst case of Classification

Figure 1: Comparison of ROC curves for different cases of classification using simulated datasets

Table 1 shows the differences and the importance of proposed GHN ROC model as compared to the existing ROC models with different simulation studies. The first case is better case of classification and the accuracy measure AUC is found to be 90% for the GHN ROC Curve when the data of both populations follows a generalized half normal distribution (KS test statistic values are given in the table). Whereas, when the shape parameter is suppressed, the proposed ROC model reduces to the half normal ROC model and this case has an accuracy of 74% and the data follows half normal distribution. The interesting fact observed is that even though the data of healthy ($D = 0.5631, p - value < 2.2e^{-16}$) and diseased ($D = 0.8461, p - value < 2.2e^{-16}$) populations do not follow the normal distribution, the accuracy is found to be 92%. This means that the Binormal ROC model is over estimating

the accuracy when the data does not satisfy the distributional properties. This is the reason that one must check for the distributional assumptions when you have the data in hand first (This type of situation can be seen in the next section with APACHE II score). Further, the corresponding ROC Curves are drawn in Figure 1a with better accuracy of classification where the curves are nearer to the top right corner of the ROC plot.

The moderate case of classification is considered (Table 1) and the GHN ROC curve (78% of accuracy) is clearly superior than the other two models half normal (72%) and Binormal ROC models (68%). Here also, the distributional properties are verified with the help of KS test statistic and found that when the data follows generalized half normal distribution, the accuracy is higher than the other two models when the data do not follow normality. Similar kind of phenomenon can be seen in Figure 1b, where the curves explain the moderate case of classification.

Finally, the worst case of classification is also considered where the parameters have the higher values in healthy population than the diseased population and the results are placed in Table 1 and Figure 1c. In this experiment also, it is found that the proposed ROC model is better than the Binormal ROC model when the data deviates from normality ($Healthy:\ D = 0.7160,\ p-value < 2.2e^{-16}$ & $Diseased:\ D = 0.6176,\ p-value < 2.2e^{-16}$).

Further, the optimal threshold, Youden's index,false positive rate and true positive rate at the corresponding optimal threshold are also computed and depicted in the Table 1. The optimal threshold is the value or score which divides the data into one of the two possible cases with a good amount of accuracy with lesser misclassification rate. These are computed for all the cases of classification along with the FPR and TPR at that particular optimal threshold.

4.2. Comparison of ROC Curves - real dataset

In this section, two real datasets are used to illustrate the proposed methodology and comparison is made with the existing ROC models and the results are as follows. The APACHE II (Acute Physiology and chronic Health Evaluation II) and SAPS III (Simplified Acute Physiology Score) datasets are considered to explain the proposed methodology and its significance over other ROC models like Binormal and HNROC models. The Tables 2 &

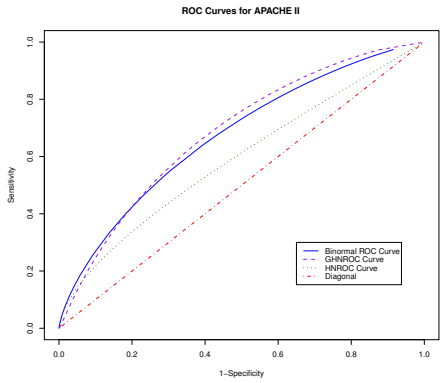


Figure 2: Comparison of ROC curves - APACHE II dataset

3 consists of optimal threshold, FPR, TPR, J and AUC along with the KS test statistics and

their significance values. The real data set is about the ICU scoring system namely APACHE II (Balaswamy and Vardhan, 2015) which is used to predict the status of the patient *i.e.* dead or alive. This is commonly used score which is derived from 11 physiological variables, the Glasgow coma (scores) and the patient's age and chronic health status. A total of 111 patients of which 66(59.46%) are alive and 45(40.54%) dead are present in this study. Further, the GHN ROC curve is plotted and the computations are done with respect to the proposed ROC model and compared with the existing ROC models like Binormal and HN ROC curves. When this data of both alive (D = 0.11785, p-value = 0.3185) and dead (D = 0.089239, p-value = 0.8661) populations follows Generalized Half Normal distribution, the accuracy of the test is 68.3% with the optimal threshold of 26, which classifies the data as abnormal as abnormal about 65% (TPR). Further, it is noticed that the accuracy is lesser in other models like Binormal (67.2%) and HN ROC curve (58.9%) than the GHN ROC model, which means the proposed GHN ROC model is performing better than the existing ROC models when the data follows generalized half normal distribution other than the normal and half normal distributions. Finally, the ROC curves are plotted to show the discrimination ability of the proposed ROC curve with the existing ROC models and is depicted in the Figure 2. The real data set is about the ICU scoring system SAPS III (Balaswamy and Vardhan,

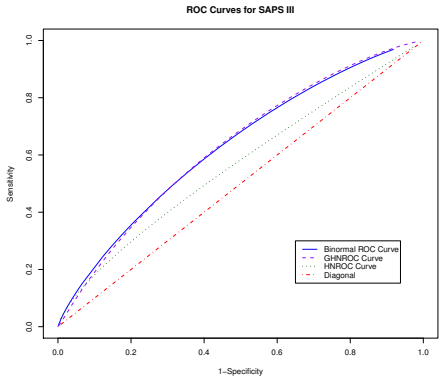


Figure 3: Comparison of ROC curves - SAPS III dataset

2022) III and issued to predict the life status of a subject who is admitted to ICU. The data consists of a total of 111 respondents of which 66(59.46%) are alive and 45(40.54%) dead. In above Table 3, comparison of three ROC curve have been done, when the data follows GHN distribution for both the populations ($normal\ population:\ D = 0.11431,\ p-value = 0.3544$ & $abnormal\ population:\ D = 0.13555,\ p-value = 0.38$) and it is also seen that the data do not follow the normal distribution ($normal\ population:\ D = 0.99997,\ p-value < 2.2e^{-16}$ & $abnormal\ population:\ D = 0.97778,\ p-value < 2.2e^{-16}$). Using the proposed methodology we have used this scoring variable to predict the mortality of patients in ICU. From the obtained result, it is observed that discriminatory ability of generalized half normal distribution (63.07%), and Binormal distribution (63.03%) is almost same *i.e.* 63% whereas when data follow half normal distribution discriminatory ability of the diagnostic test is less *i.e.* 56%. The interesting fact observed is that even though the data doesn't follow the normal distribution, the Binormal ROC curve is providing the similar accuracy with the proposed GHN ROC curve, this means that the Binormal ROC curve is over estimating the accuracy with the optimal threshold of 30, which provides only of 58.5% of true positive rates whereas the GHN ROC curve provides the optimal threshold of 26 with the higher

true positive rates of 63.6%, *i.e.*, the GHN ROC curve is more accurately classifying the data than the existing models when the data follows that particular generalized half normal distribution.

Figure 3, depicts the three ROC models for SAPS III dataset and the GHN ROC curve is slightly higher than the Binormal Roc curve and better than the HN ROC curve with the accuracy of 63%.

4.3. Optimal thresholds and confidence intervals for the ROC curve

In this section, the optimal thresholds are estimated by using different methods and are explained for the ease of medical practitioner. Further, the confidence intervals are also constructed for the proposed GHN ROC curve along with the Z test statistic for the area under the curve (AUC). Here, the effect of sample size on the proposed ROC curve is also be discussed. Three different classification situations (better, moderate and worst) are considered over different sample sizes. The entire simulations and the results are carried out using R programming and a bootstrap methodology is also used for the proposed ROC methodology. The results are as follows. Table4 (Better case) consists of optimal thresholds; FPR and TPR at that particular optimal threshold along with the confidence intervals of AUC and its Z statistic for testing the hypothesis. These optimal thresholds are evaluated using four different methods (J, ER, CZ and IU) and the results are also evaluated at various sample sizes. From these four methods of obtaining an optimal threshold at each and every sample size, the Youden's index method and IU methods are almost same with respect to the better classification scenario with AUC of more than 90%. The optimal threshold can be considered either from method J or IU, since their corresponding true positive rate (sensitivity) is higher than compared to the other methods, which means misclassification rate can be reduced using these methods with higher accuracy. Further, the Z statistic is found to be higher ($Z > 1.96$), *i.e.*, the curves obtained at this combination are significant enough to explain the accuracy of a test.

The confidence intervals are constructed for the considered combination of parameters (Better case) at various sample sizes and are depicted in Figure 4. The optimal threshold identified by method J and IU are also highlighted in the diagram with its corresponding FPR and TPR. Here, one can see the effect of sample sizes clearly, *i.e.*, as the sample size increase, the confidence intervals become closer to each other. Therefore, the accurate results may be obtained with the higher sample sizes.

The moderate case of classification scenario is also considered with $\sigma_0 = 1.5, \sigma_1 = 2.4, \alpha_0 = 0.9 and \alpha_1 = 2.5$ and the results with respect to the optimal threshold and AUC with its confidence intervals are also reported (Table 5). In this situation also, the optimal threshold can be identified by the methods of J and IU, since their sensitivity is higher than the other methods at each and every sample size. The AUC is observed to be more than 70% and the Z value is found be rejected ($Z > 1.96$), this means that the ROC curves are good enough to explain the extent of correct classification with the corresponding optimal thresholds. Further, it is to note that the optimal threshold can be obtained from method J or IU in both the cases of better and moderate case of classification scenarios. The confidence intervals are constructed for the considered moderate case at various sample sizes and are depicted in Figure 5. The optimal threshold identified by method J and IU are

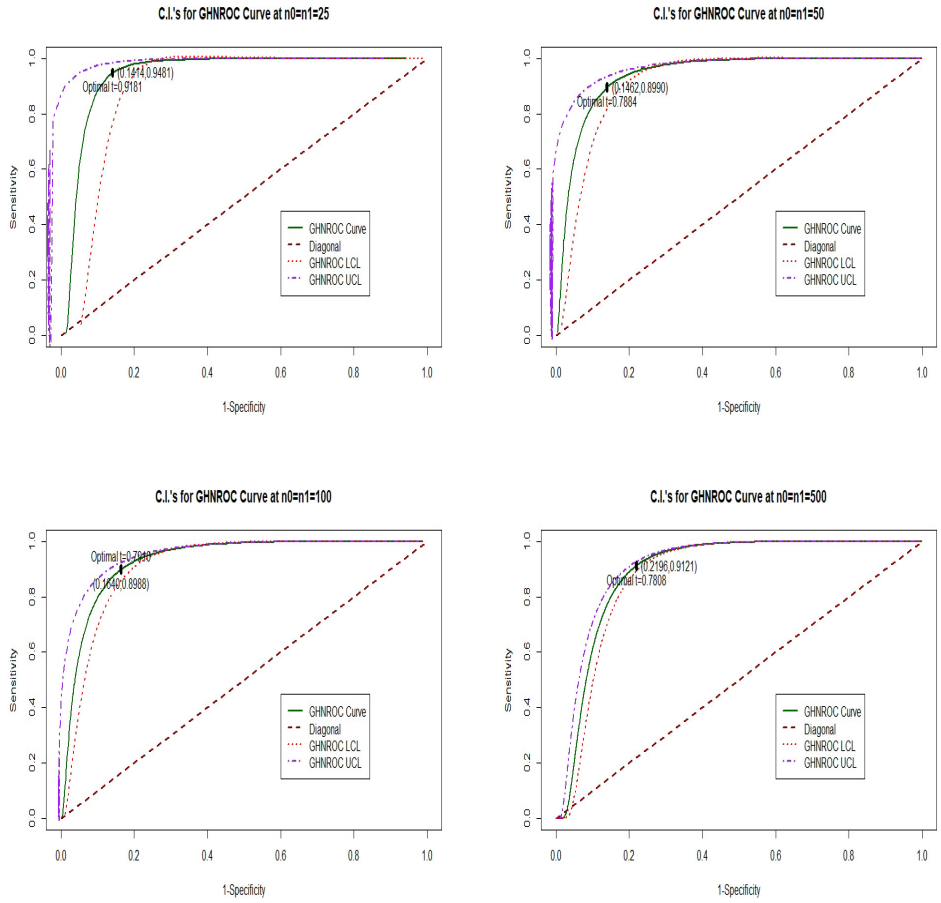


Figure 4: Confidence intervals for GHN ROC curve with its optimal threshold - better case

also highlighted in the diagram with its corresponding FPR and TPR. Here, as the sample size increase, the confidence intervals become closer to each other.Finally, the worst case classification scenario is considered (Table 6) to obtain the optimal threshold and thereby its accuracy. Though this scenario is of not at all useful in reality; the results are carried out to check the methods of obtaining optimal thresholds at various sample sizes. The very interesting factor observed here is that the ER method is found to be better with moderate amount of TPR and reasonably FPR as compared to the other methods. Even though, the sensitivity of method J is higher, but the corresponding FPR is also higher, where it should be minimum. Further, the accuracy is below 50% with the hypothesis is found to be insignificant stating that the curve obtained at this combination is not useful for future classification.

The confidence intervals are constructed for the considered worst case at various sample sizes and are depicted in Figure 6. The optimal threshold identified by method ER is highlighted in the diagram with its corresponding FPR and TPR.

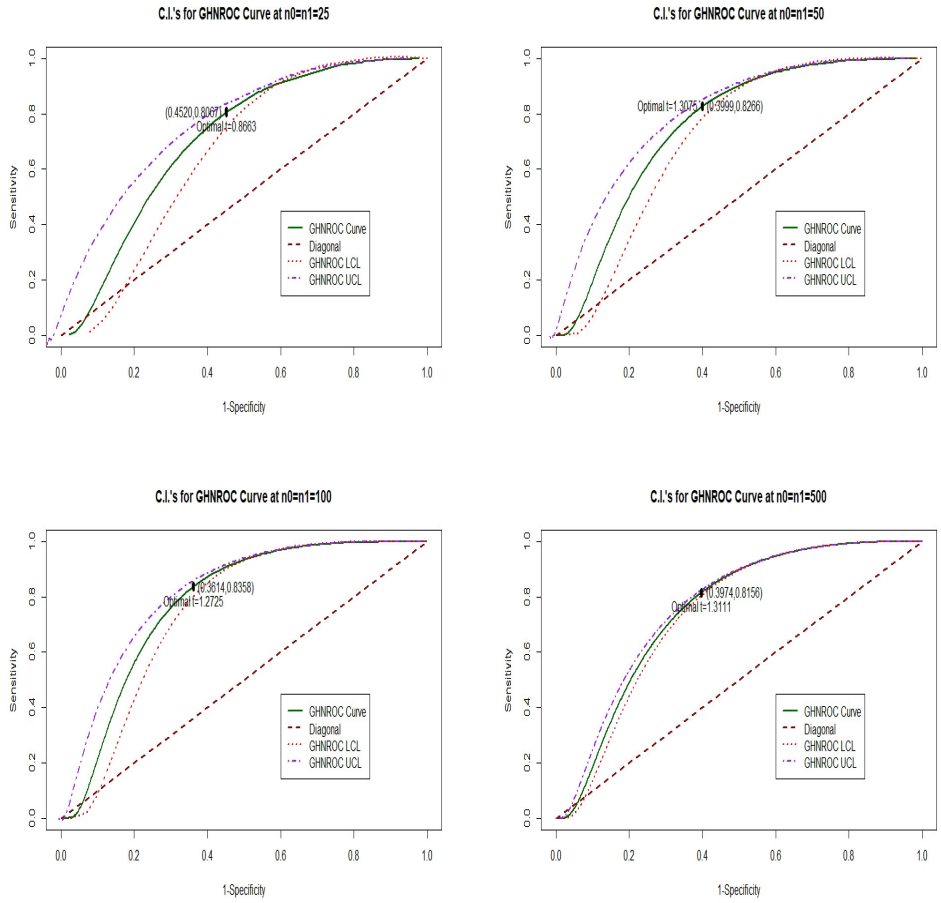


Figure 5: Confidence intervals for GHN ROC curve with its optimal threshold - moderate case

5. Conclusion

The Receiver Operating Characteristic (ROC) curves are useful in detecting the optimal threshold of medical diagnostic test with good extent of correct classification and accuracy. Therefore, on working with real datasets, the knowledge on distributional based ROC curves will be quite useful. Keeping this in mind, the ROC curve for generalized half normal distribution is proposed and the properties are verified. Further, extensive simulation studies are done with respect to the proposed ROC model and this model is also compared with the existing ROC models like Binormal and HN ROC models to show the proposed ROC model is better with skewed data of generalized half normal distribution. The real datasets (APACHE II and SAPS III) are used to demonstrate the behaviour of the proposed ROC curve in the results section. The accuracy measure for the proposed method using SAPS III dataset is higher (63%) than the AUC of SAPS III dataset (56%) proposed by Dashina and Vishnu Vardhan (2023) and that ROC model has incorrect mathematical expressions, which misleads the results. Therefore, this model is more useful than any other when the data is of generalized half normal distribution.

As, the optimal threshold identification is most important in classification, therefore

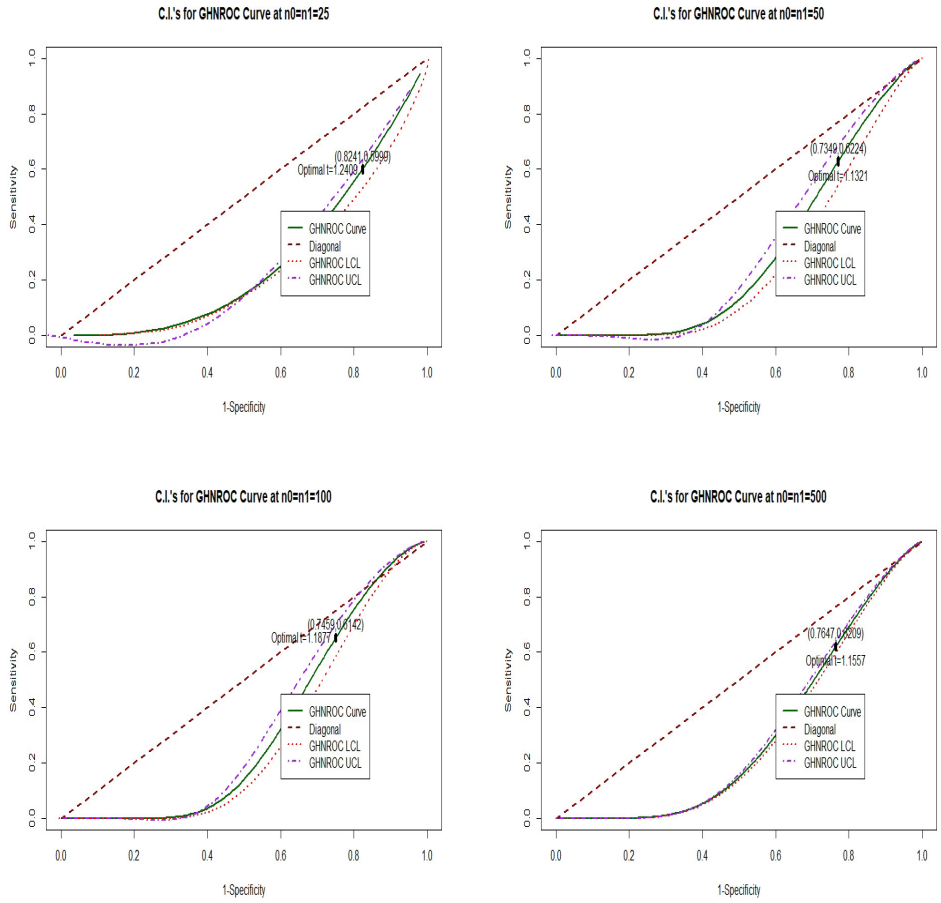


Figure 6: Confidence intervals for GHN ROC curve with its optimal threshold - worst case

four different methods are used to identify the optimal threshold with better accuracy. The interesting results observed is that the methods J and IU are found to be similar though their mathematical formulae are different at various sample sizes (Better and Moderate cases of classification scenario). But, the ER method is found to be good in case of worst classification situation, though this case is not at all considerable. Further, the proposed GHN ROC curve is found to be better with respect to the existing ROC curves when the data is of skewed in nature and follow the generalized half normal distribution.Further, the confidence intervals are also constructed for the ROC curve at various sample sizes and the AUC is also tested with the chance line 50%. Also, it is suggested that among the four methods of optimal threshold, one can consider J or IU methods with equal importance. In order to obtain the best optimal threshold, the usual method of equating densities is always not possible as in this case (no closed form solution). Therefore, we have suggested considering the sensitivity value and their corresponding specificity values to be higher. Wherever, these values are higher, that particular threshold will be of good choice with greater accuracy.

Acknowledgements

The first author would like to thank the Department of Community Medicine, Govt. Bundelkhand Medical College, Sagar, Madhya Pradesh for supporting her to carry out the research work. Also, the authors are grateful to the Editor and reviewers for their guidance, valuable comments and suggestions to improve the paper.

Conflict of interest

The authors do not have any financial or non-financial conflict of interest to declare for the research work included in this article.

References

Balaswamy S. and Vishnu Vardhan R. (2022). Estimation of the area under the ROC Curve for non Normal Data. *Communication in Statistics - Case Studies, Data Analysis and Applications*, **8**, 393-406.

Balaswamy S. and Vishnu Vardhan R. (2023). AUC Estimation and ROC model comparison in the perspective of generalized exponential distribution. *International Journal of Statistics and Reliability Engineering*, **10**, 346-351.

Balaswamy S. and Vishnu Vardhan R. (2015). Interface between the ratio β with area under the ROC Curve and Kullback-Leibler divergence under the combination of half Normal and Rayleigh distributions. *American Journal of Biostatistics*, **5**, 69-77.

Balaswamy S., Vishnu Vardhan R., and Sarma K. V. S. (2015). The Hybrid ROC (HROC) curve and its divergence measures for binary classification. *International Journal of Statistics in Medical Research*, **4**, 94-102.

Dashina P. and Vishnu Vardhan R. (2023). Estimation of AUC of bi-Generalized half-Normal ROC curve. *Statistics and Applications*, **21**, 59-71.

Dorfman D. D. and Alf Jr E. (1968). Maximum likelihood estimation of parameters of signal detection theory - a direct solution. *Psychometrika*, **33**, 117-124.

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

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

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

Ilker Unal. (2017). Defining an optimal cut-point value in ROC analysis: an alternative approach. *Computational and Mathematical Methods in Medicine*, Art.Id:3762651, doi:10.1155/2017/3762651.

Kannan Arunima S. and Vishnu Vardhan R. (2023). Estimation of area under the multiclass ROC for non Normal data, *Statistics and Applications*, **21**, 113-121.

Kannan Arunmima S. and Vishnu Vardhan R. (2022). Estimation of area under the ROC curve in the framework of gamma mixtures. *Communications in Statistics Case Studies, Data Analysis and Applications*, **8**, 714-727.

Lusted L.B. (1971). Signal detectability & medical decision making. *Science*, **171**, 1217-1219.

Tang L. L. and Balakrishnan N. (2011). A random-sum Wilcoxon statistic and its application to analysis of ROC and LROC data. *Journal of Statistical Planning and Inference*, **141**, 335-344.

Zou K. H., Hall W. J., and Shapiro D. E. (1997). Smooth nonparametric receiver operating characteristic (ROC) curves for continuous diagnostic tests. *Statistics in Medicine*, **16**, 2143-2156.

Appendix

Table 1: Comparison of ROC curves for different cases of classification using simulated datasets

Experiment	ROC Curve	σ_0	σ_1	α_0	α_1	μ_0	μ_1	Optimal Threshold	FPR	TPR	J	AUC	KS Test (Healthy)	KS Test (Diseased)
Better Case	GHN ROC Curve	1.4463	3.5406	1.4888	2.5065	-	-	2.0457	0.0938	0.8004	0.7066	0.9053	D = 0.0364, p=0.5183	D = 0.0302, p=0.7517
	HNROC Curve	1.4848	3.7447	-	-	-	-	2.2148	0.1339	0.5524	0.4186	0.7481	D = 0.0403, p=0.3887	D = 0.0465, p=0.2287
	Binormal ROC Curve	0.6242	1.0603	-	-	1.1661	2.9710	2.0183	0.0861	0.8155	0.7295	0.9288	$D = 0.5631, p < 2.2e^{-16}$	$D = 0.8461, p < 2.2e^{-16}$
Moderate Case	GHN ROC Curve	0.7587	1.8173	0.4911	1.5118	-	-	0.6257	0.3630	0.8419	0.4789	0.7816	D = 0.0356, p=0.5485	D = 0.0215, p=0.9745
	HNROC Curve	0.7931	1.8117	-	-	-	-	1.1346	0.1531	0.5300	0.3769	0.7283	D = 0.0245, p=0.9234	D = 0.0260, p=0.8854
	Binormal ROC Curve	1.1596	0.7860	-	-	0.7691	1.4589	0.7693	0.4999	0.8099	0.3099	0.6888	$D = 0.5, p < 2.2e^{-16}$	$D = 0.5836, p < 2.2e^{-16}$
Worst Case	GHN ROC Curve	1.9976	1.4890	2.4662	2.0943	-	-	3.2434	0.0010	0.0000	-0.0010	0.2864	D = 0.0225, p=0.9607	D = 0.0326, p=0.6619
	HNROC Curve	1.9358	1.4847	-	-	-	-	6.4408	0.0010	0.0000	-0.0009	0.4162	D = 0.0364,p=0.5187	D = 0.0369, p=0.501
	Binormal ROC Curve	0.5952	0.5047	-	-	1.6822	1.2250	3.2434	0.0044	0.0000	-0.0043	0.2790	$D = 0.7160, p < 2.2e^{-16}$	$D = 0.6176, p < 2.2e^{-16}$

Table 2: Results of ROC curves for the APACHE II dataset

ROC Curve	σ_0	σ_1	α_0	α_1	μ_0	μ_1	Optimal Threshold	FPR	TPR	J	AUC	KS Test (Healthy)	KS Test (Diseased)
GHN ROC Curve	29.411	42.9109	1.0555	1.5772	-	-	26	0.3799	0.6500	0.2700	0.6836	D = 0.11785, p = 0.3185	D = 0.089239, p = 0.8661
HNROC Curve	28.8184	38.4924	-	-	-	-	33	0.2521	0.3912	0.1391	0.5896	D = 0.13462, p = 0.1827	D = 0.20792, p = 0.0408
Binormal ROC Curve	17.0215	17.6888	-	-	23.3486	34.2889	28	0.3923	0.6389	0.2465	0.6720	D = 0.98485, $p < 2.2e^{-16}$	D = 1, $p < 2.2e^{-16}$

Table 3: Results of ROC curves for the SAPS III dataset

ROC Curve	σ_0	σ_1	α_0	α_1	μ_0	μ_1	Optimal Threshold	FPR	TPR	J	AUC	KS Test (Healthy)	KS Test (Diseased)
GHN ROC Curve	32.6943	41.9543	1.1832	1.5636	-	-	26.0000	0.4457	0.6360	0.1903	0.6307	D = 0.11431, p = 0.3544	D = 0.13555, p = 0.38
HNROC Curve	30.9450	38.0462	-	-	-	-	34.0000	0.2719	0.3715	0.0996	0.5646	D = 0.10829, p = 0.4213	D = 0.2031, p = 0.04883
Binormal ROC Curve	17.6210	17.6201	-	-	25.5303	33.8222	30.0000	0.3999	0.5859	0.1860	0.6303	D = 0.99997, $p < 2.2e^{-16}$	D = 0.97778, $p < 2.2e^{-16}$

Table 4: Intrinsic measures of GHN ROC curve using the methods of optimal threshold at different sample sizes - better case ($\sigma_0 = 0.5, \sigma_1 = 1.8, \alpha_0 = 0.5$ *and* $\alpha_1 = 2.5$)

Method	Sample Size	Status	Optimal Threshold	FPR	TPR	Value of method	AUC (LCL, UCL)	Z Statistic for AUC
J	25	1	0.9182	0.1414	0.9481	0.8067	0.9443	19.397
ER	25	1	0.9828	0.1286	0.9343	0.1444		
CZ	25	0	0.9493	0.1351	0.9418	0.8145	(0.8994, 0.9892)	
IU	25	1	0.9182	0.1414	0.9481	0.0919		
J	50	0	0.7885	0.1463	0.8990	0.7528	0.9339	24.2761
ER	50	1	0.8231	0.1365	0.8879	0.1766		
CZ	50	0	0.7885	0.1463	0.8990	0.7675	(0.8989, 0.9690)	
IU	50	0	0.7885	0.1463	0.8990	0.1182		
J	100	1	0.7911	0.1640	0.8988	0.7348	0.9249	30.8472
ER	100	0	0.8641	0.1428	0.8744	0.1902		
CZ	100	0	0.8180	0.1559	0.8902	0.7514	(0.8979, 0.9519)	
IU	100	1	0.7911	0.1640	0.8988	0.1238		
J	500	1	0.7808	0.2196	0.9122	0.6926	0.900	37.7962
ER	500	1	0.8920	0.1899	0.8763	0.2266		
CZ	500	0	0.8185	0.2090	0.9009	0.7126	(0.8793, 0.9208)	
IU	500	0	0.8185	0.2090	0.9009	0.1100		

Table 5: Intrinsic measures of GHN ROC curve using the methods of optimal threshold at different sample sizes - moderate case ($\sigma_0 = 1.5, \sigma_1 = 2.4, \alpha_0 = 0.9$ *and* $\alpha_1 = 2.5$)

Method	Sample Size	Status	Optimal Threshold	FPR	TPR	Value of method	AUC (LCL, UCL)	Z Statistic for AUC
J	25	1	0.8663	0.4521	0.8067	0.3547	0.7153	2.948
ER	25	1	1.1190	0.3592	0.7003	0.4678		
CZ	25	0	0.9501	0.4192	0.7733	0.4491	(0.5721, 0.8585)	
IU	25	1	1.1190	0.3592	0.7003	0.0818		
J	50	0	1.3075	0.4000	0.8266	0.4267	0.7539	4.4455
ER	50	1	1.5305	0.3263	0.7418	0.4161		
CZ	50	1	1.4413	0.3546	0.7779	0.5021	(0.6419, 0.8659)	
IU	50	1	1.5305	0.3263	0.7418	0.0951		
J	100	1	1.2725	0.3614	0.8359	0.4745	0.7766	7.2527
ER	100	1	1.4658	0.3064	0.7695	0.3834		
CZ	100	1	1.3689	0.3331	0.8043	0.5364	(0.7019, 0.8514)	
IU	100	1	1.4658	0.3064	0.7695	0.0956		
J	500	1	1.3112	0.3975	0.8157	0.4182	0.7497	14.3671
ER	500	0	1.5022	0.3380	0.7470	0.4222		
CZ	500	1	1.4373	0.3575	0.7716	0.4958	(0.7156, 0.7837)	
IU	500	1	1.4968	0.3396	0.7491	0.0900		

Table 6: Intrinsic measures of GHN ROC curve using the methods of optimal threshold at different sample sizes - worst case ($\sigma_0 = 2, \sigma_1 = 1.5, \alpha_0 = 2.3$ *and* $\alpha_1 = 3$)

Method	Sample Size	Status	Optimal Threshold	FPR	TPR	Value of method	AUC (LCL, UCL)	Z Statistic for AUC
J	25	0	2.5139	0.0348	0.0000	-0.0347	0.2693	-3.8058
ER	25	0	1.2409	0.8241	0.5999	0.9161		
CZ	25	1	1.4317	0.7258	0.4258	0.1168	(0.1505, 0.3881)	
IU	25	0	1.5985	0.6182	0.2719	0.1162		
J	50	1	0.4999	0.9500	0.9646	0.0146	0.2997	-3.7384
ER	50	1	1.1321	0.7349	0.6224	0.8263		
CZ	50	0	1.1972	0.7040	0.5615	0.1662	(0.2246, 0.4140)	
IU	50	0	1.4125	0.5931	0.3449	0.0856		
J	100	1	0.5197	0.9601	0.9688	0.0087	0.3126	-4.7832
ER	100	1	1.1877	0.7459	0.6142	0.8397		
CZ	100	1	1.2497	0.7161	0.5551	0.1576	(0.2358, 0.3894)	
IU	100	0	1.4805	0.5938	0.3190	0.0872		
J	500	1	0.3895	0.9786	0.9836	0.0050	0.3081	-11.498
ER	500	1	1.1557	0.7647	0.6210	0.8535		
CZ	500	1	1.2344	0.7291	0.5488	0.1487	(0.2753, 0.3408)	
IU	500	0	1.4827	0.6032	0.3056	0.0912		

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 23, No. 1, 2025 (New Series), pp 321–334
https://www.ssca.org.in/journal

Nonparametric Estimation of Extropy-Related Measures with Length-Biased Data

R. Dhanya Nair[1] **and E. I. Abdul Sathar**[2]
[1]*Department of Statistics, University College, Thiruvananthapuram - 695034, India*
[2]*Department of Statistics, University of Kerala, Thiruvananthapuram - 695581, India*

Received: 09 May 2024; Revised: 20 June 2024; Accepted: 30 June 2024

Abstract

Nonparametric estimators for extropy-related measures using length-biased data are proposed in this paper. The proposed estimators exhibit desirable properties, including consistency and asymptotic normality, which have been established. Furthermore, the precision of these estimators is assessed through the utilization of both simulated and real data sets, thereby validating their effectiveness in practical scenarios.

Key words: Extropy; Length-biased; Kernel estimator; Nonparametric estimation.

AMS Subject Classifications: 62N05, 62N02, 62G05.

1. Introduction

Length-biased sampling is a widely used technique for collecting lifetime data, primarily due to its cost-effectiveness and convenience. Unlike random sampling, length-biased sampling selects observations from the population of interest with probability proportional to their length. This approach finds significant applications in survival analysis, particularly when the onset time of diseases is unknown. In such scenarios, individuals who survive longer are more likely to be included in the sample, resulting in length-biased survival data. The phenomenon of length-bias was first noticed by Wicksell (1925) while investigating cell samples under a microscope. In his research, he noticed that only the cells that were larger than a particular size were visible in the microscope, leading to the study of a length-biased sample of cells. However, there are many other applications of length-biased data that make it crucial to understand the properties of this type of data. For instance, length-biased data arise in the study of diverse phenomena, such as ageing, epidemiology, and genetics. Therefore, exploring various aspects of length-biased data is essential for researchers and practitioners in fields such as medical research, public health, and social sciences.

Consider a random variable X with a probability density function (pdf), distribution function, and survival function denoted by f, F, and $\bar{F}$, respectively. Suppose a sample of size n is drawn from this population using a length-biased sampling technique, where the probability of including an observation in the sample is proportional to its size, volume,

Corresponding Author: E.I. Abdul Sathar
Email: sathare@gmail.com

length, or survival time. In other words, observations that are larger, longer, or have a longer survival time have a higher probability of being sampled. The resulting sample is length-biased, and the observed sample can be regarded as drawn from a distribution with pdf g given by

$$g(y) = \frac{y\, f(y)}{\mu}, \;\; y \geq 0 \; and \; \mu \text{ is the mean of the population.} \tag{1}$$

One crucial problem is the nonparametric estimation of functionals of the distribution function F or pdf f based on a length-biased sample $Y_i, 1 \leq i \leq n$. This paper aims to propose nonparametric estimators for extropy-related measures of a population, using a length-biased sample drawn from it. Furthermore, the properties of these estimators are thoroughly investigated. The proposed estimators are useful in various fields, such as information theory, economics, and statistical physics, where the analysis of length-biased data is required. The study of these estimators' properties can aid in better understanding and utilizing length-biased data in practical applications.

Shannon's entropy, introduced by Shannon (1948), is one of the most widely used measures for assessing the uncertainty associated with a random variable. For a discrete random variable X taking values $\{x_1, x_2, x_3, ..., x_N\}$ with probability mass function (pmf) $\mathbf{p} = (p_1, p_2, p_3, ..., p_N)$, such that $N > 1$ is finite, Shannon's entropy is defined as

$$H(X) = -\sum_{i=1}^{N} p_i \log p_i. \tag{2}$$

Because equation (2) can be rewritten as $\mathcal{H}(\mathbf{p}) = E(-log\, \mathbf{p})$, the discrete entropy $\mathcal{H}(\mathbf{p})$ can be thought of as quantifying the average information content of X. That is, the entropy of a probability distribution is just the expected value of the information in the distribution. The entropy measure has far-reaching applications in many areas such as financial analysis, data compression, statistics and information theory. Lad *et al.* (2012) observed that the entropy measure on its own do not provide complete summary of the information in a distribution. This observation was substantiated in the context of its application in the logarithmic scoring rule, widely considered to be an eminent proper scoring rule used extensively for assessing and comparing sequential forecast distributions. The expected logarithmic score of a pmf $\mathbf{p}$ is in fact $-H(X)$, called negentropy. Lad pointed out that the logarithmic scoring function provides an incomplete assessment as it is a function only of the actual observation value of a quantity, ignoring other possible but unobserved values. To address this issue, a complementary scoring rule needs to be monitored concomitantly with the log score and this led to the expanded version of the logarithmic score, termed as the total log score. As a pair, the two complementary scores constitute the total logarithmic score and both components of the total log score are relevant to the assessment of forecasting distribution. Moreover, the expectation of the total log score equals the negentropy plus the negextropy of the distribution, where negextropy is the negative of a measure of a probability distribution suggested to be called as the extropy of the distribution by Lad *et al.* (2015) and is defined as follows.

For a discrete random variable X, the complementary dual of entropy, called extropy

is defined as

$$J(X) = -\sum_{i=1}^{N} (1 - p_i) \log (1 - p_i).$$

The complementary of H and J arises from the fact that

$$J(\mathbf{p}) = (N - 1) \left[H(\mathbf{q}) - \log(N - 1) \right].$$

That is, the extropy of a pmf $\mathbf{p} = (p_1, p_2, p_3, ..., p_N)$ equals a location and scale transform of the entropy of another pmf $\mathbf{q} = \left(\frac{1-p_1}{N-1}, \frac{1-p_2}{N-1}, \frac{1-p_3}{N-1}, ..., \frac{1-p_N}{N-1} \right)$. The duality of entropy/extropy is a formal mathematical property of the pair of functions. For more details, one may refer Lad *et al.* (2015) and Lad *et al.* (2018).

As in entropy, extropy is interpreted as a measure of the amount of uncertainty represented by the distribution for X. Both entropy and extropy share many properties. They are invariant with respect to permutations of their mass functions and with respect to monotonic transformations. Moreover, the maximum extropy distribution is the uniform distribution and extropy satisfies Shannon's first and second axioms. As to differences in the two measures, the scale of the maximum entropy measure is unbounded as N increases while the scale of the maximum extropy is bounded by 1. It is evident that when $N = 2$, the entropy and extropy are identical. However, when $N > 2$, the measure bifurcates to yield distinct paired measurements $(H(X), J(X))$. As companions, these two measures relate as do the positive and negative images of a photographic film and they contribute together to characterizing the information in a distribution in much the same way. When the entropy is calculated for any assemblage such as the heat distribution for a galaxy of stars, a companion calculation of the extropy would allow us to complete our understanding of the variation inherent in its empirical distribution. An axiomatic characterization and several intriguing properties of this new measure was considered by Lad *et al.* (2015) and the results provided links to other notable information functions whose relation to entropy have not been recognized.

In the continuous context, a natural analog of discrete Shannon entropy for a probability density function f is called differential entropy and is defined as

$$H(X) = -\int_{0}^{\infty} f(x) \, \log \, f(x) \, dx.$$

The definition of differential entropy appears to be a natural extension of the Shannon entropy for discrete variables, defined in equation (2), to continuous variables. However, Shannon's differential entropy measure for a continuous density is actually derived from the limit of a linear translation of the discrete entropy measure. In order to define extropy for a continuous density, Lad *et al.* (2015) used the same procedure as the one followed by Shannon in defining differential entropy. Lad *et al.* (2015) noted that when the range of possibilities for X increases (as a result of larger N), the extropy measure $-\sum_{i=1}^{N} (1 - p_i) \log (1 - p_i)$ can be closely approximated by $1 - \frac{1}{2} \sum_{i=1}^{N} p_i^2$, which led to the definition of differential extropy.

Extropy of a non-negative absolutely continuous random variable X with pdf $f(x)$ is defined as

$$J(X) = -\frac{1}{2}\int_0^\infty f^2(x)\ dx = -\frac{1}{2}E(f(X)). \tag{3}$$

Here E denotes the expected value operator.

Differential entropy and extropy are obtained as the limit of a linear transformation of their corresponding discrete measures. The dual complementarity of extropy with entropy for continuous densities is derived in the context of relative entropy, also known as Kullback-Leibler divergence.

Through various illustrations Lad *et al.* (2012) showed that the extropies of the distributions do appear to provide interpretable complementary understandings of the character of distributions, already well-known to be summarised in a different dimension by their entropies.The total log score for densities is also better identified with the bivariate measure (negentropy, negextropy). Extropy can also be used to compare the uncertainties of two random variables. If the extropy of X is less than that of another random variable Y, that is, $J(X) \leq J(Y)$, then X is said to have more uncertainty than Y. By simultaneously considering entropy and extropy measures, researchers and practitioners can gain a more comprehensive understanding of the information and uncertainty within a given distribution. This broader perspective enables better-informed decision-making and more efficient utilization of statistical models in a range of applications. For further studies on extropy, one may also refer Noughabi and Jarrahiferiz (2019), Tahmasebi and Toomaj (2020), Buono *et al.* (2023) and Sathar and Nair (2024).

Additionally, to capture the uncertainty of a random variable which has already survived for some time, Qiu and Jia (2018) suggested the measure residual extropy. The residual extropy, denoted as $J(X;t)$, is defined as

$$J(X;t) = -\frac{1}{2\ (1-F(t))^2}\int_t^\infty f^2(x)dx \tag{4}$$

Furthermore, Krishnan *et al.* (2020) introduced a measure called past extropy, which computes the uncertainty associated with the past lifetime of a component that failed before a specific time. Past extropy of a random life time X is of course the extropy of the random variable $[X|X \leq t]$ and is given by

$$\bar{J}(X;t) = -\frac{1}{2\ F(t)^2}\int_0^t f^2(x)dx. \tag{5}$$

For a non-negative rv X having a survival function $\bar{F}$, an alternative measure of extropy based on the survival function of a rv called survival extropy (SE) has been proposed by Sathar and Nair (2021) which is defined as

$$J_s(X)\ = -\frac{1}{2}\int_0^\infty \bar{F}^2(x)\,dx.$$

The survival extropy of the random variable $[X - t|X \geq t]$ called dynamic survival extropy (DSE), was also considered by Sathar and Nair (2021) and is defined as

$$J_s(X;t) = -\frac{1}{2}\int_t^\infty \frac{\bar{F}^2(x)}{\bar{F}^2(t)}dx = -\frac{1}{2}\int_t^\infty \frac{(1-F(x))^2}{(1-F(t))^2}dx. \tag{6}$$

It is worth noting that the SE and DSE have a close relationship with well-known economic measures such as the Gini index and statistical quantities including L-moments. These connections have been extensively studied by Nair and Sathar (2022) and Nair and Sathar (2023). These insights further contribute to the interpretation and application of the SE and DSE measures, offering valuable connections to economic analysis and statistical modeling.

These alternative measures of extropy, namely residual extropy, past extropy, and survival extropy, complement Shannon's entropy and offer additional perspectives on the uncertainty and information content of a random variable. These measures find applications in various fields, including reliability analysis, survival modeling, risk assessment, economics, finance, and actuarial science, where the analysis of time-dependent uncertainty is of paramount importance. By utilizing these measures, researchers and practitioners can gain deeper insights into the temporal aspects and survival behavior of random variables in practical scenarios. In this study, we introduce nonparametric estimators for extropy related measures of the population based on a length-biased data drawn from it. Length-biased sampling has proven to be valuable in various fields, and in Section 2, we present our proposed estimator for dynamic survival extropy (DSE). We also examine the asymptotic properties of the proposed estimator to ensure its reliability. Furthermore, in Section 3, we discuss the nonparametric estimation of residual and past extropy, and analyze their asymptotic properties. Finally, in Section 4, a simulated study and real-data analysis have been carried out to illustrate the precision of the estimators. By employing these empirical investigations, we showcase the accuracy and effectiveness of the estimators in practical settings. This empirical validation adds credibility to the proposed methodology and confirms its utility in real-world scenarios.

2. Nonparametric estimation of DSE using length-biased sample

This section proposes a nonparametric estimator for the DSE of a random variable X using a length-biased sample of size n drawn from X. Due to the use of a probability proportional to size (PPS) sampling scheme, the observed sample $Y_1, Y_2, Y_3, ..., Y_n$ cannot be treated as independent and identically distributed (iid) samples from X. Consequently, existing estimators of extropy measures based on a random sample from the population cannot be applied. Instead, a different estimator suitable for length-biased data needs to be considered. To this end, it is worth noting that the observed length-biased sample can be regarded as iid observations from the distribution of a random variable Y with a pdf $g(y)$ given by equation (1). Building upon this insight, Cox (1969) proposed an empirical estimator for the distribution function $F(x)$ in the length-biased setup. The estimator is

given by

$$F_n(x) = \frac{\sum_{i=1}^{n} Y_i^{-1} I(Y_i \leq x)}{\sum_{i=1}^{n} Y_i^{-1}}, \tag{7}$$

where $I(.)$ is the indicator random variable of the event specified in parentheses. It has been demonstrated by Chaubey *et al.* (2010) that as $n \to \infty$, the empirical estimator $F_n(x)$ converges almost surely to the true distribution function $F(x)$, as shown in equation (8). Furthermore, the estimator converges in distribution to a normal distribution, as expressed in equation (9).

$$\sup_{x \in R+} |F_n(x) - F(x)| \xrightarrow{a.s} 0, \text{ as } n \to \infty. \tag{8}$$

and

$$\sqrt{n}\,(F_n(x) - F(x)) \xrightarrow{D} N(0, \delta^2(x)), \tag{9}$$

where $\delta^2(x) = \mu\left\{\int_0^x t^{-1} f(t)dt - 2F(x)\int_0^x t^{-1} f(t)dt + F^2(x)\int_0^\infty t^{-1} f(t)dt\right\}$.

Also, as $n \to \infty$,

$$E(F_n(x)) = F(x) \text{ and } Var(F_n(x)) = \frac{\delta^2(x)}{n}. \tag{10}$$

Therefore, we can obtain a nonparametric estimator of DSE of X by substituting the estimator given in equation (7) into equation (6). The resulting estimator for DSE is given by

$$\hat{J}_s(X;t) = -\frac{1}{2}\int_t^\infty \frac{(1 - F_n(x))^2}{(1 - F_n(t))^2} dx. \tag{11}$$

Now let's examine the asymptotic properties of the proposed estimator. For simplifying the notation, we define the following terms:

$$a_n(t) = \int_t^\infty \bar{F}_n^2(x)\,dx,\ m_n(t) = \bar{F}_n^2(t),\ a(t) = \int_t^\infty \bar{F}^2(x)\,dx \text{ and } m(t) = \bar{F}^2(t).$$

Thus, the estimator $\hat{J}_s(X;t)$ can be expressed as

$$\hat{J}_s(X;t) = -\frac{1}{2}\frac{a_n(t)}{m_n(t)}, \text{ while the true DSE } J_s(X;t) \text{ is given by } J_s(X;t) = -\frac{1}{2}\frac{a(t)}{m(t)}.$$

Result 1:

$$\lim_{n\to\infty} |\hat{J}_s(X;t) - J_s(X;t)| = 0 \ a.s.$$

Moreover, mean square error (MSE) of $\hat{J}_s(X;t)$ tends to 0 as $n \to \infty$.

Proof: Using Taylor series expansion,

$$\bar{F}_n^2(t) = \bar{F}^2(t) + \left(\bar{F}_n(t) - \bar{F}(t)\right) 2\bar{F}(t) + o\left(\bar{F}_n(t) - \bar{F}(t)\right)^2.$$

It follows that

$$m_n(t) - m(t) = \left(\bar{F}_n(t) - \bar{F}(t)\right) 2\bar{F}(t) + o\left(\bar{F}_n(t) - \bar{F}(t)\right)^2.$$

Similarly, we obtain

$$a_n(t) - a(t) \simeq 2 \int_t^\infty \bar{F}(x) \ \left(\bar{F}_n(x) - \bar{F}(x)\right) dx.$$

Now,

$$\frac{a_n(t)}{m_n(t)} - \frac{a(t)}{m(t)} \simeq \frac{m(t)\left[a_n(t) - a(t)\right] - a(t)\left[m_n(t) - m(t)\right]}{m^2(t)}.$$

Hence,
$\hat{J}_s(X;t) - J_s(X;t)$

$$\simeq -\frac{1}{m(t)} \int_t^\infty \bar{F}(x) \left(\bar{F}_n(x) - \bar{F}(x)\right) \, dx \, + \frac{a(t)}{m^2(t)} \left(\bar{F}_n(t) - \bar{F}(t)\right) \bar{F}(t). \tag{12}$$

By using the almost sure convergence of $F_n(x)$ given in equation (8), we obtain

$$\lim_{n\to\infty} |\hat{J}_s(X;t) - J_s(X;t)| = 0 \ a.s.$$

Additionally, from equations (12) and (10), it can be easily seen that the bias and variance of $\hat{J}_s(X;t)$ tends to 0 as $n \to \infty$. Hence, as $n \to \infty$, MSE of $\hat{J}_s(X;t) \to 0$.

□

Next, we discuss the asymptotic normality of our estimator.

Result 2: $\hat{J}_s(X;t) - J_s(X;t)$ is asymptotically normal with mean 0 and variance

$$\frac{1}{n\,\bar{F}^4(t)} \left[\int_t^\infty \bar{F}^2(x)\,\delta^2(x)dx + \frac{a^2(t)\delta^2(t)}{\bar{F}^2(t)}\right].$$

Proof: Using equation (10), as $n \to \infty$,

$$E(\bar{F}_n(x) - \bar{F}(x)) = 0 \text{ and } Var(\bar{F}_n(x)) = \tfrac{\delta^2(x)}{n}.$$

Hence, from equation (12), we obtain the following.

As $n \to \infty$, $E(\hat{J}_s(X;t) - J_s(X;t)) = 0$ and

$$Var(\hat{J}_s(X;t) - J_s(X;t)) = \tfrac{1}{n\,\bar{F}^4(t)} \left[\int_t^\infty \bar{F}^2(x)\,\delta^2(x)dx + \tfrac{a^2(t)\delta^2(t)}{\bar{F}^2(t)}\right].$$

Again, using equation (9), we have,

$$\sqrt{n}\,(\bar{F}_n(x) - \bar{F}(x)) \xrightarrow{D} N(0, \delta^2(x))$$

Hence from equation (12), it follows that $\hat{J}_s(X;t) - J_s(X;t)$ is asymptotically normal. This completes the proof. □

In a similar manner, a nonparametric estimator for dynamic failure extropy (DFE) proposed by Nair and Sathar (2020) can be obtained.
The DFE of X is defined as

$$J_f(X;t) = -\frac{1}{2}\int_0^t \frac{F^2(x)}{F^2(t)}\,dx.$$

By plugging in the estimator given by equation (7) into the above equation, we can obtain the nonparametric estimator of DFE under length-biased setup, which is as follows.

$$\hat{J}_f(X;t) = -\frac{1}{2}\int_0^t \frac{F_n^2(x)}{F_n^2(t)}\,dx. \tag{13}$$

Consistency and asymptotic normality of this estimator can be proved by proceeding in a similar manner as in Result 1 and 2.

3. Nonparametric estimation of residual and past extropies for length-biased sample

In this section, we focus on the nonparametric estimation of residual and past extropies defined by equations (4) and (5), respectively. To obtain the estimators of residual and past extropies using length-biased data, we utilize equation (7) and the kernel density estimator proposed by Jones (1991). By smoothing the estimator given in equation (7), Jones (1991) derived a new kernel density estimator given by

$$\hat{f}(x) = \frac{\sum_{i=1}^{n} \frac{1}{Y_i h}\, k\left(\frac{x-Y_i}{h}\right)}{\sum_{i=1}^{n} Y_i^{-1}}, \tag{14}$$

where k is the kernel function and $h = h_n$ is the band-width. The bias, variance and asymptotic properties of this estimator was obtained by Guillamon *et al.* (1998) as follows.

$$Bias(\hat{f}(x)) = \frac{1}{2}\,h^2\,\mu_2(k)\,f''(x) + o(h^2) \text{ and } \; Var(\hat{f}(x)) = \frac{1}{n\,h}\,\mu\,x^{-1}\,f(x)\,C_k + o\left(\frac{1}{nh}\right), \tag{15}$$

where $\mu_2(k) = \int_{-\infty}^{\infty} u^2 k(u)\,du$, $C_k = \int_{-\infty}^{\infty} k^2(u)\,du$ and $f''(x)$ is the 2^{nd} derivative of f with respect to x.

Also,

$$\sqrt{n\,h}\,(\hat{f}(x) - f(x)) \xrightarrow{D} N(0, \mu\, x^{-1}\, f(x)\, C_k). \tag{16}$$

Now, we propose a nonparametric estimator of residual extropy under length-biased set up.

Definition 1: A nonparametric kernel estimator for $J(X;t)$ shall be defined as

$$\hat{J}(X;t) = -\frac{1}{2}\left[\frac{\int_t^\infty \hat{f}^2(x)dx}{(1-F_n(t))^2}\right]. \tag{17}$$

In order to simplify the notations, define

$$p_n(t) = \int_t^\infty \hat{f}^2(x)dx, \quad p(t) = \int_t^\infty f^2(x)dx,$$

so that equation (17) can be written as $\hat{J}(X;t) = -\frac{1}{2}\left[\frac{p_n(t)}{m_n(t)}\right]$.

By using Taylor's series expansion, we get

$$p_n(t) - p(t) = 2\int_t^\infty f(x)\ \left(\hat{f}(x) - f(x)\right) dx + o\Big(\hat{f}(x) - f(x)\Big)^2.$$

Proceeding in a similar manner as in Section 2, we obtain

$$\hat{J}(X;t) - J(X;t)$$

$$\simeq -\frac{1}{m(t)}\int_t^\infty f(x)\left(\hat{f}(x) - f(x)\right)\, dx\, + \frac{p(t)}{m^2(t)}\left(\bar{F}_n(t) - \bar{F}(t)\right)\bar{F}(t).$$

The asymptotic normality of $\hat{J}(X;t)$ can now be easily obtained on using equations (16) and (9). Furthermore, using equation (15), we observe that the MSE of $\hat{J}(X;t)$ tends to 0 as $n \to \infty$, and thus the estimator $\hat{J}(X;t)$ is strongly consistent.
Similarly, a consistent and asymptotically normal nonparametric estimator for $\bar{J}(X;t)$ under length-biased set up shall be defined as

$$\hat{\bar{J}}(X;t) = -\frac{1}{2\ F_n^2(t)}\int_0^t \hat{f}^2(x)dx.$$

4. Data analysis

To demonstrate the accuracy of the presented nonparametric estimators, we first apply the proposed methods to the simulated data sets. We generate length-biased samples from beta distribution of first kind with parameters $\alpha = 2$ and $\gamma = 4$. The bias and MSE

of the suggested estimators of DSE and DFE given by equations (11) and (13) respectively, are computed for certain values of t, and the results obtained are presented in Tables 1 and 2. It can be observed from the tables that the bias and MSE are negligible. This indicates that the estimators perform well in accurately capturing the extropy measures. Figures 1 and 2 display plots of the actual and estimated values of DSE and DFE of the population for simulated data. Both graphs clearly show that the estimated values closely align with the actual values. Notably, even with a sample size of $n = 30$, the estimated values for DFE are very close to the actual values, highlighting the effectiveness of the proposed estimators. Furthermore, we computed the theoretical and estimated values of residual and past extropies using the Gaussian kernel function. These values, along with the bias and MSE, are presented in Tables 3 and 4. The results from these tables indicate that the estimators of residual and past extropies also perform well, further validating the reliability of the proposed nonparametric estimators. Overall, the results obtained from the simulations demonstrate the precision and accuracy of the nonparametric estimators proposed in this study.

Table 1: Bias and MSE of the estimator of DSE for simulated data

	$n = 50$		$n = 100$	
t	Bias	MSE	Bias	MSE
0.4	0.00147	0.00009	-0.00180	0.00004
0.5	-0.00269	0.00007	0.00029	0.00002
0.6	0.00009	0.00005	-0.00363	0.00003
0.7	-0.00961	0.00031	-0.00626	0.00018

Table 2: Bias and MSE of the estimator of DFE for simulated data

	$n = 50$		$n = 100$	
t	Bias	MSE	Bias	MSE
0.4	-0.00501	0.00038	-0.00029	0.00004
0.5	0.00239	0.00021	-0.00113	0.00005
0.6	0.00388	0.00011	-0.00173	0.00011
0.7	0.00044	0.00047	-0.00379	0.00014

Table 3: Theoretical and estimated values of residual extropy together with its bias and MSE for simulated data

		$n = 50$			$n = 100$		
t	Theory	Estimate	Bias	MSE	Estimate	Bias	MSE
0.4	-1.62017	-1.71256	-0.09239	0.08763	-1.69013	-0.06996	0.01232
0.5	-2.02822	-2.11318	-0.08496	0.09544	-2.16395	-0.13573	0.00642
0.6	-2.62262	-2.72248	-0.09986	0.05238	-2.68369	-0.06107	0.01003
0.7	-3.59451	-3.65942	-0.06491	0.03416	-3.63571	-0.04120	0.00237

Next, we consider the empirical estimator of DSE and DFE, which were proposed by Sathar and Nair (2021) and Nair and Sathar (2020), respectively. These empirical estimators

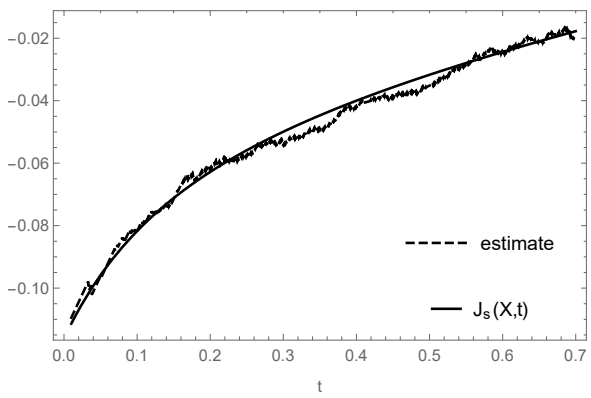


Figure 1: Plots of actual and estimated values of DSE using a simulated sample of size $n = 100$

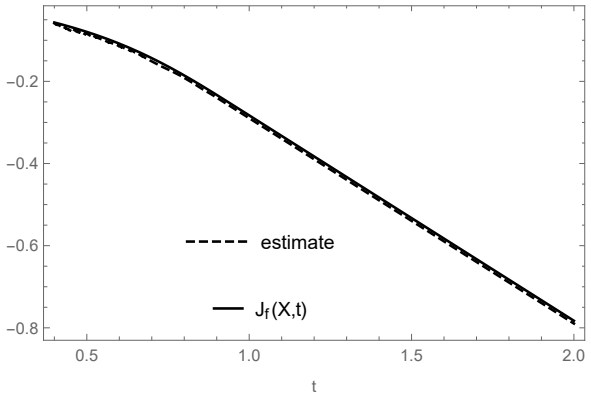


Figure 2: Plots of actual and estimated values of DFE using a simulated sample of size $n = 30$

Table 4: Theoretical and estimated values of past extropy together with its bias and MSE for simulated data

		$n = 50$			$n = 100$		
t	Theory	Estimate	Bias	MSE	Estimate	Bias	MSE
0.4	-1.38686	-1.35128	0.03558	0.00483	-1.38890	-0.00204	0.00311
0.5	-1.09420	-1.15662	-0.06242	0.00468	-1.10053	-0.00633	0.00265
0.6	-0.92836	-0.98423	-0.05587	0.00957	-0.94362	-0.01526	0.00403
0.7	-0.84123	-0.88562	-0.04439	0.00348	-0.84563	-0.00439	0.00114

are based on an iid sample from the population. The empirical dynamic survival extropy and dynamic failure extropy estimators are respectively as follows.

$$J_s(\hat{\bar{F}}_n; t) = -\frac{1}{2}\int_t^\infty \left[\frac{\hat{\bar{F}}_n(x)}{\hat{\bar{F}}_n(t)}\right]^2 dx \tag{18}$$

and

$$J_f(\hat{F}_n; t) = -\frac{1}{2}\int_0^t \left[\frac{\hat{F}_n(x)}{\hat{F}_n(t)}\right]^2 dx, \tag{19}$$

where $\hat{\bar{F}}_n(x) = \frac{1}{n}\sum_{i=1}^n I(X_i > x)$, $\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^n I(X_i \le x)$, with I being the indicator function.

To investigate the performance of the empirical estimators defined by equations (18) and (19) when applied to a length-biased sample, we compare the actual values of DSE and DFE of the population with the estimated values obtained using these estimators. The results are displayed in Figures 3 and 4. Analyzing Figures 1 to 4, we observe that the deviation between actual and estimated values is more when the empirical estimators are used instead of the proposed estimators. This suggests that the estimators defined by equations (18) and (19) are suitable when an iid sample is available from the population whereas for the length-biased sample, the estimators defined by equations (11) and (13) should be employed.

In summary, the comparison of the estimators highlights the importance of choosing the appropriate estimator based on the characteristics of the sample. The proposed nonparametric estimators are specifically tailored for length-biased data and demonstrate superior accuracy in estimating extropy measures when applied to length-biased samples, as evidenced by the smaller deviations between the actual and estimated values.

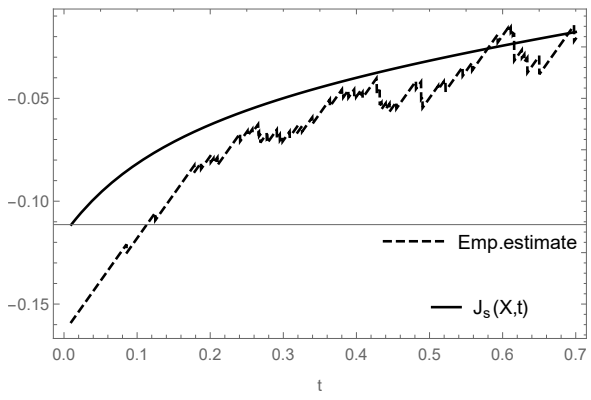


Figure 3: Plots of actual and estimated values of DSE using the empirical estimator for a simulated sample

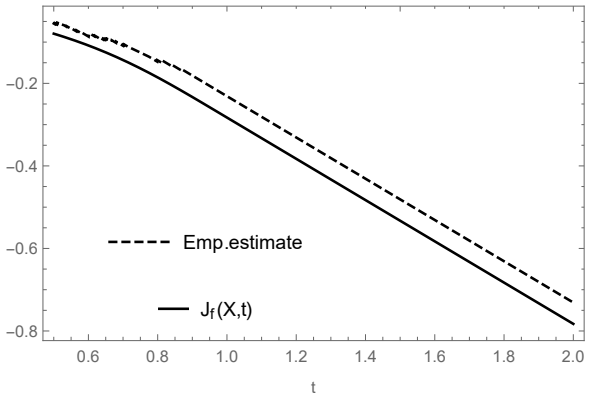


Figure 4: Plots of actual and estimated values of DFE using the empirical estimator for a simulated sample

To further assess the performance of the proposed estimators defined by equations (11) and (13), we apply them to a real-world scenario using a data that was previously investigated by Helu *et al.* (2020). The data set consists of 70 failure times of aircraft windshields, from which a sample of size 50 is drawn with probability proportional to size. The best-fitted distribution to the original data set is the Gamma distribution with parameters $\alpha = 7.75$ and $\beta = 0.285$. We plot the theoretical and estimated values of DSE and DFE for the real data in Figures 5 and 6, respectively. Upon analysis of the plots, we observe that the estimated values are remarkably close to the actual values. This indicates that the proposed estimators perform well in real-world circumstances. The accuracy of the estimators in estimating the extropy measures for the length-biased sample demonstrates their reliability and applicability in practical scenarios.

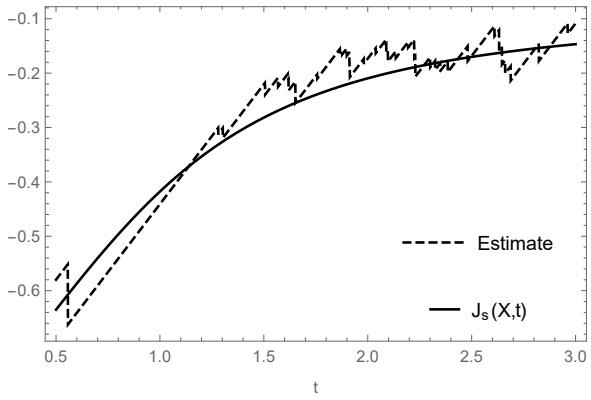


Figure 5: Plots of actual and estimated values of DSE for real data

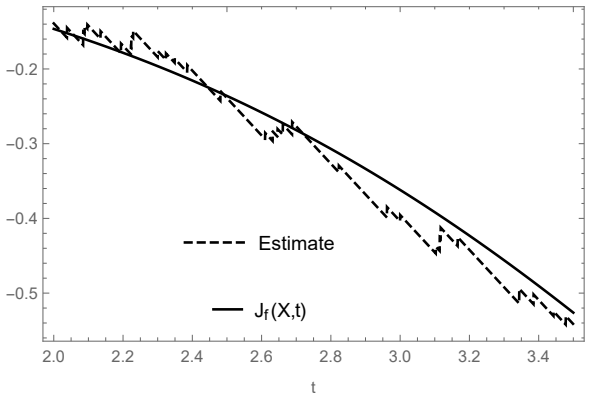


Figure 6: Plots of actual and estimated values of DFE for real data

5. Conclusions

This work proposes nonparametric estimators for extropy-related measures under length-biased sampling. The consistency and asymptotic normality of the proposed estimators are established, demonstrating their reliability in estimating these measures. The performance of the estimators is evaluated using both simulated and real data sets. The simulation results provide strong evidence of the accuracy and precision of the proposed estimators. The negligible bias and mean squared error observed in the estimators confirm their ability to closely approximate the true values of the extropy-related measures. Furthermore, the analysis of a real data set reinforces the practical utility of the proposed estimators. By evaluating the extropy-related measures using the real data, it is evident that the estimators perform well in real-life scenarios. This highlights the applicability of the estimators in various domains, such as reliability analysis, survival analysis, and engineering, where accurate estimation of extropy-related measures is crucial for making informed decisions and understanding complex systems.

In summary, this work contributes valuable nonparametric estimators for extropy-related measures under length-biased sampling. The established properties of consistency and asymptotic normality, coupled with the demonstrated accuracy in both simulated and real data settings, make these estimators highly reliable tools for researchers and practitioners. The availability of such estimators facilitates the estimation of extropy-related measures, enabling deeper insights into the dynamics of failure and survival processes in diverse fields of study.

Acknowledgements

The authors would like to express their gratitude to the editor-in-chief and the reviewers for their valuable feedback and constructive comments. Their insights and suggestions have played a significant role in enhancing the quality of this article.

Conflict of interest

The authors do not have any financial or non-financial conflict of interest to declare for the research work included in this article.

References

Buono, F., Kamari, O., and Longobardi, M. (2023). Interval extropy and weighted interval extropy. *Ricerche di Matematica*, **72**, 283-298.

Chaubey, Y. P., Sen, P. K., and Li, J. (2010). Smooth density estimation for length-biased data. *Journal of the Indian Society of Agricultural Statistics*, **64**, 145–155.

Cox, D.R. (1969). Some sampling problems in technology. In *New Developments in Survey Sampling*, Wiley-Interscience, New-York. 506–527.

Guillamon, A., Navarro, J., and Ruiz, J. M. (1998). Kernel density estimation using weighted data. *Communications in Statistics - Theory and Methods*, **27**, 2123–2135.

Helu, A., Samawi, H., and Rochani, H. (2020). Kernel density estimation based on progressive type-II censoring. *Journal of the Korean Statistical Society*, **49**, 475–498.

Jones, M. C. (1991). Kernel density estimation for length-biased data. *Biometrika*, **78**, 511–519.

Krishnan, A. S., Sunoj, S. M., and Nair, N. U. (2020). Some reliability properties of extropy for residual and past lifetime random variables. *Journal of the Korean Statistical Society*, **49**, 457–474 .

Lad, F., Sanfilippo, G., and Agro, G. (2012). Completing the logarithmic scoring rule for assessing probability distributions. *AIP Conference Proceedings*, **1490**, 13-30.

Lad, F., Sanfilippo, G., and Agro, G. (2015). Extropy: complementary dual of entropy. *Statistical Science*, **30**, 40-58.

Lad, F., Sanfilippo, G., and Agro, G. (2018). The duality of entropy/extropy, and completion of the Kullback information complex. *Entropy*, **20**, 593.

Nair, R. D. and Sathar, E. I. A. (2020). On dynamic failure extropy. *Journal of the Indian Society for Probability and Statistics*, **21**, 287-313.

Nair, R. D. and Sathar, E. I. A. (2022). A study on some properties of dynamic survival extropy and its relation to economic measures. *Stochastic and Quality Control*, **37**, 65-74. doi:10.1515/eqc-2021-0050.

Nair, R. D. and Sathar, E. I. A. (2023). Some useful results related to various measures of extropy and their interrelationship. *Statistics and Probability Letters*, **193**, doi: 10.1016/j.spl.2022.109729.

Noughabi, H. A., and Jarrahiferiz, J. (2019). On the estimation of extropy. *Journal of Non-parametric Statistics*, **31**, 88-99.

Qiu, G. and Jia, K. (2018). The residual extropy of order statistics. *Statistics and Probability Letters*, **133**, 15-22.

Sathar, E. I. A. and Nair, R.D. (2021). On dynamic survival extropy. *Communications in Statistics-Theory and Methods*, **50**, 1295-1313.

Sathar, E. I. A. and Nair, R.D. (2024). Properties of extropy and its weighted version for doubly truncated random variables. *Ricerche di Matematica*, https://doi.org/10.1007/s11587-024-00863-8.

Shannon, C. E. (1948). A mathematical theory of communication. *Bell System Technical*, **27**, 379-423 and 623-656.

Tahmasebi, S. and Toomaj, A. (2020). On negative cumulative extropy with applications. *Communications in Statistics - Theory and Methods*, **51**, 5025–5047.

Wicksell, S. D. (1925). The corpuscle problem: A mathematical study of a biometric problem. *Biometrika*, **17**, 84–99.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 23, No. 1, 2025 (New Series), pp 335–346
https://www.ssca.org.in/journal

Mathematical Model for Spread of COVID-19 Virus using Fractional Order Approach

Gajanan S. Solanke[1] **and Deepak B. Pachpatte**[2]
[1]*Department of First Year Engineering, CSMSS Chh. Shahu College of Engineering, Chhatrapati Sambhajinagar (Aurangabad), M.S., India-431002*
[2]*Department of Mathematics, Dr. Babasaheb Ambedkar Marathwada University,Chhatrapati Sambhajinagar (Aurangabad), M.S, India-431004*

Received: 19 December 2023; Revised: 19 May 2024; Accepted: 15 July 2024

Abstract

The COVID-19 pandemic has exhibited unprecedented and adverse effects on global health and living patterns. To understand and predict the spread of any disease, the Susceptible-Infectious-Recovered (SIR) model has been extensively used. However, the SIR model is failing to accurately predict the dynamics of complex systems. Therefore, this research proposes the development of a modified fractional order derivative for modelling the COVID-19 epidemic. The total population is divided into four classes with exclusive consideration of quarantined individuals. Attributes such as positivity, boundedness of solution, and stability of a model at disease-free equilibrium are thoroughly studied. The obtained results are utilized to predict the progression of COVID-19 through modelling.

Key words: Mathematical Model; Epidemic model; COVID-19 virus; Fractional calculus; Stability.

AMS Subject Classifications: 26A33, 34A30, 92-10, 00A71, 34D20.

1. Introduction

Predicting the projection of COVID-19 till remains a challenge that demands the integration of epidemiological models, statistical analyses, and real-time data streams. The COVID-19 pandemic has wrought profound and multifaceted effects on the world, touching nearly every aspect of human life. Economically, it has triggered widespread disruptions, leading to job losses, business closures, and supply chain bottlenecks, exacerbating inequalities and pushing millions into poverty. Socially, it has imposed isolation, disrupted education, and strained healthcare systems to their limits, with far-reaching implications for mental health and well-being. Therefore modelling of spread of COVID-19 is much necessary to predict the spread disease. There are various modelling approaches to forecast the spread of disease such as SIR, statistical, machine learning.

Mathematical modeling of epidemics was first introduced by W.O. Kermack and A.G

Corresponding Author: Gajanan S. Solanke
Email: gssolanke@gmail.com

McKendrick in year 1927 Kermack and McKendrick (1927). Since then many Mathematicians, Life sciences scientist and Medical professionals have used such type of model for studying the dynamics of various infectious diseases. By developing the model helps in predicting the infections. In Silva and Torres (2014) authors have developed a mathematical model on TB-HIV syndemic and treatment, in this model they divided total population into ten partitions and discussed about the positivity, boundedness of the solution and also discussed the stability. In Diethelm (2013); Solanke and Pachpatte (2019, 2021) the mathematical models on TB, dengue and Swine flu diseases have been studied.

COVID-19 virus is the one of family members of coronavirus, which are RNA viruses can be mild to lethal. It is harmful because of the risk factor, as some strains can kill up to 30 percent of affected people. As of June 22, 2021, there have been 180101870 confirmed cases and 3902501 deaths worldwide. Fever, dry cough, dyspnea, diarrhoea, sore throat and other symptoms are prevalent. It can infect cats, dogs, camels, and horses, among other animals web (a,b). The first incidence of COVID-19 was discovered in China in 2019. The first case of COVID-19 in India was detected by a person travelling from China to Kerala in the last week of January 2020.

Recently many researchers all over the world have started developing the Models for studying the behaviour of COVID-19. In Brandenburg (2020); Vega (2020) authors have developed a SIR model on COVID-19 for piecewise quadratic growth and discussed about the lockdowns due to increased infections of COVID-19 disease. Also in Shaikh *et al.* (2020) authors have discussed a mathematical model on COVID-19 formed by using fractional order derivative and also discussed the stability using the Laplace transform method. The SEIR model on COVID-19 disease have developed and studied local stability and global stability by some researchers Ssematimba *et al.* (2021); Tiwari *et al.* (2020); Wang *et al.* (2020). Some authors have developed the model by including class of quarantine or isolated population Krishna (2020); Mnganga and Zachariah (2020); Peter *et al.* (2021); Tanga *et al.* (2020), in Lina *et al.* (2020) different classes are provided for cumulative cases and deaths occurred due to COVID-19. The environmental changes such as population, air changes due to COVID-19 disease and effect of lockdown occurred due to COVID-19 Kerimray *et al.* (2020); Xu *et al.* (2020). In Makade *et al.* (2020), they have discussed about the most influential parameter for the spread of COVID-19 disease. Spread of COVID-19 active infection cases in three countries India, Italy and United States Of America(USA) have studied Pachpatte *et al.* (2021). Some logistic models, dynamic models and on spatial density are also developed on COVID-19 Abusam *et al.* (2020); Adekunle *et al.* (2020); Alzahrani *et al.* (2021); Al-Khani *et al.* (2020); Vaz and Torres (2021); Zaitri *et al.* (2021).

Simple mathematical model to investigate the transmission and regulation of the novel coronavirus disease (COVID-19) from human to human has been done in Ahmed *et al.* (2021). The researchers used mathematical epidemiology principles to model, how people are exposed to and infected with the disease, as well as their possible future recovery. Both the ordinary differential equation (ODE) and the fractional differential equation were used in the mathematical study. It is critical for health practitioners and the rest of the world to understand and predict infected individuals in order to plan for citizens' health concerns and to control the spread rate with limited supply. The simulation's data is based on the spread of disease in Nigeria.

Riyapan *et al.* (2021) have proposed and examined nonlinear mathematical model in terms of understanding the dynamics of the COVID-19 epidemic in Thailand. The formulated model's equilibrium point was determined. The basic reproduction number pertaining to the model was also calculated using the next generation matrix approach. Haq *et al.* (2023) have created a vaccination model by including the vaccine class and other factors that are crucial for immunizing those who are susceptible.

Motivated by the above work, a mathematical model for COVID-19 disease containing fractional order derivative and the properties about their solution are studied.

2. Mathematical model

2.1. Preliminaries

Now, in this section, some basic terminology that will come in useful during our discussions.

Definition 1: Podlubny (1999); Zhou (2014). The fractional calculus in classical form is given by the Riemann–Liouville integral which can be defined as

$$ {}_aD_t^{-\tau}(u(t)) = {}_a\, I_t^{\tau}(u(t)) = \frac{1}{\Gamma(\tau)} \int_a^t (t-\varsigma)^{\tau-1} u(\varsigma) d\varsigma, \tag{1} $$

where $t > a$.

Definition 2: Shaikh *et al.* (2020); Podlubny (1999); Zhou (2014). The Caputo fractional derivative operator of order τ where $\tau \geq 0$ and $n \in N \cup \{0\}$ can be given as

$$ D_t^{\tau}(u(t)) = \frac{1}{\Gamma(n-\tau)} \int_0^t (t-\varsigma)^{n-\tau-1} \frac{d^n}{dt^n} u(\varsigma) d\varsigma, \tag{2} $$

where $n-1 \leq \tau < n$.

2.2. Mathematical model

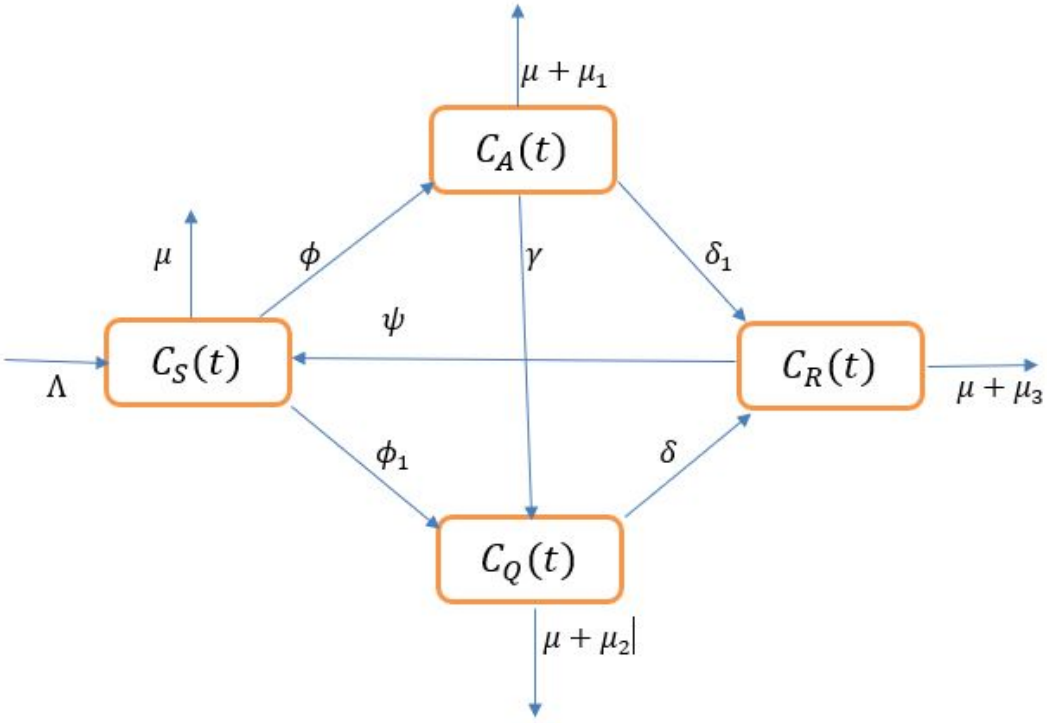
In this section present a model with class of quarantined individuals. Suppose that the entire population is organised into four classes, each of which is mutually exclusive, meaning that no individual may be assigned to more than one. Divided the population in four classes as three partitions are not sufficient to study the asymptomatic individuals and for this fifth partition is not required. Use $N(t)$ for total population is function of time t. Define four classes as
$C_s(t)$- Class of individuals Susceptible at time t,
$C_Q(t)$-Class of individuals Quarantined time t,
$C_A(t)$- Class of individuals asymptomatic (infected individuals not having symptoms) and not quarantined at time t,
$C_R(t)$- Class of individuals recovered at time t with or without medical treatment.

Total population $N(t)$ is given by
$N(t) = C_s(t) + C_A(t) + C_Q(t) + C_R(t)$.

We denote Λ is the birth rate, μ is the natural death rate, μ_1 is the death rate of asymptomatic people due to COVID-19, similarly μ_2 is the death rate of quarantine people

due to COVID-19. Some deaths are occurred in recovered class due to post COVID-19 diseases at the rate of μ_3. In this consider that very less number of asymptotic individuals recover without treatment means most of them have to pass through the quarantine class. Here β_2 denotes effective contact rate with infected people, which comprises 2 parameters as ϕ is the rate at which the people get infected but not having symptoms (not quarantined) and ϕ_1 is the rate at which people are infected and having symptoms and they are quarantined. The asymptomatic peoples recovers at the rate of δ_1 without any treatment and they are quarantined at the rate γ when they becomes symptomatic and tested positive. Quarantined people recovers at the rate of δ and recovered people will become susceptible at the rate ψ after some time due loss of immunity.

Above system can be represented in model format as follows:



$$\frac{dC_S(t)}{dt} = \Lambda + \psi C_R(t) - (\phi + \phi_1 + \mu)C_S(t), \tag{3}$$

$$\frac{dC_A(t)}{dt} = \phi C_S(t) - (\gamma + \delta_1 + \mu + \mu_1)C_A(t), \tag{4}$$

$$\frac{dC_Q(t)}{dt} = \phi_1 C_S(t) + \gamma C_A(t) - (\delta + \mu + \mu_2)C_Q(t), \tag{5}$$

$$\frac{dC_R(t)}{dt} = \delta C_Q(t) + \delta_1 C_A(t) - (\psi + \mu + \mu_3)C_R(t), \tag{6}$$

with initial conditions

$$C_S(0) \geq 0, C_A(0) \geq 0, C_Q(0) \geq 0, C_R(0) \geq 0. \tag{7}$$

Construct a mathematical model using fractional derivative.

The above system can written using Caputo fractional derivative operator as follows

$$D_t^\tau(C_S(t)) = \Lambda + \psi C_R(t) - (\phi + \phi_1 + \mu)C_S(t), \tag{8}$$

$$D_t^\tau(C_A(t)) = \phi C_S(t) - (\gamma + \delta_1 + \mu + \mu_1)C_A(t), \tag{9}$$

$$D_t^\tau(C_Q(t)) = \phi_1 C_S(t) + \gamma C_A(t) - (\delta + \mu + \mu_2)C_Q(t), \tag{10}$$

$$D_t^\tau(C_R(t)) = \delta C_Q(t) + \delta_1 C_A(t) - (\psi + \mu + \mu_3)C_R(t), \tag{11}$$

with the initial conditions

$$C_S(t) \geq 0, C_A(t) \geq 0, C_Q(t) \geq 0, C_R(t) \geq 0. \tag{12}$$

2.3. Properties of model

Now in this section we study the positivity and boundedness properties of the solution of the system. Let $\{(C_S, C_A, C_Q, C_R) \in \mathbb{R}_+^4\}$ be any solution of system 8 - 11 with initial conditions 12. Now, let us assume the region $\omega = \{(C_S, C_A, C_Q, C_R) \in \mathbb{R}_+^4 : 0 \leq N(t) \leq \frac{\Lambda}{\mu}\}$.

Now we prove positivity of system 8 - 11 with initial conditions 12 in our next theorem.

Theorem 1: Let $\{(C_S, C_A, C_Q, C_R) \in \mathbb{R}_+^4\}$ be any solution of system 8 - 11 with initial conditions 12. Consider

$$\omega = \left\{(C_S, C_A, C_Q, C_R) \in \mathbb{R}_+^4 : 0 \leq N(t) \leq \frac{\Lambda}{\mu}\right\}, \tag{13}$$

then $C_S(0) \geq 0, C_A(0) \geq 0, C_Q(0) \geq 0, C_R(0) \geq 0$.

Proof: We will prove our result by contradiction, suppose on contrary that for some point $\tilde{t} > 0$, the $C_A(t) = 0$, *i.e.* $C_A(\tilde{t}) = 0$ and $C_S(t) \geq 0, C_Q(t) \geq 0, C_R(t) \geq 0$ (given).

Then from equation 9 we have

$$D_t^\tau(C_A(t)) > 0, \tag{14}$$

which is not true.

Thus, $C_A(t) \geq 0$ for all $t > 0$.

Similarly, one can prove that, $C_S(t) \geq 0$, $C_Q(t) \geq 0$, $C_R(t) \geq 0$ for all time $t > 0$. □

Now proof of a boundedness of system 8 - 11 with initial conditions 12 is in next theorem.

Theorem 2: If $N(t)$ is the total population given by

$$N(t) = C_S(t) + C_A(t) + C_Q(t) + C_R(t),$$

then

$$D_t^\tau(N(t)) \leq \Lambda - \mu N(t), \tag{15}$$

where, Λ- birth rate, μ- death rate.

Proof: Since,

$$N(t) = C_S(t) + C_A(t) + C_Q(t) + C_R(t),$$

from the equations of system 8 - 11, we have

$$\begin{aligned}
D_t^\tau(N(t)) =& D_t^\tau(C_S(t)) + D_t^\tau(C_A(t)) + D_t^\tau(C_Q(t)) + D_t^\tau(C_R(t)),\\
=& \Lambda + \psi C_R(t) - (\phi + \phi_1 + \mu)C_S(t) + \phi C_S(t) - (\gamma + \delta_1 + \mu + \mu_1)C_A(t)\\
&+ \phi_1 C_S(t) + \gamma C_A(t) - (\delta + \mu + \mu_2)C_Q(t) + \delta C_Q(t) + \delta_1 C_A(t)\\
&- (\psi + \mu + \mu_3)C_R(t),\\
=& \Lambda + \psi C_R(t) - \phi C_S(t) - \phi_1 C_S(t) - \mu C_S(t)) + \phi C_S(t) - \gamma C_A(t)\\
&- \delta_1 C_A(t) - \mu C_A(t) - \mu_1 C_A(t) + \phi_1 C_S(t) + \gamma C_A(t) - \delta C_Q(t)\\
&- \mu C_Q(t) - \mu_2 C_Q(t) + \delta C_Q(t) + \delta_1 C_A(t) - \psi C_R(t) - \mu C_R(t)\\
&- \mu_3 C_R(t),\\
=& \Lambda - \mu C_S(t)) - \mu C_A(t) - \mu_1 C_A(t) - \mu C_Q(t) - \mu_2 C_Q(t) - \mu C_R(t)\\
&- \mu_3 C_R(t).
\end{aligned}$$

Since $N(t) = C_S(t) + C_A(t) + C_Q(t) + C_R(t)$,

$$D_t^\tau(N(t)) = \Lambda - \mu N(t) - \mu_1 C_A(t) - \mu_2 C_Q(t) - \mu_3 C_R(t) \ \leq \ \Lambda - \mu N(t).$$

Therefore, conclude that $N(t)$ is bounded for all $t > 0$ and every solution of system 8 - 11 with initial conditions 12 is bounded. □

3. Stability analysis

In this section stability of the system 8 - 11 with initial conditions 12 is studied. Now give some basic definitions of stability analysis given in Remsing (2006).

Definition 3: Remsing (2006) An equilibrium state $x = 0$ is said to be stable, if for any positive scalar ϵ there exists a positive scalar δ such that $||x(t_0)|| < \delta$ implies $||x(t)|| < \epsilon$ for all $t \geq t_0$.

Definition 4: Remsing (2006) An equilibrium state $x = 0$ is said to be asymptotically stable, if it is stable and if in addition $x(t) \to 0$ as $t \to \infty$.

The system 8 - 11 with initial conditions 12 is said to have disease free equilibrium (No disease) if

$$\Sigma_o = (C_{S0}, C_{A0}, C_{Q0}, C_{R0}) = \left(\frac{\Lambda}{\mu + \phi + \phi_1}, 0, 0, 0\right). \tag{16}$$

The Endemic Equilibrium is given by

$$\Sigma_* = (C_{S*}, C_{A*}, C_{Q*}, C_{R*}), \tag{17}$$

with $C_{A*} > 0$, $C_{Q*} > 0$, $C_{R*} > 0$ for $R_0 > 1$, where R_0 is the basic reproduction number for the system 8 - 11 with initial conditions 12.

The basic reproduction number is the average number of new infections due to a single individual when in contact with susceptible population.Silva and Torres (2014). Now in our next theorem gives the result on the stability of the system 8 - 11 with initial conditions 12.

Theorem 3: The disease free equilibrium Σ_0 is locally asymptotically stable if $R_0 < 1$.

Proof: The disease-free equilibrium Σ_0 is locally asymptotically stable for $R_0 < 1$, if all the eigenvalues of the Jacobian Matrix of the system of equations 8 - 11 here denoted by $M_T(\Sigma_0)$ computed at the disease free equilibrium Σ_0, given by 16 have negative real parts Remsing (2006); Benerjee (2014).

The Jacobian Matrix of the system of equations 8 - 11 at disease-free equilibrium is given by

$$J_0 = \begin{pmatrix} -\phi - \phi_1 - \mu & 0 & 0 & 0 \\ \phi & -\gamma - \delta_1 - \mu - \mu_1 & 0 & 0 \\ \phi_1 & \gamma & -\delta - \mu - \mu_2 & 0 \\ 0 & \delta_1 & \delta & -\psi - \mu - \mu_3 \end{pmatrix}.$$

The eigenvalues of this matrix are the roots of the equation $|J_0 - \lambda I| = 0$, consider

$$|J_0 - \lambda I|$$

$$= \begin{pmatrix} -\phi - \phi_1 - \mu - \lambda & 0 & 0 & 0 \\ \phi & -\gamma - \delta_1 - \mu - \mu_1 - \lambda & 0 & 0 \\ \phi_1 & \gamma & -\delta - \mu - \mu_2 - \lambda & 0 \\ 0 & \delta_1 & \delta & -\psi - \mu - \mu_3 - \lambda \end{pmatrix},$$

$$= (-\phi - \phi_1 - \mu - \lambda) \begin{pmatrix} -\gamma - \delta_1 - \mu - \mu_1 - \lambda & 0 & 0 \\ \gamma & -\delta - \mu - \mu_2 - \lambda & 0 \\ \delta_1 & \delta & -\psi - \mu - \mu_3 - \lambda \end{pmatrix},$$

$$= (-\phi - \phi_1 - \mu - \lambda)\Big\{(\gamma - \delta_1 - \mu - \mu_1, -\lambda)\Big[(-\delta - \mu - \mu_2 - \lambda)(-\psi - \mu - \mu_3 - \lambda)\Big]\Big\},$$

$$= -(\lambda + \phi + \phi_1 + \mu)\Big\{ - (\lambda + \gamma + \delta_1 + \mu + \mu_1)\Big[(\lambda + \delta + \mu + \mu_2)(\lambda + \psi + \mu + \mu_3)\Big]\Big\},$$

$$= (\lambda + \phi + \phi_1 + \mu)\Big\{(\lambda + \gamma + \delta_1 + \mu + \mu_1)\Big[(\lambda + \delta + \mu + \mu_2)(\lambda + \psi + \mu + \mu_3)\Big]\Big\}.$$

Let

$$\begin{aligned}
&\phi + \phi_1 + \mu = a,\\
&\gamma + \delta_1 + \mu + \mu_1 = b,\\
&\delta + \mu + \mu_2 = c,\\
&\psi + \mu + \mu_3 = d.
\end{aligned}$$

Therefore

$$\begin{aligned}
|J_0 - \lambda I| &= (\lambda + a)\Big\{(\lambda + b)\Big[(\lambda + c)(\lambda + d)\Big]\Big\},\\
&\Rightarrow (\lambda + a)(\lambda + b)(\lambda + c)(\lambda + d) = 0,\\
&\Rightarrow \lambda = -a, \lambda = -b, \lambda = -c, \lambda = -d.
\end{aligned}$$

Since all the parameters are greater than or equal to 0, $a \geq 0$, $b \geq 0$, $c \geq 0$, $d \geq 0$. Thus, all the eigenvalues of this Jacobian Matrix are negative.

The system of equations 8 - 11 with initial conditions 12 is stable. Thus a mathematical model 8 - 11 with initial conditions 12 is asymptotically stable at disease free equilibrium for $R_0 < 1$. □

4. Results and discussions

In this section verification of results with data of COVID-19 of India to this system has been done. The plot shows the curves for the Asymptomatic, Quarantine, Recovered population for COVID-19 disease in India. Here the parameters are as below, which are obtained from real data, MATLAB software is used for the numerical solution of the model.

Table 1: Parameters

Parameter	**Value**	**Parameter**	**Value**
Λ	0.0000563447	μ	0.0000194123
ϕ	0.000000125927	ϕ_1	0.00000029383
ψ	0.9999805877	μ_1	0.0118594279
δ_1	0.9881165471	γ	0.000046128
δ	0.8813863087	μ_2	0.118594279
μ_3	0	N	1380004385
$C_A(0)$	1	$C_Q(0)$	1
$C_R(0)$	0	.	.

For prediction of the spread of infection, the important parameters are β_2(effective contact rate with infected people), ϕ(the rate at which people got infected and not having any symptoms), ϕ_1(the rate at which people got infected and and having symptoms). For control of spread of infection the value of these parameter should be minimum.

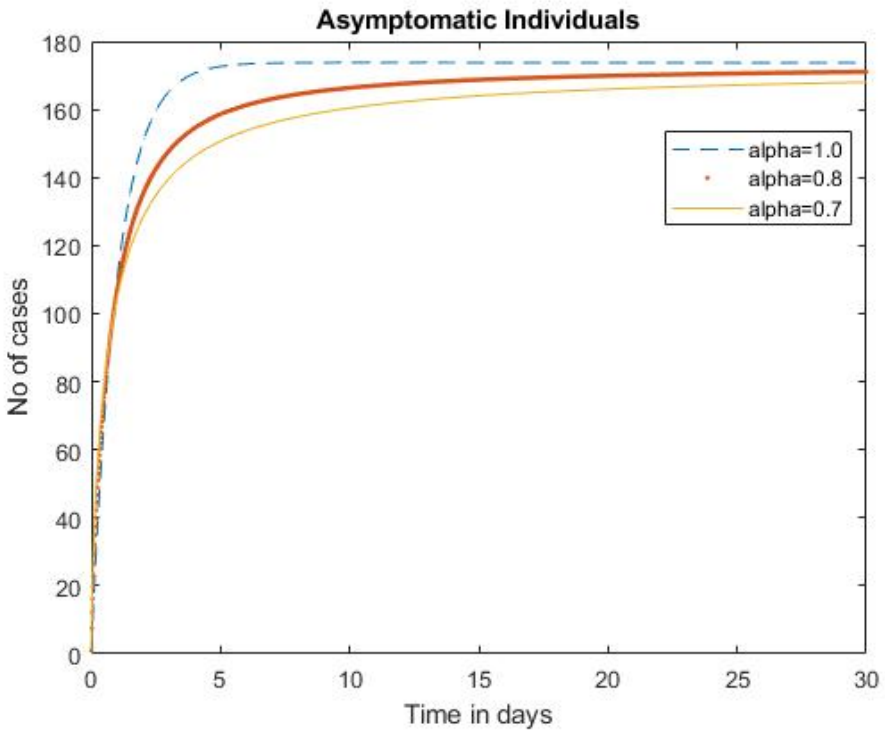


Figure 1: Graph for asymptomatic population

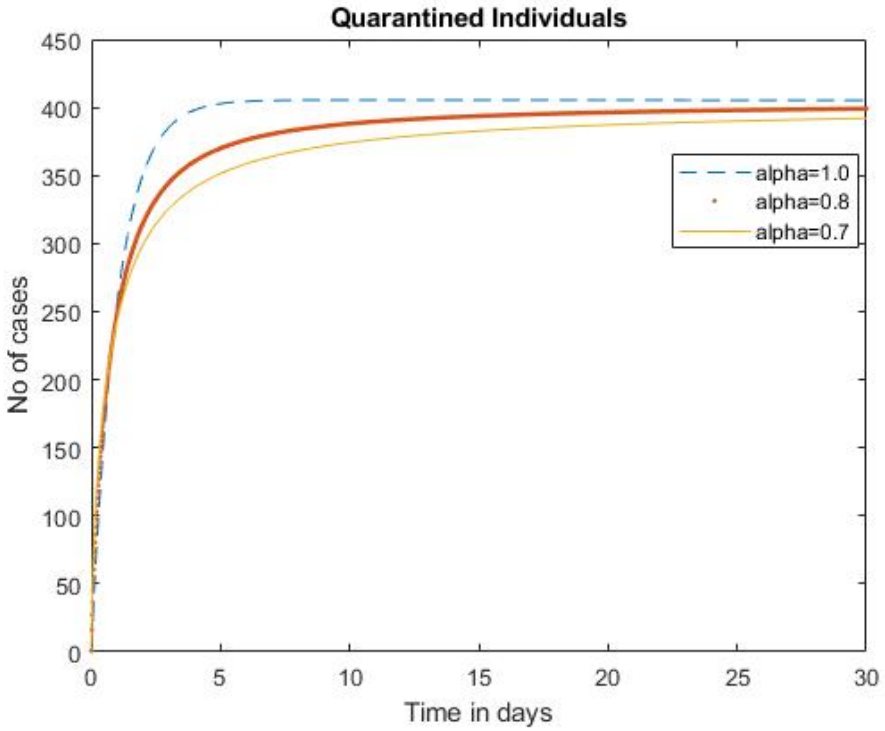


Figure 2: Graph for quarantined population

5. Conclusion

This research studied the mathematical model for the spread of COVID-19 disease using fractional derivatives successfully. The results regarding positivity, boundedness of

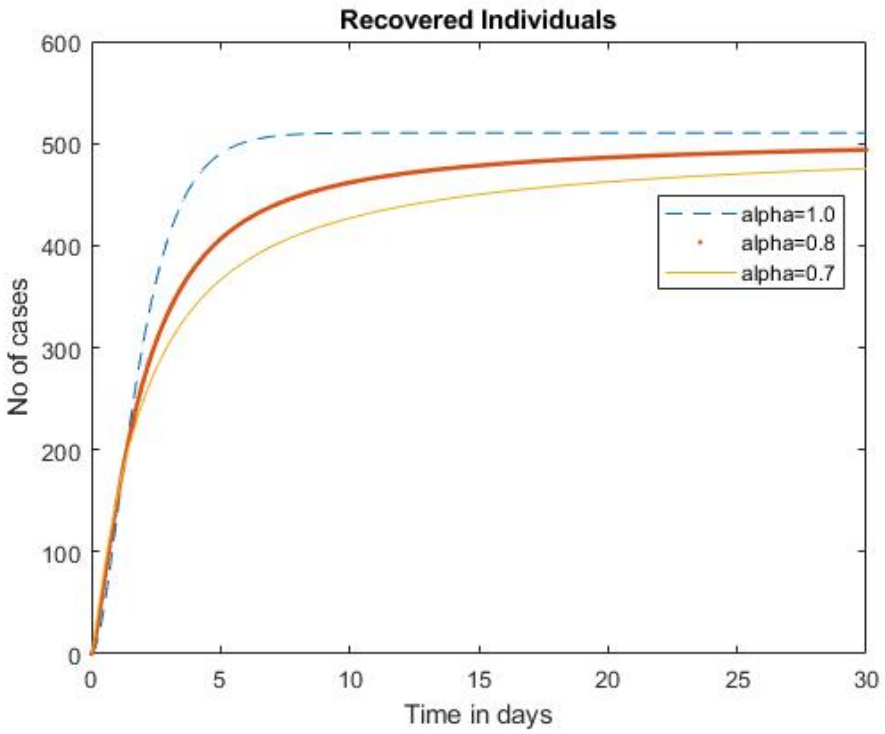


Figure 3: Graph for recovered population

the solution, and stability at equilibrium were obtained and analysed. The mathematical model's results demonstrate better performance compared to the conventional SIR model.

Conflict of interest

The authors do not have any financial or non-financial conflict of interest to declare for the research work included in this article.

References

Abusam, A., Abusam, R., and Al-Anzi, B. (2020). Adequacy of logistic models for describing the dynamics of COVID-19 pandemic. *Infectious Disease Modelling*, **5**, 536–542.

Adekunle, I., Onanuga, A., Akinola, O., and Ogunbanjo, O. (2020). Modelling spatial variations of coronavirus disease (COVID-19) in Africa. *Science of The Total Environment*, **729**, 138998.

Ahmed, I., Modu, G., Yusuf, A., Kumam, P., and Yusuf, I. (2021). A mathematical model of coronavirus disease (COVID-19) containing asymptomatic and symptomatic classes. *Results Physics*, **21**, 103776.

Al-Khani, A., Khalifa, M., Almazrou, A., and Saqui, N. (2020). The SARS-CoV-2 pandemic course in Saudi Arabia: a dynamic epidemiological model. *Infectious Disease Modelling*, **5**, 766–771.

Alzahrani, E., El-Dessoky, M., and Baleanu, D. (2021). Modeling the dynamics of the novel coronavirus using Caputo-Fabrizio derivative. *Alexandria Engineering Journal*, **60**, 4651–4662.

Benerjee, S. (2014). *Mathematical Modeling Models, Analysis and Applications*. CRC Press, Taylor and Francis Group.

Brandenburg, A. (2020). Piecewise quadratic growth during the 2019 novel coronavirus epidemic. *Infectious Disease Modelling*, **5**, 681–690.

Diethelm, K. (2013). A fractional calculus based model for the simulation of an outbreak of dengue fever. *Nonlinear Dynamics.*, **71**, 613–619.

Haq, I., Ullah, N., Ali, N., and Nisar, K. (2023). A new mathematical model of COVID-19 with quarantine and vaccination. *Mathematics*, **11**, 142.

Kerimray, A., Baimatova, N., Ibragimovaa, O., Bukenov, B., Kenessov, B., Plotitsyn, P., and Karaca, F. (2020). Assessing air quality changes in large cities during COVID-19 lockdowns: The impacts of traffic-free urban conditions in almaty, kazakhstan. *Science of The Total Environment*, **730**, 139179.

Kermack, W. and McKendrick, A. (1927). A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society A - Journal*, **115**, 700–721.

Krishna, M. (2020). Mathematical modelling on diffusion and control of COVID-19. *Infectious Disease Modelling*, **5**, 588–597.

Lina, Q., Zhaob, S., Gaod, D., Loue, Y., Yangf, S., Musae, S., Wangb, M., Caig, Y., Wangg, W., Yangh, L., and Hee, D. (2020). A conceptual model for the coronavirus disease 2019 (COVID-19) outbreak in wuhan, china with individual reaction and governmental action. *International Journal of Infectious Diseases*, **93**, 211–220.

Makade, R., Chakrabarti, S., and Jamil, B. (2020). Real-time estimation and prediction of the mortality caused due to COVID-19 using particle swarm optimization and finding the most influential parameter. *Infectious Disease Modelling*, **5**, 728–772.

Mnganga, J. and Zachariah, N. (2020). Mathematical model of COVID-19 transmission dynamics and control strategies. *International Journal of Advanced Research in Computer Science*, **11**.

Pachpatte, D., Solanke, G., and Tidke, H. (2021). Analysis and prediction of active infection cases for COVID-19 virus in India, Italy and USA. *Stochastic Modeling and Applications*, **25**, 65–72.

Peter, O., Qureshi, S., Yusuf, A., Al-Shomrani, M., and Idowu, A. (2021). A new mathematical model of COVID-19 using real data from Pakistan. *Results Physics*, **24**, 104098.

Podlubny, I. (1999). *Fractional Differential Equations, Mathematics in Science and Engineering*. Academic Press, USA.

Remsing, C. (2006). *Linear Control*. Rhodes University Grahamstown 6140, South Africa.

Riyapan, P., Shuaib, S., and Intarasit, A. (2021). A mathematical model of COVID-19 pandemic: a case study of Bangkok, Thailand. *Computational and Mathematical Methods in Medicine*, **2021**, 6664483.

Shaikh, A., Shaikh, I., and Nisar, K. (2020). A mathematical model of COVID-19 using fractional derivative: outbreak in India with dynamics of transmission and control. *Advances in Difference Equations*, **373**, .

Silva, C. and Torres, D. (2014). Modeling TB-HIV syndemic and treatment. *Journal of Applied Mathematics*, **1**, 248407.

Solanke, G. and Pachpatte, D. (2019). A fractional order differential equation model for tuberculosis. *AIP Conference Proceedings*, **2061**, 020007–1–5.

Solanke, G. and Pachpatte, D. (2021). Modelling for outbreak of swine flu using fractional derivative. *Stochastic Modeling and Applications*, **25**, 49–57.

Ssematimba, A., Nakakawa, J., Ssebuliba, J., and Mugisha, J. (2021). Mathematical model for COVID-19 management in crowded settlements and high-activity areas. *International Journal of Dynamics and Control*, **9**, 1358–1369.

Tanga, B., Xiaa, F., Tangc, S., Bragazzib, N., Lib, Q., Suna, X., Liangc, J., Xiaoa, Y., and Wua, J. (2020). The effectiveness of quarantine and isolation determine the trend of the COVID-19 epidemics in the final phase of the current outbreak in China. *International Journal of Infectious Diseases*, **95**, 288–293.

Tiwari, V., Deyal, N., and Bisht, N. (2020). Mathematical modeling based study and prediction of COVID-19 epidemic dissemination under the impact of lockdown in India. *Frontiers in Physics*, **8**, 586899.

Vaz, S. and Torres, D. (2021). A discrete-time compartmental epidemiological model for COVID-19 with a case study for Portugal. *Axioms*, **10**, 314.

Vega, D. (2020). Lockdown, one, two, none, or smart. modeling containing COVID-19 infection. a conceptual model. *Science of The Total Environment*, **730**, 138917.

Wang, T., Wu, Y., Lau, J., Yu, Y., Liu, L., Li, J., Zhang, K., Tong, W., and Jiang, B. (2020). A four-compartment model for the COVID-19 infection—implications on infection kinetics, control measures, and lockdown exit strategies. *Precision Clinical Medicine*, **3**, 104–112.

Ministry of Health and Family Welfare, Government of India, https://www.mohfw.gov.in (as on 22nd june 2021). .

World Health Orginization(WHO), https://www.who.int (as on 22nd june 2021). .

Xu, H., Yanb, C., Fuc, Q., Xiao, K., Yua, Y., Hane, D., Wang, W., and Cheng, J. (2020). Possible environmental effects on the spread of COVID-19 in China. *Science of The Total Environment*, **731**, 139211.

Zaitri, M., Bibi1, M., and Torres, D. (2021). Optimal control to limit the spread of COVID-19 in Italy. *Kuwait Journal of Science*, **2**, 1–14.

Zhou, Y. (2014). *Basic Theory of Fractional Differential Equations*. World Scientific.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 23, No. 1, 2025 (New Series), pp 347–364
https://www.ssca.org.in/journal

Neutrosophic Marshall Olkin Extended Burr-XII Distribution: Theoretical Framework and Applications with Multiple Survival Time Data Sets

Shakila Bashir[1], **Bushra Masood**[1] **and Muhammad Aslam**[2]
[1]*Department of Statistics, Forman Christian College (A Chartered University) Lahore, 54600, Pakistan*
[2]*Department of Statistics, Faculty of Science, King Abdulaziz University, Jeddah 21551, Saudia Arabia*

Received: 04 March 2024; Revised: 12 June 2024; Accepted: 31 July 2024

Abstract

In the analysis of complex data sets, selecting an appropriate distribution is crucial for real-life applications. Common probability distributions often fail to provide adequate results when dealing with imprecise, uncertain, or vague data. To address these complexities and achieve more accurate results, a neutrosophic probability distribution called the neutrosophic Marshall-Olkin extended Burr-XII distribution has been developed. This study aims to introduce a lifetime distribution capable of handling indeterminate data. Various properties of the proposed distribution are discussed. The maximum likelihood method, in terms of neutrosophic parameters, is utilized to estimate these parameters. A simulation study is conducted to validate the estimated neutrosophic parameters. Finally, two real-life data sets are analyzed to demonstrate the potential of the NMOE Burr-XII distribution, highlighting its superior efficiency and adaptability compared to classical distributions when dealing with indeterminate survival time data.

Key words: Neutrosophic statistics; Simulations; Burr-XII; Marshall-Olkin.

AMS Subject Classifications: 62K05, 05B05

1. Introduction

The Burr-XII distribution is significant in lifetime and survival data analysis. Shao *et al.* (2004) investigated models for the extended three-parameter Burr type XII distribution and applied it to model severe events, such as flood frequency. Rodriguez (1977) examined the adaptability of the Burr type XII distribution, which has been widely used in various scientific fields, including actuarial science, forestry, ecotoxicology, dependability, and survival analysis. Marshall and Olkin (1997) introduced a parameter to create a new family of distributions that are more flexible and cover a broader range of behaviors than previous distributions, known as extended distributions. Al-Saiari *et al.* (2014) further extended this

Corresponding Author: Muhammad Aslam
Email: aslam_ravian@hotmail.com

by adding one parameter to the Marshall-Olkin Extended (MOE) Burr-XII distribution, resulting in the Marshall Olkin extended Burr-XII distribution.

Neutrosophic statistics, initially introduced in 1995 and further developed by Smarandache (2014) explored the nature, origin, and application of neutralities. Neutrosophic logic is a special form of fuzzy logic. The neutrosophic statistics is more efficient than the classical statistics and interval-statistics see Smarandache (2022). While classical statistics rely on definite data, neutrosophic statistics handle partial, imprecise, ambiguous, or indeterminate data. The two fields coincide when indeterminacy is zero Chen *et al.* (2017). Neutrosophic statistics provide more accurate results by differentiating between those who partially and fully belong to a dataset. When all data and inference techniques are determined, both classical and neutrosophic statistics occur simultaneously. Neutrosophic statistics (NS) offer several advantages over interval statistics. In probability distributions, NS employs thick functions, formed by the intersections of curves, which may or may not be depicted as intervals Smarandache (2014). The neutrosophic probability distribution (NPD) for an event (x) comprises three curves: $NPD(x) = [T(x), I(x), F(x)]$, where $T(x)$ represents the probability of event E occurring, $I(x)$ denotes the indeterminate probability of E occurring or not, and $F(x)$ signifies the probability of E not occurring. These functions T(x),I(x), and F(x) can take on classical or neutrosophic (unclear, approximate, thick) forms depending on the specific application, and their sum ranges from 0 to 3 Smarandache (2013). Many researchers have developed neutrosophic probability distributions. For example, Fawzi *et al.* (2019) introduced the neutrosophic Weibull distribution and its related family, including the neutrosophic Weibull, Neutrosophic Rayleigh, neutrosophic inverse Weibull, and neutrosophic three- and six-parameter Weibull, as well as the Neutrosophic beta distribution. Rao (2023) developed the neutrosophic Log-logistic distribution, while Khan *et al.* (2021b) introduced the neutrosophic Gamma Distribution. Duan *et al.* (2021) presented the neutrosophic exponential distribution, and Khan *et al.* (2021a) proposed the Neutrosophic Beta distribution. Albassam *et al.* (2023) explored some basic properties of the neutrosophic Weibull Distribution with applications to wind speed in uncertain environments. Nayana *et al.* (2022) proposed the DUS Neutrosophic Weibull Distribution, and Eassa *et al.* (2023) introduced the neutrosophic generalized Pareto Distribution, modeling it on public debt in Egypt. Khan *et al.* (2021c) developed the Neutrosophic Rayleigh model for indeterminate data and also created V charts, neutrosophic run length, and Neutrosophic power curves for the proposed model. Sherwani *et al.* (2021) introduced new entropy measures for the Weibull Distribution under neutrosophic data, and Granados *et al.* (2022) applied both continuous and discrete probability distributions to Neutrosophic data. According to Granados *et al.* (2022), fuzzy logic is a special case of Neutrosophic logic, which generalizes fuzzy logic.

The article is structured as follows: Section 2 outlines the development of the novel Neutrosophic Marshall extended Burr-XII distribution, including graphical representation. Sections 3 and 4 discuss various properties of the proposed density. Section 5 focuses on the estimation of unknown parameters and simulation studies. Section 6 presents applications of the proposed model. Section 7 provides a discussion on these applications, and Section 8 offers concluding remarks.

2. Development of neutrosophic Marshall Olkin extended Burr-XII distribution

In this section, we will introduce the neutrosophic Marshall Olkin Extended Burr-XII distribution.

2.1. Marshall Olkin extended Burr-XII model

Burr type distributions are extensively used in life data and survival analysis. Adding more parameters to the Burr-XII distribution enhances its flexibility and appeal. Consequently, this study selects the Marshall-Olkin extended Burr-XII model for the development of a Neutrosophic model. The cumulative distribution function (CDF) and the probability density function (PDF) of the Marshall-Olkin Extended Burr-XII (MOE Burr-XII) distribution are as follows.

$$F(x;\alpha,\beta,\gamma) = \frac{1-\left(1+x^{\beta}\right)^{-\gamma}}{1-(1-\alpha)\left(1+x^{\beta}\right)^{-\gamma}}, x,\alpha,\beta,\gamma > 0 \tag{1}$$

and

$$f\left(x;\alpha,\beta,\gamma\right) = \frac{\alpha\beta\gamma x^{\beta-1}\left(1+x^{\beta}\right)^{-\gamma-1}}{\left[1-(1-\alpha)\left(1+x^{\beta}\right)^{-\gamma}\right]^{2}}, x,\alpha,\beta,\gamma > 0 \tag{2}$$

2.2. Neutrosophic random variable

Rao *et al.* (2023) discussed the extension of classical statistics called neutrosophic statistics. In classical statistics, we work with specific or predefined values. In contrast, neutrosophic statistics involves selecting values or data from a population within an unpredictable environment. For instance, when recording the temperature of a place, we might not be able to capture a precise value, such as 35°C. Instead, the value could have an uncertainty range, like 35°C to 38°C. The information in this context can be confusing, inaccurate, doubtful, partial, or even unknown.

Assuming the neutrosophic random variable $X_N = X_L + I_N X_L$, where $I_N \in [I_L, I_U]$ wherever $I_N X_L$ is the indeterminate and $I_N \in [I_L, I_U]$ is the indeterminacy. It is importance to notice that the neutrosophic random variable is the extension of the classical random variable specifically when $I_L = 0$ the neutrosophic random variable converts into classical random variable. According to his, the properties of the expectation of the neutrosophic random variable $X_N = X_L + I_N X_L = (1 + I_N)X_L$ is defined as:

Aslam and Albassam (2024) explored the mean properties of the neutrosophic random variable $X_N = X_L + X_L I_N$, defined as:

1. $E\left(X_N\right) = E\left(X_L + X_L I_N\right) = \left(1 + I_N\right) E\left(X_L\right) = \left(1 + I_N\right)\mu$
2. $E\left(X_N + t\right) = E\left[\left(X_L + X_L I_N\right) + t\right] = \left(1 + I_N\right)\mu + t$ here t is a constant.
3. $E\left(sX_N + t\right) = E\left[s\left(X_L + X_L I_N\right) + t\right] = s\left(1 + I_N\right)\mu + t$ here s and t are constant.
4. $E\left(X_N + Y_N\right) = \left(1 + I_N\right)\mu_X + \left(1 + I_N\right)\mu_Y$

Now, the variance properties of the neutrosophic random variables are as follows:

1. $V(X_N) = V(X_L + X_L I_N) = (1 + I_N)^2 V(X_L) = (1 + I_N)^2 \sigma^2$
2. $V(tX_N) = t^2 V(X_L + X_L I_N) = t^2 (1 + I_N)^2 \sigma^2$
3. $V(X_N + Y_N) = (1 + I_N)^2 \sigma_X^2 + (1 + I_N)^2 \sigma_Y^2 + 2 I_N Cov(X_N, Y_N)$
4. V $(sX_N + tY_N) = s^2 (1 + I_N)^2 \sigma_X^2 + t^2 (1 + I_N)^2 \sigma_Y^2 + 2st I_N Cov(X_N, Y_N)$
5. If we have two independent variables, X_N and Y_N:
   $V(X_N + Y_N) = (1 + I_N)^2 \sigma_X^2 + (1 + I_N)^2 \sigma_Y^2$

Let suppose the random variable X arose from the Marshall Olkin extended Burr-XII distribution with the CDF and PDF given in equations 1 and 2, we consider that the neutrosophic statistical number N, and $I_N \in [I_L, I_U]$ is an interval of indeterminacy. If the neutrosophic variable $X_N = X_L + I_N X_L$, generates the neutrosophic values of data. According to this, the neutrosophic variable is defined as: $X_N = X_L + I_N X_L = (1 + I_N) X_L$ here indeterminate and determined parts are described by X_L and $I_N X_L$ respectively.

If the random variable in terms of neutrosophic statistic $X_N \in (1 + I_N) X_L$ follows the Marshall Olkin Extended Burr-XII (NMOE Burr-XII) then by using the equations 1 and 2, the PDF and CDF of the neutrosophic Marshall Olkin Extended Burr-XII (NMO Burr-XII) distribution are developed as given below.

$$f_N(x_N; \alpha, \beta, \gamma) = \frac{\alpha\beta\gamma (1 + I_N) [(1 + I_N) x_L]^{\beta - 1} \left[1 + \{(1 + I_N) x_L\}^{\beta}\right]^{-\gamma - 1}}{\left[1 - (1 - \alpha) [1 + \{(1 + I_N) x_L\}^{\beta}]^{-\gamma}\right]^2}, x_N, \alpha, \beta, \gamma > 0 \tag{3}$$

Similarly, the CDF of the NMOE Burr-XII distribution is,

$$F_N(x_N; \alpha, \beta, \gamma) = \int_0^x f_N(x_N; \alpha, \beta, \gamma)\, dx$$

$$F_N(x_N; \alpha, \beta, \gamma) = \frac{\left[1 - \left[1 + \{(1 + I_N) x_L\}^{\beta}\right]^{-\gamma}\right]}{\left[1 - (1 - \alpha) [1 + \{(1 + I_N) x_L\}^{\beta}]^{-\gamma}\right]} \tag{4}$$

Special cases of NMOE Burr-XII distribution.

1. For $\alpha = 1$, the NMOE Burr-XII becomes Neutrosophic Burr-XII distribution.
2. For $\beta = 1$, NMOE Burr-XII becomes the Neutrosophic Marshal Olkin Extended Lomax distribution.

To prove that equation 3 is density and equation 4 is CDF, the following theorems are given.

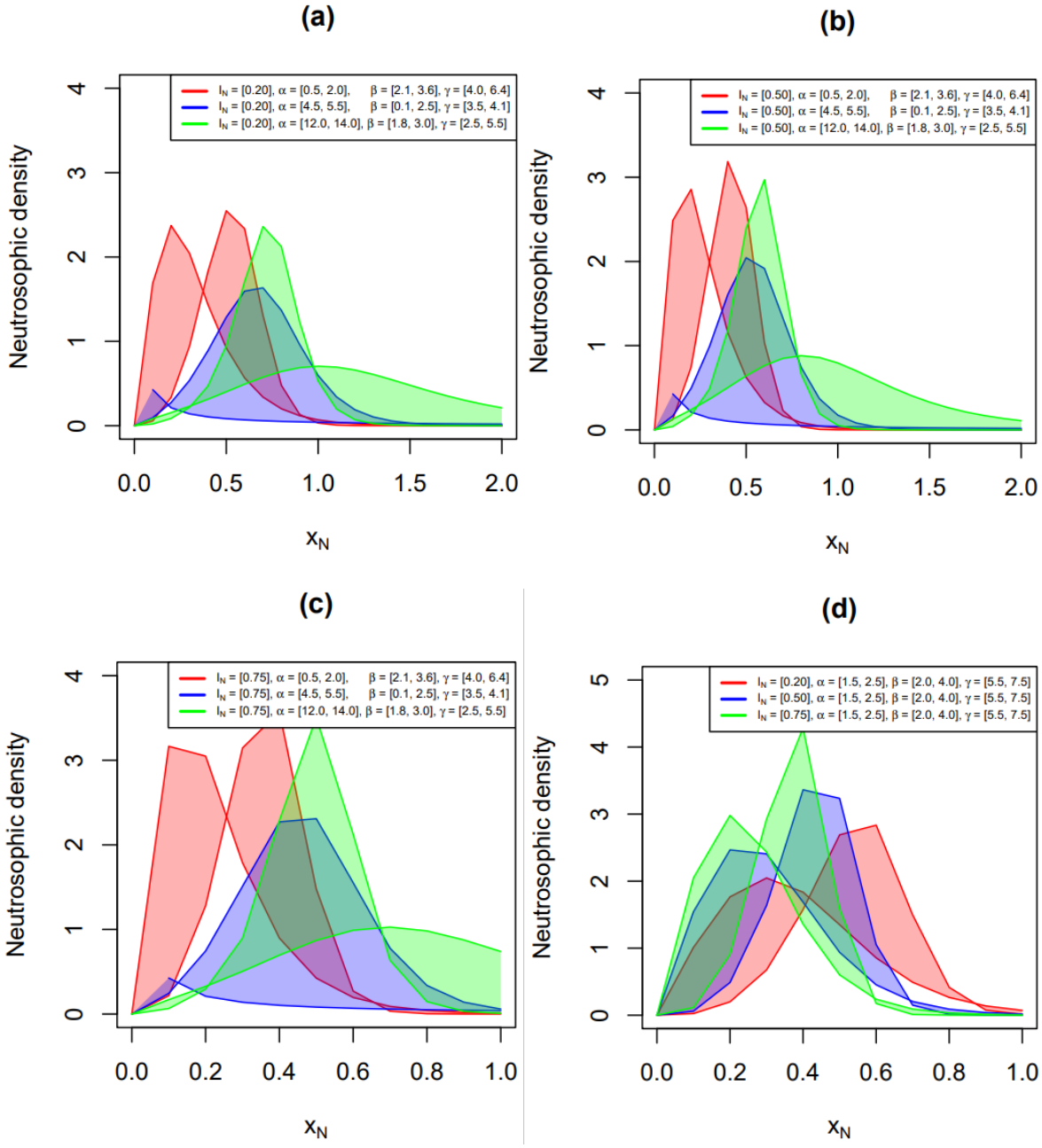


Figure 1: Density plots for the NMOE burr-XII distribution for different values of I_N, and parameters

Theorem 1: Consider $X_N \in (1 + I_N)X_L$ here indeterminate and determined parts are described by X_L and $I_N X_L$ respectively; suppose X_N follows the function given in equation 3 is a valid density function.

Proof: The random variable X follows the NMOE Burr-XII distribution in equation 3 then

$$\int_0^{\infty} \frac{\alpha\beta\gamma\left[(1+I_N)\,x_L\right]^{\beta-1}\left[1+\left\{(1+I_N)\,x_L\right\}^{\beta}\right]^{-\gamma-1}}{\left[1-(1-\alpha)\left[1+\left\{(1+I_N)\,x_L\right\}^{\beta}\right]^{-\gamma}\right]^2}(1+I_N)\;dx = 1$$

Let $\left[1+\{(1+I_N)\,x_L\}^{\beta}\right]^{-\gamma}=u$, and after some simplifications we get the

$$\alpha\int_0^1 \frac{1}{[1-(1-\alpha)u]^2}\,du=1$$

Again transform $[1-(1-\alpha)u]=z$, and simplifying it we get,

$$\frac{\alpha}{1-\alpha}\int_\alpha^1 \frac{1}{z^2}\,dz=1$$

The above integral is equal to one. Hence it is proved that equation 3 is a valid density function.

Theorem 2: Let the random variable $X_N \in (1+I_N)X_L$ here indeterminate and determined parts are described by X_L and $I_N X_L$ follows the NMOE Burr-XII distribution then the CDF given in equation 4 is a valid distribution function.

Proof: consider the random variable $X_N \in (1+I_N)X_L$ follows the CDF given in equation 4 then, it is proved that:

$$F(0)=0$$
$$F(\infty)=\infty$$

Hence the equation 4 is a valid distribution function. The graphical representation of the NMOE Burr-XII distribution is displayed below for different values of the parameters and varying I_N, Here, β & γ are the shape parameters, while α is the scale parameter. Figure 1 illustrates that the density is clearly unimodal.

3. Neutrosophic reliability measures

In this section, we develop several properties related to lifetime analysis, including survival analysis and the hazard function. The survival function is defined as the probability that an event or observation in survival data occurs after a specified time point. The survival function for the NMOE Burr-XII distribution is given as follows.

$$S_N\left(x_N;\alpha,\beta,\gamma\right)=\frac{\alpha\left[1+\{(1+I_N)\,x_L\}^{\beta}\right]^{-\gamma}}{\left[1-(1-\alpha)\left[1+\{(1+I_N)\,x_L\}^{\beta}\right]^{-\gamma}\right]} \tag{5}$$

The hazard rate function is a fundamental concept in survival analysis, which examines time-to-event data. The hazard rate function (HRF) for the NMOE Burr-XII distribution is derived as follows.

$$h_N\left(x_N;\alpha,\beta,\gamma\right)=\frac{\beta\gamma\,(1+I_N)\,\{(1+I_N)\,x_L\}^{\beta-1}}{\left[1-(1-\alpha)\left[1+\{(1+I_N)\,x_L\}^{\beta}\right]^{-\gamma}\right]\left[1+\{(1+I_N)\,x_L\}^{\beta}\right]} \tag{6}$$

Figure 2 presents the HRF shapes with various values of parameters and with different I_N. HRF of the NMOE Burr-XII distribution exhibits monotone increasing trend.

The cumulative hazard rate function for the NMOE Burr-XII distribution is

$$H\left(x_N,\alpha,\beta,\gamma\right)=-ln\left[\frac{\alpha\left[1+\left\{\left(1+I_N\right)x_L\right\}^{\beta}\right]^{-\gamma}}{\left[1-\left(1-\alpha\right)\left[1+\left\{\left(1+I_N\right)x_L\right\}^{\beta}\right]^{-\gamma}\right]}\right]$$

The reversed hazard rate function for the NMOE Burr-XII distribution is

$$r\left(x_N,\alpha,\beta,\gamma\right)=\frac{\alpha\beta\gamma\left(1+I_N\right)\left[\left(1+I_N\right)x_L\right]^{\beta-1}\left[1+\left\{\left(1+I_N\right)x_L\right\}^{\beta}\right]^{-\gamma-1}}{\left[1-\left(1-\alpha\right)\left[1+\left\{\left(1+I_N\right)x_L\right\}^{\beta}\right]^{-\gamma}\right]\left[1-\left[1+\left\{\left(1+I_N\right)x_L\right\}^{\beta}\right]^{-\gamma}\right]}$$

In the context of neutrosophic reliability measures, censoring can be accommodated by incorporating neutrosophic sets to handle the uncertainty and indeterminacy associated with censored data. This approach allows for a more flexible representation of reliability metrics, where traditional binary logic (failure or survival) is extended to include degrees of membership, indeterminacy, and non-membership, thus providing a nuanced way to account for incomplete information due to censoring.

4. Some statistical properties of neutrosophic Marshall Olkin extended Burr-XII distribution

This section explores various statistical properties of the NMOE Burr-XII distribution, including the mean, variance, quantile function, skewness, and kurtosis.
The mean of the neutrosophic MOE Burr-XII distribution is derived as

$$\mu_N=E\left[\left(1+I_N\right)X_L\right]=\left(1+I_N\right)E\left(X_L\right) \tag{7}$$

Where,

$$E\left(X_L\right)=E\left(X\right)=\int_0^{\infty}x\frac{\alpha\beta\gamma x^{\beta-1}\left(1+x^{\beta}\right)^{-\gamma-1}}{\left[1-\left(1-\alpha\right)\left(1+x^{\beta}\right)^{-\gamma}\right]^2}\,dx$$

The above expression does not have a closed form, so we can determine its numerical values by substituting the parameter values.

Similarly, the variance of the neutrosophic MOE Burr-XII distribution is obtained as

$$\sigma^2=Var\left[\left(1+I_N\right)X_L\right]=\left(1+I_N\right)^2Var\left(X_L\right) \tag{8}$$

The variance also does not have a closed form. Therefore, we can determine its numerical values by substituting the parameter values.

Another important statistical property of the NMOE Burr-XII distribution is the quantile function, which is crucial for the Monte Carlo simulation approach. This function is also useful for generating random numbers from the probability distribution model. The quantile function of the NMOE Burr-XII distribution is derived as follows.

$$Q_N(p)=F_N^{-1}\left(X_p\right)$$

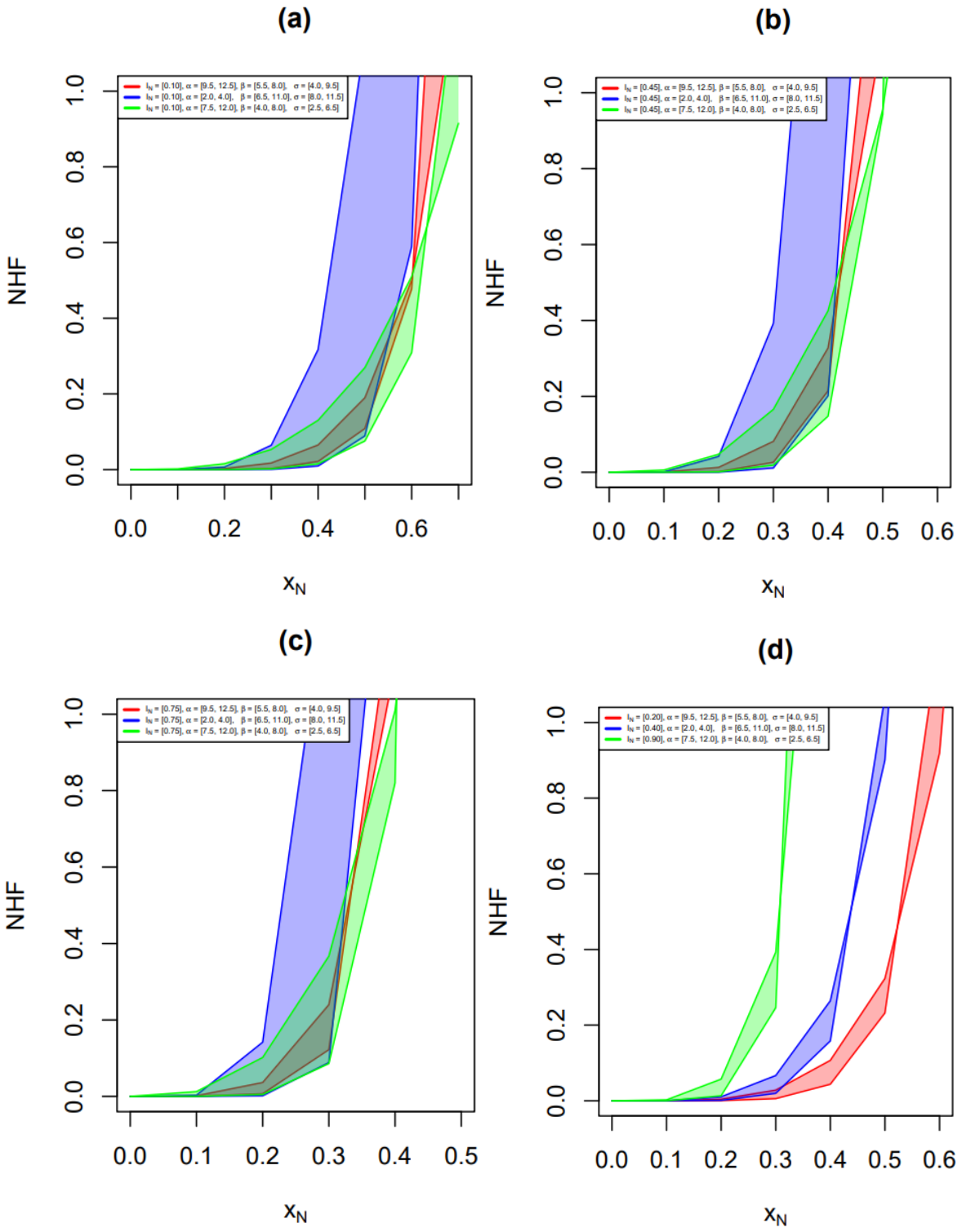


Figure 2: HRF plots for the NMOE Burr-XII distribution for different values of I_N, and parameters

$$Q_N(p) = \frac{\left[\left(\frac{1-p}{1-p(1-\alpha)}\right)^{-\frac{1}{\gamma}} - 1\right]^{\frac{1}{\beta}}}{(1+I_N)} \tag{9}$$

The median, first quartile, third quartile and Inter quartile range (IQR) for proposed distribution are calculated as $Median = Q_N(0.5)$, First quartile $= Q_N(0.25)$, Third quartile$= Q_N(0.75)$ and $IQR = Q_N(0.75) - Q_N(0.25)$.

Neutrosophic Measure of Skewness and Kurtosis based on the Quantile function for NMOE Burr-XII distribution are given as follows,

$$SK_N = \frac{Q_N(6/8) - 2Q_N(4/8) + Q_N(2/8)}{Q_N(6/8) - Q_N(2/8)} \tag{10}$$

and

$$K_N = \frac{Q_N(7/8) - Q_N(5/8) + Q_N(3/8) - Q_N(1/8)}{Q_N(6/8) - Q_N(2/8)} \tag{11}$$

5. Parameter estimation

In this section, we discuss the estimation of unknown parameters for the NMOE Burr-XII distribution using the method of maximum likelihood estimator (MLE).

Maximum likelihood estimation method

Given the observed data, this method is used to find the parametric values of the proposed distribution. Suppose that $(1+I_N)X_N1, (1+I_N)X_N2, \ldots, (1+I_N)X_Nn$, be a neutrosophic random samples of NMOE Burr-XII distribution then log-likelihood function is derived as:

The loglikelihood function is:

$$\begin{aligned} l(\alpha,\beta,\gamma) &= log(\alpha) + log(\beta) + log(\gamma) + log(1+I_N) + (\beta-1)\sum_{i=1}^{n} log(1+I_N)x_i \\ -(\gamma+1)log\sum_{i=1}^{n}\left[1+\{(1+I_N)x\}^{\beta}\right] &- 2log\sum_{i=1}^{n}\left[1-(1-\alpha)\left[1+\{(1+I_N)\,x_L\}^{\beta}\right]^{-\gamma}\right] \end{aligned} \tag{12}$$

To find the values of parameters, obtain the derivative of the above expression with respect to α, β and γ.

$$\frac{\partial l}{\partial \alpha} = \frac{1}{\alpha} - \frac{\left[1+\{(1+I_N)x\}^{\beta}\right]^{-\gamma}}{\left[1-(1-\alpha)\left[1+\{(1+I_N)\,x_L\}^{\beta}\right]^{-\gamma}\right]} \tag{13}$$

$$\begin{aligned} \frac{\partial l}{\partial \beta} = \frac{1}{\beta} + log\{(1+I_N)x\} - \frac{(\gamma+1)\{(1+I_N)x\}^{\beta} log\{(1+I_N)x\}}{[1+\{(1+I_N)x\}^{\beta}]} \\ - \frac{2\gamma log\{(1+I_N)x+1\}\left[-\left[-(\alpha-1)\{(1+I_N)x+1\}^{\beta}\right]^{-\gamma}\right]}{1-(1-\alpha)\left[1+\{(1+I_N)x\}^{\beta}\right]^{-\gamma}} \end{aligned} \tag{14}$$

$$\frac{\partial l}{\partial \gamma} = \frac{1}{\gamma} - log\{1+\{(1+I_N)x\}^{\beta}\} - \frac{2\left[1+\{(1+I_N)x\}^{\beta}\right]^{-\gamma} log\left[1+\{(1+I_N)x\}^{\beta}\right]}{1-(1-\alpha)\left[1+\{(1+I_N)x\}^{\beta}\right]^{-\gamma}} \tag{15}$$

5.1. Simulation study

In this section, we conduct a Monte Carlo simulation study to evaluate the performance of the estimated parameters for the NMOE Burr-XII distribution. We assess the performance of the neutrosophic Maximum Likelihood estimator using the neutrosophic average biased (AB_N) and the neutrosophic root mean square error (RMSE).

$$AB_N = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{\theta}_{N_i} - \theta_N\right)$$

and

$$RMSE_N = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{\theta}_{N_i} - \theta_N\right)^2$$

In R software, a Monte Carlo simulation with varying sample sizes and fixed values of the Neutrosophic parameters $\alpha = [0.01, 0.2]$, $\beta = [2.0, 2.7]$ and $\gamma = [1.2, 1.8]$ is conducted. The NMOE Burr-XII is used to build an imprecise dataset with $\alpha = [0.01, 0.2]$, $\beta = [2.0, 2.7]$ and $\gamma = [1.2, 1.8]$, and simulation is replicated $N = 10000$ times with sample sizes of $n = 50, 100, 300, 500$, respectively. The performance of the neutrosophic Maximum Likelihood estimators is then computed and shown in Tables 1, 2, 3 and 4. In the tables from 1-4, it is observed as the sample size increases the MSE, MRE and bias is decreasing for all parameters. Moreover, comparing estimated results when the I_N has been calculated from respective parameters *i.e.* $I_{N_{parameters}}$ with when $I_N = 0$. Then the MSE, Bias and MRE for I_{N_α} given in table 1, I_{N_β} given in table 2, and I_{N_γ} given in table 3, are less as compared to when $I_N = 0$, given in table 4.

Table 1: Parameter's bias, average bias, mean square error (MSE), and mean relative error (MRE) for $\alpha_N = [0.01, 0.09], \beta_N = [0.5, 1.5], \gamma_N = [0.05, 1.5]$ and $I_{N_\alpha} = 0.89$ calculated from α_N

Sizes	MLE Estimates	$\alpha_N = [0.01, 0.09]$	$\beta_N = [0.5, 1.5]$	$\gamma_N = [0.05, 1.5]$
50	Bias	[0.1641, 1.2994]	[0.4957, 1.4747]	[0.2716, 2.5005]
	Average Bias	[0.1566, 1.2434]	[0.0487,0.1905]	[0.2339, 1.4319]
	MSE	[20.8938, 196.1348]	[0.0038, 0.0635]	[0.2364, 5.1005]
	MRE	[15.6634, 17.8155]	[0.0974, 0.1270]	[4.6771, 0.9546]
100	Bias	[0.0358, 0.5668]	[0.4973, 1.4839]	[0.1589, 2.0133]
	Average Bias	[0.0285, 0.5085]	[0.0339, 0.1324]	[0.1218, 0.9201]
	MSE	[0.0036, 89.8824]	[0.0018, 0.0301]	[0.0517, 1.9413]
	MRE	[2.8496, 5.6503]	[0.0678, 0.0882]	[2.4361, 0.6134]
300	Bias	[0.0170, 0.1080]	[0.4996,1.4960]	[0.0824, 1.6468]
	Average Bias	[0.0098, 0.0427]	[0.0189, 0.0731]	[0.0459, 0.4588]
	MSE	[0.0003, 0.0040]	[0.0006, 0.0085]	[0.0055, 0.3739]
	MRE	[0.9829, 0.4742]	[0.0379, 0.0487]	[0.9182, 0.3058]
500	Bias	[0.0137, 0.1006]	[0.4998, 1.4971]	[0.0672, 1.5890]
	Average Bias	[0.0064, 0.0313]	[0.0149, 0.0559]	[0.0302, 0.3473]
	MSE	[0.0001, 0.0019]	[0.0004, 0.0050]	[0.0021, 0.2086]
	MRE	[0.6396, 0.3487]	[0.0297, 0.0373]	[0.6055, 0.2315]

Table 2: Parameter's bias, average bias, mean square error (MSE), and mean relative error (MRE) for $\alpha_N = [0.01, 0.09], \beta_N = [0.5, 1.5], \gamma_N = [0.05, 1.5]$ and $I_{N_\beta} = 0.67$ calculated from β_N

Sizes	MLE Estimates	$\alpha_N = [0.01, 0.09]$	$\beta_N = [0.5, 1.5]$	$\gamma_N = [0.05, 1.5]$
50	Bias	[0.1553, 1.4602]	[0.4958, 1.4721]	[0.2834, 2.5290]
	Average Bias	[0.1477, 1.4042]	[0.0501, 0.1899]	[0.2453, 1.4662]
	MSE	[30.9636, 232.3428]	[0.0041, 0.0641]	[0.2568, 5.3880]
	MRE	[14.7735, 15.6019]	[0.1001, 0.1266]	[4.9058, 0.9774]
100	Bias	[0.0350, 0.3235]	[0.4979, 1.4878]	[0.1562, 1.9594]
	Average Bias	[0.0278, 0.2666]	[0.0342, 0.1303]	[0.1196, 0.8842]
	MSE	[0.0033, 30.7377]	[0.0019, 0.0280]	[0.0483, 1.7041]
	MRE	[2.7784, 2.9620]	[0.0683, 0.0868]	[2.3922, 0.5895]
300	Bias	[0.0167, 0.1083]	[0.4999, 1.4970]	[0.0809, 1.6536]
	Average Bias	[0.0095, 0.0429]	[0.0192, 0.0725]	[0.0448, 0.4618]
	MSE	[0.0003, 0.0041]	[0.0006, 0.0085]	[0.0054, 0.3837]
	MRE	[0.9519, 0.4764]	[0.0383, 0.0484]	[0.8950, 0.3079]
500	Bias	[0.0139, 0.1009]	[0.4997, 1.4974]	[0.0682, 1.5950]
	Average Bias	[0.0066, 0.0316]	[0.0147, 0.0562]	[0.0314, 0.3505]
	MSE	[0.0001, 0.0019]	[0.0003, 0.0051]	[0.0023, 0.2099]
	MRE	[0.6631, 0.3513]	[0.0293, 0.0375]	[0.6272, 0.2336]

Table 3: Parameter's bias, average bias, mean square error (MSE), and mean relative error (MRE) for $\alpha_N = [0.01, 0.09], \beta_N = [0.5, 1.5], \gamma_N = [0.05, 1.5]$ and $I_{N_\gamma} = 0.97$ calculated from γ_N

Sizes	MLE Estimates	$\alpha_N = [0.01, 0.09]$	$\beta_N = [0.5, 1.5]$	$\gamma_N = [0.05, 1.5]$
50	Bias	[0.1257, 1.4512]	[0.4954, 1.4742]	[0.2921, 2.5106]
	Average Bias	[0.1183, 1.3951]	[0.0507, 0.1901]	[0.2544, 1.4363]
	MSE	[6.8381, 231.8247]	[0.0041, 0.0638]	[0.2779, 5.1529]
	MRE	[11.8248, 15.5015]	[0.1014, 0.1267]	[5.0875, 0.9576]
100	Bias	[0.0349, 0.4358]	[0.4978, 1.4863]	[0.1563, 1.9758]
	Average Bias	[0.0276, 0.3778]	[0.0343, 0.1311]	[0.1194, 0.8831]
	MSE	[0.0031, 60.1300]	[0.0019, 0.0288]	[0.0485, 1.7537]
	MRE	[2.7619, 4.1978]	[0.0687, 0.0874]	[2.3882, 0.5888]
300	Bias	[0.0170, 0.1085]	[0.4994, 1.4952]	[0.0821, 1.6575]
	Average Bias	[0.0098, 0.0430]	[0.0187, 0.0728]	[0.0459, 0.4647]
	MSE	[0.0003, 0.0040]	[0.0006, 0.0085]	[0.0055, 0.3847]
	MRE	[0.9831, 0.4776]	[0.0374, 0.0485]	[0.9187, 0.3098]
500	Bias	[0.0138, 0.0996]	[0.4999, 1.4989]	[0.0678, 1.5792]
	Average Bias	[0.0065, 0.0313]	[0.0148, 0.0563]	[0.0307, 0.3494]
	MSE	[0.0001, 0.0019]	[0.0003, 0.0050]	[0.0022, 0.2066]
	MRE	[0.6473, 0.3478]	[0.0295, 0.0375]	[0.6144, 0.2329]

Table 4: Parameter's bias, average bias, mean square error (MSE), and mean relative rrror (MRE) for $\alpha_N = [0.01], \beta_N = [0.5], \gamma_N = [0.05]$ and $I_N = 0$

Sizes	MLE Estimates	$\alpha_N = [0.01]$	$\beta_N = [0.5]$	$\gamma_N = [0.05]$
50	Bias	0.1923	0.4954	0.2899
	Average Bias	0.1849	0.0503	0.2523
	MSE	27.206	0.0041	0.2861
	MRE	18.4847	0.1007	5.0461
100	Bias	0.0348	0.4978	0.1552
	Average Bias	0.0275	0.0344	0.1186
	MSE	0.0031	0.0019	0.0469
	MRE	2.7534	0.0687	2.3711
300	Bias	0.0171	0.4994	0.0828
	Average Bias	0.0099	0.0189	0.0465
	MSE	0.0003	0.0006	0.0056
	MRE	0.9923	0.0377	0.9292
500	Bias	0.0140	0.4998	0.0688
	Average Bias	0.0067	0.0147	0.0319
	MSE	0.0001	0.0003	0.0023
	MRE	0.6712	0.0295	0.6376

6. Applications

In this section, we apply the NMOE Burr-XII model to two real-world datasets characterized by uncertain or complex values. We aim to gauge the suitability of the NMOE Burr-XII model for such data. Various model selection methods are employed to assess the performance of the proposed distribution and compare it with other competing distributions to determine the best model. Two datasets are used in this study, that are Remission time dataset and Covid-19 dataset. The understudy datasets are presented in interval form, meaning they exhibit uncertainty in the upper bounds of their data values, rather than providing single, fixed values. This inherent uncertainty may result in insufficient information. To address this issue, the upper bounds in each dataset are calculated using the indeterminacy component I_N, thereby converting them into neutrosophic statistics. The values of I_N can be changed to $2\%, 5\%$, or 10% based on the desired degree of assurance or uncertainty. By immediately identifying and incorporating uncertainties into each dataset, this technique enables a more nuanced analysis and improves the comprehensiveness and utility of the data in medical research and decision-making across different investigations. In applications $I_N = 0.05$ is used to find the upper values of the datasets. A balance between being cautious and accommodating of data uncertainties is achieved by setting $I_N = 0.05$. It allows for considerable flexibility while maintaining a respectable degree of analytical precision.

Remission time dataset

The first data consists of a collection of 128 cancer patients' remission durations measured in months. After getting therapy, each value indicates how long a patient stayed in remission. When it comes to cancer therapy, remission is the time when the disease's symptoms and indicators are either minimal or nonexistent. This data has been taken from bladder cancer study reported by Lee and Wang (2003). The understudy data (remission

time) is available in neutrosophic form. We took the lower limit of the data from the source and estimated the upper limit by applying an indeterminacy factor of $I_N = 0.05$. This same indeterminacy factor was then used to calculate the descriptive statistics shown in 7, estimate the values of neutrosophic parameters also in Table 8, and model the proposed density in Table 10 for the remission time data. [Note: this factor can be taken any other value.].

Table 5: Remission time Dataset

0.08	2.09	3.48	4.87	6.94	8.66	13.11	23.63	0.2
2.23	3.52	4.98	6.97	9.02	13.29	0.4	2.26	3.57
5.06	7.09	9.22	13.8	25.74	0.5	2.46	3.64	5.09
[7.26, 8.2]	9.47	14.24	25.82	0.51	2.54	3.7	5.17	7.28
9.74	14.76	**[5.3,7.1]**	0.81	2.62	3.82	5.32	7.32	10.06
[12, 14.77]	32.15	2.64	3.88	5.32	7.39	10.34	14.83	34.26
0.9	2.69	4.18	5.34	7.59	10.66	15.96	36.66	1.05
2.69	4.23	5.41	7.62	10.75	16.62	43.01	1.19	2.75
4.26	5.41	7.63	**[15, 17.2]**	46.12	1.26	2.83	4.33	5.49
7.66	11.25	17.14	**[75.02, 81]**	1.35	2.87	5.62	7.87	11.64
17.36	1.4	3.02	4.34	5.71	7.93	11.79	18.1	1.46
4.4	5.85	8.26	11.98	19.13	1.76	3.25	4.5	6.25
8.37	12.02	**[1.5, 3.2]**	3.31	4.51	6.54	**[7.5, 8.2]**	12.03	20.28
2.02	3.36	6.76	12.07	21.73	2.07	3.36	6.93	8.65
12.63	22.69	—	—	—	—	—	—	—

Covid-19 dataset

The data research by Almongy *et al.* (2021) describes the duration of relief in hours for 30 patients who received analgesic medication, likely as a part of a treatment for managing COVID-19-related symptoms. The data displays a range of response times, which suggests that patients in the research group had varying responses to the medicine. The relief time data is available from the source in neutrosophic form. We considered the lower limit of the data and calculated the upper limit by using an indeterminacy value of $I_N = 0.05$. This same indeterminacy value is later used to determine the descriptive statistics in Table 7, estimate the neutrosophic parameters in Table 9, and model the proposed density in Table 11 for the relief time data.

Table 6: Covid-19 Dataset

(14.918, 15.6639)	(10.056, 11.1888)	(12.274, 12.88770)	(10.289, 10.80345)
(10.832, 11.3736)	(7.099, 7.4539)	(5.928, 6.22440)	(13.211, 13.87155)
(7.968, 8.36640)	(7.584, 7.96320)	(5.555, 5.83275)	(6.027, 6.32835)
(4.097, 4.30185)	(3.611, 3.79155)	(4.960, 5.20800)	(7.498, 7.87290)
(6.940, 7.28700)	(5.307, 5.57235)	(5.048, 5.30040)	(2.857, 2.99985)
(2.254, 2.36670)	(5.431, 5.70255)	(4.462, 4.68510)	(3.883, 4.07715)
(3.461, 3.63405)	(3.647, 3.82935)	(1.974, 2.07270)	(1.273, 1.33665)
(1.416, 1.48680)	(4.235, 4.44675)	—	—

Table 7: Descriptive statistics for both data sets from proposed density

Descriptives	Remission time data	COVID-19
Mean	[0.9010, 0.8945]	[0.8162, 0.8205]
Variance	[0.0009, 0.0007]	[1.1e-05, 2.587e-06]
Median	[0.9004, 0.8941]	[0.8169, 0.8212]
First Quartile	[0.8821, 0.8785]	[0.8141, 0.8189]
Third Quartile	[0.9191, 0.9100]	[0.8193, 0.8230]
Skewness	[0.4324, 0.7092]	[-13.4441, -121.5781]
Kurtosis	[0.7801, 0.5726]	[177.7792, 941.5738]

Table 8: ML estimates and standard errors remission time dataset

NMOE Burr-XII	α	[63.6772, 58.4272]	[59.2463, 47.0918]
	β	[59.2463, 47.0918]	[0.955, 0.9926]
	γ	[0.955, 0.9926]	[0.3767, 0.3711]
N-Burr-III	θ	[1.033, 1.0232]	[0.0601, 0.0591]
	λ	[0.0601, 0.0591]	[4.3325, 4.5161]
Burr-XII	θ	[2.3454, 2.3303]	[0.355, 0.3518]
	λ	[0.355, 0.3518]	[0.2351, 0.235]

Table 9: ML estimates and standard errors for the COVID-19 (relief time) data

NMOE Burr-XII	α	[97.7882, 118.707]	[81.9116, 102.2903]
	β	[81.9116, 102.2903]	[25.5239, 23.8676]
	γ	[25.5239, 23.8676]	[30.2665, 86.2472]
N-Burr-III	θ	[1.6581, 1.6477]	[0.1983, 0.1974]
	λ	[0.1983, 0.1974]	[10.657, 11.3512]
Burr-XII	θ	[21.4342, 15.715]	[32.6983, 23.3943]
	λ	[32.6983, 23.3943]	[0.0283, 0.0375]

From Table 7, the following results information is obtained.

- The average remission time (in months) for bladder cancer patients is between the interval by mean is [0.9010, 0.8945] and by median is [0.9004, 0.8941] with spread [0.0009, 0.0007]. From the skewness and kurtosis, it is seen that the remission time data is slightly positively skewed, and platykurtic. From the first quartile it is seen that 25% of the patients have less remission time by this interval [0.8821, 0.8785] and from third quartile it is seen that 75% of the patients have less remission time by this interval [0.9191, 0.9100].

- The average relief times for bladder cancer patients is between the interval by mean is [0.8162, 0.8205] and by median is [0.8169, 0.8212] with spread [1.1e-05, 2.587e-06]. From the skewness and kurtosis, it is seen that the relief times data is extremely negatively skewed, and leptokurtic. From the first quartile it is seen that 25% of the patients have relief times by this interval [0.8141, 0.8189] and from third quartile it is seen that 75% of the patients have less relief times by this interval [0.8193, 0.8230].

Model Selection Criteria with estimates for the remission time dataset is shown in table 8 and 10. Model Selection Criteria with estimates for the COVID-19 (relief time) data is shown in tables 9 and 11. The estimated values of the parameters in tables 8 and 9 are in interval form because the parameters are neutrosophic due the uncertainty in the data sets.

7. Comparative study

This section presents a comparative study of the proposed model using two real-life datasets. The comparison is conducted with the neutrosophic Burr-II and classical Burr-XII models.

In Tables 10 and 11, the proposed neutrosophic density NMOE-Burr-XII is modeled and compared for both datasets, with the indeterminacy component I_N set at 0.05. This value represents the uncertainty in the datasets. When I_N equals 0, the density is in its classical form, such as NMOE Burr-XII and N-Burr-III, with the test statistic criterion being the lower bound only. However, when I_N is 0.05 or any other value, the densities become neutrosophic, and the test statistic criterion is presented in interval form due to the neutrosophic nature of the data.

In table 10, the modeling of the proposed density on remission time for bladder cancer patients' data shows that NMOE Burr-XII distribution shows more flexibility over the neutrosophic Burr-III (N-Burr-III) and classical Burr-XII distributions due to the lowest values of AIC, BIC, CAID, HQIC KS test and larger p-value for the KS test. It is also observed that Burr-XII shows p-value as 0.000 which particularly shows its inadequacy for the neutrosophic data, while NMOE Burr-XII and N-Burr-III both fit the data, but NMOE Burr-XII provides very strong p-value which shows its superiority. Table 11 shows the modeling of the proposed model on the relief time for COVID-19 data, the results shows that the NMOE Burr-XII distribution shows more flexibility as compared to the N-Burr-III and classical Burr-XII distributions due to the lowest values of AIC, BIC, CAID, HQIC KS test and larger p-value for the KS test.

Furthermore, the proposed density demonstrates superior flexibility and provides evidence across all three datasets compared to the classical Burr-XII and even the N-Burr-III distribution. Importantly, it is observed that the classical Burr-XII distribution does not fit well on both neutrosophic datasets (remission time for bladder cancer and relief time for COVID-19 datasets), yielding a p-value of 0.000.

In conclusion, the neutrosophic Marshall-Olkin Extended Burr-XII distribution emerges as a valuable tool particularly in scenarios where data is indeterminate, contrasting with the classical Burr-XII distribution. Classical distributions are unsuitable for modeling indeterminate and ambiguous datasets. The two data examples discussed above fall under the neutrosophic setup because they deal with lifetime data that inherently includes elements of uncertainty and incomplete information, which are better handled using neutrosophic statistics. While classical methods rely on precise probabilities, neutrosophic methods provide a more comprehensive framework by incorporating indeterminacy and partial truth values, thus offering more robust and realistic estimates in the presence of real-world complexities. This allows for better decision-making and reliability assessments in environments where data is not perfectly exact or complete.

Table 10: Model selection criteria and parameter estimates for remission time data

Models	LL	AIC	BIC	CAIC	HQIC	K-S	*p*-value
NMOE Burr-XII	[-405.695, -407.067]	[817.390, 820.134]	[825.946, 828.690]	[817.583, 820.327]	[820.866, 823.610]	[0.033, 0.039]	[0.9989, 0.9904]
N-Burr-III	[-423.562, -431.383]	[851.123, 866.767]	[856.828, 872.471]	[851.219, 866.863]	[853.441, 869.084]	[0.107, 0.105]	[0.106, 0.121]
Burr-XII	449.983	903.9661	909.670	904.062	906.284	0.254	0.000

Table 11: Model selection criteria and parameter estimates for the COVID-19 (relief time) data

Models	LL	AIC	BIC	CAIC	HQIC	K-S	*p*-value
NMOE Burr-XII	[-77.063, -78.736]	[160.127, 163.472]	[164.330, 167.675]	[161.050, 164.395]	[161.472, 164.817]	[0.066, 0.064]	[0.9983, 0.9989]
N-Burr-III	[-80.267, -81.901]	[164.534, 167.803]	[167.337, 170.605]	[164.979, 168.247]	[165.431, 168.699]	[0.139, 0.139]	[0.563, 0.5616]
Burr-XII	-94.298	192.595	195.397	193.039	193.492	0.362	0.000

8. Conclusion

In this study, we introduce a novel model called the neutrosophic Marshall-Olkin Extended Burr-XII distribution. We demonstrate that this model is advantageous for analyzing survival and reliability datasets with indeterminacies compared to classical distributions. Various neutrosophic properties are explored, including the neutrosophic survival function, hazard function, mean, variance, mode, skewness, and kurtosis. The distribution exhibits left-skewed, right-skewed, and symmetric shapes. The hazard rate function displays a monotonically increasing trend. Parametric values are determined using the maximum likelihood method. A simulation study assesses the performance of estimators across small, medium, and large sample sizes, revealing a decrease in mean square error with increasing sample size. Additionally, the proposed NMOE Burr-XII distribution is applied to two real-life datasets with uncertain values, demonstrating its superior flexibility compared to classical distributions.

Acknowledgements

The authors are deeply thankful to the editor and the reviewers for their valuable suggestions to improve the presentation and quality of the paper.

Conflict of interest

The authors do not have any financial or non-financial conflict of interest to declare for the research work included in this article.

References

Al-Saiari, A. Y., Baharith, L., and Mousa, S. A. (2014). Marshall-Olkin extended Burr Type XII distribution. *International Journal of Statistics and Probability*, **3**, 78–84.

Albassam, M., ul Haq, M. A., and Aslam, M. (2023). Weibull distribution under indeterminacy with applications. *AIMS Mathematics*, **9**, 10745–10757.

Almongy, H. M., Almetwally, E. M., Aljohani, H. M., Alghamdi, A. S., and e, E. H. (2021). A new extended Rayleigh distribution with applications of covid-19 data. *Results in Physics*, **23**, 104012.

Aslam, M. and Albassam, M. (2024). Neutrosophic geometric distribution: Data generation under uncertainty and practical applications. *AIMS Mathematics*, **9**, 16436–16452.

Chen, J., Du, S., and Yong, R. (2017). Expressions of rock joint roughness coefficient using neutrosophic interval statistical numbers. **9**, 123.

Duan, W.-Q., Khan, Z., Gulistan, M., and Khurshid, A. (2021). Neutrosophic exponential distribution: Modeling and applications for complex data analysis. *Complexity*, **2021**.

Eassa, N. I., Zaher, H. M., and El-Magd, N. A. T. A. (2023). Neutrosophic generalized Pareto distribution. *Infinite Study*, **11**, 827–833.

Fawzi, K., Alhasan, H., Smarandache, F., and Hamza, K. (2019). Neutrosophic Weibull distribution and neutrosophic family Weibull distribution. *Neutrosophic Sets and Systems*, **28**, 191–199.

Granados, C., Das, A. K., and Das, B. (2022). Some continuous neutrosophic distributions with neutrosophic parameters based on neutrosophic random variables. *Advances in the Theory of Nonlinear Analysis and its Application*, **6**, 380–389.

Khan, R., Naeem, M., Aslam, M., and Raza, M. A. (2021a). Neutrosophic beta distribution with properties and applications. *Neutrosophic Sets and Systems*, **41**, 209–214.

Khan, Z., Al-Bossly, A., Almazah, M. M. A., and Alduais, F. S. (2021b). On statistical development of neutrosophic Gamma distribution with applications to complex data analysis. *Complexity*, **2021**, 1–8.

Khan, Z., Gulistan, M., Kausar, N., and Park, C. (2021c). Neutrosophic Rayleigh model with some basic characteristics and engineering applications. *IEEE Access*, **9**, 71277–71283.

Lee, E. T. and Wang, J. W. (2003). *Statistical Methods for Survival Data Analysis*, volume 476. John Wiley & Sons.

Marshall, A. W. and Olkin, I. (1997). A new method for adding a parameter to a family of distributions with application to the exponential and Weibull families. *Biometrika*, **84**, 641–652.

Nayana, B. M., Anakha, K. K., Chacko, V. M., Aslam, M., and Albassam, M. (2022). A new neutrosophic model using DUS-Weibull transformation with application. *Complex & Intelligent Systems*, **8**, 4079–4088.

Rao, G. S. (2023). Neutrosophic log-Logistic distribution model in complex alloy metal melting point applications. *International Journal of Computational Intelligence Systems*, **16**, 48.

Rao, G. S., Norouzirad, M., and Mazarei, D. (2023). Neutrosophic generalized exponential distribution with application. *Neutrosophic Sets and Systems*, **55**, 471–485.

Rodriguez, R. N. (1977). A guide to the Burr type XII distributions. *Biometrika*, **64**, 129–134.

Shao, Q., Wong, H., Xia, J., and Ip, W.-C. (2004). Models for extremes using the extended three-parameter Burr XII system with application to flood frequency analysis / modèles d'extrêmes utilisant le système Burr XII étendu à trois paramètres et application à l'analyse fréquentielle des crues. *Hydrological Sciences Journal*, **49**, 685–702.

Sherwani, R. A. K., Arshad, T., Albassam, M., Aslam, M., and Abbas, S. (2021). Neutrosophic entropy measures for the Weibull distribution: theory and applications. *Complex & Intelligent Systems*, **7**, 3067–3076.

Smarandache, F. (2013). *Introduction to Neutrosophic Measure, Neutrosophic Integral, and Neutrosophic Probability*. Infinite Study.

Smarandache, F. (2014). *Introduction to Neutrosophic Statistics*. Infinite Study.

Smarandache, F. (2022). Neutrosophic statistics is an extension of interval statistics, while plithogenic statistics is the most general form of statistics (second version). *International Journal of Neutrosophic Science*, **19**, 148–165.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 23, No. 1, 2025 (New Series), pp 365–383
https://www.ssca.org.in/journal

Prediction of COVID-19 Disease Progression in India under the Effect of National Lockdown

Sourish Das
Chennai Mathematical Institute, India

This is a special invited paper on request from the Chair Editor.

Received: 30 August 2024; Revised: 15 November 2024; Accepted: 20 November 2024

Abstract

In this paper, we implemented the epidemiological Susceptible-Infected-Recovered (SIR) model to estimate the basic reproduction number, $\mathcal{R}_0$, at both national and state levels in India. To the best of our knowledge, it was the first study that attempted to estimate $\mathcal{R}_0$ for India and its different states, see Das (2020). As this was the first attempt, the study used data until March 24, 2020. In the very early days of the pandemic, the data were sparse and it was difficult to conduct analysis and make forecasts. Under such circumstances, we developed a statistical machine learning model to predict future case numbers.

Our analysis showed that the situation in Punjab ($\mathcal{R}_0 \approx 16$) was critical and required immediate, aggressive intervention. We observed that the $\mathcal{R}_0$ values for Madhya Pradesh (3.37), Maharashtra (3.25), and Tamil Nadu (3.09) all exceeded 3. The $\mathcal{R}_0$ values for Andhra Pradesh (2.96), Delhi (2.82), and West Bengal (2.77) were higher than India's overall $\mathcal{R}_0$ of 2.75, as of March 4, 2020. India's $\mathcal{R}_0$ of 2.75 at that stage was comparable to that of Hubei, China during the early phase of the outbreak in December, 2019.

Our analysis indicated that India's early disease progression was similar to China. With the lockdown in place, India could have expected a number of cases comparable to, if not more than, those in China. If the lockdown had been effective, we anticipated fewer than 66,224 cases by 1 May 2020. The out-of-sample R^2 was 0.9323, and the observed number of cases on 1 May 2020 was 37,263, which was less than the predicted value, indicating the lockdown's effectiveness. All data and `R` code for this paper are available at `https://github.com/sourish-cmi/Covid19`.

Key words: Basic reproduction number; Epidemiological model; Statistical machine learning model.

AMS Subject Classifications: 62K05, 05B05

Corresponding Author: Sourish Das
Email: sourish@cmi.ac.in

1. Introduction

The World Health Organization (WHO) declared the outbreak of the novel coronavirus, COVID-19, a pandemic. It was estimated that it would take twelve to eighteen months to develop a vaccine for COVID-19 (see Ferguson *et al.* (2020)). The absence of a vaccine worsened the situation for India's already overstretched healthcare system. For example, the number of hospital beds per 1,000 population was less than one World-Bank (2021)—just one indicator of the miserable state of India's healthcare system. In the absence of a vaccine, "social distancing" was considered the optimal strategy to control the spread of the novel coronavirus Ferguson *et al.* (2020).

Aside from social distancing, widespread rapid testing and cluster testing were essential to identify infected individuals and isolate them. However, India did not have sufficient testing capacity, as widely reported in the media Biswas (2020). Although Indian scientists recently developed an affordable COVID-19 testing kit Pandey (2020), India needed a complete overhaul of its healthcare system on a war footing. In this context, India's Prime Minister Narendra Modi announced an unprecedented three-week nationwide lockdown on March 24, 2020. The purpose of the lockdown was to slow the spread of the novel coronavirus, allowing the government to pursue a multi-pronged strategy to add more beds to its hospital network, scale up production of COVID-19 testing kits, and provide personal protective equipment (PPE) for healthcare workers.

In such a grim scenario, the key question for Indian health officials was how many new confirmed cases would emerge and by what time, with the hope that the national lockdown would slow the virus's spread and buy them time to overhaul the healthcare system. However, there was uncertainty about whether the lockdown would provide the necessary slowdown of virus transmission. Even if the lockdown helped India control the virus's spread, it was not economically sustainable to extend it further, given the large number of workers employed in the informal sector as daily wage laborers. Therefore, in this policy paper, we attempted to estimate the effect of the lockdown and proposed a framework to track its effectiveness.

In this paper, we developed an epidemiological SIR model and a statistical machine learning model to predict disease progression in India. We implemented the SIR model to estimate the basic reproduction number, $\mathcal{R}_0$, at both national and state levels, to identify which states required more attention. Then, we applied the machine learning model to predict the number of cases ahead of time, so that the Indian administration could be better prepared in advance.

In Section (2), we introduced the database from which the data was downloaded and the model was built. In Section (3), we presented the methodology used to analyse and predict the data. In Section (4), we provided our analysis and prediction of the Covid-19 disease progression in India. Section (5) discusses the follow-up literature that came after this initial work, and Section (6) concluded the paper.

2. Data

In this paper, we utilised the following major databases to gather relevant data for our analysis and model development:

1. The data repository for the 2019 Novel Coronavirus, maintained by Johns Hopkins University. This globally recognized repository aggregates COVID-19 data from numerous official sources worldwide. The database is available at: `https://github.com/CSSEGISandData/COVID-19`.

2. Covid19India, a crowdsourced open-source database for India, which provides real-time updates on COVID-19 cases across Indian states and districts. This database offers a granular level of detail critical for region-specific analysis. It is available at: `https://www.covid19india.org/`.

3. Kaggle-Covid-19 in India, a dataset available on Kaggle that compiles COVID-19 data for India, including daily updates on confirmed cases, recoveries, and deaths. It also features various features like population and testing data that help enhance the predictive power of models. This dataset is available at: `https://www.kaggle.com/sudalairajkumar/covid19-in-india`.

These databases provided comprehensive and up-to-date information necessary for tracking the disease's progression and for building predictive models. By leveraging this data, we aimed to generate accurate forecasts and offer actionable insights for public health officials and policymakers.

3. Methodology

Legendary statistician Prof George Box, once said

"All models are wrong, but some are useful", see Box (1976).

Keeping this in mind, in this paper, we took a model-agnostic, two-pronged approach. The first was to understand the severity of the ground situation, and the second was to provide predictions to help health officials make informed plans. Epidemic models for infectious diseases provided insights into the dynamic behavior of disease spread. With these new insights, health officials could develop more effective intervention strategies. Moreover, such epidemic models were also used to forecast the course of the epidemic.

In addition to epidemic models, we considered statistical machine learning (SML) models, which were highly effective for prediction. Often, the interpretability of SML models was questioned. However, as we took a model-agnostic approach, we were able to use the epidemic models to understand the ground reality while adopting SML models to achieve better prediction accuracy.

3.1. SIR epidemiological model

The popular epidemic models for an infectious disease is the Susceptible, Infected, Recovered (SIR) model. The model considers a closed population. To start with, a few infected people are added to the population. It assumes that the mixing pattern is homogeneous. During the period of the sickness, the contagious people each infect on average $\mathcal{R}_0$ other people, who each then go on to infect $\mathcal{R}_0$ others, who are susceptible. The $\mathcal{R}_0$ is

popularly known as the Basic Reproduction Number. The $\mathcal{R}_0$ is the fundamental quantity of the disease progression, and higher $\mathcal{R}_0$ means, more people will tend to be infected in the course of the epidemic. The major advantage of the SIR model is it gives a number $\mathcal{R}_0$, which can be used to benchmark and compare the ground situation of different states and resource allocations can be made to those states which are hard hit. The SIR model can be described as,

$$\begin{aligned} \frac{\partial S}{\partial t} &= -\beta \frac{SI}{N} \\ \frac{\partial I}{\partial t} &= +\beta \frac{SI}{N} - \gamma I \\ \frac{\partial R}{\partial t} &= +\gamma I \end{aligned} \tag{1}$$

where S, I, and R are the number of people in the population that are susceptible, infected and recovered. The β is the transmission rate. Each susceptible person contacts β people per day; a fraction $\frac{I}{N}$ of which are infectious. Therefore $\beta \frac{SI}{N}$ move out of the susceptible group and goes into the infected group. The transmission rate is the average rate of contacts a susceptible person makes that is sufficient to transmit the infection. The parameter γ is the recovery rate, and γI is the flow out of the infected crowd and goes into the recovered group. The average duration a person spends in the infected group is $\frac{1}{\gamma}$ days. For Covid-19, $\frac{1}{\gamma}$ is around 14 days, see Ferguson *et al.* (2020).

In this paper, we followed the SIR implementation methodology as described in Towers (2012). Given $\mathcal{R}_0$, β, and γ, the implementation of the SIR model was fairly straightforward using the `deSolve` package, a solver for initial value problems of differential equations (see Soetaert *et al.* (2020)). It was known that $\mathcal{R}_0 = \frac{\beta}{\gamma}$, as noted in Brauer *et al.* (2008). We considered γ as $\frac{1}{14}$, based on Ferguson *et al.* (2020). However, we needed reliable estimates of $\mathcal{R}_0$ to implement the SIR model and predict the disease progression in India.

To estimate $\mathcal{R}_0$, we used the `R` package '`R0`', a toolbox for estimating $\mathcal{R}_0$, as described in Obadia *et al.* (2012). The time between the infection of a primary case and one of its secondary cases is referred to as the generation time, see Svensson (2007). The '`R0`' package assumed that the generation time of the infection was known and required it as input. The mean generation time for Wuhan was reported as 6.5 days Li *et al.* (2020). In this paper, we assumed the generation time followed a Gamma distribution and we estimated the mean and shape parameter of the Gamma distribution using data. Our estimated mean generation time for the Hubei province turned out to be 6.7 days, as presented in Table 2. Upon recovery from infection, we assumed that individuals were immune to re-infection in the short term, consistent with the assumption made in Ferguson *et al.* (2020).

At that time, we deployed a grid search method over the mean and shape of the Gamma distribution for the generation time process. For a particular choice of the mean (μ) and shape (κ) parameter, we generated the generation times and then, using that as input, we estimated $\mathcal{R}_0$ using the '`R0`' package in `R`. For the estimated $\mathcal{R}_0$ and γ (assumed to be 1/14), we simulated the disease progression for the period during which we observed new

incidences. We then calculated the Mean Square Error (MSE) in the following way:

$$MSE(\mu, \kappa) = \frac{1}{T}\sum_{t=1}^{T}\left(\hat{I}(t) - i_{obs}(t)\right)^2, \tag{2}$$

where $\hat{I}(t)$ was the new incidence estimated from the SIR model described in (1) at time point t, and $i_{obs}(t)$ was the actual incidence observed in the data at time point t. We estimated the mean parameter μ and shape parameter κ for which the MSE in (2) was minimized. Then, for the estimated mean and shape parameters, $\mathcal{R}_0$ was estimated using the '`R0`' package.

3.2. Statistical machine learning model

The infection rate of a typical epidemic reaches its peak and then slows down. The SIR model predicts when that peak will be reached very well because it captures the inherent dynamics of the epidemic. However, the SIR model is not as helpful for short and medium-term predictions. We also need short and medium-term predictions to forecast cases as quickly as possible so that health officials can make appropriate decisions. Statistical Machine Learning (SML) models are popular for their prediction accuracy in the short to medium term Sambasivan *et al.* (2020). Consequently, SML and SIR models complement each other.

It is important to note that SML models do not perform well in long-term prediction, particularly when predicting when the peak will be reached. With this understanding, we developed traditional SML models rather than deep learning models. We refrained from developing deep learning models because they require a large amount of data, which is not available in epidemiology. Additionally, the literature on how to apply deep learning to small datasets is still insufficient. Therefore, we focused on developing traditional regression-based SML models for short to medium-term predictions.

As different countries or provinces have varying population levels, we considered our variable of analysis to be the number of cases per 100,000 people (also known as the Rate),

$$\text{Rate} = \frac{\text{Cases}}{\text{Population Size}} \times 100,000.$$

We then modeled the Rate as a function of time, country, and time-country interaction in the following way:

$$\begin{aligned} \ln\{\text{Rate}_{it} + 1\} &= \beta_0 + \beta_1 t + \beta_2 t^2 + \cdots + \beta_p t^p \\ &\quad + \alpha_i + \alpha_i t + \alpha_i t^2 + \cdots + \alpha_i t^p + \epsilon, \end{aligned} \tag{3}$$

where Rate_{it} represents the Rate of the i^{th} country at the t^{th} time point, α_i is the effect of the i^{th} country, $\alpha_i t$ is the linear effect of time on the Rate of the i^{th} country, and $\alpha_i t^2$ is the quadratic effect of time on the Rate of the i^{th} country. We considered the following countries in our model: (1) India, (2) China, (3) US, (4) Iran, (5) South Korea, (6) Japan, (7) Italy, (8) France, (9) Germany, and (10) Spain.

3.3. Model training strategy for India to measure the effect of the lockdown

On March 24, 2020, India announced a national lockdown. To measure the effectiveness of the lockdown, we used all data up to March 24, 2020, to train the model and

learn its parameters. Based on the trained model, we predicted the disease progression path. Since the incubation period of COVID-19 is about 14 days, it was likely that for 14 days following the start of the lockdown, the disease would follow the predicted path and then begin to deviate downward. If the new confirmed cases fell below the predicted path, we could conclude that this was due to the effect of the lockdown. On the other hand, if the disease progression stayed on the predicted path, we would know that the lockdown did not work. If the disease progression rose above the predicted path, we could conclude that the situation worsened during the lockdown.

4. Analysis and prediction

Exploratory Data Analysis (EDA) was important for developing good predictive models. In Figure (1), we plotted the cases per 100,000 (also known as the Rate) for the US, EU, and Iran. The worst-hit regions—US, EU, and Iran—had rates in the range of 70 to 250. On the other hand, disease progression among Asian countries was very different, as shown in Figure (2). The disease progression in both India and Japan was similar. We observed an exponential rise in India and Japan, but at a much lower rate than in Western nations. China was able to flatten the curve, and South Korea managed to curb the rise from exponential to linear. However, up to that point, South Korea experienced the worst rate among the four major Asian countries.

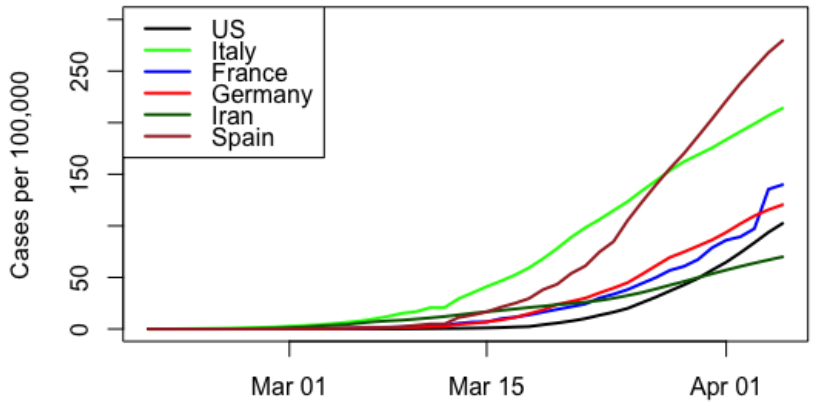


Figure 1: Cases per 100,000 in the US, EU, and Iran. The plot illustrates the rate of cases in the US, Italy, France, Germany, Iran, and Spain from early March to early April, 2020. The rates range from 70 to 250 cases per 100,000, with Spain and Italy experiencing the steepest rises, followed by Germany and the US. The data highlights the rapid escalation in Europe and Iran compared to the US during the observed period.

Prediction of Disease Progression for India from the SML model (3). The solid black vertical line in Figure (3) represented March 24, 2020. The black points to the left of

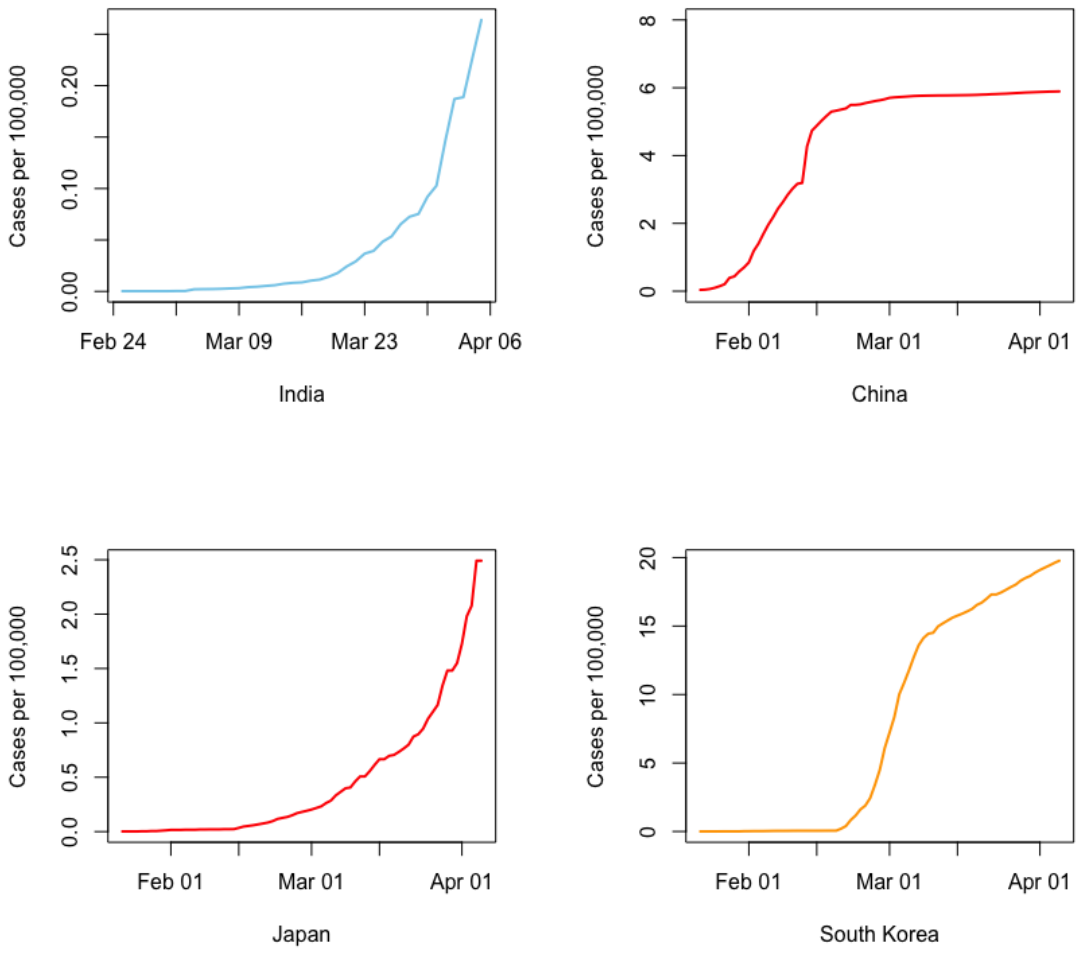


Figure 2: Cases per 100,000 in India, China, Japan, and South Korea. Note that India and Japan's cases per 100,000 are in exponential rise. However, China and South Korea were able to flatten the curve. But at different levels. China was able to flatten the curve at around 6 per 100,000 population; whereas South Korea has partially flattened its curve and increasing as a linear scale.

the vertical black line were the confirmed cases up to March 24, 2020. These black points were used in model training. The solid red line indicated the predicted path of disease progression. The blue points represented the out-of-sample test points, or the confirmed cases that appeared after March 24, 2020. As of April 7, 2020, we had not yet seen the effect of the lockdown. However, if the lockdown worked, its effect should have been visible soon. The blue points should have appeared below the predicted red line. In Table (1), we presented the actual predictions up to May 1, 2020. Had the lockdown been effective, we anticipated case numbers would stay below 66,224 by 1 May 2020. With an out-of-sample R^2 of 0.9323, the actual case count on 1 May 2020 was 37,263; below the predicted value; suggesting that the lockdown was indeed effective.

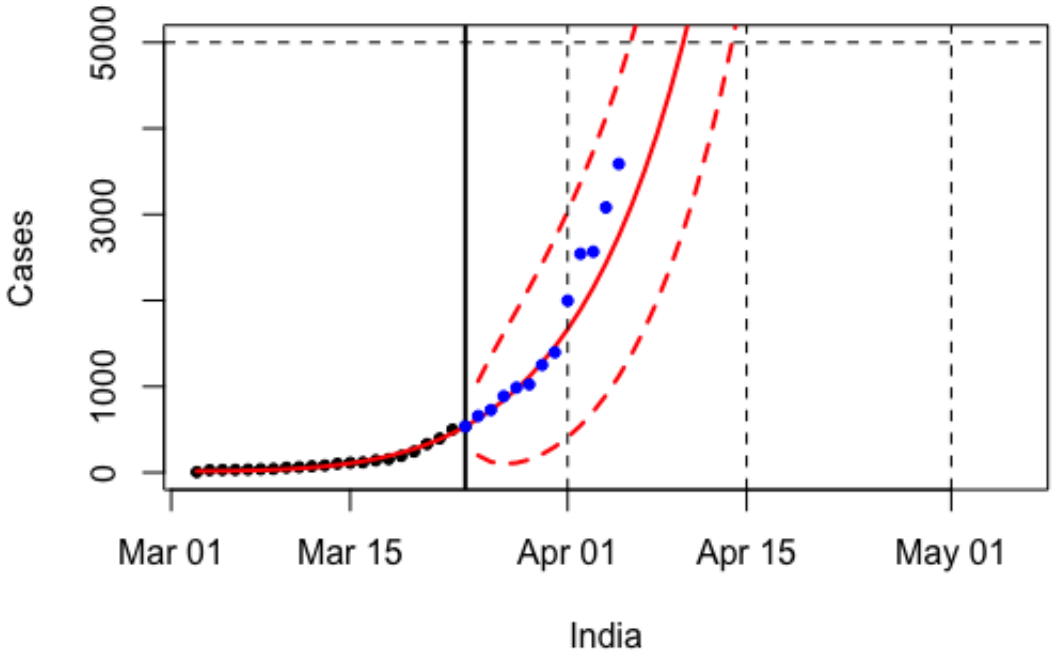


Figure 3: Predicted path of the disease progression in India. The solid black vertical line represent the 24 March 2020. The black points left of the vertical black line are confirmed cases till 24 March 2020. These black points are used in model training. The solid red line is the predicted path of the disease progression. The blue points are the out of sample test point or the confirmed cases that comes after 24 March 2020. As of 07 April 2020, we don't see the effect of lockdown. However, if lockdown works - it should shows its effect any time soon now. The blue point should appear below the predicted red line.

Comparison of $\mathcal{R}_0$ between India and China: In Table (2), the $\mathcal{R}_0$ with a 95% confidence interval for Hubei province and China was around 2.5 during the first 23 days from the start of the lockdown. India's $\mathcal{R}_0$, with a 95% confidence interval, was computed using

Table 1: The table presents the actual cases and prediction from the SML model (3). We used all the data till the 24th March 2020. Here due to space constraint, we present only 5 days interval and recent out of sample values at the daily level. Outsample $R^2 = 0.9323$. The actual values (marked in blue) were added in the current version.

	Dates	Actual Case	Prediction
1	2020-03-03	5	14.99
5	2020-03-07	34	22.42
10	2020-03-12	73	57.72
15	2020-03-17	142	158.74
20	2020-03-22	396	387.54
21	2020-03-23	499	456.29
22	2020-03-24	536	534.79
23	2020-03-25	657	624.10
24	2020-03-26	727	725.36
25	2020-03-27	887	839.85
26	2020-03-28	987	968.95
27	2020-03-29	1024	1114.20
28	2020-03-30	1251	1277.28
29	2020-03-31	1397	1460.05
30	2020-04-01	1998	1664.59
31	2020-04-02	2543	1893.20
32	2020-04-03	2567	2148.44
33	2020-04-04	3082	2433.18
34	2020-04-05	3588	2750.66
35	2020-04-06	4778	3104.50
39	2020-04-10	7599	4974.57
44	2020-04-15	12371	8838.36
49	2020-04-20	18544	15791.88
54	2020-04-25	26283	29126.81
59	2020-04-30	34867	57229.81
60	2020-05-01	37263	66223.94

two different starting points as breakouts. The first was from March 2, 2020, because the number of cases in India started rising from that day. The $\mathcal{R}_0$ for India for the first 22 days up to the lockdown was around 2.5, similar to China. However, if we used the data up to April 4, 2020, the $\mathcal{R}_0$ was around 2.75. This indicated that the situation had worsened since the lockdown, as was clear from Figure (3).

In the second approach, we considered India's breakout from January 23, 2020. In that case, if we considered the data up to March 24, 2020, the $\mathcal{R}_0$ with 95% confidence was almost 1.9, and if we considered the data up to April 4, 2020, the $\mathcal{R}_0$ was nearly 2.1. This meant that if we used the data prior to March 2, 2020, India's $\mathcal{R}_0$ appeared better. In Figure (4), we compared the incidences of Hubei and India in Figures (4:a) and (4:b). We considered the date range for Hubei from January 23, 2020, to February 14, 2020, *i.e.*, during the first 23 days of the Hubei lockdown. On the other hand, we considered the data for India from January 2, 2020, to January 24, 2020, up to the lockdown. On January 23, 2020, Hubei had

Table 2: $\mathcal{R}_0$ with a 95% confidence interval for Hubei province and China is around 2.5 during the first 23 days from the start of the lockdown. India's $\mathcal{R}_0$ with a 95% confidence interval is computed using two different starting points: one from 02-Mar-2020, as the number of cases in India started rising from that day. The $\mathcal{R}_0$ for India for the first 22 days until the lockdown is around 2.5, similar to China. However, if we use the data until 04-Apr-2020, then the $\mathcal{R}_0$ value is around 2.75. In the second approach, we consider India's breakout from 23-Jan-2020. In that case, if we consider the data until 24-Mar-2020, the $\mathcal{R}_0$ with a 95% confidence interval is almost 1.9, and if we consider data until 04-Apr-2020, the $\mathcal{R}_0$ is nearly 2.1.

	Date Range	$\mathcal{R}_0$	$\mathcal{R}_0$ Lower	$\mathcal{R}_0$ Upper	Initial Infections Considered	Mean ($\hat{\mu}$)	Shape (κ)
Hubei	23-Jan-20 to 14-Feb-20	2.53	2.50	2.57	444	6.7	0.24
China	23-Jan-20 to 14-Feb-20	2.46	2.43	2.49	548	8.7	2.7
India	02-Mar-20 to 24-Mar-20	2.52	2.35	2.71	3	5.84	6.56
India	02-Mar-20 to 04-Apr-20	2.75	2.63	2.89	3	5.41	1.10
India	23-Jan-20 to 24-Mar-20	1.87	1.78	1.97	1	2.96	1.53
India	23-Jan-20 to 04-Apr-20	2.09	2.04	2.14	1	1.25	4.98

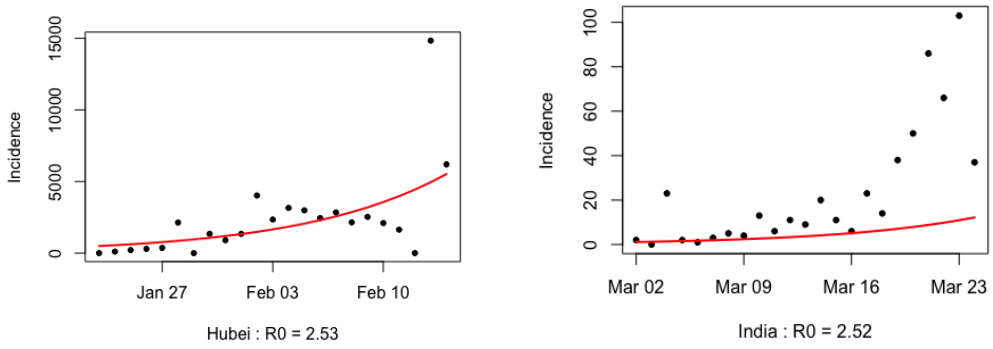
444 confirmed cases, and overall, China had 548 confirmed cases. On January 2, 2020, India had only 3 confirmed cases, whereas on the day of lockdown, *i.e.*, March 24, 2020, India had 536 confirmed cases. So, on the day when the lockdown started, both India and Hubei and/or China had a comparable number of cases.

Perhaps, we should have considered India's $\mathcal{R}_0$ to be around 2.5, similar to the early stage of COVID-19 disease progression in China. Even with the lockdown, China experienced more than 80,000 cases. Perhaps, we should have prepared for at least that many cases, if not more, in India.

State-wise $\mathcal{R}_0$: In Table (3), we presented the state-wise Basic Reproduction Number, $\mathcal{R}_0$, as of March 4, 2020. We observed that Punjab's $\mathcal{R}_0$ was the worst in the country. Punjab's high $\mathcal{R}_0 \approx 16$ was likely due to a super spreader who ignored advice to self-quarantine after returning from a trip to Italy and Germany (see BBC News (2020)). The situation in Punjab was really complicated, and serious intervention was required. In Figure (5), we presented the cases in Punjab over time. Since March 20, 2020, the number of confirmed cases increased at an unprecedented rate.

From Table (3), we saw that the $\mathcal{R}_0$ for Madhya Pradesh (3.37), Maharashtra (3.25), and Tamil Nadu (3.09) were all above 3. Clearly, the situations were complicated in these three states. The $\mathcal{R}_0$ for Andhra Pradesh (2.96), Delhi (2.82), and West Bengal (2.77) were also higher than India's overall $\mathcal{R}_0$ of 2.75. These seven states needed special attention as their $\mathcal{R}_0$ exceeded that of India. These numbers were as of April 4, 2020.

For the following states, we either did not have enough data to make inferences for $\mathcal{R}_0$, or the algorithm failed to converge: (1) Andaman and Nicobar Islands; (2) Arunachal Pradesh; (3) Chhattisgarh; (4) Goa; (5) Haryana; (6) Jharkhand; (7) Manipur; (8) Mizoram; (9) Odisha; (10) Puducherry.



(a) (b)

Figure 4: In this figure, we compare the incidences of Hubei and India in (a) and (b). We consider the date range for Hubei from 23-Jan-2020 to 14-Feb-2020, *i.e.*, during the first 23 days of Hubei lockdown. On the other hand, we considered the data for India, from the 02-Jan-2020 to 24-Jan-2020, before the lockdown. On the 23-Jan-2020, Hubei had 444 confirmed cases and overall China had 548 confirmed cases. On 02-Jan-2020, India had only 3 confirmed cases, whereas on the day of lockdown, *i.e.*, on 24-Jan-2020, India had 536 confirmed cases.

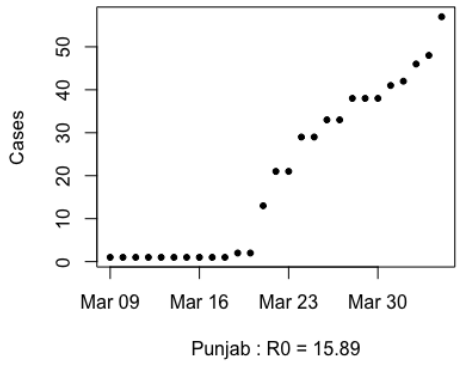


Figure 5: Confirmed cases of COVID19 in Punjab. The $\mathcal{R}_0 = 15.89$. The high $\mathcal{R}_0$ is likely due to a super spreader ignored advice to self quarantine after returning from a trip to Italy and Germany, see BBC News (2020)

Table 3: State Wise Basic Reproduction Number, $\mathcal{R}_0$, as of 04 March, 2020. Punjab's high $\mathcal{R}_0$ is likely due to a super spreader ignored advice to self quarantine after returning from a trip to Italy and Germany, see BBC News (2020)

State/UT	$\mathcal{R}_0$	Lower	Upper
Andhra Pradesh	2.96	2.56	3.45
Bihar	2.13	1.35	3.40
Chandigarh	1.14	0.89	1.48
Delhi	2.82	2.60	3.08
Gujarat	0.98	0.84	1.15
Himachal Pradesh	1.59	1.00	3.13
Jammu and Kashmir	2.02	1.69	2.48
Karnataka	2.29	1.87	2.77
Kerala	1.62	1.52	1.74
Ladakh	1.54	1.17	2.18
Madhya Pradesh	3.37	2.73	4.14
Maharashtra	3.25	2.95	3.58
Punjab	**15.89**	**4.12**	**149.27**
Rajasthan	2.45	2.25	2.67
Tamil Nadu	3.09	2.74	3.53
Telengana	2.16	1.97	2.38
Uttar Pradesh	2.30	2.10	2.52
Uttarakhand	1.33	1.13	1.61
West Bengal	2.77	2.21	3.47
India	2.75	2.63	2.89

5. Discussion

The COVID-19 pandemic has prompted extensive research to understand transmission dynamics, evaluate the impact of interventions, and forecast its trajectory. Our early-stage analysis provided a critical assessment of the severity of the situation across various Indian states. We observed that the reproduction number ($\mathcal{R}_0$) for Punjab was alarmingly high, requiring immediate and aggressive intervention. Madhya Pradesh (3.37), Maharashtra (3.25), and Tamil Nadu (3.09) also exhibited reproduction numbers above 3, indicating the need for urgent action in these states. We noted that the $\mathcal{R}_0$ values for Andhra Pradesh (2.96), Delhi (2.82), and West Bengal (2.77) exceeded India's overall $\mathcal{R}_0$ of 2.75. As of 4 March 2020, India's $\mathcal{R}_0$ was comparable to Hubei, China, during the early outbreak phase, suggesting that India could experience a similar case trajectory if effective containment measures were not implemented. Based on the assumption of lockdown efficacy, we predicted that the total cases in India might remain below 66,224 by 1 May 2020.

Subsequent studies built upon this initial analysis. Early estimates of the basic reproduction number (R_0) for India by Das (2020) placed it around 2.75, similar to China's early pandemic stage. Later, Sinha (2020) revised this estimate to approximately 1.82 by analysing time-series data of active cases in India and other countries, confirming that non-pharmaceutical interventions, such as lockdowns, were effective in reducing transmission rates but insufficient to completely halt transmission. Both Das (2020) and Sinha (2020) highlighted regional variations in COVID-19 dynamics across India. Early studies like Mittal (2020) employed Exploratory Data Analysis (EDA) to examine COVID-19 case trends in

India up to 22 April 2020, offering insights into daily and weekly case patterns, comparing trends with neighbouring and severely affected countries, and assessing India's healthcare preparedness for the pandemic.

Further descriptive studies, such as Bhatnagar *et al.* (2021), analysed COVID-19 cases in India, examining factors like age, gender, travel history, transmission type, and patient status. They found no significant correlation between age and susceptibility but observed a strong relationship between gender and transmission type. Halder *et al.* (2022) analysed mortality and recovery rates during the lockdown phases in India, revealing high correlations between active cases and both death ($R^2 = 0.8754$) and recovery rates ($R^2 = 0.9246$), though the results offered predictable insights with limited novelty.

Deo *et al.* (2020) extended the containment strategy analysis by developing a time-series SIR model to predict COVID-19 dynamics in India. Their model incorporated progressive containment measures and provided forecasts for transmission rates and daily cases under various scenarios, aligning with our early focus on timely intervention. The study by Rath *et al.* (2020) applied Linear and Multiple Linear Regression models to predict daily active COVID-19 cases in Odisha and India, achieving high accuracy (R^2 close to 1). At the state level, Tinani *et al.* (2020) explored COVID-19 modelling for hotspot states using the ARIMA model to predict cases, recoveries, and deaths across key states like Maharashtra, Delhi, and Gujarat, which corresponded with our findings on the need for focused attention on states with higher $\mathcal{R}_0$ values.

The study by Ghosh *et al.* (2020) conducted a statewise COVID-19 analysis, predicting infection trends using ensemble models and categorising states by severity to guide resource allocation, with recommended preventive measures for states with rising daily infection rates (DIR). Roy *et al.* (2021) employed ARIMA models and GIS-based spatial analysis to forecast COVID-19 prevalence in India, identifying western and southern regions as particularly vulnerable, and demonstrated ARIMA's effectiveness in epidemiological surveillance. The study by Arora *et al.* (2020) applied deep learning models, particularly LSTM variants, to predict COVID-19 case numbers in India with high accuracy (errors below 3% for daily and 8% for weekly forecasts). They categorised states into zones based on case spread and growth rates to identify hotspots, with preventive recommendations provided. Additionally, they created a website to update these predictions for authorities and researchers.

Further studies, such as Tomar and Gupta (2020), utilised data-driven methods like LSTM and curve fitting to forecast COVID-19 trends in India over a 30-day period, evaluating the effect of preventive measures and offering accurate predictions to aid health officials and administrators. Tiwari (2020) employed an SIQR model to analyse COVID-19's progression, estimating effective reproduction rates, doubling times, and infection-to-quarantine ratios, while emphasising the link between testing rates and case detection, and suggesting model enhancements for accuracy.

Recognising lockdowns' role in controlling COVID-19, Das *et al.* (2020b) proposed a Susceptible-Exposed-Infected-Recovered (SEIR) model to estimate Temporary Eradication of Spread Time (TEST) and Critical Community Size (CCS) for Indian states, supporting our initial analysis on the need for decisive action. Similarly, Kumar (2020) applied cluster analysis to identify groups within COVID-19 data across Indian states and union territories, enhancing monitoring strategies to support government and healthcare decision-making for

improved resource allocation.

Beyond epidemiology, researchers examined socioeconomic and demographic factors influencing COVID-19 outcomes. Chakravarty *et al.* (2021) analysed the impact of comorbidities, health expenditure, and life expectancy on case fatality rates across SAARC nations, underscoring the importance of targeted interventions based on local vulnerabilities, complementing our early epidemiological analysis. Broader impacts of the pandemic were explored in studies like Pyne *et al.* (2020), who assessed social vulnerabilities to guide post-pandemic recovery, particularly in India. Similarly, Dutta *et al.* (2020) analysed the economic effects of lifting or partially implementing lockdowns in Maharashtra and Gujarat, using statistical models to project future scenarios and providing additional perspectives on the socioeconomic ramifications observed in the pandemic's early stages.

Economically, Das *et al.* (2020a) examined the pandemic's effect on payment transactions in India, noting significant reductions in economic activity due to lockdowns, followed by gradual recovery in digital payments, aligning with our initial analysis of the broader economic impacts. Grover and Magan (2020) estimated Quality Adjusted Life Years (QALY) for COVID-19 patients across Indian states, offering quantitative assessments of the pandemic's impact and providing further context to our initial predictions regarding the severity of the pandemic in different regions. The study by Shruthi and Ramani (2021) analysed COVID-19's effects on financial systems, revealing that post-crisis oil market volatility impacted agricultural commodities (excluding sugar), while pre-crisis risk transmission was minimal.

Methodologically, Venkatesan (2020) addressed modelling uncertainties using back-calculation to reconstruct past infection patterns and predict future cases in India. Sarkar (2020) proposed group testing methodologies to reduce mass testing costs, particularly valuable when disease prevalence was low. These methodological refinements complemented the epidemiological insights from our early work, enhancing pandemic management approaches.

Internationally, Maleki *et al.* (2020) and Zhang *et al.* (2020) analysed COVID-19 dynamics in the U.S., with the former examining the association between comorbidities and death rates across U.S. cities, and the latter identifying change points in the pandemic's progression. These studies offered comparative insights that informed COVID-19 management in India. A different study by Gupta *et al.* (2020) investigated the influence of weather, particularly temperature and absolute humidity, on COVID-19 spread in the U.S., finding significant case increases in states with absolute humidity levels between 4 and 6 g/m^3. The results aligned with global trends and identified Indian regions potentially vulnerable to weather-driven COVID-19 transmission, underscoring weather's role in transmission risk.

In summary, our early analysis laid the foundation for subsequent research on COVID-19 in India, providing essential insights into the pandemic's progression and emphasising the need for swift, aggressive intervention in states with high $\mathcal{R}_0$ values. The extensive literature that followed expanded upon these initial findings, offering deeper insights into the epidemiological, socioeconomic, and policy dimensions of the pandemic, while validating many of our initial predictions and observations.

6. Conclusion

The conclusion of this study, aimed at predicting COVID-19 progression in India using both the SIR epidemiological model and a statistical machine learning approach, provides several key insights into the trajectory of the disease under the national lockdown. Conducted during the early phase of the pandemic, this research offers a valuable reference for understanding the dynamics of COVID-19 and implementing effective intervention strategies.

Firstly, the results underscored the critical importance of timely and aggressive interventions in mitigating the spread of COVID-19. The high basic reproduction number (R_0) observed in states such as Punjab, Madhya Pradesh, Maharashtra, and Tamil Nadu indicated the urgent need for concentrated efforts in these regions. Punjab's R_0 of 15.89 - driven by a super spreader event—highlighted the need for immediate and comprehensive containment measures. Other states, including Andhra Pradesh, Delhi, and West Bengal, also had reproduction numbers exceeding India's overall R_0, which was calculated at 2.75 as of 4 March 2020. This finding highlighted the regional disparities in COVID-19 transmission, necessitating tailored interventions to effectively curb the spread of the virus.

The study also revealed that India's disease progression mirrored that of China's early pandemic phase, particularly in terms of the R_0 values. With China's experience showing a similar reproduction number, it was evident that without a successful lockdown and containment strategy, India could have faced a similar, or even greater, number of cases. The model predicted that, if the lockdown was effective, the number of confirmed cases in India by 1 May 2020 would remain under 66,224. However, the analysis demonstrated that India's R_0 began to rise following the lockdown, indicating that while the initial lockdown slowed the virus's spread, it might not have been sufficient to halt transmission entirely.

Another significant outcome of this research was the validation of a hybrid modelling approach, where the SIR model provided accurate long-term predictions of disease dynamics, while the machine learning model excelled in short- to medium-term forecasts. This dual strategy was especially useful in understanding the immediate impacts of the lockdown, enabling public health officials to allocate resources more effectively and plan for the spread of the virus. The efficacy of the lockdown could be evaluated by comparing the actual number of cases after the lockdown to the predicted numbers based on pre-lockdown data. If new cases fell below the predicted levels, it would suggest that the lockdown was working. Conversely, if case numbers exceeded predictions, the lockdown measures would need to be reconsidered.

The use of data from multiple sources, such as Johns Hopkins University's COVID-19 repository and open-source platforms like Covid19India, was crucial in ensuring the accuracy of the model's predictions. The integration of both global and local datasets enabled a more detailed understanding of the pandemic's progression in India, a country with vast regional differences in population density, healthcare infrastructure, and socioeconomic factors. These variations were reflected in the model's predictions, which highlighted states like Kerala, with relatively lower R_0 values, indicating that local intervention efforts were somewhat successful.

However, it is important to recognise that while this study provided early estimates and predictions, the dynamics of the pandemic were rapidly evolving. Continuous data col-

lection and refinement of models would be essential to ensure that public health responses could adapt to new developments. The early prediction that India could experience over 66,000 cases by May 2020, assuming successful lockdown measures, offered a critical window for the government to expand healthcare capacity and implement more targeted interventions, such as scaling up testing, improving contact tracing, and ensuring the availability of personal protective equipment (PPE) for healthcare workers.

In conclusion, this study offered a vital early framework for understanding and predicting the spread of COVID-19 in India, delivering actionable insights for policymakers. The hybrid approach, combining epidemiological models with statistical machine learning, allowed for more accurate short- and long-term predictions, helping to shape India's pandemic response. The key takeaway is the necessity of timely, aggressive, and region-specific interventions to control the spread of infectious diseases, particularly in a country as diverse and densely populated as India. Moreover, the study emphasised the limitations of lockdowns as a long-term solution and stressed the need for a robust healthcare infrastructure and continuous policy adaptation based on real-time data.

Addendum

Prediction of Disease Progression for India: In Table (1), we presented the actual predictions up to May 1, 2020. Had the lockdown been effective, we anticipated case numbers would stay below 66,224 by 1 May 2020. With an out-of-sample R^2 of 0.9323, the actual case count on 1 May 2020 was 37,263; below the predicted value; suggesting that the lockdown was indeed effective.

Table (1) Description: The table presents the actual cases and prediction from the SML model (3). We used all the data till the 24th March 2020. The blue values were added in the current version

Acknowledgements

I sincerely thank the chief editor for inviting me to contribute this paper and for providing valuable guidance and advice. I am also deeply grateful to the reviewer for their insightful comments, helpful suggestions, and for kindly recommending numerous useful references.

Conflict of interest

The author declares no financial or non-financial conflicts of interest related to the research presented in this article.

References

Arora, P., Kumar, H., and Panigrahi, B. K. (2020). Prediction and analysis of COVID-19 positive cases using deep learning models: A descriptive case study of India. *Chaos, Solitons & Fractals*, **139**, 110017. Epub 2020 Jun 17.

BBC News (2020). Coronavirus: India 'super spreader' quarantines 40,000 people. `https://www.bbc.com/news/world-asia-india-52061915`. Accessed: 27 March 2020.

Bhatnagar, V., Poonia, R. C., Nagar, P., Kumar, S., Singh, V., Raja, L., and Dass, P. (2021). Descriptive analysis of COVID-19 patients in the context of India. *Journal of Interdisciplinary Mathematics*, **24**, 489–504.

Biswas, S. (2020). Coronavirus: Why is India testing so little? `https://www.bbc.com/news/world-asia-india-51922204`. Accessed: 20 March 2020.

Box, G. (1976). Science and statistics. *Journal of the American Statistical Association*, **71**, 791–799.

Brauer, F., Driessche, P. v. d., and Wu, J., editors (2008). *Mathematical Epidemiology*. Springer. Lecture Notes in Mathematics, 1945.

Chakravarty, S., Grover, G., and Aggarwal, S. (2021). Association of socioeconomic and demographic factors with COVID-19 related health outcomes in SAARC nations. *Statistics and Applications*, **19**, 367–386. ISSN 2454-7395 (online).

Das, A., Das, S., Jaiswal, A., and Sonthalia, T. (2020a). Impact of COVID-19 on payment transactions. *Statistics and Applications*, **18**, 239–251.

Das, S. (2020). Prediction of COVID-19 disease progression in India under the effect of national lockdown. First Version: April 07, 2020, Available from: `https://arxiv.org/pdf/2004.03147`.

Das, S., Ghosh, P., Sen, B., Pyne, S., and Mukhopadhyay, I. (2020b). Critical community size for COVID-19: A model based approach for strategic lockdown policy. *Statistics and Applications*, **18**, 181–196.

Deo, V., Chetiya, A. R., Deka, B., and Grover, G. (2020). Forecasting transmission dynamics of COVID-19 in India under containment measures- a time-dependent state-space SIR approach. *Statistics and Applications*, **18**, 157–180.

Dutta, S., Das, K., Chatterjee, K., and Chakraborty, A. (2020). What if lockdown is removed? district level predictions for Maharashtra and Gujarat. *Statistics and Applications*, **18**, 209–221.

Ferguson, N. M., Laydon, D., Nedjati-Gilani, G., Imai, N., Ainslie, K., Baguein, M., Bhatia, S., Boonyasiri, A., Cucunubá, Z., Cuomo-Dannenburg, G., *et al.* (2020). Report 9: Impact of non-pharmaceutical interventions (NPIs) to reduce COVID-19 mortality and healthcare demand. Technical report, Imperial College London, WHO Collaborating Centre for Infectious Disease Modelling, MRC Centre for Global Infectious Disease Analysis, Abdul Latif Jameel Institute for Disease and Emergency Analytics. Imperial College COVID-19 Response Team, Available from `https://www.imperial.ac.uk/media/imperial-college/medicine/sph/ide/gida-fellowships/Imperial-College-COVID19-NPI-modelling-16-03-2020.pdf`.

Ghosh, P., Ghosh, R., and Chakraborty, B. (2020). COVID-19 in India: Statewise analysis and prediction. *JMIR Public Health and Surveillance*, **6**, e20341.

Grover, G. and Magan, R. (2020). Estimation of quality adjusted life year (qaly) for different states of India during COVID-19. *Statistics and Applications*, **18**, 319–331.

Gupta, S., Raghuwanshi, G. S., and Chanda, A. (2020). Effect of weather on COVID-19 spread in the US: A prediction model for India in 2020. *Science of The Total Environment*, **728**, 138860. Epub 2020 Apr 21; Erratum in: Sci Total Environ. 2020 Dec 15;748:142577. doi: 10.1016/j.scitotenv.2020.142577.

Halder, B., Bandyopadhyay, J., and Banik, P. (2022). Statistical data analysis of risk factor associated with mortality rate by COVID-19 pandemic in India. *Modeling Earth Systems and Environment*, **8**, 511–521.

Kumar, S. (2020). Monitoring novel corona virus (COVID-19) infections in India by cluster analysis. *Annals of Data Science*, **7**, 417–425.

Li, Q., Guan, X., and Wu, P. e. (2020). Early transmission dynamics in Wuhan, China, of novel corona virus–infected pneumonia. *The New England Journal of Medicine*, **382**.

Maleki, M., McLachlan, G. J., Gurewitsch, R., Aruru, M., and Pyne, S. (2020). A mixture of regressions model of COVID-19 death rates and population comorbidities. *Statistics and Applications*, **18**, 295–306.

Mittal, S. (2020). An exploratory data analysis of COVID-19 in India. *International Journal of Engineering Research & Technology (IJERT)*, **9**, IJERTV9IS040550. `http://www.ijert.org`.

Obadia, T., Haneef, R., and Boëlle, P.-Y. (2012). The R0 package: a toolbox to estimate reproduction numbers for epidemic outbreaks. *BMC Medical Informatics and Decision Making*, **12**, 147.

Pandey, G. (2020). Coronavirus: The woman behind India's first testing kit. `https://www.bbc.com/news/world-asia-india-52083196`. Accessed: 28 March 2020.

Pyne, S., Ray, S., Gurewitsch, R., and Aruru, M. (2020). Transition from social vulnerability to resiliency vis-à-vis COVID-19. *Statistics and Applications*, **18**, 197–208.

Rath, S., Tripathy, A., and Tripathy, A. R. (2020). Prediction of new active cases of coronavirus disease (COVID-19) pandemic using multiple linear regression model. *Diabetes & Metabolic Syndrome: Clinical Research & Reviews*, **14**, 1467–1474. Epub 2020 Aug 1.

Roy, S., Bhunia, G. S., and Shit, P. K. (2021). Spatial prediction of COVID-19 epidemic using ARIMA techniques in India. *Modeling Earth Systems and Environment*, **7**, 1385–1391.

Sambasivan, R., Das, S., and Sahu, S. K. (2020). A Bayesian perspective of statistical machine learning for big data. *Computational Statistics*, **35**, 893–930. Accepted and available here :`https://link.springer.com/article/10.1007/s00180-020-00970-8`.

Sarkar, J. (2020). Reducing the number of tests for COVID-19 infection via group testing methodologies. *Statistics and Applications*, **18**, 281–294.

Shruthi, M. and Ramani, D. (2021). Statistical analysis of impact of COVID 19 on India commodity markets. *Materials Today: Proceedings*, **37**, 2306–2311. International Conference on Newer Trends and Innovation in Mechanical Engineering: Materials Science.

Sinha, S. (2020). Epidemiological dynamics of the COVID-19 pandemic in India: An interim assessment. *Statistics and Applications*, **18**, 333–350.

Soetaert, K., Petzoldt, T., and Setzer, R. W. (2020). *deSolve: Solvers for Initial Value Problems of Differential Equations*. R package version 1.28.

Svensson, A. (2007). A note on generation times in epidemic models. *Mathematical Bioscience*, **208**, 300–311.

Tinani, K., Muralidharan, K., Deshmukh, A., Patil, B., Salat, T., and Rajodia, R. (2020). Analysis and forecasting of COVID-19 cases across hotspot states of India. *Statistics and Applications*, **18**, 223–238.

Tiwari, A. (2020). Modelling and analysis of COVID-19 epidemic in India. *Journal of Safety Science and Resilience*, **1**, 135–140.

Tomar, A. and Gupta, N. (2020). Prediction for the spread of COVID-19 in India and effectiveness of preventive measures. *Science of The Total Environment*, **728**, 138762. Epub 2020 Apr 20.

Towers, S. (2012). Epidemic modelling with compartmental models using R. Available from https://sherrytowers.com/2012/12/11/simple-epidemic-modelling-with-an-sir-model/.

Venkatesan, P. (2020). A comprehensive modeling framework for estimation and prediction of COVID-19 in India. *Statistics and Applications*, **18**, 269–280.

World-Bank (2021). Hospital beds (per 1,000 people); data from the World Health Organization, supplemented by country data. license: Cc by-4.0. https://data.worldbank.org/indicator/SH.MED.BEDS.ZS?end=2021&name_desc=false&start=1960&view=chart.

Zhang, S., Xu, Z., and Peng, H. (2020). Change point modeling of Covid-19 data in the United States. *Statistics and Applications*, **18**, 307–318.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 23, No. 1, 2025 (New Series), pp 385–388
https://www.ssca.org.in/journal

Understanding Fellegi Scheme for Sample Size Three

Yumnam Menon Singh[1] **and Opendra Salam**[2]
[1]*Research Scholar, Statistics, Manipur University, Imphal, Manipur, India*
[2]*Professor of Statistics, Manipur University, Imphal, Manipur, India*

Received: 13 May 2024; Revised: 30 September 2024; Accepted: 03 October 2024

Abstract

In this short communication, we attempt to rework on Fellegi (1963) scheme for sample size 3, taking clue from Choudhry (1981) and Sinha (1973, 1974).

Key words: Sampling designs; Sampling schemes; Inclusion probabilities of first and second orders; Mixture designs.

1. Introduction

Brewer and Hanif (1983) reviewed sampling schemes with unequal probabilities without replacement and compiled several selection procedures. Among the schemes, Brewer (1963) and Fellegi (1963) schemes for $n = 2$ are described in text books such as in Hedayat and Sinha (1991) but cannot be readily extended to $n = 3$. For Fellegi scheme, Choudhry (1981) attempted to develop computational formulae using Fortran language specifically for $n = 3$ and 4. However, satisfactory techniques are not yet available. We make an attempt to extend Fellegi scheme from algebraic consideration. Our contribution in this study is essentially a follow-up of Fellegi ($n = 2$) to $n = 3$. We are able to generalize Fellegi scheme and we explain our procedure through a numerical example.

It may be noted that Choudhry (1981) made an attempt to work out a solution for $n = 3$ underlying Fellegi scheme. He did not pursue any analytical exercise to solve for the choice of $p_3(i)$ values. He used the second stage p-values ($p_2(i)$) as trial values for the third stage p-values ($p_3(i)$) and developed a Fortran programme to approximately work out stabilized third stage p-values.

2. Fellegi scheme $(N, n = 3)$

For Fellegi Scheme $(N, n = 3)$, $P(i, j, k)$ has to be chosen in such a way that at each trial, inclusion probability of i^{th} unit is p_i for all i. Hence, overall inclusion probability for i^{th} unit is $3p_i$. To achieve this, set k^{th} trial selection probability for i^{th} unit $= p_k(i)$ for

Corresponding Author: Salam Opendra Singh
Email: sopendra@manipuruniv.ac.in

$k = 1, 2, 3$; $i = 1, 2, 3, \ldots, N$ where $\sum_{i=1}^{N} p_k(i) = 1$ for each k. Then we have the expression

$$\pi_i = p_1(i) + \sum_{j(\neq i)}^{N} p_1(j)\frac{p_2(i)}{1-p_2(j)} + \sum_{k(\neq i)}^{N}\sum_{j(\neq i,k)}^{N} p_1(k)\frac{p_2(j)}{(1-p_2(k))}\frac{p_3(i)}{(1-p_3(k)-p_3(j))} = 3p_i \quad (1)$$

It may be noted that in the above, we are tacitly using the expression for $p_2(i)$ as was derived by Fellegi (1963) for the case of $n = 2$. Set $p_1(i) = p_i$ for each $i = 1, 2, \ldots, N$.
So, $p_3(i)$'s have to satisfy

$$\sum_{k(\neq i)}^{N}\sum_{j(\neq i,k)}^{N} p_1(k)\frac{p_2(j)}{(1-p_2(k))}\frac{p_3(i)}{(1-p_3(k)-p_3(j))} = p_i, \quad i = 1, 2, \ldots, N.$$

$$\Rightarrow B_i = \frac{p_i}{p_3(i)}\left[\frac{1-2p_3(i)-p_2(i)}{(1-p_2(i))(1-2p_3(i))}\right] \quad (2)$$

where $B_i = \sum_{k=1}^{N}\sum_{j=1}^{N}\frac{p_1(k)p_2(j)}{(1-p_2(k))(1-p_3(k)-p_3(j))} - \sum_{j=1}^{N}\frac{p_1(i)p_2(j)}{(1-p_2(i))(1-p_3(i)-p_3(j))}$

$$-\sum_{k=1}^{N}\frac{p_1(k)p_2(i)}{(1-p_2(k))(1-p_3(k)-p_3(i))} - \sum_{k=1}^{N}\frac{p_1(k)p_2(k)}{(1-p_2(k))(1-2p_3(k))} + \frac{2p_1(i)p_2(i)}{(1-p_2(i))(1-2p_3(i))}$$

After simplifying (2), we obtain a quadratic equation in $p_3(i)$ as

$$2B_i(1-p_2(i))p_3^2(i) - [2p_i + B_i(1-p_2(i))]p_3(i) + p_i(1-p_2(i)) = 0 \quad (3)$$

$$\text{So, } p_3(i) = \frac{(2p_i + B_i(1-p_2(i))) \pm \sqrt{(2p_i + B_i(1-p_2(i)))^2 - 8B_ip_i(1-p_2(i))^2}}{4B_i(1-p_2(i))} \quad (4)$$

Remark 1: It must be noted that the expressions in (2) and (4) basically refer to only one relation involving B_i and $p_3(i)$. A judicial choice of B_i for evaluation of $p_3(i)$ has, so far, eluded us. Therefore, we have taken up an alternative approach that refers to a choice of $p_3(i)$ as a function of p_i and $p_2(i)$ with the solo objective: Choice of $p_3(i)$ must lead to the 3rd stage $\pi_i = 3p_i$ to best possible approximation.

Remark 2: At this stage it is pertinent to note that we will be using the concept of mixture designs of the type $pD_1 + qD_2$, $0 < p,\ q < 1,\ p + q = 1$. We recall that Sinha (1973, 1974) made a similar study with the provision that one of p and q would be negative, however, satisfying the necessary condition that $pD_1(s) + qD_2(s) > 0$ for every sample 's' in the underlying design. In our study below we will follow Sinha's approach to come up with a solution.

Remark 3: This problem is simply stated and theoretical solutions are quite hard to obtain. We make attempts to minimize the gap between π_i and $3p_i$ by making suitable choice of $p_3(i)$'s. Similar problem was encountered by Sinha (1973, 1974) who had developed a mixture solution of the type: $p_3(i) = ap_i + bp_2(i)$ with choices of a and b subject to $a + b = 1$, by admitting the solutions with negative values of a or b ! Of course, the mixture has to yield all positive fractions. Our attempt is illustrated in the following example.

Example 1: $N = 6,\ p_1 = 0.25,\ p_2 = p_3 = 0.20,\ p_4 = p_5 = 0.15,\ p_6 = 0.05.$
With reference to Fellegi (1963), for the case of $n = 2$,

(i) Solve for A from the equation: $N - 2 = \sum_{i=1}^{N} \sqrt{1 - \frac{4p_i}{A}}$, where $A = \sum_{t=1}^{N} \frac{p_t}{1 - p_2(t)}$.

(ii) Solve for $p_2(i)$ from the equation: $p_2(i) = \frac{1}{2} - \sqrt{\frac{1}{4} - \frac{p_i}{A}}$.

Newton's method is used to obtain: $A = 1.24727$, and then values for $p_2(i)$ are deduced as given below in Table 1.

Table 1: Calculation of $p_2(i)$

i	1	2	3	4	5	6	Sum
p_i	0.25	0.2	0.2	0.15	0.15	0.05	1
$p_2(i)$	0.27737	0.20058	0.20058	0.13981	0.13981	0.04184	1

Keeping the possibility of one of a and b being negative, after some trial and error, we ended up with $a = 2.55$ and $b = -1.55$. The end-result is shown below.

Table 2: Computation of π_i

i	1	2	3	4	5	6	Total
π_i	0.73560 ≈0.74	0.60475 ≈0.60	0.60475 ≈0.60	0.45445 ≈0.45	0.45445 ≈0.45	0.14598 ≈0.15	2.99998 ≈3
$3p_i$	0.75	0.6	0.6	0.45	0.45	0.15	3

Table 3: Computation of $\pi_{ij} = \sum_{s \ni (i,j)} P(s)$

π_{ij}						
i j	1	2	3	4	5	6
1		0.40506	0.40506	0.28831	0.28831	0.08445
2	0.40506		0.31175	0.21547	0.21547	0.06174
3	0.40506	0.31175		0.21547	0.21547	0.06174
4	0.28831	0.21547	0.21547		0.14762	0.04202
5	0.28831	0.21547	0.21547	0.14762		0.04202
6	0.08445	0.06174	0.06174	0.04202	0.04202	

Remark 4: We can readily verify numerically for $n = 3$ that $\pi_{ik} > \pi_{jk}$ whenever $p_i > p_j$ for all $i \neq j \neq k$ and $\pi_i \pi_j > \pi_{ij}$ for all $i \neq j$.

3. Conclusion

From the above illustration it can be seen that if one can express $p_3(i)$ as a linear combination of p_i and $p_2(i)$ that is $p_3(i) = wp_i + (1 - w)p_2(i)$, with a suitable choice of w, the Fellegi scheme for $n = 3$ can be constructed in a simple way. Further research is needed to find an appropriate value of w, assuming that it can take negative values as well.

Acknowledgements

We are indeed very grateful to the Chief Editor of S&A for making a number of valuable suggestions towards revision on an earlier draft of the manuscript. His insightful comments have been extremely helpful in revising the manuscript. We also thank Professor Bikas Sinha, Retired Professor of ISI, Kolkata for his interest in this investigation.

Conflict of interest

The authors do not have any financial or non-financial conflict of interest to declare for the research work included in this article.

References

Brewer, K. R. W. (1963). A model of systematic sampling with unequal probabilities. *Australian Journal of Statistics*, **5**, 5–13.

Brewer, K. R. W. and Hanif, M. (1983). *Sampling with Unequal Probabilities*, volume 15 of *Lecture Notes in Statistics*. Springer-Verlag, New York.

Choudhry, G. H. (1981). Construction of working probabilities and joint selection probabilities for Fellegi's PPS sampling scheme. *Survey Methodology*, **7**, 93–108.

Fellegi, I. (1963). Sampling with varying probabilities without replacement: rotating and non-rotating samples. *Journal of the American Statistical Association*, **58**, 183–201.

Hedayat, A. and Sinha, B. K. (1991). *Design and Inference in Finite Population Sampling.* Wiley New York.

Sinha, B. K. (1973). On sampling schemes to realize pre-assigned sets of inclusion probabilities of first two orders. *Calcutta Statistical Association Bulletin*, **22**, 89–110.

Sinha, B. K. (1974). On sampling schemes to realize invariant pre-assigned sets of inclusion probabilities of first two orders. *Calcutta Statistical Association Bulletin*, **23**, 45–72.

Publisher
Society of Statistics, Computer and Applications
Registered Office: I-1703, Chittaranjan Park, New Delhi-110019, INDIA
Mailing Address: B-133, Ground Floor, C.R. Park, New Delhi-110019, INDIA
Tele: 011-40517662
https://ssca.org.in/
statapp1999@gmail.com
2025

Printed by : Galaxy Studio & Graphics
Mob: +91 9818 35 2203, +91 9582 94 1203