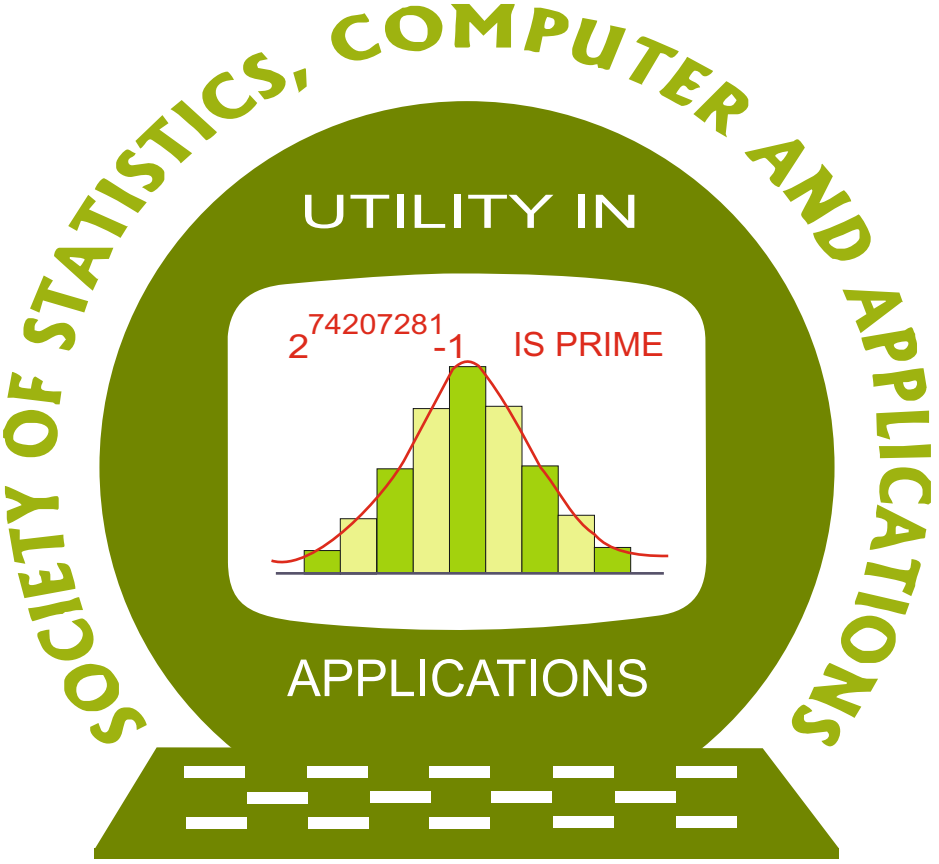
ISSN 2454-7395(online)

STATISTICS AND APPLICATIONS



FOUNDED 1998

Journal of the Society of Statistics, Computer and Applications

https://ssca.org.in/journal.html
Volume 22, No. 2, 2024 (New Series)

Society of Statistics, Computer and Applications

Council and Office Bearers

Founder President
Late M.N. Das

President
V.K. Gupta

Executive President
Rajender Parsad

Patrons

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

Vice Presidents

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

Secretary
D. Roy Choudhury

Foreign Secretary
Abhyuday Mandal

Treasurer
Ashish Das

Joint Secretaries

Aloke Lahiri, Shibani Roy Choudhury, Vishal Deo

Council Members

B. Re. Victor Babu, Banti Kumar, Imran Khan, Mukesh Kumar, Parmil Kumar
Piyush Kant Rai, Rajni Jain, Rakhi Singh, Raosaheb V. Latpate, Renu Kaul
Sapam Sobita Devi, Shalini Chandra, V. Srinivasa Rao, V.M. Chacko, Vishnu Vardhan R.

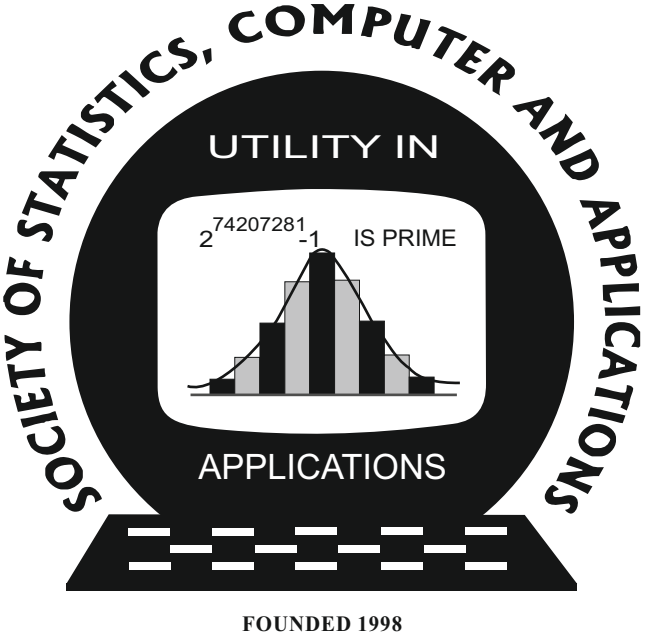
Ex-Officio Members (By Designation)

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 22, No. 2, 2024 (New Series)

Statistics and Applications

Volume 22, No. 2, 2024 (New Series)

Statistics and Applications ISSN 2454-7395 (online)
Volume 22, No. 2, 2024 (New Series)
http://www.ssca.org.in/journal

CONTENTS

Sr. No.	Title and Authors	Page Numbers
1.	Modelling Climate, COVID-19, and Reliability Data: A New Continuous Lifetime Model under Different Methods of Estimation *Arvind Pandey, Ravindra Pratap Singh, Shikhar Tyagi and Abhishek Tyagi*	1–27
2	Simple Linear Slope Estimators Based On Sample Quasi Ranges *Sharada V. Bhat and Shrinath M. Bijjargi*	29–45
3	Comparison of Cause Specific Rate Functions of Panel Count Data with Multiple Modes of Recurrence *Sankaran P. G., Ashlin Mathew P. M. and Sreedevi E. P.*	47–61
4	Price Forecasting of TOP (Tomato, Onion and Potato) Commodities using Hidden Markov-based Deep Learning Approach *G. Avinash, Ramasubramanian V., Ranjit Kumar Paul, Mrinmoy Ray, Shashi Dahiya, Mir Asif Iquebal, Samarth Godara and B. Manjunatha*	63–90
5	A New Lifetime Distribution: Statistical Inference and it's Applications *Arvind Pandey, Pawan Kumar Singh and Mahendra Saha*	91–108
6	LDPC Codes Based on New Combinatorial Designs *Shyam Saurabh and Kishore Sinha*	109–119
7	Calibration Approach for Estimating Mean of a Stratified Population in the Presence of Non-response *Manoj K. Chaudhary, Anil Prajapati and Basant K. Ray*	121–132
8	Reliability Assessment of Two-Component Series System Shock Model *S. B. Munoli and Abhijeet Jadhav*	133–143
9	A Modified Measurement Error Model for Replicated Method Comparison Data with Skewness and Heavy Tails *Jeevana Duwarahan and Lakshika S. Nawarathna*	145–167
10	Study of Priority Based Network Nodes Using Quasi Birth and Death Process *Malla Reddy Perati and Abhilash Vollala*	169–180
11	Variations of Wholesale Price of Wheat in Different States of India under COVID-19 Pandemic *Rashmi, H. P. Singh and P. K. Singh*	181–187

12	Joint Importance Measures for Repairable Multistate Systems *V. M. Chacko, Ann Sania and Amrutha M.*	189–202
13	A Comprehensive Study of the Power Modified Lindley-Geometric Distribution in the T-X Family: COVID-19 Applications *Meenu Jose and Lishamol Tomy*	203–216
14	Wavelet-ARIMA-TDNN Model for Agricultural Commodity Price Forecasting *Sathees Kumar K., Banjul Bhattacharyya, Gowthaman T. and Elakkiya N.*	217–229
15	Bayesian Analysis of Exponentiated Exponential Power Distribution under Hamiltonian Monte Carlo Method *Laxmi Prasad Sapkota and Vijay Kumar*	231–258
16	Linear Trend-Free Group Divisible Design *Longjam Roshini Chanu and K. K. Singh Meitei*	259–265
17	A Systematic Literature Review of Sustainable Probabilistic Inventory Models *Khimya Tinani and Anuja Sarangale*	267–283
18	Extropy Properties of Ranked Set Sample for Sarmanov Family of Distributions *Manoj Chacko and Varghese George*	285–305
19	Construction of Nearly Orthogonal Arrays Mappable to Tight Orthogonal Arrays of Strength Two Using Projective Geometry *Poonam Singh, Mukta D. Mazumder and Santosh Babu*	307–322
20	Power Generalized DUS Transformation of Inverse Kumaraswamy Distribution and Stress-Strength Analysis *Amrutha M. and V. M. Chacko*	323–359
	SHORTER COMMUNICATIONS	
21	Resolvability of a BIB Design of Takeuchi (1962) *Shyam Saurabh*	361–362
22	Improving Data Validation *A. K. Nigam*	363–368

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 22, No. 2, 2024 (New Series), pp 1–27
http://www.ssca.org.in/journal

Modelling Climate, COVID-19, and Reliability Data: A New Continuous Lifetime Model under Different Methods of Estimation

Arvind Pandey[1], **Ravindra Pratap Singh**[1], **Shikhar Tyagi**[2] **and Abhishek Tyagi**[3]
[1]*Department of Statistics, Central University of Rajasthan, Rajasthan-305817, India*
[2]*Department of Statistics and Data Science, Christ deemed to be University, Bangalore-560073, India*
[3]*Department of Statistics, Chaudhary Charan Singh University, Meerut-250004, India*

Received: 31 July 2023; Revised: 30 September 2023; Accepted: 15 October 2023

Abstract

In this article, a new continuous probability distribution called Arvind distribution is developed and studied. The proposed distribution has only one parameter but it exhibits a wide variety of shapes for density and hazard rate functions. A number of important distributional properties including mode, quantile function, moments, skewness, kurtosis, mean deviation, probability-weighted moments, stress-strength reliability, order statistics, reliability and hazard rate functions, Bonferroni Lorenz and Zenga curves, conditional moments, mean residual and mean past life functions, and stochastic ordering of the Arvind distribution are derived. For point estimation of the parameter of the proposed distribution, six estimation procedures including maximum likelihood, maximum product spacings, least squares, weighted least squares, Cramér-von Mises, and Anderson-Darling estimators are used. The interval estimation of the unknown parameter has also been discussed using observed Fisher's information. A vast simulation study has been conducted to examine the behaviour of different estimation procedures. Finally, the applicability of the proposed model is demonstrated by using three real-life datasets. The results of the real data analysis clearly announce that the Arvind distribution can be a better alternative to several existing models for modelling different types of data from various fields.

Key words: Arvind distribution; Maximum likelihood estimation; Maximum product spacings; Least squares estimation; Stress strength reliability

AMS Subject Classifications: 62K05, 05B05

1. Introduction

In today's competitive world, the data generated in numerous disciplines such as engineering, economics, biological sciences, actuarial sciences, *etc.* is becoming more difficult to analyze. As a consequence, for modelling such data, we require distributions that are

Corresponding Author: Ravindra Pratap Singh
Email: stats.rpsingh@gmail.com

best suited for analyzing these multi-features and complicated data. For these reasons, the invention of novel probability distributions has dominated statistical research during the last few decades. In this order, the well-known reference was Mudholkar *et al.* (1996), who described a particular generalization of the Weibull distribution and applied it to survival data. Gupta and Kundu (1999) introduced generalized exponential distribution to provide more flexibility over baseline exponential distribution. This model has decreasing and unimodal shapes of the density function and its hazard rate can take increasing, decreasing, and U-shapes. Nadarajah and Kotz (2006) proposed beta exponential distribution with decreasing and unimodal density whereas the hazard rate can exhibit decreasing and increasing shapes. Nadarajah and Haghighi (2011) developed the Nadarajah-Haghighi distribution to model increasing, decreasing, and constant hazard rate functions. Chaubey and Zhang (2015) pioneered exponentiated Chen distribution with bathtub rate hazard function. Yadav *et al.* (2021a) proposed Burr-Hatke exponential distribution to model decreasing density as well as decreasing hazard rate function. Bakouch *et al.* (2021) proposed a unit half-normal distribution with unimodal and asymmetric (left and right skewed) density and increasing hazard rate function. El-Morshedy *et al.* (2021) proposed a new generalization of the odd Weibull-G family by consolidating two notable families of distributions. Choudhary *et al.* (2021) enhanced the modified Weibull distribution with an additional parameter to provide its density and hazard rate function greater flexibility. This distribution is capable of modeling the bathtub-shaped, decreasing, increasing and the constant hazard rate function. Recently, Alsuhabi *et al.* (2022) pioneered a four-parameter distribution named the extended odd Weibull Lomax distribution. This model has increasing, and decreasing, bell shapes and unimodal shapes of the density function and its hazard rate can take increasing and decreasing shapes. They also show the applicability of this model to COVID-19 data. Promiscuous crucial literature includes Tyagi *et al.* (2022) and Agiwal *et al.* (2023).

Traditional continuous models and their modified or generalized counterparts (in the existing literature) sometimes become very restricted, for example, some models have a complex form of density and hazard functions that are difficult to handle, and a few models are limited to the model-specific type of failure rate, a non-existence of moments, a high number of parameters, an excessive amount of complexity in calculations of some characteristics, *etc.* Although some continuous distributions are less restrictive, there is still room for the construction of more flexible continuous models that may be suitable for the analysis of different types of data generated from distinct fields. With this motivation, we developed a more flexible and simpler continuous distribution called Arvind distribution. This distribution is extremely flexible compared to conventional and recently developed continuous models and we have noticed this in real data applications. Another advantage of Arvind distribution over other rival models is that it has just one parameter and therefore the expressions of this distribution are not too complicated in both analytical and computational handling.

The rest of the structure of this article is as follows. Section 2 introduced Arvind distribution and portrayed various shapes of its density. In Section 3, we have derived various imperative distributional and reliability properties of Arvind distribution with some numerical illustrations. In Section 4, different methods of estimation like maximum likelihood, maximum product spacings, ordinary and weighted least squares, Cramér-von-Mises, and Anderson-Darling have been used to estimate the unknown parameters of the proposed model. Section 4 also includes the asymptotic confidence interval (ACI) for the unknown

parameter based on Fisher's information. A detailed simulation study is presented to inspect the performance of various estimation methods in Section 5. In Section 6, the applicability of Arvind distribution has been demonstrated using three case studies from different fields over other well-known continuous models. In the end, some concluding remarks are provided in Section 7.

2. Synthesis of the Arvind distribution

Before discussing the density function and form of the proposed model, we mention the following proposition.

Proposition 1: For a random variable X with domain $(0, \infty)$, the following function is a valid cumulative distribution function (CDF).

$$F(x,\theta) = P\left(X \leq x\right) = \begin{cases} 1 - \frac{\exp(-\theta x^2)}{(1+\theta x)}; & x > 0, \theta > 0 \\ 0 \qquad\qquad\quad ; & otherwise \end{cases}, \tag{1}$$

where $\theta \in (0, \infty)$ is an unknown parameter.

Proof: Since, $x > 0$ and $\theta > 0$, therefore, we can see that $F(x,\theta) \leq 1$ and $F(x,\theta) \geq 0$. Furthermore $\lim_{x\to 0} F(x,\theta) = 0 = F(0,\theta)$ implying that $F(x,\theta)$ is continuous at 0 and a fortiori in $\mathbb{R}$. It is clear that $\lim_{x\to+\infty} F(x,\theta) = 1$. Now, for $x > 0$, we have

$$F'(x,\theta) = \frac{d}{dx}F(x,\theta) = \frac{\theta\left(1 + 2x + 2\theta x^2\right)}{\left(1+\theta x\right)^2}\ \exp(-\theta x^2)\ \geq 0,$$

implying that $F(x,\theta)$ is non-decreasing. The required properties for a valid CDF are satisfied, therefore $F(x,\theta)$ is a valid CDF. □

Based on Proposition 1, we can easily define the Arvind distribution as follows:
Definition 1: A continuous random variable X is said to follow Arvind distribution with parameter θ if its CDF is of the form (1) or it can be specified by the following probability density function (PDF)

$$f(x,\theta) = \begin{cases} \frac{\theta(1+2x+2\theta x^2)}{(1+\theta x)^2}\exp(-\theta x^2); & x > 0, \theta > 0 \\ 0 \qquad\qquad\qquad\qquad\quad ; & otherwise \end{cases}, \tag{2}$$

here, it is clear that $f(x,\theta) \geq 0$ and $\int_0^\infty f(x,\theta)dx = 1$.
Some of the possible shapes of the PDF of the Arvind distribution for a few arbitrary values of the parameter θ are portrayed in Figure 1. From this figure, we can easily see that the PDF of the Arvind distribution is versatile enough as it takes a variety of shapes for different values of θ. Also, the limiting behaviour of the PDF of Arvind distribution can be defined as

$$\lim_{x\to 0} f(x,\theta) = \theta \quad \text{and} \quad \lim_{x\to\infty} f(x,\theta) = 0.$$

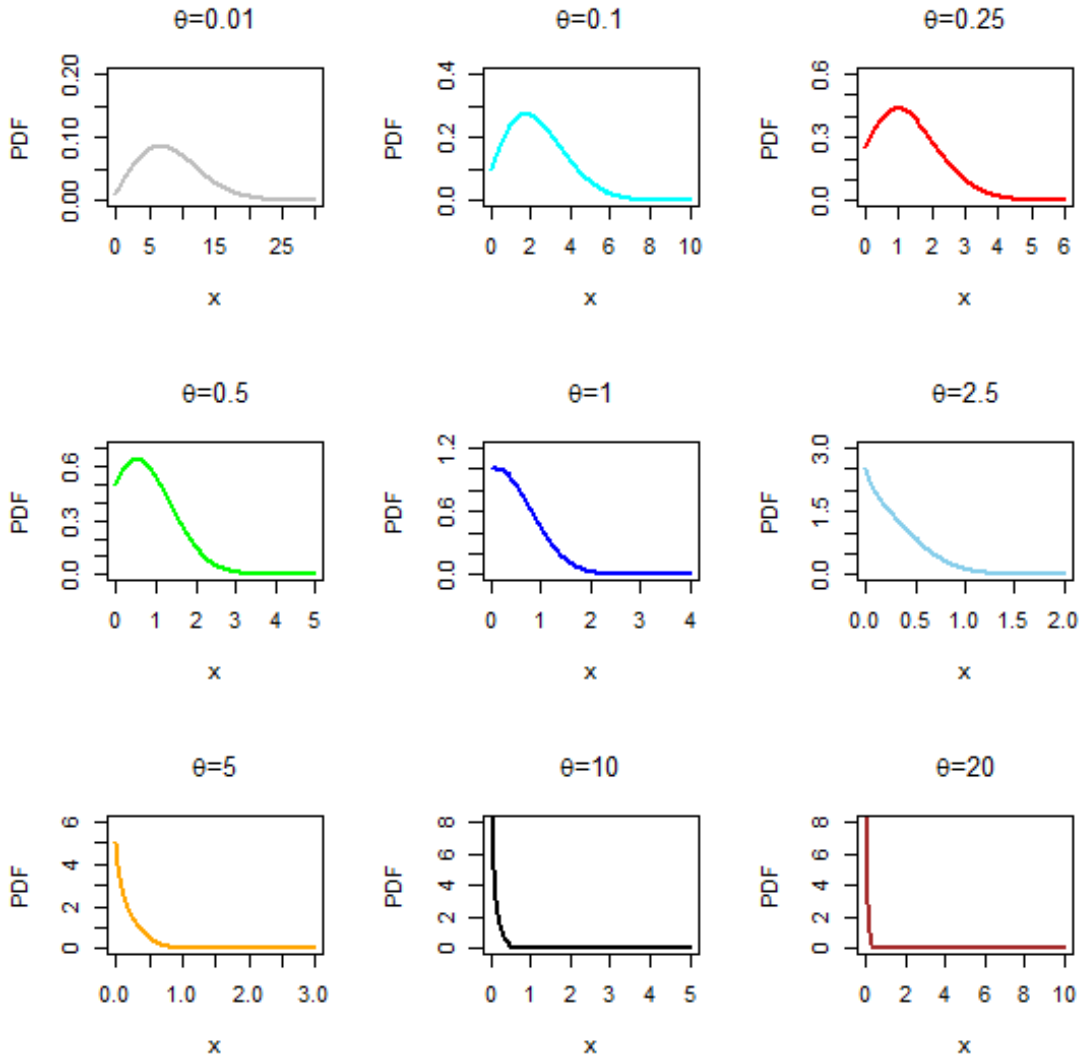


Figure 1: The PDF plots for Arvind distribution for different values of θ

3. Statistical properties of the Arvind distribution

The development of a probability distribution without discussing its statistical properties is not of much use. The Arvind distribution has many important distributional properties and some of them are presented below:

3.1. Mode

A value of a random variable that maximizes its PDF is known as a mode. In the case of Arvind distribution, the mode can be obtained by solving the following equation

$$\frac{\partial \log f(x,\theta)}{\partial x} = 0 \Rightarrow 2\theta^3x^4 + 4\theta^2x^3 + (2\theta + \theta^2)x^2 + \theta - 1 = 0.$$

The above equation cannot be solved analytically in closed form. Therefore, we have obtained the values of mode numerically for different values of θ, and these are listed in Table 1. From these tabulated values of mode, we have verified the conclusion of the PDF plot that the

proposed distribution is unimodal.

3.2. Quantile function, random number generation and median

The quantile function is an important tool to specify a probability distribution. It is very useful in random number generation and computation of positional averages like median. The quantile (Q) of the Arvind distribution can be obtained by solving the following equation

$$\exp(-\theta Q^2) - (1+\theta Q)(1-u) = 0, \tag{3}$$

where u is the uniform random variable from $U(0,1)$. The median of the proposed distribution can be computed by putting $u = 0.5$ in Equation (3). We have numerically obtained the median for different values of θ and these are listed in Table 1, and it concludes that the median goes down as θ gets up. By solving Equation (3) for Q, we can generate random numbers from the proposed distribution for different values of u from $U(0,1)$.

3.3. Moments, skewness and kurtosis

For portraying different characteristics of a probability distribution like mean, variance, skewness (Sk), and kurtosis (Kur), moments are very useful in statistical theory. Suppose X is a random variable that follows Arvind distribution with parameter θ. Then, the r^{th} raw moment can be derived as

$$\begin{aligned}
\mu'_r = E(X^r) &= \int_0^\infty x^r f(x,\theta)dx \\
&= \int_0^\infty x^r \frac{\theta(1+2x+2\theta x^2)}{(1+\theta x)^2}\exp(-\theta x^2)dx \\
&= r\int_0^\infty \frac{x^{r-1}\exp(-\theta x^2)}{(1+\theta x)}dx.
\end{aligned} \tag{4}$$

In particular, the mean and variance of the proposed model can be presented as

$$\mu = \mu'_1 = E(X) = \int_0^\infty \frac{\exp(-\theta x^2)}{(1+\theta x)}dx$$

and

$$\begin{aligned}
\mu_2 = Var(X) &= \mu'_2 - {\mu'_1}^2 \\
&= 2\textstyle\int_0^\infty \frac{x\exp(-\theta x^2)}{(1+\theta x)}dx - \left[\int_0^\infty \frac{\exp(-\theta x^2)}{(1+\theta x)}dx\right]^2,
\end{aligned}$$

respectively. Similarly, we can obtain other central moments using raw moments. From these raw moments, we can also calculate the Sk and Kur of the proposed model using the following formula:

$$Sk = \frac{E(X^4) - 3E(X^2)E(X) + 2(E(X))^3}{(Var(X))^{3/2}},$$

and

$$Kur = \frac{E(X^4) - 4E(X^2)E(X) + 6E(X^2)(E(X))^2 - 3(E(X))^4}{(Var(X))^2},$$

respectively. As we can easily observe the mean, variance, Sk, and Kur of the Arvind distribution cannot be found in closed expressions, therefore we compute them numerically for different values of θ, and these are listed in Table 1. From this table, we yield the following outcomes:

- The mean and variance of the Arvind distribution decrease as the value of θ increases.
- As the value of the coefficient of skewness based on moments is positive, the proposed model is positively skewed. Also, the Sk of the Arvind model increases as θ increases.
- From Table 1, since the value of the coefficient of kurtosis is less than 3, therefore, the proposed distribution is platykurtic and its peakedness increases as θ increases.

Table 1: Descriptive statistics for Arvind distribution for different values of θ

θ	Mode	Median	Mean	Variance	MD(μ)	MD(m)	Skewness	Kurtosis
0.1	1.77056	2.21965	2.40702	2.11585	1.17113	0.97464	0.69274	0.29880
0.5	0.52137	0.83134	0.93373	0.40649	0.51371	0.40521	0.81903	0.47350
1	0.00003	0.52237	0.60513	0.19600	0.35623	0.26777	0.91400	0.64761
1.5	0.11719	0.39208	0.46573	0.12693	0.28611	0.20687	0.98351	0.79551
2	0.26459	0.31746	0.38531	0.09288	0.24423	0.17082	1.03954	0.92660
2.5	0.42256	0.26827	0.33191	0.07271	0.21563	0.14647	1.08699	1.04560
3	0.10358	0.23306	0.29340	0.05943	0.19453	0.12869	1.12840	1.15534
4	0.24065	0.18560	0.24089	0.04308	0.16495	0.10420	1.19865	1.35382
5	0.71926	0.15482	0.20627	0.03347	0.14484	0.08796	1.25747	1.53157
10	0.07255	0.08580	0.12588	0.01503	0.09546	0.05024	1.46474	2.24050
20	0.65754	0.04588	0.07548	0.00657	0.06161	0.02741	1.71150	3.24746

3.4. Mean deviation

If we take the average absolute deviation about the mean (or median) it is known as the mean deviation about the mean (or median). Mean deviation about the mean (or median) is another important tool for measuring dispersion besides the variance. Suppose μ and m denote the mean and median, then the mean deviation about the mean (or median) can be defined as

$$MD(\zeta) = E\left|X-\zeta\right| = \int_0^{\infty} \left|x-\zeta\right| f(x,\theta)dx = 2\left\{\zeta F(\zeta,\theta) - \int_0^{\zeta} xf(x,\theta)dx\right\}, \qquad (5)$$

where $\zeta = \mu$ or m . Using the above expression with some simplification, the mean deviation about mean (or median) for the Arvind distribution can be obtained as

$$MD(\zeta) = E\left|X-\zeta\right| = 2\left\{\zeta - \int_0^{\zeta} \frac{\exp(-\theta x^2)}{(1+\theta x)}dx\right\}. \qquad (6)$$

The expression of mean deviation (6) cannot be bound up in closure form, so to measure the behaviour of mean deviation about mean (or median), we have calculated these average deviations numerically and they are listed in Table 1. This table announces that the mean deviation about the mean (or median) decreases as θ increases and the mean deviation about the median is smaller than the mean deviation about the mean as the theory claims.

3.5. Probability weighted moments

The generalization of the simple moments is known as probability-weighted moments (PWMs). They can be developed for a distribution whose ordinary moments can be derived. For the Arvind random variable X, the $(r,s)^{th}$ PWM is given by

$$\begin{aligned}\varsigma_{r,s} &= E\left[X^r F^s(x,\theta)\right] \\ &= \int_0^\infty x^r F^s(x,\theta) f(x,\theta) dx \\ &= \int_0^\infty x^r \left(1 - \frac{\exp(-\theta x^2)}{1+\theta x}\right)^s \frac{\theta(1+2x+2\theta x^2)\exp(-\theta x^2)}{(1+\theta x)^2} dx \\ &= \sum_{j=0}^{s} (-1)^j \theta \binom{s}{j} \int_0^\infty x^r \left(\frac{1}{(1+\theta x)^{j+2}} + \frac{2x}{(1+\theta x)^{j+1}}\right) \exp(-(j+1)\theta x^2) dx.\end{aligned}$$

After some simplification, the $(r,s)^{th}$ PWM of the Arvind distribution is given by

$$\varsigma_{r,s} = \sum_{j=0}^{s} (-1)^j \frac{r}{(j+1)} \binom{s}{j} \int_0^\infty \frac{x^{r-1}\exp(-(j+1)\theta x^2)}{(1+\theta x)^{j+1}} dx. \tag{7}$$

3.6. Stress-strength reliability

The probability $\varpi = P(X_2 < X_1)$ is referred to as stress-strength (S-S) reliability if the random variable X_1 represents the strength of a system under stress X_2, assuming that X_1 and X_2 are stochastically independent random variables. The S-S reliability is widely used in reliability theory, especially in engineering concepts like different structures, static fatigue of ceramic components, the aging of concrete pressure vessels, fatigue failure of aviation structures, *etc.* The research on S-S reliability models has received a lot of attention recently due to the expanded scope of S-S reliability. For more detail, see Goel and Singh (2020). In our case, suppose $X_1 \sim \text{Arvind}(\theta_1)$ and $X_2 \sim \text{Arvind}(\theta_2)$ distributions, then S-S reliability is given by

$$\begin{aligned}\varpi = P(X_2 < X_1) &= \int_0^\infty P(X_2 < X_1 | X_1 = x) f_{X_1}(x,\theta_1) dx \\ &= \int_0^\infty F_{X_2}(x,\theta_2) f_{X_1}(x,\theta_1) dx \\ &= \int_0^\infty \left(1 - \frac{\exp(-\theta_2 x^2)}{(1+\theta_2 x)}\right) \left(\frac{\theta_1(1+2x+2\theta_1 x^2)\exp(-\theta_1 x^2)}{(1+\theta_1 x)^2}\right) dx \\ &= 1 - \theta_1 \int_0^\infty \frac{(1+2x+2\theta_1 x^2)\exp(-(\theta_1+\theta_2)x^2)}{(1+\theta_2 x)(1+\theta_1 x)^2} dx. \end{aligned} \tag{8}$$

The expression (8) ϖ cannot be easily tractable in closed form. Therefore, to study the behaviour of ϖ for different values of θ_1 and θ_2, we have computed ϖ numerically. The outcomes of ϖ have been given in Table 2. From this table, we observe that

- For a fixed value of θ_1, the value of ϖ increases as θ_2 increases.

- For a fixed value of θ_2, the value of ϖ goes down as θ_1 increases.

- When the values of θ_1 and θ_2 are equal, the value of ϖ becomes 0.5.

Table 2: The S-S reliability ϖ under different values of θ_1 and θ_2

$\theta_1 \rightarrow$ $\theta_2 \downarrow$	0.1	0.5	1	1.5	2	2.5	3	4	5	10	20
0.1	0.5	0.17715	0.10231	0.0734	0.05786	0.04808	0.04133	0.03255	0.02705	0.01526	0.00862
0.5	0.82285	0.5	0.34692	0.27031	0.22365	0.19196	0.1689	0.13737	0.11661	0.06916	0.04041
1	0.89769	0.65308	0.5	0.41108	0.35211	0.30975	0.27765	0.23187	0.2005	0.12452	0.07519
1.5	0.9266	0.72969	0.58892	0.5	0.43784	0.39151	0.35543	0.30244	0.26504	0.17049	0.1057
2	0.94214	0.77635	0.64789	0.56216	0.5	0.45243	0.41462	0.35785	0.31686	0.20963	0.13287
2.5	0.95192	0.80804	0.69025	0.60849	0.54757	0.5	0.46159	0.40289	0.35974	0.24359	0.15734
3	0.95867	0.8311	0.72235	0.64457	0.58538	0.53841	0.5	0.44045	0.39601	0.27348	0.17957
4	0.96745	0.86264	0.76813	0.69756	0.64215	0.59711	0.55955	0.5	0.45448	0.32402	0.21868
5	0.97295	0.88339	0.7995	0.73496	0.68314	0.64026	0.60399	0.54552	0.5	0.36549	0.25223
10	0.98474	0.93084	0.87548	0.82951	0.79037	0.75641	0.72652	0.67598	0.63451	0.5	0.37072
20	0.99138	0.95959	0.92481	0.8943	0.86713	0.84266	0.82043	0.78132	0.74777	0.62928	0.5

We have also portrayed a 3-D plot for ϖ under different values of θ_1 and θ_2 in Figure 2, this plot also announces that ϖ can take a variety of values from small to large for distinct values of θ_1 and θ_2.

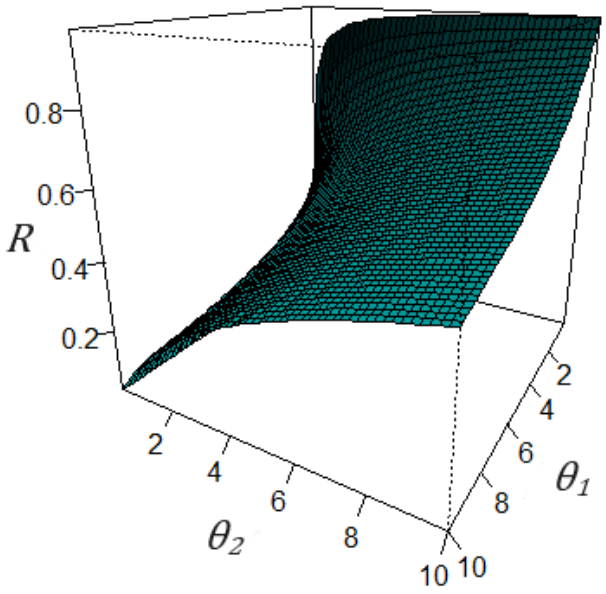


Figure 2: A 3-D plot for ϖ under different values of θ_1 and θ_2

3.7. Order statistics

Let $X_1, X_2, ..., X_n$ be a random sample of size n generated from Arvind(θ) distribution and $X_{1:n} \leq X_{2:n} \leq ..... \leq X_{n:n}$ denotes the corresponding order statistics. Then, the

PDF and CDF, respectively, of i^{th} order statistics are given as

$$\begin{aligned} f_{i:n}(x,\theta) &= \frac{n!}{(i-1)!\,(n-i)!}[F(x,\theta)]^{i-1}[1-F(x,\theta)]^{n-i}f(x,\theta) \\ &= \frac{n!\theta(1+2x+2\theta x^2)\exp(-\theta(n-i+1)x^2)(1+\theta x-\exp(-\theta x^2))^{i-1}}{(i-1)!\,(n-i)!(1+\theta x)^{n+1}}, \end{aligned} \tag{9}$$

and

$$\begin{aligned} F_{i:n}(x,\theta) &= \sum_{r=i}^{n}\binom{n}{r}[F(x,\theta)]^{r}[1-F(x,\theta)]^{n-r} \\ &= \sum_{r=i}^{n}\sum_{j=0}^{n-r}(-1)^{j}\binom{n}{r}\binom{n-r}{j}[F(x,\theta)]^{j+r} \\ &= \sum_{r=i}^{n}\sum_{j=0}^{n-r}(-1)^{j}\binom{n}{r}\binom{n-r}{j}\left(1-\frac{\exp(-\theta x^2)}{1+\theta x}\right)^{j+r}. \end{aligned} \tag{10}$$

In particular, by putting $i = 1$ and $i = n$, respectively, we can find the PDF and CDF of minimum and maximum order statistics. For odd sample size n, we can obtain the PDF and CDF of the median order statistics by setting $i = \frac{n+1}{2}$.

3.8. Reliability and hazard rate functions

The reliability function (RF) $R(x,\theta)$ and hazard rate function (HRF) $h(x,\theta)$ of the Arvind(θ) distribution, respectively, are given by

$$R(x,\theta) = P\,(X > x) = \frac{\exp(-\theta x^2)}{(1+\theta x)}; \qquad x \geq 0, \theta > 0, \tag{11}$$

$$h(x,\theta) = \frac{f(x,\theta)}{R(x,\theta)} = \frac{\theta\,(1+2x+2\theta x^2)}{(1+\theta x)}; \qquad x \geq 0, \theta > 0. \tag{12}$$

We have plotted the hazard rate for different values of θ in Figure 3. From this figure, we can easily observe that the HRF of the Arvind distribution can increase, decrease, and U-shaped. Also, the limiting behaviour of the HRF can be stated as:

$$\lim_{x\to 0} h(x,\theta) = \theta \quad \text{and} \quad \lim_{x\to\infty} h(x,\theta) = \infty.$$

Also, the cumulative and reverse hazard rate (RHR) functions of the Arvind distribution, respectively, are given by

$$H(x,\theta) = \theta x^2 + \log(1+\theta x); \qquad x \geq 0, \theta > 0, \tag{13}$$

$$RHR(x,\theta) = \frac{\theta(1+2x+2\theta x^2)\exp(-\theta x^2)}{(1+\theta x)(1+\theta x-e^{-\theta x^2})}; \quad x \geq 0, \theta > 0. \tag{14}$$

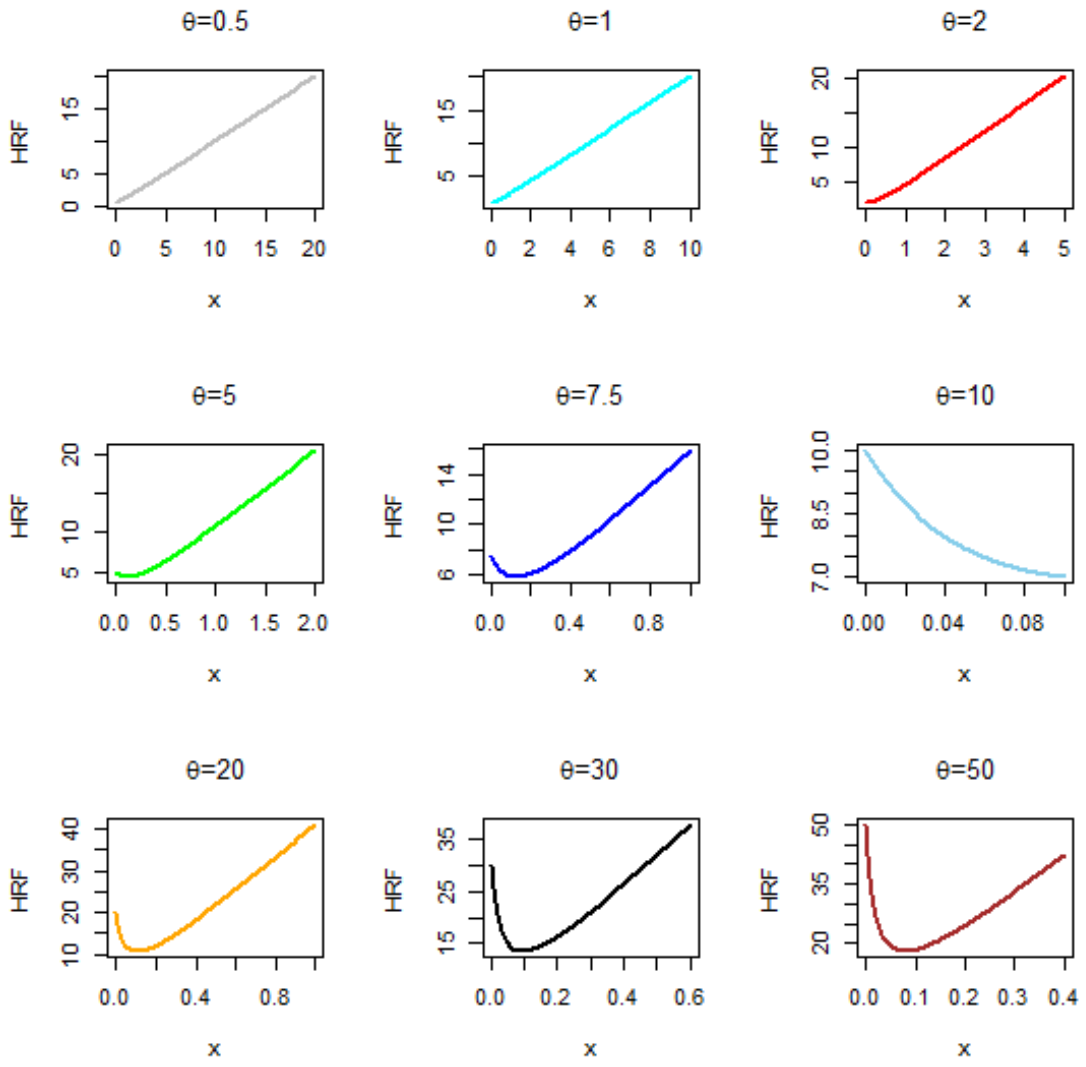


Figure 3: The HRF plot of Arvind distribution for different values of θ

3.9. Inequality measures: Lorenz, Bonferroni and Zenga curves

The Lorenz, Bonferroni, and Zenga curves are the most often used inequality measures in the literature (Lorenz (1905), Bonferroni (1930), and Zenga (2007)). These three curves can be defined using simply the population mean and the means of certain subgroups. Inequality curves are useful because they may be used to create a variety of orderings that allow for distribution comparisons based on inequality. Such comparisons within the same model make it possible to comprehend how distributional parameters influence inequality. For the Arvind distribution, the Lorenz, Bonferroni, and Zenga curves, respectively, are obtained as

$$L(p) = \frac{1}{\mu}\int_0^q xf(x)dx = \frac{1}{\mu}\left[-\frac{q\exp(-\theta q^2)}{1+\theta q} + \int_0^q \frac{\exp(-\theta x^2)}{(1+\theta x)}dx\right], \tag{15}$$

$$B(p) = \frac{L(p)}{p}, \tag{16}$$

$$Z(p) = \frac{p - L(p)}{p(1 - L(p))}, \tag{17}$$

where $q = F^{-1}(p)$ which can be computed numerically using Equation (3). Table 3 lists numerical values for Lorenz, Bonferroni, and Zenga curves of Arvind distribution for different values of $q = F^{-1}(p)$ and θ.

Table 3: Values for Lorenz, Bonferroni, and Zenga curves of Arvind distribution for a variety of θ

$\theta \rightarrow$	0.5			1			2			5		
$p \downarrow$	$L(p)$	$B(p)$	$Z(p)$	$L(p)$	$B(p)$	$Z(p)$	$L(p)$	$B(p)$	$Z(p)$	$L(p)$	$B(p)$	$Z(p)$
0.05	0.00260	0.05191	0.95055	0.00207	0.04132	0.96066	0.00165	0.03298	0.96862	0.00124	0.02487	0.97635
0.1	0.01013	0.10126	0.90793	0.00827	0.08267	0.92497	0.00670	0.06704	0.93926	0.00512	0.05118	0.95370
0.15	0.02229	0.14863	0.87079	0.01860	0.12398	0.89262	0.01533	0.10219	0.91179	0.01184	0.07893	0.93211
0.2	0.03891	0.19453	0.83808	0.03310	0.16548	0.86309	0.02769	0.13843	0.88610	0.02165	0.10824	0.91149
0.25	0.05983	0.23933	0.80908	0.05178	0.20710	0.83619	0.04395	0.17579	0.86209	0.03481	0.13923	0.89181
0.3	0.08500	0.28334	0.78323	0.07469	0.24895	0.81167	0.06429	0.21430	0.83968	0.05167	0.17225	0.87286
0.35	0.11439	0.32683	0.76012	0.10190	0.29113	0.78929	0.08885	0.25385	0.81891	0.07243	0.20694	0.85499
0.4	0.14801	0.37002	0.73942	0.13349	0.33374	0.76891	0.11791	0.29478	0.79949	0.09748	0.24371	0.83798
0.45	0.18591	0.41314	0.72088	0.16960	0.37689	0.75037	0.15163	0.33697	0.78154	0.12723	0.28272	0.82183
0.5	0.22820	0.45640	0.70433	0.21037	0.42074	0.73358	0.19028	0.38055	0.76501	0.16206	0.32412	0.80660
0.55	0.27501	0.50002	0.68964	0.25600	0.46545	0.71848	0.23413	0.42569	0.74988	0.20248	0.36814	0.79228
0.6	0.32652	0.54419	0.67679	0.30674	0.51124	0.70502	0.28353	0.47255	0.73618	0.24892	0.41487	0.77906
0.65	0.38306	0.58932	0.66567	0.36293	0.55836	0.69324	0.33891	0.52140	0.72396	0.30212	0.46479	0.76690
0.7	0.44497	0.63567	0.65641	0.42500	0.60714	0.68323	0.40079	0.57256	0.71334	0.36278	0.51826	0.75601
0.75	0.51277	0.68369	0.64920	0.49353	0.65804	0.67518	0.46988	0.62650	0.70455	0.43187	0.57582	0.74662
0.8	0.58718	0.73397	0.64441	0.56936	0.71170	0.66947	0.54713	0.68392	0.69796	0.51064	0.63830	0.73913
0.85	0.66931	0.78742	0.64283	0.65371	0.76907	0.66687	0.63398	0.74586	0.69433	0.60094	0.70699	0.73426
0.9	0.76101	0.84556	0.64620	0.74860	0.83178	0.66914	0.73278	0.81420	0.69531	0.70570	0.78411	0.73357
0.95	0.86601	0.91159	0.65984	0.85826	0.90343	0.68131	0.84824	0.89289	0.70581	0.83075	0.87447	0.74167

3.10. Conditional moments

In the context of lifetime models, it is also useful to have a knowledge of the expression $E(X^r|X > x)$. This expression is called the r^{th} conditional moment of the random variable 'X'. The computation of mean deviations around the mean and the median, as well as the mean residual life function (See, Section 3.11) are all areas in which the conditional moments find widespread usage. The r^{th} conditional moment of a random variable following Arvind(θ) distribution can be obtained as

$$E(X^r|X > x) = \frac{1}{1 - F(x, \theta)} \Lambda_r(x, \theta), \tag{18}$$

where

$$\Lambda_r(x, \theta) = \int_x^\infty v^r f(v, \theta) dv = \frac{x^r \exp(-\theta x^2)}{(1 + \theta x)} + r \int_x^\infty \frac{v^{r-1} \exp(-\theta v^2)}{1 + \theta v} dv.$$

3.11. Mean residual life

The expected value of the remaining lifetimes after a fixed time point x, is called the mean residual life (MRL) function. Since it is representative of the aging mechanism, the MRL function is put to considerable use in a broad range of fields, including reliability

engineering, survival analysis, and biological research. For Arvind(θ) distribution, it can be derived as

$$MRL(x,\theta) = E(X-x|X>x) = \frac{1}{1-F(x,\theta)}\int_x^\infty vf(v,\theta)dv - x$$
$$= \frac{1+\theta x}{\exp(-\theta x^2)}\int_x^\infty \frac{\exp(-\theta v^2)}{1+\theta v}dv. \tag{19}$$

From the above expression of MRL, we can easily observe that the MRL is an application of conditional moments and it can be obtained by putting $r = 1$ in Equation (18).

3.12. Mean past life

The expected time elapsed from the failure of a system given that its lifetime is less than or equal to a time point $x(x \geq 0)$ is referred to as the mean past life (MPL) function. Similar to the MRL function, the MPL function has applications in a vast array of fields, such as actuarial research, forensic science, reliability theory, and survival analysis. The expression of the MPL function for Arvind(θ) distribution can be developed as

$$MPL(x,\theta) = E(x-X|X\leq x) = x - \frac{1}{F(x,\theta)}\int_0^x vf(v,\theta)dv$$
$$= \frac{1}{F(x,\theta)}\left[x - \int_0^x \frac{\exp(-\theta v^2)}{1+\theta v}dv\right]. \tag{20}$$

3.13. Stochastic ordering

It is crucial to compare two or more random variables indicating the state of things in two or more circumstances. In the situation of two random variables that are independent, stochastic orderings are extremely advantageous. For two independent random variables Y and Z if $F_Y(y) \geq F_Z(y)$ for all y, Y is said to be stochastically smaller than Z *i.e.* $Y \leq_{st} Z$. Similarly, we can define stochastic ordering in terms of hazard rate, mean residual life, and likelihood ratio functions as

- hazard rate order ($Y \leq_{hr} Z$) if $h_Y(y) \geq h_Z(y)$ for all y.
- mean residual life order ($Y \leq_{mrl} Z$) if $MRL_Y(y) \leq MRL_Z(y)$ for all y.
- likelihood ratio order ($Y \leq_{lr} Z$) if $f_Y(y)/f_Z(y)$ decreases in y.

The following implications Shaked and Shanthikumar (2007) are well-known

$$Y \leq_{lr} Z \Rightarrow Y \leq_{hr} Z \Rightarrow Y \leq_{mrl} Z$$
$$\Downarrow$$
$$Y \leq_{st} Z.$$

The Arvind distributions are ordered with respect to the strongest "likelihood ratio" ordering as we can easily observe from the following theorem.

Theorem 1: Let Y and Z be two independent random variables form Arvind(θ_1) and Arvind(θ_2) distributions, respectively. If $\theta_1 > \theta_2$ then $Y \leq_{lr} Z$ and hence $Y \leq_{hr} Z$, $Y \leq_{mrl} Z$ and $Y \leq_{st} Z$.

Proof: Firstly, we observe that

$$\frac{f_Y(y,\theta_1)}{f_Z(y,\theta_2)} = \frac{\theta_1(1+2y+2\theta_1 y^2)(1+\theta_2 y)^2 \exp(-\theta_1 y^2)}{\theta_2(1+2y+2\theta_2 y^2)(1+\theta_1 y)^2 \exp(-\theta_2 y^2)}, \quad y > 0.$$

Since, for $\theta_1 > \theta_2$,

$$\frac{d}{dx}\log\left(\frac{f_Y(y,\theta_1)}{f_Z(y,\theta_2)}\right) = 2(\theta_2-\theta_1)\left[\frac{\begin{array}{l}2+(6+\theta_1+\theta_2)y+(6(1+\theta_2)+\theta_1(6+\theta_2))y^2 \\ +2(\theta_1^2+\theta_2(5+\theta_2)+\theta_1(5+3\theta_2))y^3 \\ +2(2\theta_2^2+\theta_1^2(2+\theta_2)+\theta_1\theta_2(9+\theta_2))y^4 \\ +8\theta_1\theta_2(\theta_1+\theta_2)y^5+4\theta_1^2\theta_2^2 y^6\end{array}}{(1+2y+2\theta_1 y^2)(1+2y+2\theta_2 y^2)(1+\theta_1 y)(1+\theta_2 y)}\right] < 0,$$

i.e. $f_Y(y)/f_Z(y)$ is decreasing in y. It implies that $Y \leq_{lr} Z$. The rest of the ordering is a direct consequence of the results provided by Shaked and Shanthikumar (2007). □

4. Parameter estimation of the Arvind distribution

Under this section, the estimation of the unknown parameter of the proposed model has been discussed using six different classical approaches, namely, the method of maximum likelihood, maximum product spacings, ordinary and weighted least squares, Cramér-von-Mises, and Anderson Darling method of estimation. These methods are briefly discussed as follows,

4.1. Maximum likelihood estimation

Suppose $\underset{\sim}{X} \equiv X_1, X_2, ..., X_n$ be a random sample of size n from the Arvind distribution. Then, the log-likelihood (logL) function can be written as

$$\log L = n\log(\theta) - \theta\sum_{i=1}^{n} x_i^2 + \sum_{i=1}^{n} log\left(1+2x_i+2\theta x_i^2\right) - 2\sum_{i=1}^{n} log\left(1+\theta x_i\right). \tag{21}$$

To find out the maximum likelihood estimator (MLE) of θ, the normal equation is given by

$$\frac{\partial \log L}{\partial \theta} = \frac{n}{\theta} - \sum_{i=1}^{n}\left[x_i^2 + \frac{2}{(1+\theta x_i)} - \frac{2x_i^2}{(1+2x_i+2\theta x_i^2)}\right] = 0. \tag{22}$$

The solution of Equation (22) yields the MLE of θ. Unfortunately, the above normal equation cannot be solved analytically. Therefore, we can use numerical iteration procedures such as Newton-Raphson (NR) through the open-source programming language R.

4.2. Observed Fisher's information and asymptotic confidence interval

The observed Fisher's information for Arvind(θ) distribution is specified by

$$I_o(\hat{\theta}) = -\frac{\partial^2 \log L}{\partial \theta^2} | \theta = \hat{\theta},$$

where the second-order derivative of the log-likelihood function (21) with respect to θ is given by

$$\frac{\partial^2 \log L}{\partial \theta^2} = -\frac{n}{\theta^2} - \sum_{i=1}^{n} \left[-\frac{2x_i}{(1+\theta x_i)^2} + \frac{4x_i^4}{(1+2x_i+2\theta x_i^2)^2} \right].$$

Using this Fisher's information, the asymptotic variance of $\hat{\theta}$ can be obtained as

$$Var_o(\hat{\theta}) = \frac{1}{I_o(\hat{\theta})}.$$

Under some regularity conditions, the sampling distribution of $(\hat{\theta} - \theta)/\sqrt{Var_o(\hat{\theta})}$ can be approximated by a standard normal distribution. The large-sample $100 \times (1-\alpha)\,\%$ confidence interval (also called ACI) for θ is given by

$$[\hat{\theta}_L, \hat{\theta}_U] = \hat{\theta} \mp z_{\alpha/2} \sqrt{Var_o(\hat{\theta})}.$$

Using simulation, we can estimate the coverage probability $P\left[\left|\frac{(\theta-\hat{\theta})}{\sqrt{Var_o(\hat{\theta})}}\right| \leq z_{\alpha/2}\right]$, here z_p is such that $p = \int_{z_p}^{\infty} (1/\sqrt{2\pi}) e^{-z^2/2} dz$.

4.3. Maximum product of spacings method of estimation

As an alternative to the approach of maximum likelihood, the maximum product spacing (MPS) method was developed by Cheng and Amin (1979) for estimating the unknown parameters of continuous univariate distributions. Cheng and Amin (1983) proved that this technique is just as efficient as the maximum likelihood estimation and that it is consistent under more general conditions. Suppose $X_1, X_2, ..., X_n$ be a random sample from Arvind distribution $F(x, \theta)$, and $X_{1:n} \leq X_{2:n} \leq ... \leq X_{n:n}$ be the corresponding ordered values. Based on this random sample, let us define the uniform spacings as

$$D_i(\theta) = F(x_{i:n}, \theta) - F(x_{i-1:n}, \theta), \qquad i = 1, 2, ..., n,$$

where $F(x_{0:n}, \theta) = 0$, $F(x_{n+1:n}, \theta) = 1$, and $\sum_{i=1}^{n+1} D_i(\theta) = 1$. The MPS estimator $\hat{\theta}_{MPS}$ of the parameter θ is determined by maximizing the geometric mean of the spacings with respect to θ, or, evenly, by maximizing the following function

$$H(\theta) = \frac{1}{n+1} \sum_{i=1}^{n+1} \log(D_i(\theta)).$$

The estimator $\hat{\theta}_{MPS}$ can also be obtained by solving the following non-linear equation,

$$\frac{1}{n+1}\sum_{i=1}^{n+1}\frac{1}{D_i(\theta)}\left[\xi(x_{i:n},\theta)-\xi(x_{i-1:n},\theta)\right]=0, \tag{23}$$

where $\xi(x_{i:n},\theta)=\left(\frac{x_i e^{-\theta x_i^2}}{1+\theta x_i}\right)\left(x_i+\frac{1}{1+\theta x_i}\right)$.

4.4. Ordinary and weighted least squares estimation

Swain *et al.* (1988) firstly introduced regression-based estimators called ordinary least squares (OLS) and weighted least squares (WLS) estimators for estimating the unknown parameters of the beta distribution. These two methods are based on the combination of the non-parametric and parametric distribution functions. Suppose $X_1, X_2, ..., X_n$ be a random sample from Arvind distribution, and $X_{1:n} \leq X_{2:n} \leq ... \leq X_{n:n}$ be the corresponding ordered values. Then, the OLS estimator of θ, say $\hat{\theta}_{OLS}$ can be derived by minimizing the following function with respect to θ

$$V(\theta)=\sum_{i=1}^{n}\left[F\left(x_{i:n},\theta\right)-\frac{i}{n+1}\right]^2.$$

Alternatively, we can obtain the OLS estimator of θ by solving the following expression for θ,

$$\sum_{i=1}^{n}\left[F\left(x_{i:n},\theta\right)-\frac{i}{n+1}\right]\xi(x_{i:n},\theta)=0. \tag{24}$$

The WLS estimator of θ, say $\hat{\theta}_{WLS}$ can be found by minimizing the following equation,

$$W(\theta)=\sum_{i=1}^{n}\frac{(n+1)^2(n+2)}{i(n-i+1)}\left[F\left(x_i,\theta\right)-\frac{i}{n+1}\right]^2.$$

The WLS estimator $\hat{\theta}_{WLS}$ can also be obtained by solving the following non-linear equation with respect to θ,

$$\sum_{i=1}^{n}\frac{(n+1)^2(n+2)}{i(n-i+1)}\left[F\left(x_{i:n},\theta\right)-\frac{i}{n+1}\right]\xi(x_{i:n},\theta)=0, \tag{25}$$

where $\xi(x_{i:n},\theta)$ is defined in Section 4.3.

4.5. Cramér-von-Mises estimation

Cramer-von-Mises type minimum distance estimator is a widely used minimum distance estimator since the empirical data suggests that the bias of this estimator is lower than that of the other minimum distance estimators. In our case, the Cramér-von Mises (CVM) minimum distance estimator of θ can be obtained by minimizing, the following function:

$$C(\theta)=\frac{1}{12n}\left[F\left(x_i,\theta\right)-\frac{2i-1}{2n}\right]^2.$$

Moreover, we can obtain the CVM estimator of θ by solving the following equation for θ,

$$\sum_{i=1}^{n}\left[F\left(x_{i:n},\theta\right)-\frac{2i-1}{2n}\right]\xi(x_{i:n},\theta)=0, \tag{26}$$

where $\xi(x_{i:n},\theta)$ is already defined in Section 4.3.

4.6. Anderson-Darling method of estimation

The Anderson-Darling estimator (ADE) is another sort of minimum distance estimator that utilizes Anderson–Darling statistics. The ADE of θ, say $\hat{\theta}_{ADE}$, can be obtained by minimizing the following function with respect to θ,

$$A\left(\theta\right)=-n-\frac{1}{n}\sum_{i=1}^{n}\left(2i-1\right)\Big\{\log F\left(x_{i:n},\theta\right)+\log\bar{F}\left(x_{n+1-i:n},\theta\right)\Big\}.$$

The estimator $\hat{\theta}_{ADE}$ can also be achieved by simplifying the following nonlinear equation

$$\sum_{i=1}^{n+1}\left(2i-1\right)\left[\frac{\xi(x_{i:n},\theta)}{F(x_{i:n},\theta)}-\frac{\xi(x_{n+1-i:n},\theta)}{\bar{F}(x_{n+1-i:n},\theta)}\right]=0, \tag{27}$$

where $\xi(x_{i:n},\theta)$ is given in Section 4.3.

5. A Monte Carlo simulation study

This section showcases the behaviour of different estimation procedures for estimating the unknown parameter of the Arvind distribution. For this purpose, we have performed an empirical experiment which utilizes the following steps:

1. Generate 2,500 samples of size n =10, 20, 40, 60, 80, and 100 from Arvind distribution with $\theta=0.5,1.0,2.0$, and 4.0. For sample generation, Equation (3) has been used.
2. Calculate the MLE, MPS, OLS, WLS, CVM, and AD estimators for the 2,500 samples, say $\hat{\theta}_j; j=1,2,...,2,500$. Also, compute the 95% ACI for the above-generated samples.
3. Determine the expected value (EV), mean-squared error (MSE), and average bias (AB) for all point estimators, whereas, for 95% ACI, we compute the average lower confidence limit (ALCL), average upper confidence limit (AUCL), average width (AW), and coverage probability, *i.e.*,

$$EV=\frac{1}{2500}\sum_{j=1}^{2500}\hat{\theta}_j,\;\; MSE=\frac{1}{2500}\sum_{j=1}^{2500}\Big(\hat{\theta}_j-\theta\Big)^2,\;\; AB=\frac{1}{2500}\sum_{j=1}^{2500}\Big(\hat{\theta}_j-\theta\Big),$$

$$ALCL=\frac{1}{2500}\sum_{j=1}^{2500}LCL_j,\;\; AUCL=\frac{1}{2500}\sum_{j=1}^{2500}UCL_j,$$

$$AW=\frac{1}{2500}\sum_{j=1}^{2500}(UCL_j-LCL_j),\;\; CP=\frac{1}{2500}\sum_{j=1}^{2500}I_j(LCL_j<\theta_j<UCL_j),$$

where LCL_j and UCL_j denotes the upper and lower confidence limit for the j^{th} sample, respectively and $I_j(\bullet)$ is the indicator function takes value 1 if $LCL_j < \theta < UCL_j$ otherwise 0.

The simulation study was conducted using the R software and the codes are available upon request. Various classical estimates of θ with their MSE, AB are listed in Table 4. On the other hand, Table 5 contains the ALCL, AUCL, AW, and CP for 95% ACI.
From this empirical study, the following outcomes have been noted:

- We found that the average bias and MSE of all estimators approach zero for large n, indicating that the parameter estimates are consistent and asymptotically unbiased.
- The performance of all of the estimating techniques is satisfactory. However, in the overall comparison, for the proposed model, MPS is the most favourable estimation procedure while CVM is the least favourable estimation method.
- Additionally, the hierarchy of the best estimation technique among the numerous methods taken into consideration for estimating the parameter of the proposed distribution, as determined by the MSE, is as follows:
$$MPS \rightarrow MLE \rightarrow WLSE \rightarrow ADE \rightarrow LSE \rightarrow CVM$$
$$(HighlyPreferable \rightarrow Less\ Preferable)$$
- Except for MPS, all classical point estimators overestimate the parameter of the proposed model.
- From Table 5, we can simply conclude that ACI performed well. Even with a small sample size, for all values of θ, the asymptotic intervals computed here are able to sustain nominal levels of coverage probability. Furthermore, when we increase the sample size n, the AW of the ACI diminishes.

6. Application of Arvind distribution

The fitting capabilities of the Arvind distribution are shown in this section using three real datasets. We have used three distinct datasets from different areas. The detailed summary and graphical representation of these datasets can be found in Table 6 and Figure 4, respectively. The fitting of the proposed model has been compared with that of numerous well-known conventional and recently developed models. A list of the rival models can be found in Table 7. The fitted models' parameters have been estimated using MLE estimation for comparison's sake. Based on -logL, the Akaike information criterion (AIC), the corrected Akaike information criterion (CAIC), the Bayesian information criterion (BIC), and Kolmogorov-Smirnov (KS) statistic with the related P-value, the model comparison has been carried out. The open-source program R has been used to do the necessary calculations. The datasets along with their fitting summary are as follows:

Table 4: Various classical point estimates for different values of n

θ	n	MLE			MPS			OLS			WLS			CVM			ADE		
		EV	MSE	AB	EV	MSE	AB	EV	MSE	AB	EV	MSE	AB	EV	MSE	AB	EV	MSE	AB
0.5	10	0.5663	0.0550	0.0663	0.4990	0.0390	-0.0010	0.5510	0.0770	0.0510	0.5470	0.0710	0.0470	0.5600	0.0780	0.0600	0.5420	0.0540	0.0420
	20	0.5313	0.0199	0.0313	0.4900	0.0160	-0.0100	0.5240	0.0260	0.0240	0.5220	0.0240	0.0220	0.5290	0.0270	0.0290	0.5200	0.0210	0.0200
	40	0.5148	0.0085	0.0148	0.4897	0.0076	-0.0103	0.5109	0.0112	0.0109	0.5102	0.0101	0.0102	0.5133	0.0113	0.0133	0.5095	0.0096	0.0095
	60	0.5099	0.0053	0.0099	0.4913	0.0049	-0.0087	0.5068	0.0070	0.0068	0.5065	0.0063	0.0065	0.5085	0.0071	0.0085	0.5060	0.0061	0.0060
	80	0.5081	0.0040	0.0081	0.4931	0.0037	-0.0069	0.5057	0.0053	0.0057	0.5057	0.0048	0.0057	0.5070	0.0053	0.0070	0.5052	0.0046	0.0052
	100	0.5062	0.0031	0.0062	0.4936	0.0030	-0.0064	0.5043	0.0041	0.0043	0.5043	0.0037	0.0043	0.5052	0.0041	0.0052	0.5038	0.0036	0.0038
1	10	1.1423	0.2458	0.1423	1.0025	0.1722	0.0025	1.1112	0.3376	0.1112	1.1027	0.3136	0.1027	1.1285	0.3398	0.1285	1.0899	0.2397	0.0899
	20	1.0668	0.0872	0.0668	0.9812	0.0701	-0.0188	1.0520	0.1151	0.0520	1.0476	0.1051	0.0476	1.0614	0.1171	0.0614	1.0438	0.0935	0.0438
	40	1.0315	0.0368	0.0315	0.9794	0.0326	-0.0206	1.0235	0.0485	0.0235	1.0219	0.0438	0.0219	1.0283	0.0490	0.0283	1.0204	0.0415	0.0204
	60	1.0210	0.0228	0.0210	0.9825	0.0210	-0.0175	1.0150	0.0303	0.0150	1.0143	0.0272	0.0143	1.0182	0.0306	0.0182	1.0131	0.0263	0.0131
	80	1.0171	0.0172	0.0171	0.9860	0.0161	-0.0140	1.0123	0.0227	0.0123	1.0121	0.0204	0.0121	1.0147	0.0229	0.0147	1.0111	0.0199	0.0111
	100	1.0131	0.0133	0.0131	0.9870	0.0127	-0.0130	1.0093	0.0177	0.0093	1.0093	0.0158	0.0093	1.0112	0.0178	0.0112	1.0083	0.0155	0.0083
2	10	2.3080	1.1200	0.3080	2.0184	0.7785	0.0184	2.2417	1.4897	0.2417	2.2242	1.3867	0.2242	2.2752	1.4966	0.2752	2.1962	1.0748	0.1962
	20	2.1441	0.3889	0.1441	1.9662	0.3100	-0.0338	2.1137	0.5094	0.1137	2.1044	0.4659	0.1044	2.1321	0.5172	0.1321	2.0953	0.4139	0.0953
	40	2.0680	0.1629	0.0680	1.9594	0.1433	-0.0406	2.0518	0.2140	0.0518	2.0484	0.1934	0.0484	2.0613	0.2162	0.0613	2.0447	0.1827	0.0447
	60	2.0453	0.1008	0.0453	1.9649	0.0921	-0.0351	2.0328	0.1332	0.0328	2.0313	0.1196	0.0313	2.0392	0.1342	0.0392	2.0286	0.1154	0.0286
	80	2.0367	0.0755	0.0367	1.9716	0.0702	-0.0284	2.0269	0.0999	0.0269	2.0264	0.0899	0.0264	2.0317	0.1005	0.0317	2.0242	0.0873	0.0242
	100	2.0281	0.0584	0.0281	1.9733	0.0552	-0.0267	2.0204	0.0776	0.0204	2.0202	0.0694	0.0202	2.0242	0.0780	0.0242	2.0181	0.0680	0.0181
4	10	4.6605	5.0300	0.6605	4.0686	3.5644	0.0686	4.5206	6.5489	0.5206	4.4866	6.1365	0.4866	4.5832	6.5622	0.5832	4.4271	4.8272	0.4271
	20	4.3121	1.7657	0.3121	3.9425	1.3947	-0.0575	4.2468	2.2763	0.2468	4.2273	2.0916	0.2273	4.2815	2.3042	0.2815	4.2081	1.8653	0.2081
	40	4.1480	0.7350	0.1480	3.9208	0.6414	-0.0792	4.1137	0.9582	0.1137	4.1068	0.8691	0.1068	4.1318	0.9661	0.1318	4.0983	0.8203	0.0983
	60	4.0986	0.4527	0.0986	3.9296	0.4110	-0.0704	4.0721	0.5914	0.0721	4.0688	0.5324	0.0688	4.0843	0.5950	0.0843	4.0629	0.5147	0.0629
	80	4.0790	0.3363	0.0790	3.9421	0.3108	-0.0579	4.0590	0.4435	0.0590	4.0578	0.3997	0.0578	4.0683	0.4457	0.0683	4.0530	0.3886	0.0530
	100	4.0603	0.2581	0.0603	3.9448	0.2427	-0.0552	4.0447	0.3428	0.0447	4.0442	0.3070	0.0442	4.0521	0.3441	0.0521	4.0393	0.3004	0.0393

Table 5: Classical confidence intervals for different values of n

θ	n	ALCL	AUCL	AW	CP	θ	n	ALCL	Upper	Width	CP
0.5	10	0.1829	0.9497	0.7668	95.52	2.0	10	0.5978	4.0183	3.4205	95.33
	20	0.2787	0.7839	0.5052	95.57		20	1.0360	3.2530	2.2170	95.65
	40	0.3423	0.6874	0.3451	95.19		40	1.3160	2.8200	1.5030	95.25
	60	0.3705	0.6493	0.2788	95.41		60	1.4400	2.6510	1.2120	95.37
	80	0.3878	0.6283	0.2405	95.43		80	1.5150	2.5590	1.0440	95.36
	100	0.3991	0.6133	0.2142	95.25		100	1.5634	2.4928	0.9294	95.21
1	10	0.3371	1.9475	1.6104	95.48	4.0	10	1.0020	8.3190	7.3160	95.02
	20	0.5403	1.5933	1.0531	95.61		20	1.9600	6.6650	4.7050	95.47
	40	0.6731	1.3899	0.7168	95.29		40	2.5600	5.7360	3.1760	95.13
	60	0.7317	1.3103	0.5786	95.41		60	2.8200	5.3770	2.5560	95.37
	80	0.7677	1.2665	0.4989	95.52		80	2.9780	5.1790	2.2010	95.35
	100	0.7910	1.2352	0.4443	95.21		100	3.0810	5.0390	1.9580	95.29

Table 6: Summary of dataset I, II, and III

Dataset No.	Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum	SD
I	0.3200	0.9150	1.4700	1.6750	2.0870	4.7500	1.0006
II	2.8870	5.3290	8.1440	9.8870	13.8360	23.3940	5.8567
III	0.1100	0.7175	1.2350	1.5427	1.9425	4.7300	1.1276

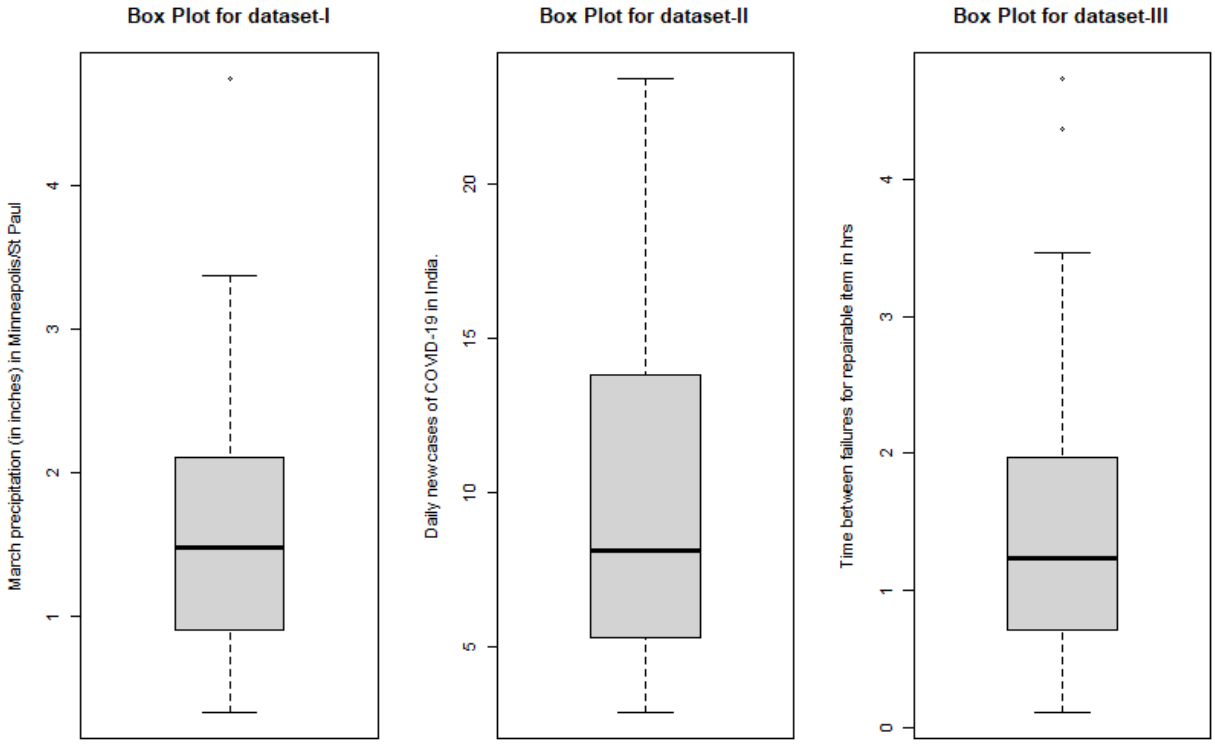


Figure 4: Graphical representation of dataset I, II, and III using boxplot

Table 7: The Competitive models

Model	Abbreviation	Parameter(s)	Author(s)
Lindley	L	θ	Lindley (1958)
Inverse Lindley	IL	θ	Sharma *et al.* (2015)
Inverted Modified Lindley	IML	θ	Chesneau *et al.* (2020)
Exponential	E	λ	-
Inverted Exponential	IE	β	Lin *et al.* (1989)
Inverse Rayleigh	IR	σ	Voda (1972)
Inverse Xgamma	IXG	θ	Yadav *et al.* (2021b)
Inverted Gamma	IG	α, β	Lin *et al.* (1989)
Inverse Weibull	IW	η, β	Khan *et al.* (2008)
Inverted Nadarajah–Haghighi	INH	λ, α	Tahir *et al.* (2018)
Inverted Topp-Leone	ITL	θ	Hassan *et al.* (2020)
Burr-Hatke Exponential	BHE	λ	Yadav *et al.* (2021a)
Maxwell Distribution	M	θ	Bekker and Roux (2005)
Laplace Distribution	La	μ, b	Kotz *et al.* (2001)
Inverse Lomax Distribution	ILo	α, β	Kleiber (2004)
Exponential Poisson Distribution	EP	λ, β	Kuş (2007)
Rayleigh	R	σ	Siddiqui (1962)

Dataset (I): The first real dataset represents thirty successive values of March precipitation (in inches) in Minneapolis/St Paul Yousef *et al.* (2023). The data values are:
0.77, 1.74, 0.81, 1.2, 1.95, 1.2, 0.47, 1.43, 3.37, 2.2, 3, 3.09, 1.51, 2.1, 0.52, 1.62, 1.31, 0.32, 0.59, 0.81, 2.81, 1.87, 1.18, 1.35, 4.75, 2.48, 0.96, 1.89, 0.9, 2.05.

Under dataset I, the fitting of the Arvind distribution is compared with L, IL, IML, E, IE, IR, IXG, IG, IW, and INH models. Table 8 lists the MLEs of the unknown parameters (standard errors (SEs) between parentheses) with the values of the –logL, AIC, BIC, CAIC, and KS statistic with associated P-value. Table 8 demonstrates that the suggested model has the lowest -logL, AIC, BIC, CAIC, and KS statistic as well as the highest P-value, hence, the Arvind distribution is superior to a number of competing models for this dataset. Figure 5 depicts the density and empirical vs fitted CDF plots for the proposed model with respect to dataset I. This graph also indicates that the Arvind distribution closely resembles the pattern of this real data.

Dataset (II): The second application takes into account the daily new cases of COVID-19 that have been reported in India. The data can be accessed at `https://www.worldometers.info/coronavirus/country/india/` and describes the daily new cases (in thousands) that occurred between the 16th of March 2021 and the 16th of April 2021. The data values are as follows:
28869, 35838, 39643, 40950, 43815, 40611, 47264, 53419, 59069, 62291, 62631, 68206, 56119, 53158, 72182, 81441, 89019, 92998, 103793, 96557, 115269, 126315, 131893, 144829, 152682, 169914, 160694, 185248, 199509, 216850, 233943.

To facilitate fitting, this dataset has been divided by 10000. The Arvind distribution's fit to this COVID data is compared to the L, IL, ITL, IXG, BHE, M, La, ILo, EP, and INH models. Table 9 summarizes the MLEs of the parameters (SEs between parentheses) as well

Table 8: The goodness-of-fit statistics for various models under dataset I

Model	MLE (SEs)	-logL	AIC	BIC	CAIC	KS	P-value
Arvind	0.1928 (0.0367)	39.7202	81.4403	82.8415	81.5832	0.0899	0.9685
L	0.9096 (0.1247)	43.1437	88.2875	89.6886	88.4303	0.1882	0.2383
IL	1.5835 (0.2267)	45.2212	92.4424	93.8436	92.5852	0.2279	0.0886
IML	1.247 (0.1906)	43.8683	89.7366	92.5390	90.1810	0.1975	0.1925
E	0.5971 (0.1090)	45.4744	92.9488	94.3500	93.0917	0.2352	0.0723
IE	1.1405 (0.2083)	46.2726	94.5452	95.9464	94.6881	0.2538	0.0420
IR	0.9267 (0.0846)	44.1365	90.2730	91.6740	90.4160	0.9360	0.0640
IXG	1.9440 (0.2680)	46.9850	95.9701	97.3713	96.1129	0.2632	0.0313
IG	2.5928 (0.6306), 2.9599 (0.7944)	40.3072	84.6144	87.4168	85.0589	0.1380	0.6174
IW	1.0163 (0.1273), 1.5495 (0.2026)	41.9170	87.8340	90.6364	88.2785	0.1523	0.4896
INH	3.0625 (2.8279), 0.2647 (0.2975)	44.5344	93.0689	95.8713	93.5133	0.1961	0.1989

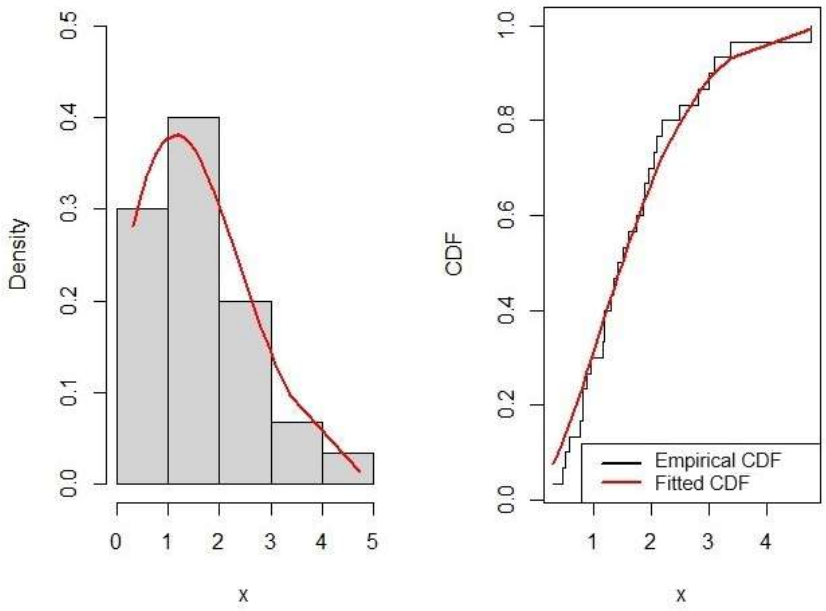


Figure 5: Histogram and the empirical vs fitted CDF under datasets I

as the -logL, AIC, BIC, CAIC, and KS statistic with its P-value. The developed model has the lowest -logL, AIC, BIC, CAIC, and KS statistic as well as the maximum P-value, as shown in Table 9; as a consequence, the Arvind distribution surpasses other competing models for this dataset. Figure 6 depicts the density and empirical vs fitted CDF for the proposed model under dataset II. This figure also reveals that the Arvind distribution closely follows the actual data pattern.

Table 9: The goodness-of-fit statistics for various models under dataset II

Model	MLE (SEs)	-logL	AIC	BIC	CAIC	KS	P-value
Arvind	0.0071 (0.0013)	95.0043	192.0085	193.4425	192.1465	0.1432	0.5032
L	0.1863 (0.0238)	96.7713	195.5426	196.9766	195.6805	0.1662	0.3216
IL	7.8937 (1.2790)	101.7100	205.4199	206.8539	205.5578	0.2593	0.0250
ITL	0.6169 (0.1108)	117.8740	237.7481	239.1820	237.8860	0.4094	<0.0001
IXG	8.4800 (1.3980)	102.7350	207.4706	208.9046	207.6086	0.2726	0.0158
BHE	0.0552 (0.1108)	103.7590	209.5174	210.9514	209.6554	0.2831	0.0108
M	0.0229 (0.0034)	97.5042	197.0083	198.4423	197.1463	0.2450	0.0401
La	8.1441 (0.0020), 4.6943 (0.8431)	100.4290	204.8579	207.7259	205.2865	0.1632	0.3434
ILo	0.6935 (0.2864), 10.6855 (3.7119)	102.8453	209.6907	212.5587	210.1193	0.2681	0.0186
EP	0.1010 (0.0182), 2.904e-07(0.01461)	102.0284	208.0567	210.9247	208.4853	0.2715	0.0165
INH	14.7742 (6.8420), 0.3258 (0.1251)	96.9953	197.9905	200.8585	198.4191	0.2028	0.1353

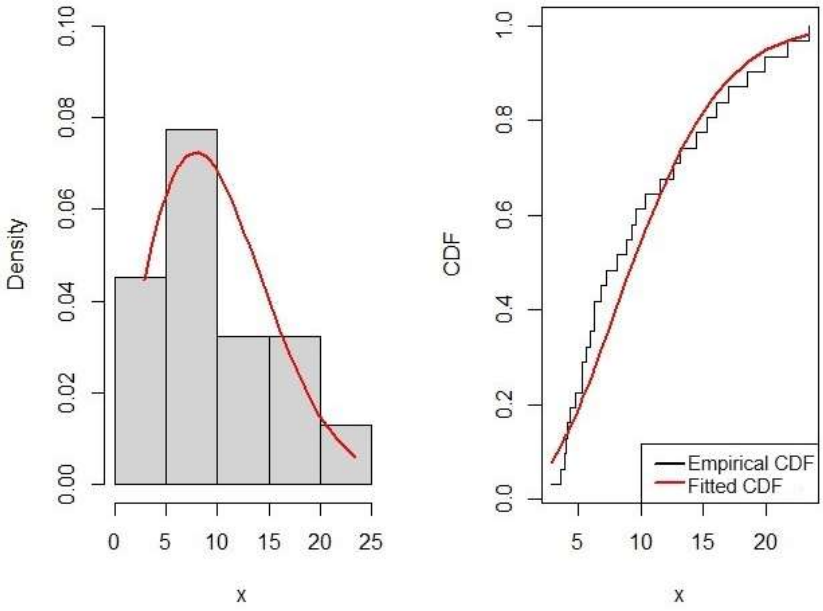


Figure 6: Histogram and the empirical vs fitted CDF under datasets II

Dataset (III): The third dataset includes the time between failures for repairable items, Murthy *et al.* (2004). The data values are as follows:
1.43, 0.11, 0.71, 0.77, 2.63, 1.49, 3.46, 2.46, 0.59, 0.74, 1.23, 0.94, 4.36, 0.40, 1.74, 4.73, 2.23, 0.45, 0.70, 1.06, 1.46, 0.30, 1.82, 2.37, 0.63, 1.23, 1.24, 1.97, 1.86, 1.17.

The Arvind distribution is fitted to this data and compared to the L, IL, E, R, IXG, La, ILo, IG, IW, and INH models. Table 10 depicts the MLEs of the parameters (SEs between parentheses) with the different fitting measures. The suggested model has the lowest -logL, AIC, BIC, CAIC, and KS statistics, as well as the greatest P-value, as shown in Table 10; as a result, the Arvind distribution excels other competing models for this dataset. Figure 7

shows the density and empirical vs fitted CDF for the Arvind model under dataset II. This graphic also demonstrates how well the Arvind distribution fits the actual data pattern.

Table 10: The goodness-of-fit statistics for various models under dataset III

Model	MLE (SEs)	-logL	AIC	BIC	CAIC	KS	P-value
Arvind	0.2042 (0.0388)	40.9532	83.9063	85.3075	84.0492	0.1205	0.7763
L	0.9761 (0.1345)	41.5473	85.0946	86.4958	85.2374	0.1406	0.5931
IL	1.1605 (0.1619)	46.9329	95.8658	97.2670	96.0087	0.1412	0.5885
E	0.6484 (0.1184)	43.0054	88.0108	89.4120	88.1536	0.1845	0.2586
R	1.3434 (0.1226)	42.9183	87.8366	89.2378	87.9794	0.1865	0.2479
IXG	1.4160 (0.1892)	48.9037	99.8073	101.2085	99.9502	0.1556	0.4615
La	1.2374 (8.3889), 0.8074 (0.1474)	44.37386	92.74771	95.55011	93.19216	0.12375	0.7478
ILo	0.11873 (0.05041), 7.73475 (2.62162)	46.01338	96.02677	98.82916	96.47121	0.18931	0.2325
IG	1.4209 (0.3325), 1.1271 (0.3152)	45.5074	95.0147	97.8171	95.4591	0.1576	0.4452
IW	0.7665 (0.1388), 1.0730 (0.1314)	46.3756	96.7512	99.5536	97.1957	0.1338	0.6557
INH	0.8517 (0.2348), 1.0347 (0.5133)	46.3701	96.7402	99.5426	97.1846	0.1786	0.2942

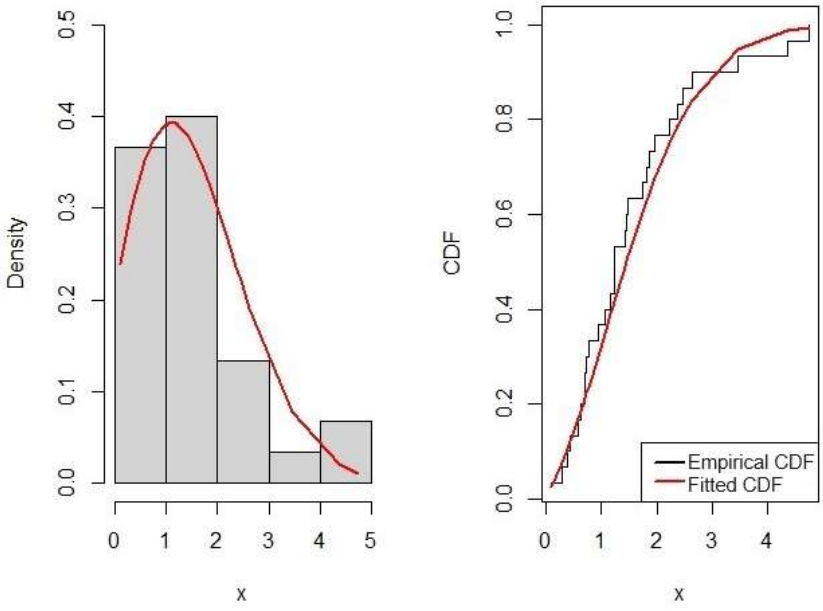


Figure 7: Histogram and the empirical vs fitted CDF under datasets III

6.1. Other classical estimates for datasets I, II, and III

Here, we estimate the Arvind distribution's unknown parameter using several different approaches that have been used in this article. Under I, II, and III datasets, we also obtain the 95% ACI for the unknown parameter θ. Table 11 includes estimates for θ from MLE,

MPS, OLS, WLS, CVM, and AD estimation methods along with respective SEs, and 95% ACI. In order to compare various approaches, Table 11 also contains the KS statistic and corresponding P-value for all approaches. From Table 11, we can easily see the reverse trend from the simulation section, as MPS is the least favourable estimation method for datasets I, II, and III.

Table 11: Classical estimates for dataset I, II, and III

Dataset	Methods	Estimate	SEs	KS	P-value	ACI (Width)
I	MLE	0.1928	0.0367	0.0899	0.9685	[0.1209, 0.2647] (0.1439)
	MPS	0.1824	0.1941	0.0952	0.9484	
	OLS	0.1920	0.0039	0.0892	0.9707	
	WLS	0.1878	0.0051	0.0851	0.9816	
	CVM	0.1927	0.0048	0.0899	0.9684	
	ADE	0.1904	0.0401	0.0877	0.9751	
II	MLE	0.0071	0.0013	0.1432	0.5032	[0.0046, 0.0096] (0.0050)
	MPS	0.0067	0.0068	0.1571	0.3877	
	OLS	0.0081	0.0003	0.1114	0.7960	
	WLS	0.0075	0.0004	0.1304	0.6209	
	CVM	0.0081	0.0003	0.1107	0.8026	
	ADE	0.0075	0.0015	0.1303	0.6220	
III	MLE	0.2042	0.0388	0.1205	0.7763	[0.1281, 0.2803] (0.1522)
	MPS	0.1910	0.2029	0.1427	0.5745	
	OLS	0.2417	0.0050	0.0698	0.9986	
	WLS	0.2364	0.0066	0.0709	0.9982	
	CVM	0.2425	0.0049	0.0702	0.9985	
	ADE	0.2292	0.0499	0.0815	0.9886	

7. Concluding remarks

A new lifetime model named Arvind distribution has been developed for modelling different types of data. The suggested model's PDF and HRF have a variety of forms that make it possible to analyze a broad range of real data. Its impressive statistical properties have been derived. Six different estimation methods namely the maximum likelihood, maximum product spacings, ordinary and weighted least square, Cramér-von Mises, and Anderson-Darling are discussed for estimating the unknown parameter. The asymptotic confidence interval has also been provided for the unknown parameter. An extensive simulation study has been performed to study the performance of the considered methods of estimations. This study suggests that methods of maximum product spacings and maximum likelihood are highly preferable whereas Cramér-von Mises is the least preferable method of estimation for the proposed model.

The goodness-of-fit of the proposed distribution has been explained with three real datasets from different fields and the fits of the proposed model have been found quite satisfactory over other existing lifetime models like Lindley, inverse Lindley, inverted modified Lindley, inverse Xgamma, inverse gamma, inverse Weibull, inverted Nadarajah-Haghighi, Burr-Hatke Exponential, *etc.* As a result, we may draw the conclusion that the proposed model may be utilized as a substitute for several well-known current models to analyze data

produced from diverse fields. In the future, we will extend this work by implementing censoring and different stress-strength models in the Arvind Distribution under various classical and non-classical estimation procedures.

Funding

No funds, grants, or other support were received.

Data availability

The authors confirm that the data supporting the study's findings are presented in the manuscript.

Conflict of interest

The authors state that they do not have any conflicts of interest.

Acknowledgements

The authors express their deep gratitude to the respected editor and reviewer for the valuable time and effort put into enhancing the overall quality of this manuscript.

References

Agiwal, V., Tyagi, S., and Chesneau, C. (2023). Bayesian and frequentist estimation of stress-strength reliability from a new extended burr xii distribution: Accepted: March 2023. *REVSTAT-Statistical Journal*, .

Alsuhabi, H., Alkhairy, I., Almetwally, E. M., Almongy, H. M., Gemeay, A. M., Hafez, E., Aldallal, R., and Sabry, M. (2022). A superior extension for the lomax distribution with application to covid-19 infections real data. *Alexandria Engineering Journal*, **61**(12), 11077–11090.

Bakouch, H. S., Nik, A. S., Asgharzadeh, A., and Salinas, H. S. (2021). A flexible probability model for proportion data: Unit-half-normal distribution. *Communications in Statistics: Case Studies, Data Analysis and Applications*, **7**(2), 271–288.

Bekker, A. and Roux, J. (2005). Reliability characteristics of the maxwell distribution: A bayes estimation study. *Communications in Statistics-Theory and Methods*, **34**(11), 2169–2178.

Bonferroni, C. (1930). *Elmenti di statistica generale*. Libreria Seber, Firenze.

Chaubey, Y. P. and Zhang, R. (2015). An extension of chen's family of survival distributions with bathtub shape or increasing hazard rate function. *Communications in Statistics-Theory and Methods*, **44**(19), 4049–4064.

Cheng, R. and Amin, N. (1979). Maximum product-of-spacings estimation with applications to the lognormal distribution. *Math report*, **791**.

Cheng, R. and Amin, N. (1983). Estimating parameters in continuous univariate distributions with a shifted origin. *Journal of the Royal Statistical Society: Series B (Methodological)*, **45**(3), 394–403.

Chesneau, C., Tomy, L., Gillariose, J., and Jamal, F. (2020). The inverted modified lindley distribution. *Journal of Statistical Theory and Practice*, **14**(3), 46.

Choudhary, N., Tyagi, A., and Singh, B. (2021). A flexible bathtub-shaped failure time model: Properties and associated inference. *Statistica*, **81**(1), 65–92.

El-Morshedy, M., Alshammari, F. S., Tyagi, A., Elbatal, I., Hamed, Y. S., and Eliwa, M. S. (2021). Bayesian and frequentist inferences on a type i half-logistic odd weibull generator with applications in engineering. *Entropy*, **23**(4), 446.

Goel, R. and Singh, B. (2020). Estimation of p (y¡ x) for modified weibull distribution under progressive type-ii censoring. *Life Cycle Reliability and Safety Engineering*, **9**, 227–240.

Gupta, R. D. and Kundu, D. (1999). Theory & methods: Generalized exponential distributions. *Australian & New Zealand Journal of Statistics*, **41**(2), 173–188.

Hassan, A. S., Elgarhy, M., and Ragab, R. (2020). Statistical properties and estimation of inverted topp-leone distribution. *J. Stat. Appl. Probab*, **9**(2), 319–331.

Khan, M. S., Pasha, G., and Pasha, A. H. (2008). Theoretical analysis of inverse weibull distribution. *WSEAS Transactions on Mathematics*, **7**(2), 30–38.

Kleiber, C. (2004). Lorenz ordering of order statistics from log-logistic and related distributions. *Journal of Statistical Planning and Inference*, **120**(1-2), 13–19.

Kotz, S., Kozubowski, T., and Podgórski, K. (2001). *The Laplace distribution and generalizations: a revisit with applications to communications, economics, engineering, and finance*. Number 183. Springer Science & Business Media.

Kuş, C. (2007). A new lifetime distribution. *Computational Statistics & Data Analysis*, **51**(9), 4497–4509.

Lin, C., Duran, B., and Lewis, T. (1989). Inverted gamma as a life distribution. *Microelectronics Reliability*, **29**(4), 619–626.

Lindley, D. V. (1958). Fiducial distributions and bayes' theorem. *Journal of the Royal Statistical Society. Series B (Methodological)*, , 102–107.

Lorenz, M. O. (1905). Methods of measuring the concentration of wealth. *Publications of the American statistical association*, **9**(70), 209–219.

Mudholkar, G. S., Srivastava, D. K., and Kollia, G. D. (1996). A generalization of the weibull distribution with application to the analysis of survival data. *Journal of the American Statistical Association*, **91**(436), 1575–1583.

Murthy, D. P., Xie, M., and Jiang, R. (2004). *Weibull Models*. John Wiley & Sons.

Nadarajah, S. and Haghighi, F. (2011). An extension of the exponential distribution. *Statistics*, **45**(6), 543–558.

Nadarajah, S. and Kotz, S. (2006). The beta exponential distribution. *Reliability engineering & system safety*, **91**(6), 689–697.

Shaked, M. and Shanthikumar, J. G. (2007). *Stochastic orders*. Springer.

Sharma, V. K., Singh, S. K., Singh, U., and Agiwal, V. (2015). The inverse lindley distribution: a stress-strength reliability model with application to head and neck cancer data. *Journal of Industrial and Production Engineering*, **32**(3), 162–173.

Siddiqui, M. M. (1962). Some problems connected with rayleigh distributions. *Journal of Research of the National Bureau of Standards D*, **66**, 167–174.

Swain, J. J., Venkatraman, S., and Wilson, J. R. (1988). Least-squares estimation of distri-

bution functions in johnson's translation system. *Journal of Statistical Computation and Simulation*, **29**(4), 271–297.

Tahir, M., Cordeiro, G. M., Ali, S., Dey, S., and Manzoor, A. (2018). The inverted nadarajah–haghighi distribution: estimation methods and applications. *Journal of Statistical Computation and Simulation*, **88**(14), 2775–2798.

Tyagi, S., Kumar, S., Pandey, A., Saha, M., and Bagariya, H. (2022). Power xgamma distribution: Properties and its applications to cancer data. *International Journal of Statistics and Reliability Engineering*, **9**(1), 51–60.

Voda, V. G. (1972). On the inverse rayleigh distributed random variable. *Rep. Statis. App. Res. JUSE*, **19**(4), 13–21.

Yadav, A. S., Altun, E., and Yousof, H. M. (2021a). Burr–hatke exponential distribution: a decreasing failure rate model, statistical inference and applications. *Annals of Data Science*, **8**, 241–260.

Yadav, A. S., Maiti, S. S., and Saha, M. (2021b). The inverse xgamma distribution: statistical properties and different methods of estimation. *Annals of Data Science*, **8**, 275–293.

Yousef, M. M., Alsultan, R., and Nassr, S. G. (2023). Parametric inference on partially accelerated life testing for the inverted kumaraswamy distribution based on type-ii progressive censoring data. *Math. Biosci. Eng*, **20**, 1674–1694.

Zenga, M. (2007). Inequality curve and inequality index based on the ratios between lower and upper arithmetic means. *Statistica & Applicazioni*, **5**(1), 3–27.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 22, No. 2, 2024 (New Series), pp 29–45
http://www.ssca.org.in/journal

Simple Linear Slope Estimators Based On Sample Quasi Ranges

Sharada V. Bhat and Shrinath M. Bijjargi
Department of Statistics,
Karnatak University, Dharwad - 580 003, India.

Received: 01 October 2023; Revised: 25 October 2023; Accepted: 28 October 2023

Abstract

The slope parameter in simple linear regression measures the change in mean of distribution of response variable for unit change in predictor variable. Some estimators based on sample quasi ranges of predictor variables are proposed. The mean and variance of the proposed estimators are derived. The relative efficiencies among the proposed estimators are obtained. Also, these estimators are compared with the estimators available in the literature. Few datasets are considered to illustrate the fitting of simple linear regression using proposed estimators and comparing their performances.

Key words: Simple linear regression; Sample quasi ranges; Unbiased estimators; Variance; Slope parameter; Relative efficiency.

AMS Subject Classifications: 62G05, 62J05

1. Introduction

Regression analysis helps in understanding the nature and strength of the relationship among two or more variables. Linear regression model is helpful in modeling the relationship among response variable (y) and the predictor variable (x). This model is used by economists to relate variables such as consumption, savings with income; by environmental scientists to relate environmental factors such as temperature, pollution levels with ecosystem or public health; psychologists to relate human behavior with mental health and stress levels with academic performances, *etc.* In addition, it is used in various domains of studies like finance, marketing, real estate, pharmaceuticals, clinical trials, national development, education and many others. The least square estimator is widely used in linear regression to estimate the slope parameter. The literature reveals that the method of least squares was due to Legendre (1805). Gauss (1809) claimed that he had been using the procedure since 1795. Harter (1974), Stigler (1986) and Hald (1998) noticed that, "Euler (1749) and Mayer (1750) independently developed a method known as method of averages" for fitting a linear equation to observed data. Their method deals with arranging the predictor variables in descending order and grouping them into as many numbers of existing parameters.

Corresponding Author: Sharada V. Bhat
Email: bhat_sharada@yahoo.com

Bose (1938) proposed three estimators based on method of successive differences, method of differences at half range and method of range as alternative to least square estimator for slope parameter in simple linear regression, when predictor variables are equidistant. Wald (1940) observed that the efficiency of slope estimator will be maximum when x_i's are arranged in ascending order. Nair and Shrivastava (1942) generalized procedure of Bose (1938) to method of group averages to improve the relative efficiency of the estimators.

Liu and Preve (2016) proposed estimators to slope parameter in simple linear regression based on robust measures of location, *viz.* median and trimmed mean. The focus is on the case where predictor variable is assumed to be stochastic, having symmetric stable distribution and error having distribution either symmetric stable or a normal mixture. Cliff and Billy (2017) developed simple averaging method based on the average of successive slopes. Prabowo *et al.* (2020) simplified this method and investigated its performance. Singthongchai *et al.* (2021) developed improved simple averaging method replacing median in the place of mean in the method due to Cliff and Billy (2017). Jlibene *et al.* (2021) studied the least square estimator when the error has uniform distribution. Yao *et al.* (2021) proposed best linear unbiased estimators using moving extremes ranked set sampling.

Bhat and Bijjargi (2023) proposed estimation procedures generalizing the methods due to Bose (1938) including some adaptive estimators, in the presence of unequal distances among predictor variables. Among the methods proposed, the method of differences among ordered predictors lying equally on either side of the half range outperforms all other estimators. Basically, this estimator is based on quasi ranges. The immediate quest that arises is, whether the estimator is improved by taking some weights to quasi ranges. To investigate this fact, we develop few estimators based on different types of weights to quasi ranges.

In this paper, we propose some estimators for slope parameter of simple linear regression model based on sample quasi ranges given in Govindarajulu (2007). Suppose $x_i\ ,\ 1 \leq i \leq n$ are arranged in ascending order of magnitude, $x_{(i)}$ is the i^{th} order statistic, then, for $n = 2m$, the j^{th} quasi range, $j = 1,\ 2,\ \cdots,\ m-1$ is defined as the range of $(n-2j)$ sample values. Suppose q_j is the j^{th} quasi range, then q_j is given by $q_j = x_{(n-j)} - x_{(j+1)}$. We observe that, $q_0 = x_{(n)} - x_{(1)}$ is the range of n observations. Mosteller (2006) and Harter (1959) used quasi ranges to estimate population standard deviation.

The proposed estimators are given in section 2, their mean and variance are derived in section 3 and their performance using relative efficiency is investigated in section 4. The simple linear regression using proposed estimators along with their performances are illustrated through examples in section 5. Section 6 contains conclusions.

2. Estimators based on sample quasi ranges

Consider the simple linear regression model,

$$y_i = \alpha + \beta x_i + e_i\ , \qquad 1 \leq i \leq n \tag{1}$$

where, y_i is response variable, x_i is predictor variable, e_i is independent and identically distributed random error from distribution with zero mean and finite variance σ^2. Here, α is intercept parameter and β is slope parameter to be estimated from the data to explore the linear relation between x_i and y_i. The slope parameter β represents the change in mean of

distribution of y for unit change in x. The least square estimator of β is given by

$$\widehat{\beta}_{ul} = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sum_{i=1}^{n}(x_i - \overline{x})^2}, \tag{2}$$

where, $\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$ and $\overline{y} = \frac{\sum_{i=1}^{n} y_i}{n}$.

Among the methods proposed by Bose (1938), the estimator $\widehat{\beta}_{eh}$ obtained by method of differences at half range outperforms other estimators and is given by

$$\widehat{\beta}_{eh} = \frac{\sum_{i=1}^{m}(y_{m+i} - y_i)}{m^2 d}, \tag{3}$$

where d is distance among ordered x_i.

In case of unequal distances among predictor variables, estimator due to Bhat and Bijjargi (2023) based on method of distances among ordered observations lying equally on either side of half range outperforms other proposed estimators and is given by

$$\widehat{\beta}_{ud} = \frac{\sum_{i=1}^{m}\left(y^*_{m+i} - y^*_{m-i+1}\right)}{\sum_{i=1}^{m}\left(x_{(m+i)} - x_{(m-i+1)}\right)}. \tag{4}$$

Here, y^*_i is y observation corresponding to $x_{(i)}$, i^{th} order statistic. $\widehat{\beta}_{ud}$ reduces to the method of differences at half range given by

$$\widehat{\beta}_{uh} = \frac{\sum_{i=1}^{m}\left(y^*_{m+i} - y^*_i\right)}{\sum_{i=1}^{m}\left(x_{(m+i)} - x_{(i)}\right)}. \tag{5}$$

Also, when distances among ordered predictor variables are equal, $\widehat{\beta}_{ud} = \widehat{\beta}_{uh}$ reduces to $\widehat{\beta}_{eh}$.

We propose estimators $\widehat{\beta}_k,\ k = 1,\ 2,\ \cdots,\ 6$ using quasi ranges respectively based on the weights $w_k,\ k = 1,\ 2,\ \cdots,\ 6$. Representing arbitrary weight by $a_{ki},\ k = 1,\ \cdots,\ 6,\ i = 1,\ \cdots,\ m$, w_1 is given by $a_{1i} = \frac{1}{m-i+1}$, w_2 by $a_{2i} = \frac{1}{i}$, w_3 by $a_{3i} = \frac{m-i+1}{\sum_{i=1}^{m} m-i+1}$, $a_{4i} = \frac{i}{\sum_{i=1}^{m} i}$, $a_{5i} = m - i + 1$ and $a_{6i} = i$. We see that, a_{1i}, a_{4i} and a_{6i} relatively assign heavier weights to quasi range with extreme order statistics, where as, a_{2i}, a_{3i} and a_{5i} assign lower weights. That is, a_{1i}, a_{4i} and a_{6i} assign highest weight to q_0, relatively lesser weight to q_1, q_2, $\cdots$ and q_{m-1}. Similarly, a_{2i}, a_{3i} and a_{5i} assign lowest weight to q_0, relatively heavier weights to q_1, q_2, $\cdots$ and q_{m-1}. As efficiency and robustness are vital to estimators, the motivation for assigning various weights to the quasi ranges is to develop adaptive estimators in terms of efficiency and robustness. In the presence of several estimators, researcher seeks efficient estimator that closely estimates the parameter, whereas, robust estimator is sought to estimate the parameter sensibly in the presence of outliers in the data.

The weights employed to propose various estimators are given in detail in Table 1.

Table 1: Weights for proposition of estimators

range/quasi ranges	y values	w_1	w_2	w_3	w_4	w_5	w_6
$q_0 = x_{(n)} - x_{(1)}$	$y_n^* - y_1^*$	1	$\frac{1}{m}$	$\frac{1}{\sum_{i=1}^{m}(m-i+1)}$	$\frac{m}{\sum_{i=1}^{m} i}$	1	m
$q_1 = x_{(n-1)} - x_{(2)}$	$y_{n-1}^* - y_2^*$	$\frac{1}{2}$	$\frac{1}{m-1}$	$\frac{2}{\sum_{i=1}^{m}(m-i+1)}$	$\frac{m-1}{\sum_{i=1}^{m} i}$	2	$m-1$
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.
$q_{m-2} = x_{(m+2)} - x_{(m-1)}$	$y_{m+2}^* - y_{m-1}^*$	$\frac{1}{m-1}$	$\frac{1}{2}$	$\frac{m-1}{\sum_{i=1}^{m}(m-i+1)}$	$\frac{2}{\sum_{i=1}^{m} i}$	$m-1$	2
$q_{m-1} = x_{(m+1)} - x_{(m)}$	$y_{m+1}^* - y_m^*$	$\frac{1}{m}$	1	$\frac{m}{\sum_{i=1}^{m}(m-i+1)}$	$\frac{1}{\sum_{i=1}^{m} i}$	m	1

The proposed estimators are given by

$$\widehat{\beta}_1 = \frac{\sum_{i=1}^{m}\left(\frac{y_{m+i}^* - y_{m-i+1}^*}{m-i+1}\right)}{\sum_{i=1}^{m}\left(\frac{x_{(m+i)} - x_{(m-i+1)}}{m-i+1}\right)}, \tag{6}$$

$$\widehat{\beta}_2 = \frac{\sum_{i=1}^{m}\left(\frac{y_{m+i}^* - y_{m-i+1}^*}{i}\right)}{\sum_{i=1}^{m}\left(\frac{x_{(m+i)} - x_{(m-i+1))}}{i}\right)}, \tag{7}$$

$$\widehat{\beta}_3 = \frac{\sum_{i=1}^{m}\frac{(m-i+1)\left(y_{m+i}^* - y_{m-i+1}^*\right)}{\sum_{i=1}^{m}(m-i+1)}}{\sum_{i=1}^{m}\frac{(m-i+1)\left(x_{(m+i)} - x_{(m-i+1)}\right)}{\sum_{i=1}^{m}(m-i+1)}}, \tag{8}$$

$$\widehat{\beta}_4 = \frac{\sum_{i=1}^{m}\frac{i\left(y_{m+i}^* - y_{m-i+1}^*\right)}{\sum_{i=1}^{m} i}}{\sum_{i=1}^{m}\frac{i\left(x_{(m+i)} - x_{(m-i+1)}\right)}{\sum_{i=1}^{m} i}}, \tag{9}$$

$$\widehat{\beta}_5 = \frac{\sum_{i=1}^{m}(m-i+1)\left(y_{m+i}^* - y_{m-i+1}^*\right)}{\sum_{i=1}^{m}(m-i+1)\left(x_{(m+i)} - x_{(m-i+1)}\right)} = \widehat{\beta}_3 \tag{10}$$

and

$$\widehat{\beta}_6 = \frac{\sum_{i=1}^{m} i\left(y_{m+i}^* - y_{m-i+1}^*\right)}{\sum_{i=1}^{m} i\left(x_{(m+i)} - x_{(m-i+1)}\right)} = \widehat{\beta}_4. \tag{11}$$

To obtain these estimators under the situation that the predictor variables are equidistant, $d\sum_{i=1}^{m}(2i-1)$ is substituted in the place of $\sum_{i=1}^{m}\left(x_{(m+i)} - x_{(m-i+1)}\right)$.

For odd number of sample sizes, i.e. $n = 2m + 1$, the middle pair of observation, $(y^*_{m+1}, x_{(m+1)})$ is not considered. When the distances among $x_{(i)}$'s are unequal and the weights are equal, the estimators $\widehat{\beta}_k,\ k = 1,\ 2,\ 3,\ 4$ reduce to $\widehat{\beta}_{ud}$ and to $\widehat{\beta}_{eh}$ when distances are equal.

3. Mean and variance of the proposed estimators

In this section, the mean of the proposed estimators and their variances are obtained. The mean of $\widehat{\beta}_1$ is given by

$$
\begin{aligned}
E\left(\widehat{\beta}_1\right) &= E\left(\frac{\sum_{i=1}^{m}\left(\frac{y^*_{m+i}-y^*_{m-i+1}}{m-i+1}\right)}{\sum_{i=1}^{m}\left(\frac{x_{(m+i)}-x_{(m-i+1)}}{m-i+1}\right)}\right)\\
&= \frac{1}{\sum_{i=1}^{m}\left(\frac{x_{(m+i)}-x_{(m-i+1)}}{m-i+1}\right)} E\left(\sum_{i=1}^{m}\left(\frac{y^*_{m+i}-y^*_{m-i+1}}{m-i+1}\right)\right)\\
&= \frac{1}{\sum_{i=1}^{m}\left(\frac{x_{(m+i)}-x_{(m-i+1)}}{m-i+1}\right)} \sum_{i=1}^{m}\frac{1}{m-i+1} E\left(y^*_{m+i}-y^*_{m-i+1}\right)\\
&= \frac{1}{\sum_{i=1}^{m}\left(\frac{x_{(m+i)}-x_{(m-i+1)}}{m-i+1}\right)} \sum_{i=1}^{m}\frac{\beta}{m-i+1}\left(x_{(m+i)}-x_{(m-i+1)}\right)\\
&= \frac{1}{\sum_{i=1}^{m}\left(\frac{x_{(m+i)}-x_{(m-i+1)}}{m-i+1}\right)} \sum_{i=1}^{m}\beta\left(\frac{x_{(m+i)}-x_{(m-i+1)}}{m-i+1}\right)
\end{aligned}
$$

$$
E\left(\widehat{\beta}_1\right) = \beta. \tag{12}
$$

Hence, $\widehat{\beta}_1$ is an unbiased estimator of β.

The variance of $\widehat{\beta}_1$ is given by

$$
\begin{aligned}
V\left(\widehat{\beta}_1\right) &= V\left(\frac{\sum_{i=1}^{m}\left(\frac{y^*_{m+i}-y^*_{m-i+1}}{m-i+1}\right)}{\sum_{i=1}^{m}\left(\frac{x_{(m+i)}-x_{(m-i+1)}}{m-i+1}\right)}\right)\\
&= \frac{1}{\left[\sum_{i=1}^{m}\left(\frac{x_{(m+i)}-x_{(m-i+1)}}{m-i+1}\right)\right]^2} V\left(\sum_{i=1}^{m}\left(\frac{y^*_{m+i}-y^*_{m-i+1}}{m-i+1}\right)\right)\\
&= \frac{1}{\left[\sum_{i=1}^{m}\left(\frac{x_{(m+i)}-x_{(m-i+1)}}{m-i+1}\right)\right]^2} \left(\sum_{i=1}^{m}\frac{1}{(m-i+1)^2}\ V(y^*_{m+i}-y^*_{m-i+1})\right)\\
&= \frac{1}{\left[\sum_{i=1}^{m}\left(\frac{x_{(m+i)}-x_{(m-i+1)}}{m-i+1}\right)\right]^2} \left(\sum_{i=1}^{m}\frac{1}{(m-i+1)^2}\ 2\sigma^2\right)
\end{aligned}
$$

$$
V\left(\widehat{\beta}_1\right) = \frac{2\sigma^2\sum_{i=1}^{m}\frac{1}{(m-i+1)^2}}{\left[\sum_{i=1}^{m}\left(\frac{x_{(m+i)}-x_{(m-i+1)}}{m-i+1}\right)\right]^2} \tag{13}
$$

Under equidistant $x_{(i)}$'s,

$$
V\left(\widehat{\beta}_1\right) = \frac{2\sigma^2\sum_{i=1}^{m}\frac{1}{(m-i+1)^2}}{\left[d\ \sum_{i=1}^{m}\left(\frac{2i-1}{m-i+1}\right)\right]^2} \tag{14}
$$

Similarly, we observe that, all the proposed estimators are unbiased estimators of β and they have different variances. The variances of $\widehat{\beta}_k$, $k = 1,\ 2,\ 3,\ 4.$ for equal and unequal distances among $x_{(i)}$'s are furnished in Table 2.

Table 2: Variance of $\widehat{\boldsymbol{\beta}}_{\boldsymbol{k}}$, $\boldsymbol{k = 1,\ 2,\ 3,\ 4}$

Estimator	$V\left(\widehat{\beta}_k\right)$ under unequal distances	$V\left(\widehat{\beta}_k\right)$ under equal distances
$\widehat{\beta}_1$	$\frac{2\sigma^2 \sum_{i=1}^{m} \frac{1}{(m-i+1)^2}}{\left[\sum_{i=1}^{m}\left(\frac{x_{(m+i)}-x_{(m-i+1)}}{m-i+1}\right)\right]^2}$	$\frac{2\sigma^2 \sum_{i=1}^{m} \frac{1}{(m-i+1)^2}}{\left[d\ \sum_{i=1}^{m}\left(\frac{2i-1}{m-i+1}\right)\right]^2}$
$\widehat{\beta}_2$	$\frac{2\sigma^2 \sum_{i=1}^{m} \frac{1}{i^2}}{\left[\sum_{i=1}^{m}\left(\frac{x_{(m+i)}-x_{(m-i+1)}}{i}\right)\right]^2}$	$\frac{2\sigma^2 \sum_{i=1}^{m} \frac{1}{i^2}}{\left[d\ \sum_{i=1}^{m}\left(\frac{2i-1}{i}\right)\right]^2}$
$\widehat{\beta}_3$	$\frac{2\ \sigma^2 \sum_{i=1}^{m}(m-i+1)^2}{\left[\sum_{i=1}^{m}(m-i+1)\left(x_{(m+i)}-x_{(m-i+1)}\right)\right]^2}$	$\frac{48\ \sigma^2}{d^2 n(n+1)(n+2)}$
$\widehat{\beta}_4$	$\frac{2\ \sigma^2 \sum_{i=1}^{m} i^2}{\left[\sum_{i=1}^{m} i\left(x_{(m+i)}-x_{(m-i+1)}\right)\right]^2}$	$\frac{48\ (n+1)\sigma^2}{d^2 n(n+2)(2n-1)^2}$

When the distances among $x_{(i)}$'s are equal, the least square estimator given in (2) reduces to

$$\widehat{\beta}_{el} \quad = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{d^2 \frac{n(n^2-1)}{12}} \tag{15}$$

and its variance given by

$$V\left(\widehat{\beta}_{el}\right) = \frac{12\ \sigma^2}{d^2 n\left(n^2-1\right)} \tag{16}$$

When $d = 1$, $\sigma = 1$, $V(\widehat{\beta}_k)$ is computed from Table 2 for various values of n and are given in Table 3 and plotted in Figure 1.

Table 3: $\boldsymbol{V\left(\widehat{\beta}_k\right)}$, $\boldsymbol{k = 1, 2, 3, 4}$ for various values of $\boldsymbol{n}$

n	$V\left(\widehat{\beta}_4\right)$	$V\left(\widehat{\beta}_1\right)$	$V\left(\widehat{\beta}_3\right)$	$V\left(\widehat{\beta}_2\right)$	$V\left(\widehat{\beta}_{el}\right)$	$V\left(\widehat{\beta}_{eh}\right)$
6	0.057851	0.058299	0.142857	0.15680	0.057143	0.074074
8	0.024000	0.024638	0.066667	0.081333	0.023810	0.031250
10	0.012188	0.012810	0.036364	0.049158	0.012121	0.016000
14	0.004409	0.004879	0.014286	0.023236	0.004396	0.005831
18	0.002068	0.002409	0.007018	0.013380	0.002064	0.002743
22	0.001131	0.001384	0.003953	0.008650	0.001129	0.001503
26	0.000684	0.000877	0.002442	0.006033	0.000684	0.000910
30	0.000445	0.000595	0.001613	0.004440	0.000445	0.000593
40	0.000188	0.000276	0.000697	0.002409	0.000188	0.000250
50	0.000096	0.000154	0.000362	0.001506	0.000096	0.000128
70	0.000035	0.000064	0.000134	0.000746	0.000035	0.000047
100	0.000012	0.000026	0.000047	0.000356	0.000012	0.000016

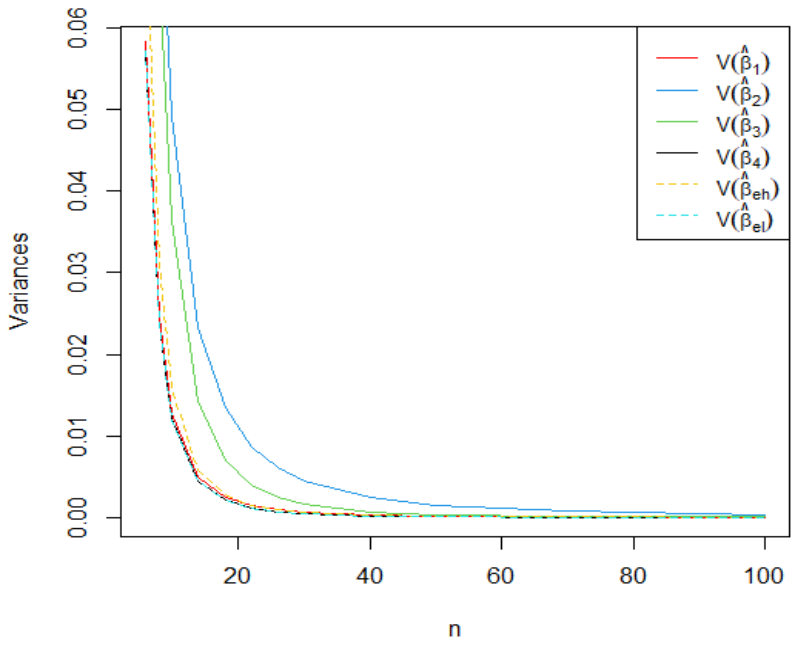


Figure 1: Variance of various slope estimators for different values of n

From Table 3 and Figure 1, we observe that, for all n, $V\left(\widehat{\beta}_4\right) < V\left(\widehat{\beta}_1\right) < V\left(\widehat{\beta}_3\right) < V\left(\widehat{\beta}_2\right)$. Among the proposed estimators, $\widehat{\beta}_4$ has minimum variance and for $n > 22$, $V\left(\widehat{\beta}_4\right)$ is equal to $V(\widehat{\beta}_{el})$. The $V\left(\widehat{\beta}_4\right)$ is less than $V\left(\widehat{\beta}_{eh}\right)$ and for $n \leq 30$, $V\left(\widehat{\beta}_1\right)$ is less than $V\left(\widehat{\beta}_{eh}\right)$. Also, for increasing value of n, $V\left(\widehat{\beta}_k\right)$, $k = 1, 2, 3, 4$ is decreasing.

4. Performance of the proposed estimators

In this section, we study the performance of proposed estimators using relative efficiencies. The relative efficiency (RE) of two estimators, namely, A and B is given by

$$RE(A, B) = \frac{V(B)}{V(A)}. \tag{17}$$

We conclude that, A is better than B in terms of its performance if $RE(A, B) > 1$. The RE among proposed estimators for both cases where in predictor variables have unequal distance and equal distance are derived and given in Table 4. A comparison among $\widehat{\beta}_k$, $k = 1, 2, 3, 4$ is carried out in Table 5 in terms of computed values of RE for various values of n when $x_{(i)}$'s are equidistant. Using Table 5, RE of $\widehat{\beta}_4$ with respect to (wrt) $\widehat{\beta}_1$, $\widehat{\beta}_2$, $\widehat{\beta}_3$, RE of $\widehat{\beta}_1$ wrt $\widehat{\beta}_2$, $\widehat{\beta}_3$ and RE of $\widehat{\beta}_3$ wrt $\widehat{\beta}_2$ are given in Figure 2.

From Table 5 and Figure 2, it is observed that, $RE\left(\widehat{\beta}_4, \widehat{\beta}_2\right) > RE(\widehat{\beta}_4, \widehat{\beta}_3) > RE(\widehat{\beta}_4, \widehat{\beta}_1)$ and $RE\left(\widehat{\beta}_1, \widehat{\beta}_2\right) > RE\left(\widehat{\beta}_1, \widehat{\beta}_3\right)$. Hence, $\widehat{\beta}_4$ is performing better than all other proposed estimators, $\widehat{\beta}_1$ performs better than $\widehat{\beta}_2$, $\widehat{\beta}_3$ and $\widehat{\beta}_3$ outperforms $\widehat{\beta}_2$. Also, RE of $\widehat{\beta}_4$ wrt $\widehat{\beta}_1$, $\widehat{\beta}_2$, $\widehat{\beta}_3$, RE of $\widehat{\beta}_1$ wrt $\widehat{\beta}_2$ and RE of $\widehat{\beta}_3$ wrt $\widehat{\beta}_2$ increases for increasing values of n, whereas, RE of $\widehat{\beta}_1$ wrt $\widehat{\beta}_3$ decreases for $n > 14$. As $\widehat{\beta}_4$ outperforms $\widehat{\beta}_1$, $\widehat{\beta}_2$ and $\widehat{\beta}_3$, we compute $RE(\widehat{\beta}_4, \widehat{\beta}_{el})$

Table 4: RE among proposed $\widehat{\beta}_k$, $k = 1,\ 2,\ 3,\ 4$

	For unequal distances among $x_{(i)}$'s	For equal distances among $x_{(i)}$'s
$RE(\widehat{\beta}_1, \widehat{\beta}_2)$	$\frac{\left[\sum_{i=1}^{m}\left(\frac{x_{(m+i)}-x_{(m-i+1)}}{m-i+1}\right)\right]^2}{\left[\sum_{i=1}^{m}\left(\frac{x_{(m+i)}-x_{(m-i+1)}}{i}\right)\right]^2}$	$\frac{\left[\sum_{i=1}^{m}\left(\frac{2i-1}{m-i+1}\right)\right]^2}{\left[\sum_{i=1}^{m}\left(\frac{2i-1}{i}\right)\right]^2}$
$RE(\widehat{\beta}_1, \widehat{\beta}_3)$	$\frac{\left[\sum_{i=1}^{m}\left(\frac{x_{(m+i)}-x_{(m-i+1)}}{m-i+1}\right)\right]^2 \sum_{i=1}^{m}(m-i+1)^2}{\left[\sum_{i=1}^{m}(m-i+1)\left(x_{(m+i)}-x_{(m-i+1)}\right)\right]^2 \sum_{i=1}^{m}\frac{1}{(m-i+1)^2}}$	$\frac{48\ \left[\sum_{i=1}^{m}\left(\frac{2i-1}{m-i+1}\right)\right]^2}{n(n+1)(n+2)\sum_{i=1}^{m}\frac{1}{(m-i+1)^2}}$
$RE(\widehat{\beta}_1, \widehat{\beta}_4)$	$\frac{\left[\sum_{i=1}^{m}\left(\frac{x_{(m+i)}-x_{(m-i+1)}}{m-i+1}\right)\right]^2 \sum_{i=1}^{m} i^2}{\left[\sum_{i=1}^{m} i\left(x_{(m+i)}-x_{(m-i+1)}\right)\right]^2 \sum_{i=1}^{m}\frac{1}{(m-i+1)^2}}$	$\frac{24(n+1)\left[\sum_{i=1}^{m}\left(\frac{2i-1}{m-i+1}\right)\right]^2}{n(n+2)(2n-1)^2\sum_{i=1}^{m}\frac{1}{(m-i+1)^2}}$
$RE(\widehat{\beta}_2, \widehat{\beta}_3)$	$\frac{\left[\sum_{i=1}^{m}\left(\frac{x_{(m+i)}-x_{(m-i+1)}}{i}\right)\right]^2 \sum_{i=1}^{m}(m-i+1)^2}{\left[\sum_{i=1}^{m}(m-i+1)\left(x_{(m+i)}-x_{(m-i+1)}\right)\right]^2 \sum_{i=1}^{m}\frac{1}{i^2}}$	$\frac{24\left[\sum_{i=1}^{m}\left(\frac{2i-1}{i}\right)\right]^2}{n(n+1)(n+2)\sum_{i=1}^{m}\frac{1}{i^2}}$
$RE(\widehat{\beta}_2, \widehat{\beta}_4)$	$\frac{\left[\sum_{i=1}^{m}\left(\frac{x_{(m+i)}-x_{(m-i+1)}}{i}\right)\right]^2 \sum_{i=1}^{m} i^2}{\left[\sum_{i=1}^{m} i\left(x_{(m+i)}-x_{(m-i+1)}\right)\right]^2 \sum_{i=1}^{m}\frac{1}{i^2}}$	$\frac{24(n+1)\left[\sum_{i=1}^{m}\left(\frac{2i-1}{i}\right)\right]^2}{n(n+2)(2n-1)^2\sum_{i=1}^{m}\frac{1}{i^2}}$
$RE(\widehat{\beta}_3, \widehat{\beta}_4)$	$\frac{\left[\sum_{i=1}^{m}(m-i+1)\left(x_{(m+i)}-x_{(m-i+1)}\right)\right]^2}{\left[\sum_{i=1}^{m} i\left(x_{(m+i)}-x_{(m-i+1)}\right)\right]^2}$	$\frac{(n+1)^2}{(2n-1)^2}$

Table 5: RE among proposed estimators for various n

n	$RE(\widehat{\beta}_4, \widehat{\beta}_1)$	$RE(\widehat{\beta}_4, \widehat{\beta}_2)$	$RE(\widehat{\beta}_4, \widehat{\beta}_3)$	$RE(\widehat{\beta}_1, \widehat{\beta}_2)$	$RE(\widehat{\beta}_1, \widehat{\beta}_3)$	$RE(\widehat{\beta}_3, \widehat{\beta}_2)$
6	1.007729	2.710394	2.469380	2.689600	2.450440	1.097598
8	1.026578	3.388911	2.777778	3.301130	2.705850	1.220003
10	1.050994	4.033234	2.983472	3.837540	2.838730	1.351845
14	1.106672	5.270092	3.240021	4.762070	2.927710	1.626545
18	1.165107	6.469979	3.393396	5.553050	2.912470	1.906650
22	1.223496	7.649354	3.495281	6.251810	2.856780	2.188423
26	1.280902	8.815233	3.567861	6.881760	2.785450	2.470600
30	1.337024	9.971084	3.622270	7.457780	2.709190	2.752773
40	1.471540	12.835320	3.712642	8.722070	2.523000	3.456978
50	1.598491	15.673980	3.768181	9.804740	2.357310	4.159215
70	1.834458	21.303790	3.832739	11.613810	2.089310	5.558644
100	2.155869	29.691210	3.882138	13.772790	1.800680	7.648769

and $RE(\widehat{\beta}_4, \widehat{\beta}_{eh})$ for various values of n and furnish in Table 6.

From Table 6, we notice that, $RE(\widehat{\beta}_4, \widehat{\beta}_{eh}) > 1$, increases as n increases, stabilizes at 1.3333 and $RE(\widehat{\beta}_4, \widehat{\beta}_{el}) \cong 1$ for increasing values of n.

5. Illustration

In this section, we illustrate the performance of $\widehat{\beta}_k$, $k = 1,\ 2,\ 3,\ 4$ through some examples considered in literature. We compute $\widehat{\beta}_*$ and its variance, where $\widehat{\beta}_*$ is any estimator of β. Also, we compute $RE(\widehat{\beta}_4, \widehat{\beta}_*)$. To fit the simple linear regression model given in (1),

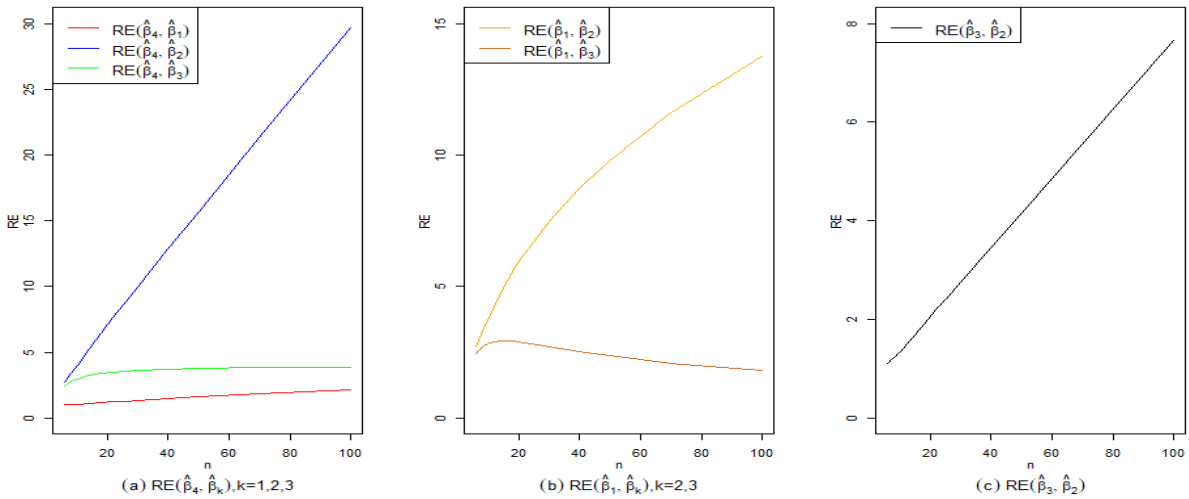


Figure 2: RE among proposed estimators

Table 6: RE of $\widehat{\beta}_4$ wrt $\widehat{\beta}_{el}$ and $\widehat{\beta}_{eh}$

n	$RE(\widehat{\beta}_4, \widehat{\beta}_{el})$	$RE(\widehat{\beta}_4, \widehat{\beta}_{eh})$
6	0.987755	1.280423
8	0.992063	1.302083
10	0.994490	1.312727
14	0.996923	1.322449
18	0.998045	1.326619
22	0.998650	1.328782
26	0.999012	1.330046
30	0.999246	1.330848
40	0.999565	1.331921
50	0.999718	1.332424
70	0.999853	1.332866
100	0.999927	1.333103

the intercept parameter α is estimated using various $\widehat{\beta}_*$,

$$\widehat{\alpha}_* = \overline{y} - \widehat{\beta}_* \overline{x} \tag{18}$$

and

$$\widehat{\alpha}'_* = \widetilde{y} - \widehat{\beta}_* \widetilde{x}, \tag{19}$$

where $\widetilde{x}$, $\widetilde{y}$ are median of x and y values respectively. Using various estimators, the regression lines are fitted.

Example 1: The data due to Anscombe (1973) taken from R software consists of four datasets known as Anscombe's quartet. Here, we consider the data of third quartet given in Table 7.

Using equation (2), (5), (6), (7), (8) and (9), $\widehat{\beta}_{ul}, \widehat{\beta}_{uh}$, $\widehat{\beta}_1$, $\widehat{\beta}_2$, $\widehat{\beta}_3$ and $\widehat{\beta}_4$ and their variances are computed. The relative efficiency of $\widehat{\beta}_4$ wrt other estimators are computed.

Table 7: Third quartet due to Anscombe (1973)

x	10	8	13	9	11	14	6	4	12	7	5
y	7.46	6.77	12.74	7.11	7.81	8.84	6.08	5.39	8.15	6.42	5.73

Table 8: Computed values of $\widehat{\beta}_*$, V($\widehat{\beta}_*$) and $RE(\widehat{\beta}_4, \widehat{\beta}_*)$ for example 1

$\widehat{\beta}_*$	Value of $\widehat{\beta}_*$	V($\widehat{\beta}_*$) (in 10^{-2})	$RE(\widehat{\beta}_4, \widehat{\beta}_*)$
$\widehat{\beta}_{ul}$	0.49972	0.9090 σ^2	1
$\widehat{\beta}_{uh}$	0.48700	1.1111 σ^2	1.22222
$\widehat{\beta}_1$	0.46727	0.9660 σ^2	1.06353
$\widehat{\beta}_2$	0.45175	2.9272 σ^2	3.21994
$\widehat{\beta}_3$	0.46700	2.2448 σ^2	2.46939
$\widehat{\beta}_4$	0.49972	0.9090 σ^2	-

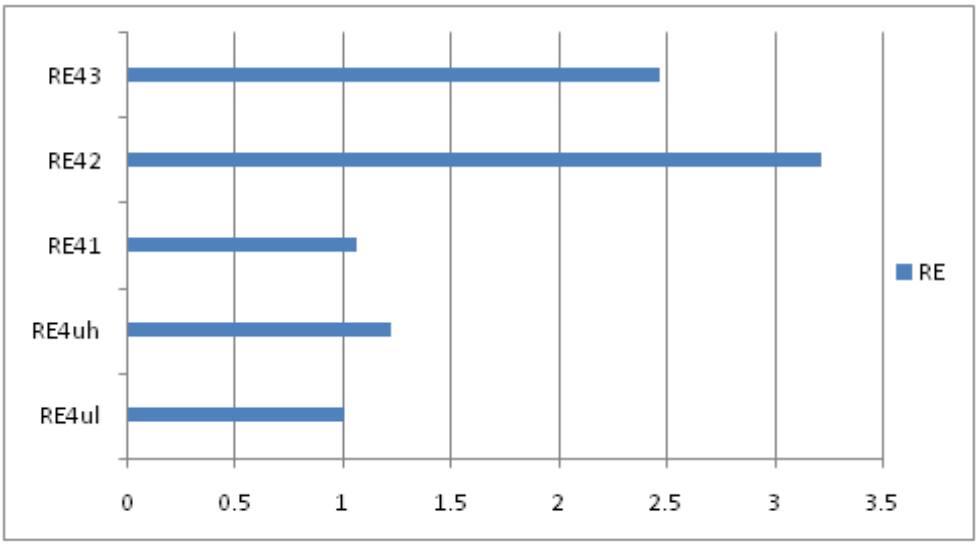


Figure 3: RE of $\widehat{\beta}_4$ wrt $\widehat{\beta}_*$

From Figure 3 and computed V($\widehat{\beta}_*$) given in Table 8, it is observed that, performance of $\widehat{\beta}_4$ and $\widehat{\beta}_{ul}$ are equivalent. Also, $\widehat{\beta}_4$ and $\widehat{\beta}_1$ are better than $\widehat{\beta}_{uh}$. From Figure 4(a) and 4(b), it is observed that, all the regression lines fitted using various $\widehat{\beta}_*$ show slight change in their slopes. In Figure 4(b), as α is estimated using $\widehat{\alpha}'_*$, we see a shift in the intercept and the outlier present in the data has not influenced the regression lines where as the influence of outlier observation is evident in Figure 4(a).

Example 2: This example is due to Montgomery *et al.* (2021) and is given in Table 9. The dataset explains, the shear strength (Y_{i}) of bond between two types of propellant used to manufacture a rocket motor and age in weeks (X_{i}) of the batch of propellant.

From Figure 5 and Table 10, it is observed that, the performance of $\widehat{\beta}_4$ and $\widehat{\beta}_{ul}$ is almost identical. Also, $\widehat{\beta}_1$ and $\widehat{\beta}_4$ are performing better than $\widehat{\beta}_{uh}$. From Figure (6), we observe that, various regression lines fitted using $\widehat{\alpha}_*$ differ in their intercepts than those fitted using $\widehat{\alpha}'_*$. In both cases $\widehat{\beta}_4$ and $\widehat{\beta}_{ul}$ are the lines of best fit.

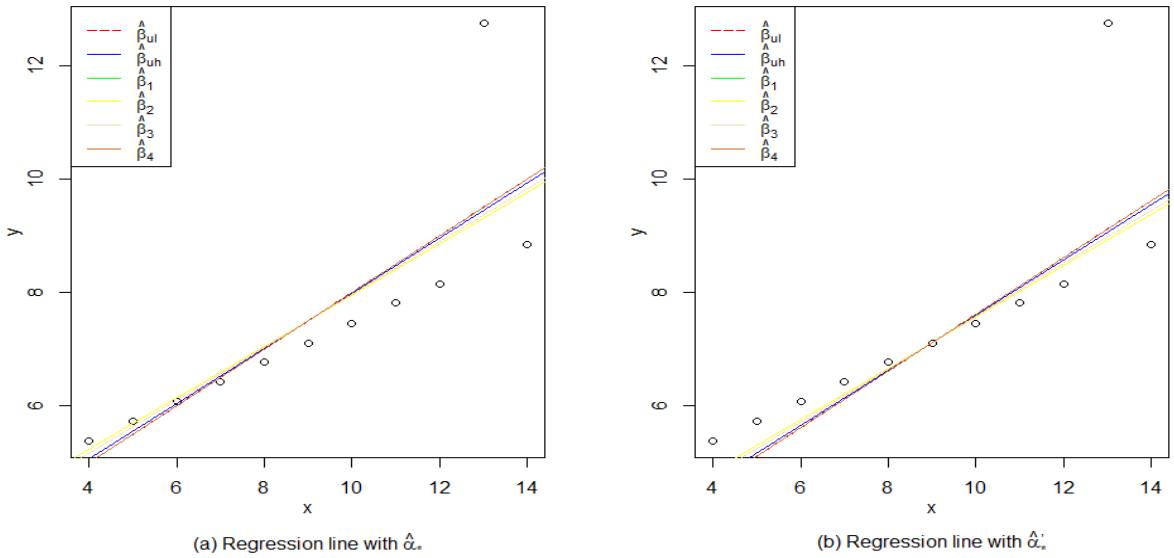


Figure 4: The fitted regression lines using $\widehat{\alpha}_*$ and $\widehat{\alpha}'_*$

Table 9: Data due to Montgomery *et al.* (2021)

Sr. no.	y	x	Sr. no.	y	x
1	2158.7	15.5	11	2165.2	13
2	1678.15	23.75	12	2399.55	3.75
3	2316	8	13	1779.8	25
4	2061.3	17	14	2336.75	9.75
5	2207.5	5.5	15	1765.3	22
6	1708.3	19	16	2353.5	18
7	1784.7	24	17	2414.4	6
8	2575	2.5	18	2200.5	12.5
9	2357.9	7.5	19	2654.2	2
10	2256.7	11	20	1753.7	21.5

Table 10: Computed values of $\widehat{\beta}_*$, $\mathbf{V}(\widehat{\beta}_*)$ and $\boldsymbol{RE}(\widehat{\beta}_4, \widehat{\beta}_*)$ for example 2

$\widehat{\beta}_*$	Value of $\widehat{\beta}_*$	$\mathrm{V}(\widehat{\beta}_*)$ (in 10^{-2})	$RE(\widehat{\beta}_4, \widehat{\beta}_*)$
$\widehat{\beta}_{ul}$	-35.9	0.09037 σ^2	0.994988
$\widehat{\beta}_{uh}$	-34.62457	0.11788 σ^2	1.297978
$\widehat{\beta}_1$	-36.31487	0.11260 σ^2	1.240828
$\widehat{\beta}_2$	-33.28090	0.63480 σ^2	6.989736
$\widehat{\beta}_3$	-32.74453	0.29373 σ^2	3.234026
$\widehat{\beta}_4$	-35.6700	0.09082 σ^2	-

Example 3: The dataset studied by Graybill and Iyer (1994) is considered. The variable y is average systolic blood pressure (BP) at 8 A.M. over two weeks and x is age of individuals ranging 21 to 70 years. The dataset is given in Table 11.

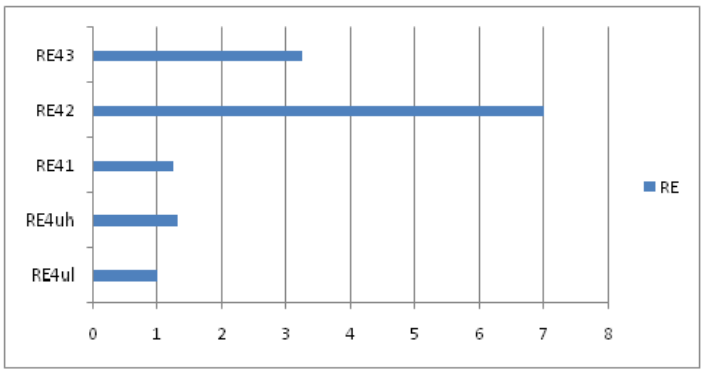
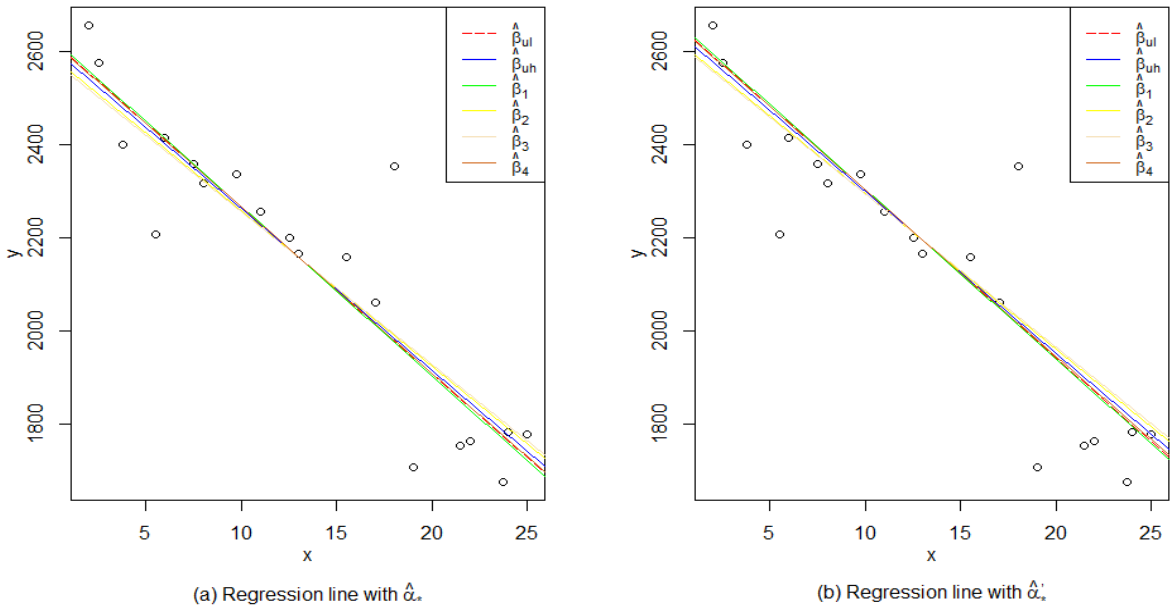


Figure 5: RE of $\widehat{\beta}_4$ wrt $\widehat{\beta}_*$



(a) Regression line with $\hat{\alpha}_*$ (b) Regression line with $\hat{\alpha}'_*$

Figure 6: The fitted regression lines using $\widehat{\alpha}_*$ and $\widehat{\alpha}'_*$

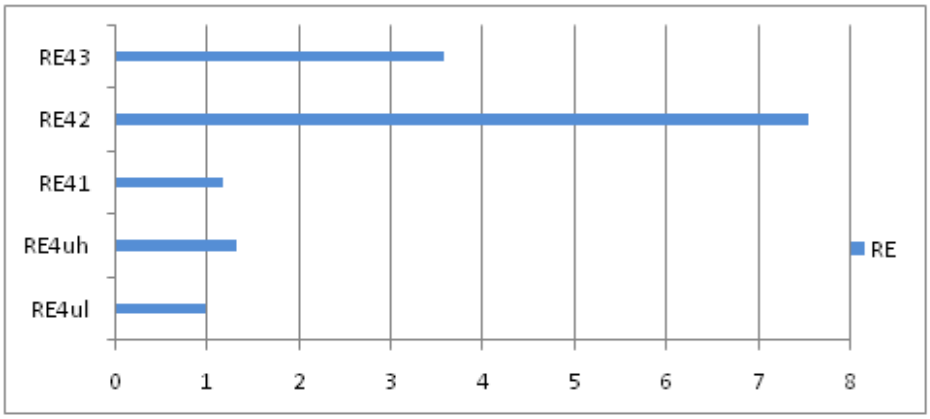


Figure 7: RE of $\widehat{\beta}_4$ wrt $\widehat{\beta}_*$

From computed values of $\widehat{\beta}_*$, $V(\widehat{\beta}_*)$ given in Table 12 and Figure 7, it is observed that, all the values of $\widehat{\beta}_*$ are nearly same. $\widehat{\beta}_4$ performs better than $\widehat{\beta}_1$, $\widehat{\beta}_2$, $\widehat{\beta}_3$, $\widehat{\beta}_{uh}$ and is

Table 11: Data due to Graybill and Iyer (1994)

Sr. no.	x	y	Sr. no.	x	y
1	34	116	13	40	135
2	26	112	14	34	126
3	51	151	15	67	172
4	58	161	16	23	100
5	34	122	17	47	139
6	40	129	18	42	135
7	31	119	19	61	163
8	57	158	20	38	128
9	56	144	21	57	159
10	53	150	22	66	177
11	29	111	23	42	135
12	50	148	24	53	149

Table 12: Computed values of $\widehat{\beta}_*$, $\mathrm{V}(\widehat{\beta}_*)$ and $RE(\widehat{\beta}_4, \widehat{\beta}_*)$ for example 3

$\widehat{\beta}_*$	Value of $\widehat{\beta}_*$	$\mathrm{V}(\widehat{\beta}_*)$ (in 10^{-2})	$RE(\widehat{\beta}_4, \widehat{\beta}_*)$
$\widehat{\beta}_{ul}$	1.60900	0.02771 σ^2	0.98823
$\widehat{\beta}_{uh}$	1.59288	0.03749 σ^2	1.33702
$\widehat{\beta}_1$	1.61958	0.03296 σ^2	1.17546
$\widehat{\beta}_2$	1.59735	0.21211 σ^2	7.56455
$\widehat{\beta}_3$	1.56162	0.10074 σ^2	3.59272
$\widehat{\beta}_4$	1.60938	0.02804 σ^2	-

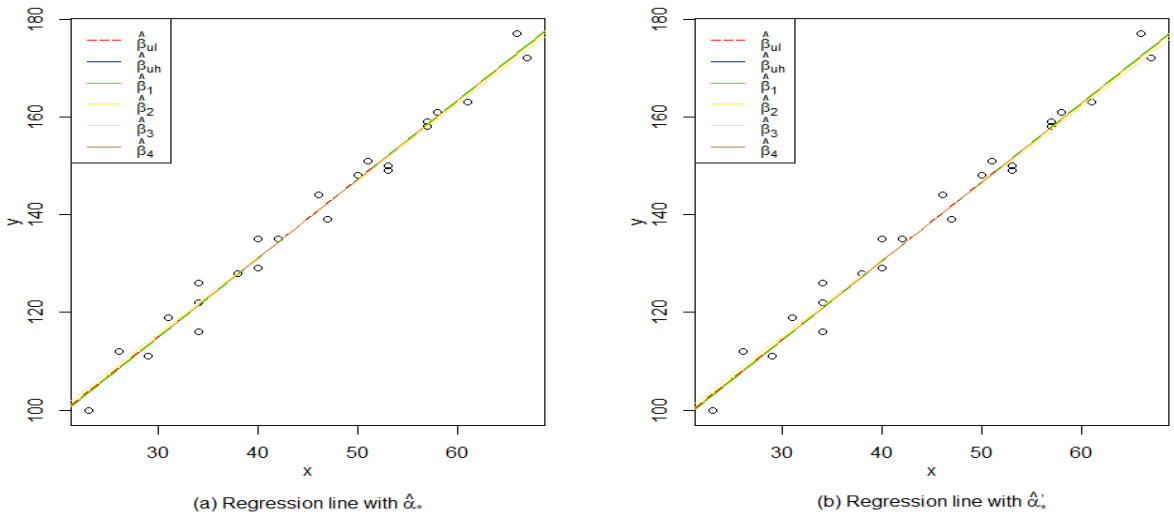


Figure 8: The fitted regression lines using $\widehat{\alpha}_*$ and $\widehat{\alpha}'_*$

almost equivalent to $\widehat{\beta}_{ul}$. Also, $\widehat{\beta}_1$ performs better than $\widehat{\beta}_{uh}$, $\widehat{\beta}_2$ and $\widehat{\beta}_3$. From Figure 8(a) and 8(b), we observe that, all the regression lines plotted using various $\widehat{\beta}_*$, $\widehat{\alpha}_*$ and $\widehat{\alpha}'_*$ are

identical.

Example 4: The data is taken from from `nseindia.com` and `bseindia.com`. It explains daily closing price (x) of index NIFTY50 from National Stock Exchange (NSE) and daily closing price (y) of index SENSEX50 from Bombay Stock Exchange (BSE). The data consists of 988 observations of 4 years from 2017 to 2020. Here, we furnish the values of estimators, their variances and relative efficiencies along with fitting of regression lines using various estimators.

Table 13: Computed values of $\widehat{\beta}_*$, V($\widehat{\beta}_*$) and $RE(\widehat{\beta}_4, \widehat{\beta}_*)$ for example 4

$\widehat{\beta}_*$	Value of $\widehat{\beta}_*$	V($\widehat{\beta}_*$) (in 10^{-10})	$RE(\widehat{\beta}_4, \widehat{\beta}_*)$
$\widehat{\beta}_{ul}$	1.05800	8.593147 σ^2	0.94391
$\widehat{\beta}_{uh}$	1.06240	13.97219 σ^2	1.53477
$\widehat{\beta}_1$	1.00837	38.82940 σ^2	4.26520
$\widehat{\beta}_2$	1.09918	3754.035 σ^2	412.3605
$\widehat{\beta}_3$	1.06689	57.23892 σ^2	6.28739
$\widehat{\beta}_4$	1.06061	9.10377 σ^2	-

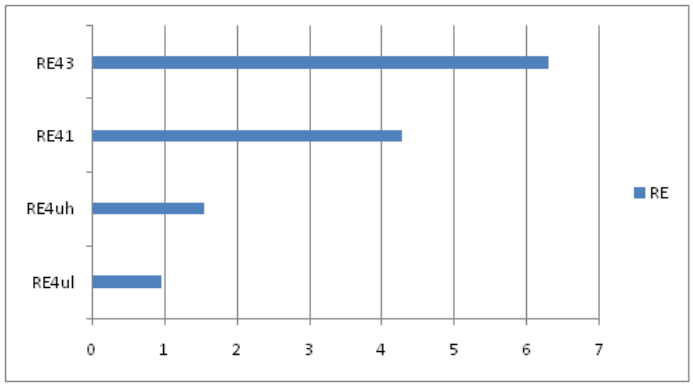


Figure 9: RE of $\widehat{\beta}_4$ wrt $\widehat{\beta}_*$

From Table 13, the computed values of $\widehat{\beta}_*$ and $V(\widehat{\beta}_*)$, $RE(\widehat{\beta}_4, \widehat{\beta}_2)$ is too high to record in Figure 9. The proposed estimator, $\widehat{\beta}_4$ performs better than $\widehat{\beta}_1$, $\widehat{\beta}_3$ and $\widehat{\beta}_{uh}$. From Figure 10, we notice that all the regression lines fitted either using $\widehat{\alpha}_*$ or $\widehat{\alpha}'_*$ seem to be the same as number of observations are very large.

6. Conclusions

- Some estimators based on quasi ranges are proposed for slope parameter of simple linear regression model, $y_i = \alpha + \beta x_i + e_i, \quad i = 1, 2, \cdots, n$.
- Among the proposed estimators, *viz.* $\widehat{\beta}_k$, $k = 1, 2, \cdots, 6$ based on weighted sample quasi ranges, $\widehat{\beta}_5$ reduces to $\widehat{\beta}_3$ and $\widehat{\beta}_6$ reduces to $\widehat{\beta}_4$.
- When equal weights are assigned to each quasi range, all the proposed estimators reduce to $\widehat{\beta}_{ud}$.

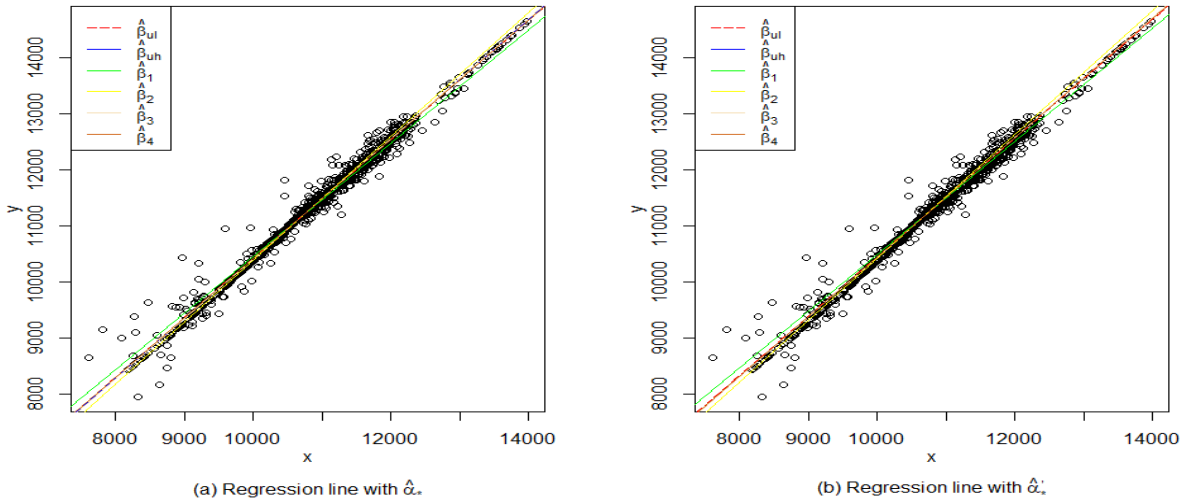


Figure 10: The fitted regression lines using $\hat{\alpha}_*$ and $\hat{\alpha}'_*$

- For equal weights and equidistant $x_{(i)}$'s, all the proposed estimators reduce to $\widehat{\beta}_{eh}$, due to Bose (1938).
- All the proposed estimators are unbiased estimators of slope parameter β.
- The variance of proposed estimators is decreasing with the increasing values of n.
- Among the estimators proposed, $\widehat{\beta}_4$ outperforms $\widehat{\beta}_1$, $\widehat{\beta}_2$, $\widehat{\beta}_3$; $\widehat{\beta}_1$ outperforms $\widehat{\beta}_2$, $\widehat{\beta}_3$ and $\widehat{\beta}_3$ outperforms $\widehat{\beta}_2$.
- $RE\left(\widehat{\beta}_4, \widehat{\beta}_1\right)$, $RE(\widehat{\beta}_4, \widehat{\beta}_2)$, $RE(\widehat{\beta}_4, \widehat{\beta}_3)$, $RE(\widehat{\beta}_1, \widehat{\beta}_2)$ and $RE(\widehat{\beta}_3, \widehat{\beta}_2)$ increase as n increases, but $RE(\widehat{\beta}_1, \widehat{\beta}_3)$ increases upto $n = 14$ and decreases for $n > 14$.
- $\widehat{\beta}_4$ outperforms $\widehat{\beta}_{uh}$, due to Bhat and Bijjargi (2023) and its performance is equivalent to least square estimate $\widehat{\beta}_{ul}$.
- As a_{4i} and a_{1i} assign relatively heavier weights to quasi ranges with extreme order statistics, the estimators $\widehat{\beta}_4$ based on a_{4i} and $\widehat{\beta}_1$ based on a_{1i} are relatively more efficient than other estimators.
- $\widehat{\beta}_2$ based on a_{2i} and $\widehat{\beta}_3$ based on a_{3i} exhibit robustness to outliers if present in the data, since a_{2i} and a_{3i} assign lower weights to quasi ranges with extreme order statistics.

Acknowledgements

Authors thank the unknown referee for his useful comments and second author acknowledges Karnatak University, Dharwad for the award of University Research Scholarship.

References

Anscombe, F. J. (1973). Graphs in statistical analysis. *The American Statistician*, **27**, 17–21.

Bhat, S. V. and Bijjargi, S. M. (2023). Nonparametric approach to estimation in linear regression. *International Journal of Current Research*, **15**, 24419–24424.

Bose, S. S. (1938). Relative efficiencies of regression coefficients estimated by the method of finite differences. *Sankhyā: The Indian Journal of Statistics*, **3**, 339–346.

Cliff, K. R. and Billy, K. M. (2017). Estimation of the parameters of a linear regression system using the simple averaging method. *Global Journal of Pure and Applied Mathematics*, **13**, 7749–7758.

Euler, L. (1749). Recherches sur la question des inégalités du mouvement de saturne et de jupiter. *Pièce qui ont remporté le prix de l'académie royale des sciences*, **3**, 1–123.

Gauss, C. F. (1809). Theoria motus corporum coelestum. *Werke*, **3**, 1–123.

Govindarajulu, Z. (2007). *Nonparametric Inference*. World Scientific.

Graybill, F. A. and Iyer, H. K. (1994). *Regression Analysis: Concepts and Applications*. Duxbury Press.

Hald, A. (1998). *A History of Mathematical Statistics from 1750 to 1930*. Wiley Interscience.

Harter, H. L. (1959). The use of sample quasi-ranges in estimating population standard deviation. *The Annals of Mathematical Statistics*, **30**, 980–999.

Harter, W. L. (1974). The method of least squares and some alternatives: Part i. *International Statistical Review/Revue Internationale de Statistique*, **42**, 147–174.

Jlibene, M., Taoufik, S., and Benjelloun, S. (2021). *Analysis of Least square estimator for simple Linear Regression with a uniform distribution error*. arXiv preprint arXiv:2111.04200 –arxiv.org.

Legendre, A. M. (1805). *Mémoire sur les opérations trigonométriques: dont les résultats dépendent de la figure de la terre*. F. Didot.

Liu, X. and Preve, D. (2016). Measure of location-based estimators in simple linear regression. *Journal of Statistical Computation and Simulation*, **86**, 1771–1784.

Mayer, J. T. (1750). Abhandlung über die umwälzung des monds und seine axe. *Kosmographische Nachrichten un Sammlungen*, **1**, 52–183.

Montgomery, D. C., Peck, E. A., and Vining, G. G. (2021). *Introduction to Linear Regression Analysis*. John Wiley & Sons.

Mosteller, F. (2006). *On Some Useful "Inefficient" Statistics*. Springer.

Nair, K. and Shrivastava, M. (1942). On a simple method of curve fitting. *Sankhyā: The Indian Journal of Statistics*, **6**, 121–132.

Prabowo, A., Sugandha, A., Tripena, A., Mamat, M., Firman, S., and Budiono, R. (2020). A new method to estimate parameters in the simple regression linear equation. *Mathematics and Statistics*, **8**, 75–81.

Singthongchai, j., Thongmual, N., and Nitisuk, N. (2021). An improved simple averaging approach for estimating parameters in simple linear regression model. *Mathematics and Statistics*, **9**, 939–946.

Stigler, S. M. (1986). *The History of Statistics: The Measurement of Uncertainty Before 1900*. Harvard University Press.

Wald, A. (1940). The fitting of straight lines if both variables are subject to error. *The Annals of Mathematical Statistics*, **11**, 284–300.

Yao, D.-S., Chen, W.-X., and Long, C.-X. (2021). Parametric estimation for the simple linear regression model under moving extremes ranked set sampling design. *Applied Mathematics-A Journal of Chinese Universities*, **36**, 269–277.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 22, No. 2, 2024 (New Series), pp 47–61
http://www.ssca.org.in/journal

Comparison of Cause Specific Rate Functions of Panel Count Data with Multiple Modes of Recurrence

Sankaran P. G.[1], **Ashlin Mathew P. M.**[2] **and Sreedevi E. P.**[1]
[1]*Department of Statistics, Cochin University of Science and Technology, Cochin*
[2]*Department of Statistics, St. Thomas College (Autonomous), Thrissur*

Received: 11 September 2022; Revised: 09 November 2023; Accepted: 15 November 2023

Abstract

Panel count data refer to the data arising from studies concerning recurrent events where study subjects are observed only at distinct time points. If these study subjects are exposed to recurrent events of several types, we obtain panel count data with multiple modes of recurrence. In the present paper, we propose a nonparametric test to compare cause specific rate functions of panel count data with more than one mode of recurrence. We carry out simulation studies to evaluate the performance of the test statistic in a finite sample setup. The proposed test is illustrated using two real-life panel count data sets, one arising from a medical follow-up study on skin cancer chemo prevention trial and the other on a warranty database for a fleet of automobiles.

Key words: Cause specific rate functions; Chi-Square test; Kernel estimation; Panel count data; Recurrent events.

AMS Subject Classifications: 62N01, 62N03

1. Introduction

Panel count data arise from longitudinal studies on recurrent events where each subject is observed only at discrete time points. In many situations, continuous observation is impossible due to cost, feasibility or other practical considerations. As a result, the number of occurrence of the events between consecutive observation times are only available; the exact recurrence times remain unknown (Kalbfleisch and Lawless (1985); Sun and Tong (2009); Zhao *et al.* (2011)). Panel count data is also termed interval count data or interval censored recurrent event data (Lawless and Zhan (1998); Thall and Lachin (1988)). In panel count data, the number of observation times and observation time points may vary for each subject. If each subject is observed only once, the number of recurrences of the event up to the observation time is only available. This special case of panel count data is commonly known as current status data.

The standard methods in the analysis of panel count data are focused on the rate function or the mean function of the underlying recurrent event process. Thall and Lachin

Corresponding Author: Sankaran P. G.
Email: sankaran.pg@gmail.com

(1988) and Lawless and Zhan (1998) considered the analysis of panel count data using rate functions. An estimator for the mean function based on isotonic regression theory was developed by Sun and Kalbfleisch (1995). Wellner and Zhang (2000) discussed likelihood based nonparametric estimation methods for the mean function and proposed a nonparametric maximum likelihood estimator (NPMLE) and a nonparametric maximum pseudo-likelihood estimator (NPMPLE) for the same. They also showed that NPMPLE is exactly the one studied in Sun and Kalbfleisch (1995). Some recent research works in this area include Zhou *et al.* (2017), Xu *et al.* (2018), Wang *et al.* (2019), Jiang *et al.* (2020) and Wang and Lin (2020) among others.

When an individual (subject) in the study is exposed to the risk of recurrence due to several types of events at each point of observation, we obtain panel count data with multiple modes of recurrence. Such data naturally arise from survival and reliability studies where the interest is focused on the recurrence of competing events which can be observed only at discrete time points. For example, consider the data on skin cancer chemo prevention trial discussed in Sun and Zhao (2013). The cancer recurrences of 290 patients with a history of non-melanoma skin cancers are observed at different monitoring times. The types of cancers are classified into basal cell carcinoma and squamous cell carcinoma and the recurrences due to both types of cancers at each monitoring time are observed for each individual. Covariate information on age, gender, number of prior tumours and DFMO status is also observed for each individual. As a result, we obtain panel count data with multiple modes of recurrence. A detailed analysis of the data is given in Section 4.

Even though recurrent event data exposed to multiple modes of recurrence is studied by many authors in literature (Cook and Lawless, 2007), panel count data with multiple modes of recurrence is less explored in literature. Sreedevi and Sankaran (2021) derived an expression for the cause specific mean functions and developed a nonparametric test for comparing the effect of different causes on recurrence times based on the developed estimators. Sankaran *et al.* (2020) considered non parametric estimation of cause specific rate functions and studied their properties. When study subjects are exposed to multiple modes of recurrence, it is important to test whether the effect of different modes are identical on the lifetime (Gray (1988)). Many authors including Aly *et al.* (1994) and Sankaran *et al.* (2010) addressed the above testing problem for right censored data. When the current status data is only available, Sreedevi *et al.* (2012) developed a test for independence of time to failure and cause of failure. Comparison of cumulative incidence functions of current status data with continuous and discrete observation times is studied by Sreedevi *et al.* (2014) and Sreedevi *et al.* (2019) respectively. Even though current status data can be considered as a special case of panel count data, the estimation procedures are different for both data types and the aforementioned tests cannot be used in the present situation.

The test proposed by Sreedevi and Sankaran (2021) can be used for comparing the mean functions of panel count data with more than one recurrence mode. However, there are several advantages in using the rate functions for the analysis of panel count data compared to the mean functions. Often, we assume that the mean function follows a non-homogeneous Poisson process, but this assumption is not required for analysing rate functions directly. In addition, rate functions are not constrained by the non decreasing property as of mean functions and hence it is easy to understand the changing recurrence patterns with rate functions. This motivated us to propose a test to compare the cause specific rate functions

proposed by Sankaran *et al.* (2020).

The paper is organized as follows. In Section 2, we discuss the estimation of the cause specific rate functions and then propose a non parametric test to compare the rate functions of panel count data with multiple modes of recurrence. We also discuss the asymptotic properties of the proposed test statistic. In Section 3, we report the results of the simulation study conducted to evaluate the performance of the proposed test in finite samples. We illustrate the practical usefulness of the method by applying it to two real data sets in Section 4. Finally, Section 5 summarizes the major conclusions of the study with a discussion on future works.

2. Inference procedures

We study cause specific rate functions and their properties in detail in this section. Further, a non parametric test for comparing cause specific rate functions is presented.

2.1. Cause specific rate functions

Consider a study on n individuals from a homogeneous population who are exposed to the recurrent events due to $\{1, 2, ..., J\}$ possible causes. Assume that the event process is observed only at a sequence of random monitoring times. Consequently, the counts of the event recurrences due to each cause in between the observation times are only available; the exact recurrence times remain unknown. As a result, we observe the cumulative number of recurrences up to every observation time due to each cause. Define a counting process $N_j = \{N_j(t); t \geq 0\}$ where $N_j(t)$ denotes the number of recurrences of the event due to cause j up to time t. Define $\mu_j(t) = E(N_j(t))$ as the mean function of the recurrent event process due to cause j which are termed as cause specific mean functions. Define $r_j(t)dt = d\mu_j(t) = EdN_j(t)$ as the rate function of the recurrent event process due to cause j, for $j = 1, 2, ..., J$. $r_j(t)$ is referred to as the cause specific rate function. By studying cause specific rate functions, one can easily understand the difference in recurrence patterns due to various causes (modes) of recurrence.

In panel count data, we can note that the number of observation times as well as observation time points may be different for each individual. Let M_i be an integer-valued random variable denoting the number of observation times for $i = 1, 2, .., n$. Also let $T_{i,p}$ denote the p^{th} observation time for i^{th} individual for $p = 1, 2, ..M_i$ and $i = 1, 2, .., n$. Assume that the number of recurrences due to different causes is independent of the number of observation times as well as observation time points. Let $N^j_{i,p}$ denote the number of recurrences of the event observed for i^{th} individual due to cause j , for $p = 1, 2, ..., M_i$, $i = 1, 2, ..., n$ and $j = 1, 2, ..., J$. Now we observe n independent and identically distributed copies of $\{M_i, T_{i,p}, N^1_{i,p}, ..., N^J_{i,p}\}$, $p = 1, 2, ..., M_i$. The observed data will be of the form $\{m_i, t_{i,p}, n^1_{i,p}, ..., n^J_{i,p}\}$, $p = 1, 2, ..., m_i$ and $i = 1, 2, ..., n$.

Sankaran *et al.* (2020) introduced various estimators for cause specific rate functions and established their practical utility through numerical illustrations. The empirical estima-

tors for the cause specific rate functions $r_j(t)$'s are defined as

$$\widehat{r_j(t)} = \frac{\sum_{i=1}^{n}\left[\sum_{p=1}^{m_i}\frac{(n_{i,p}^{j}-n_{i,p-1}^{j})I(t_{i,p}<t\leq t_{i,p-1})}{(t_{i,p}-t_{i,p-1})}\right]}{\sum_{i=1}^{n} I(t\leq t_{i,p})} \qquad j=1,2,...,J. \tag{1}$$

In this definition, the numerator gives the average number of recurrences due to cause j and the denominator is the number of individuals at risk at time t. Hence the estimators $\widehat{r_j(t)}$'s are the average of rate functions due to cause j, over all individuals. The cause specific mean functions can be directly estimated from Equation (1). When $J = 1$, Equation (1) reduces to the empirical estimator of the rate function given in Sun and Zhao (2013) and the expression is given by

$$\widehat{r(t)} = \frac{\sum_{i=1}^{n}\left[\sum_{p=1}^{m_i}\frac{(n_{i,p}-n_{i,p-1})I(t_{i,p}<t\leq t_{i,p-1})}{(t_{i,p}-t_{i,p-1})}\right]}{\sum_{i=1}^{n} I(t\leq t_{i,p})} \tag{2}$$

where $n_{i,p}$ denotes the number of recurrences of the event observed for i^{th} individual due to all possible modes of recurrence up to time p, for $p = 1, 2, ..., M_i$, $i = 1, 2, ..., n$. By definition, $\widehat{r(t)} = \sum_{j=1}^{J}\widehat{r_j(t)}$. In practice, the estimators of cause specific rate functions presented in Equation (1) change only at the observed time points. Accordingly, Sankaran *et al.* (2020) proposed a smoothed version of the estimators of cause specific rate functions using kernel estimation techniques and also studied the asymptotic properties.

Let $K(t)$ be a non-negative kernel function symmetric about $t = 0$ with $\int_{-\infty}^{\infty} K(t)dt = 1$. Also, let $h_n > 0$ be the bandwidth parameter. Let $b_1 < b_2 < ... < b_l$ are the distinct observed time points in the set $\{T_{i,p},\ p = 1, 2, ..., M_i,\ i = 1, 2, ..., n\}$. Define $\widehat{r_{qj}} = \widehat{r_j(b_q)}$, for $q = 1, 2, ..., l$, $j = 1, 2, ..., J$. Now, the kernel estimators of $r_j(t)$'s are given as

$$\widehat{r_j^*(t)} = \sum_{q=1}^{l} w_q(t)\widehat{r_{qj}} \qquad j=1,2,...,J. \tag{3}$$

where

$$w_q(t) = \frac{w_q^*(t,h_n)}{\sum_{u=1}^{l} w_u^*(t,h_n)} \qquad q=1,2,...,l.$$

and

$$w_q^*(t,h_n) = h_n^{-1}K\left(\frac{t-b_q}{h_n}\right)$$

with

$$K(t) = (2\pi)^{-1/2}\exp(-t^2/2).$$

The smoothed estimators $\widehat{r_j^*(t)}$ of the cause specific rate functions are weighted average of $\widehat{r_j(t)}$'s. Smoothed estimators of overall rate functions can also be constructed in a similar way (Sun and Zhao (2013)). Clearly, $\widehat{r^*(t)} = \sum_{j=1}^{J}\widehat{r_j^*(t)}$, where $\widehat{r^*(t)}$ is the kernel estimator of the overall rate function. In practice, the bandwidth h_n for which the MSE is minimum is selected to employ smoothing.

The asymptotic properties of the estimators $\widehat{r_j^*(t)}$'s are studied and derived in Sankaran *et al.* (2020). Without loss of generality, assume that the kernel function $K(x)$ satisfies the following mild regularity conditions.
C1 : $K(x)$ is bounded ie $\sup\{K(x), x \in R\} < \infty$
C2 : $|xK(x)| \to 0$ as $|x| \to \infty$
C3 : $K(x)$ is symmetric about 0, ie $K(-x) = K(x)$, $x \in R$
Also suppose that, as $n \to \infty$ the bandwidth parameter h_n satisfies the conditions (i) $h_n \to 0$ (ii) $nh_n \to \infty$ and (iii) $nh_n^2 \to \infty$.
Under the assumptions C1, C2 and C3, Sankaran *et al.* (2020) showed that for fixed t, the estimators $\widehat{r_j^*(t)}$'s are asymptotically normal with mean $\lambda_j(t) = E(\widehat{r_j^*(t)})$ and standard deviation $\sigma_j(t) = \text{s.d}(\widehat{r_j^*(t)})$ for $j = 1, 2, ..., J$.

2.2. Test statistic

In this study, we focus on comparing the cause specific rate functions due to various recurrence modes. This may be helpful in selecting the appropriate treatment for a group of patients in a clinical study or to evaluate a newly introduced system in reliability experiments. To develop a test statistic, we now consider the hypothesis,

$$H_0 : r_j(t) = r_{j\prime}(t) \text{ for all } t > 0,\ j \neq j\prime = 1, 2, ..., J$$

against

$$H_1 : r_j(t) \neq r_{j\prime}(t) \text{ for some } t > 0 \text{ and } \quad j \neq j\prime = 1, 2, ..., J. \tag{4}$$

Since $r(t) = \sum_{j=1}^{J} r_j(t)$, the above hypothesis can also be written as

$$H_0 : r_j(t) = \frac{r(t)}{J} \text{ for all } t > 0, \quad j \neq j\prime = 1, 2, ..., J$$

against

$$H_1 : r_j(t) \neq \frac{r(t)}{J} \text{ for some } t > 0 \text{ and } \quad j \neq j\prime = 1, 2, ..., J. \tag{5}$$

To test H_0 against H_1, we choose $\widehat{r_j^*(t)}$ as the smoothed estimators for the cause specific rate functions defined in Equation (3). A smoothed estimator for the overall rate function $r(t)$ specified in Equation (2) is constructed by omitting the information on the mode of recurrence. Let $\widehat{r^*(t)}$ denote smoothed estimator of the overall rate function. A similar procedure of estimating the overall mean function by ignoring the cause of recurrence information is used in Sreedevi and Sankaran (2021) for comparing cause specific mean functions.

To develop a test statistic for comparing cause specific rate functions, consider the function

$$v_j(t) = \int_0^t w(u) \left[\widehat{r_j^*(u)} - \frac{\widehat{r^*(u)}}{J} \right] \mathrm{d}u \quad \text{for all } j = 1, 2, \ldots, J \tag{6}$$

where $w(.)$ is an appropriate data dependent weight function which is used to compensate the effect of censoring. The weight functions are also employed to increase the efficiency of

the test statistic and to set it asymptotically distribution free (Pepe and Mori (1993)). The function $v_j(.)$ is similar to the one proposed by Sreedevi and Sankaran (2021) to compare the cause specific mean functions of panel count data. Now to test the null hypothesis given in Equation (4), we propose the test statistic

$$Z(\tau) = v'(\tau)\hat{\sum}(\tau)^{-1}v(\tau) \tag{7}$$

where τ is the largest monitoring time in the study and $v(\tau) = (v_1(\tau), \ldots, v_k(\tau))'$; $\sum(\tau)^{-1}$ is the generalized inverse $\hat{\sum}(\tau)$, where $\hat{\sum}(\tau)$ is a consistent estimator of $\sum(\tau)$, the variance-covariance matrix of $v(\tau)$. The matrix $\sum(\tau)$ involves variances of $\widehat{r_j^*(\tau)}$ and $\widehat{r(\tau)}$ and covariances between $\widehat{r_j^*(\tau)}$ and $\widehat{r_{j'}^*(\tau)}$ for $j \neq j' = 1, 2, \ldots, J$ and that between $\widehat{r_j^*(\tau)}$ and $\widehat{r(\tau)}$. The bootstrap procedure is used to find the estimate of the variance-covariance matrix, since the expression for $\sum(\tau)$ is complex. To find the asymptotic distribution of $Z(\tau)$ given in Equation(7), consider the quantity

$$v_j(t) = \int_0^t w(u)\left[\widehat{r_j^*(u)} - \frac{\widehat{r^*(u)}}{J}\right]\mathrm{d}u \quad \text{for all } j = 1, 2, \ldots, J$$

which can be written as

$$\begin{aligned} v_j(t) = &\int_0^t w(u)\left[\widehat{r_j^*(u)} - r_j(u)\right]\mathrm{d}(u) + \int_0^t w(u)\left[r_j(u) - \frac{r(u)}{J}\right]\mathrm{d}u \\ &+ \int_0^t w(u)\left[\frac{r(u)}{J} - \frac{\widehat{r^*(u)}}{J}\right]\mathrm{d}u, \quad j = 1, 2, \ldots J \end{aligned}$$

Now under H_0, $r_j(t) = r(t)/J$ for all t, we get

$$v_j(t) = \int_0^t w(u)\left[\widehat{r_j^*(u)} - r_j(u)\right]\mathrm{d}u + \int_0^t w(u)\left[\frac{r(u)}{J} - \frac{\widehat{r^*(u)}}{J}\right]\mathrm{d}u, \quad j = 1, 2, \ldots, J$$

Now from the asymptotic properties of the kernel estimators of cause specific rate functions discussed in Sankaran *et al.* (2020) it follows that, under H_0 for any $\tau > 0$, the limiting distribution of $v(\tau) = (v_1(\tau), \ldots, v_J(\tau))'$ is a $J-$ variate normal with mean zero vector and variance-covariance matrix $\sum(\tau)$, where τ is the largest monitoring time in the study. Accordingly, under the regularity conditions stated in Section 2.1, the quadratic form $Z(\tau)$ follows a χ^2 distribution with $(J-1)$ degrees of freedom. We reject H_0, if $Z(t) \geq \chi^2_{\alpha,(J-1)}$ where $\chi^2_{\alpha,(J-1)}$ is the ordinate value of chi-square distribution with $(J-1)$ degrees of freedom at α level.

3. Simulation studies

We conduct simulation studies to evaluate the performance of the proposed test statistic in finite samples. The situation with two modes of recurrence is considered here. We generate panel count data of the form $\{m_i, t_{i,p}, n^1_{i,p}, n^2_{i,p}\}$ for $p = 1, 2, ..., m_i$ and $i = 1, 2, ..., n$ to carry out simulation. The number of observation times m_i for each individual is generated

Table 1: Empirical Type I error and power of the test in percentage for the weight functions $w(.) = 1, w(.) = n$ and $w(.) = \widehat{r^*(t)}$

		n					n		
$(\theta_1, \theta_2, \theta_3)$	α	100	200	500	$(\theta_1, \theta_2, \theta_3)$	α	100	200	500
				$w(t) = 1$					
(1,1,1)	5	5.8	5.4	5.1	(1,1,2)	5	5.6	5.2	4.9
	1	2	1.7	1.3		1	1.7	1.4	1.1
(1,2,1)	5	65.8	71.4	79.5	(1,2,2)	5	66.8	74.8	80.7
	1	63.7	67.2	73.1		1	65.2	73.1	75.2
(1,3,1)	5	74.5	81.9	86.4	(1,3,2)	5	81.5	87.7	92.4
	1	73.0	78.6	83.1		1	79.4	85.6	91.6
(1,4,1)	5	90.3	92.1	97.2	(1,4,2)	5	96.5	98.2	99.9
	1	87.4	91.8	94.5		1	96.8	98.2	99.1
(1,5,1)	5	98.9	100	100	(1,5,2)	5	100	100	100
	1	98.4	99.7	100		1	99.8	100	100
				$w(t) = n$					
(1,1,1)	5	4.5	4.7	5.2	(1,1,2)	5	4.4	4.8	5.1
	1	2	1.7	1.3		1	1.4	1.3	0.9
(1,2,1)	5	67.1	73.2	78.4	(1,2,2)	5	70.4	79.5	84.7
	1	66.7	69.2	74.1		1	68.1	74	79
(1,3,1)	5	79.6	83.9	86.4	(1,3,2)	5	85.2	89.3	94.7
	1	73.0	78.6	83.1		1	80.5	87.2	93.7
(1,4,1)	5	94.3	98.1	99.9	(1,4,2)	5	99.9	100	100
	1	87.4	96.8	97.2		1	99.8	99.9	100
(1,5,1)	5	100	100	100	(1,5,2)	5	100	100	100
	1	100	100	100		1	99.8	100	100
				$w(t) = \widehat{r^*(t)}$					
(1,1,1)	5	4.7	5.2	5	(1,1,2)	5	5.5	4.8	5.1
	1	0.7	1.2	0.9		1	1.3	1.2	1
(1,2,1)	5	73.2	81.0	85.7	(1,2,2)	5	76.9	84.1	85.4
	1	71.1	78.9	84.3		1	71.0	77.2	84.2
(1,3,1)	5	89.5	92.5	98.4	(1,3,2)	5	88.8	91.4	97.5
	1	83.2	88.6	96.9		1	85.0	87.3	96.0
(1,4,1)	5	99.9	100	100	(1,4,2)	5	100	100	100
	1	99.7	100	100		1	99.8	99.8	100
(1,5,1)	5	100	100	100	(1,5,2)	5	100	100	100
	1	100	100	100		1	100	100	100

from a discrete uniform distribution $U(1,10)$ for $i = 1, 2, ..., n$. Thus the maximum number of observations for each individual is restricted up to 10. Then we generate gap times between each observation from uniform distribution $U(0,5)$. The discrete observation time points $t_{i,p}$ for $p = 1, 2, ..., m_i$ and $i = 1, 2, ..., n$ are generated using the above-mentioned time gaps. A bivariate Poisson distribution with parameters $(\theta_1, \theta_2, \theta_3)$ is employed to generate recurrent processes $n^1_{i,p}$ and $n^2_{i,p}$.

The joint mass function of the bivariate Poisson distribution with parameters $(\theta_1, \theta_2, \theta_3)$ is given by

$$f(x,y) = \exp\{-(\theta_1+\theta_2+\theta_3)\}\frac{{\theta_1}^x}{x!}\frac{{\theta_2}^y}{y!}\sum_{k=0}^{min(x,y)}\binom{x}{k}\binom{y}{k}k!\left(\frac{\theta_3}{\theta_1\theta_2}\right)^k. \tag{8}$$

The marginal distribution of X and Y is Poisson distribution with $E(X) = \theta_1 + \theta_3$, $E(Y) = \theta_2 + \theta_3$ and $\text{cov}(X,Y) = \theta_3$ gives a measure of dependence between random variables X and Y. Sankaran *et al.* (2020) used a similar procedure to generate panel count data with multiple modes of recurrence.

In the above simulation framework, if we set $\theta_1 = \theta_2$ and assign a non-zero value for θ_3, it corresponds to a situation where the cause specific rate functions are identical. Accordingly, the null hypothesis H_0 will be true. When the difference between θ_1 and θ_2 increases, the difference between the two rate functions also increases which results in a situation where the null hypothesis is false. Hence the parameter combination with $\theta_1 = \theta_2$ gives the type I error of the test and all other choices of parameter combinations give the power of the proposed test. We carry out simulation studies for different combinations of $(\theta_1, \theta_2, \theta_3)$ to calculate the empirical type I error and power of the test. For this purpose, observations of different sample sizes n = 100 or n = 200 or n = 500 are simulated and the process is repeated 1000 times. We employ three different choices of weight functions similar to Sreedevi and Sankaran (2021) which are (i) $w(t) = 1$, (ii) $w(t) = n$, the number of individuals in the study and (iii) $w(t) = \widehat{r^*(t)}$, the smoothed estimator of overall rate function.

Table 1 gives the type I error and the power of the proposed test statistic in percentage for significance level $\alpha = 0.05$ and $\alpha = 0.01$. From Table 1, we can see that type I error of the test approaches the chosen significance level. The test is efficient in terms of power also. Also, as the difference between θ_1 and θ_2 increases, the power of the test also increases.

4. Data analysis

The proposed inference procedures are illustrated using two real-life data sets in this section.

4.1. Skin cancer chemo prevention trial data

We consider the data arising from the skin cancer chemo prevention trial given in Sun and Zhao (2013) for demonstration. The study was conducted to test the effectiveness of the DFMO (DIfluromethylornithire) drug in reducing new skin cancers in a population with a history of non-melanoma skin cancers, basal cell carcinoma and squamous cell carcinoma.

The data consists of 290 patients with a history of non-melanoma skin cancers. The observation and follow-up times differ for each patient. The data has the counts of two types of recurring events basal cell carcinoma and squamous cell carcinoma which we treat here as two modes of recurrence (Sreedevi and Sankaran (2021)).

Table 2: Test statistic values of the proposed test for different weight functions

Weight function	Test statistic	p -value
1	26.97	$< .0005$
n	31.92	$< .0005$
$\widehat{r^*(.)}$	37.74	$< .0005$

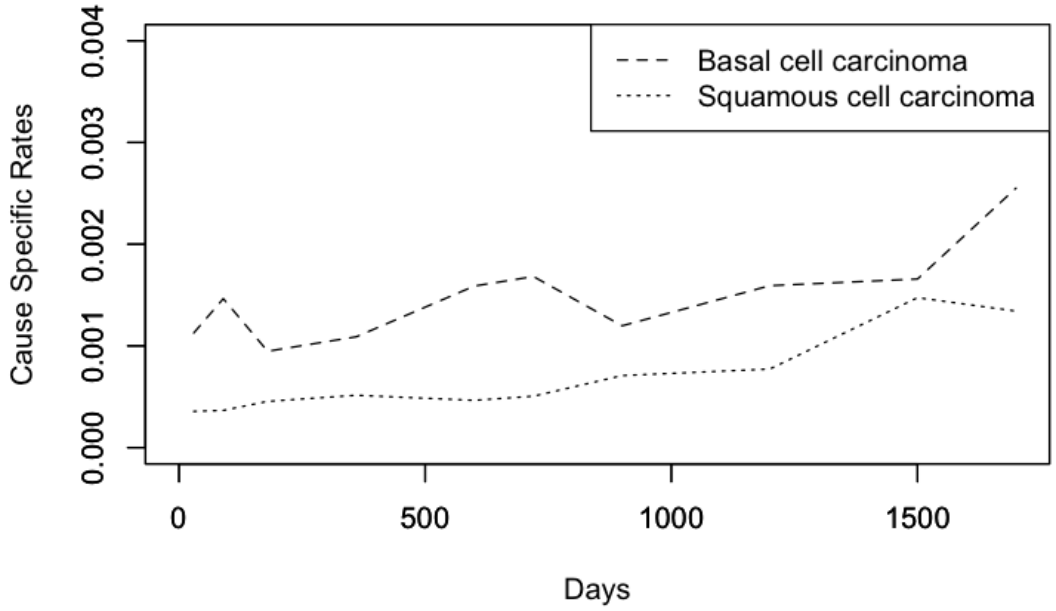


Figure 1: Kernel estimates of cause specific rate functions due to basal cell carcinoma and squamous cell carcinoma for $h_n = 1.76$

In the data set, the number of observations on an individual varies from 1 to 17 and the time of observation varies from 12 to 1766 days. The cause specific rate functions due to basal cell carcinoma and squamous cell carcinoma are estimated using Equation (3). Further, the proposed procedures are applied to evaluate the test statistic. Table 2 gives the chi-square test statistic values of the proposed test for different weight functions. From the value of the test statistic, it is clear that we reject the null hypothesis and conclude that the rate functions due to basal cell carcinoma and squamous cell carcinoma are significantly different.

The plots of the kernel estimators with bandwidth parameter value $h_n = 1.76$ is given in Figure 1. The bandwidth value $h_n = 1.76$ is chosen, which minimizes the MSE of the

estimates, $\widehat{r_j^*(t)}$ for $j = 1, 2$.

From Figure 1, it can be noted that the recurrence rate of basal cell carcinoma is greater than the recurrence rate of squamous cell carcinoma at all time points, which clearly indicates the rejection of H_0. Since the rate functions are not monotonic, the change points of recurrence patterns can also be easily identified from the graph.

4.2. Automobile warranty claims data

We apply the proposed methods to the automobile warranty claims data studied in Somboonsavatdee and Sen (2015). The data set comprises the recurrent failure history of a fleet of automobiles. The outcome of interest is the repeated mileages at failure for multiple vehicles of a certain model and make, obtained from the warranty claim database which also includes the labour code associated with the failure. In the data, the source and specifics are masked for de-identification purposes.

Table 3: Test statistic values of the proposed test for different weight functions

Weight function	Test statistic	p -value
1	49.15	$< .0005$
n	68.96	$< .0005$
$\widehat{r^*(.)}$	79.55	$< .0005$

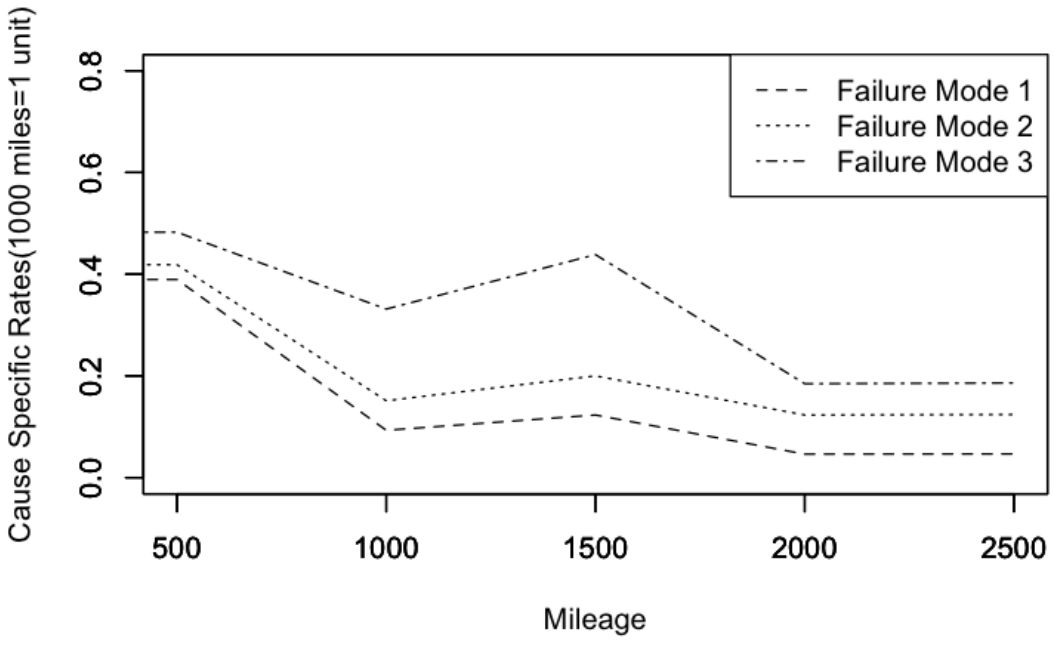


Figure 2: Kernel estimates of cause specific rate functions due to three modes of recurrences for $h_n = 1.67$

The database consists of the recurrent failure history of 456 vehicles subjected to type I censoring at 3000 miles. Fourteen different labor codes of the warranty claims of each vehicle were recorded with mileage at filing. Due to the absence of a specific description of the component associated with labor code, the grouping was determined on the basis of the rate of failures. The fourteen individual labor codes were combined into three broad groups of failure modes FM1, FM2 and FM3, where FM1 comprises labor codes with shape parameters ranging between 0.2 and 0.36, FM2 covers labor codes with shape parameter estimates between 0.4 and 0.55, whereas FM3 combines the remaining codes that have the slowest rate of growth with shape parameter estimates varying between 0.7 and 0.93. Table IV in Somboonsavatdee and Sen (2015) presents the data on 172 vehicles that have at least one documented record of warranty claim for repair.

We observed the recurrent failure history data at 1000, 2000 and 3000 mileages at which the number of failures due to each mode is noted, thereby making the recurrent event data as a panel count data with multiple modes of recurrence. The complete data set used in our study is given in Table 4 in Appendix.

Table 3 gives the chi-square test statistic values of the proposed test for different weight functions for automobile warranty data. For all choices of weight functions, we reject the null hypothesis and conclude that the rate functions due to the three modes of failure are significantly different. The plots of the kernel estimators with bandwidth parameter value $h_n = 1.67$ is given in Figure 2. The bandwidth value $h_n = 1.67$ is chosen as it minimizes the MSE of the estimates. From Figure 2, it can be noted that the recurrence rates of each mode of recurrence (FM1, FM2 and FM3) are distinct at all observed miles, which clearly indicates the rejection of H_0.

5. Conclusion

In the present paper, we developed non parametric inference procedures for the analysis of panel count data with multiple modes of recurrence based on cause specific rate functions. We proposed a test statistic to test the equality of cause specific rate functions. A simulation study was carried out by generating the data from a bivariate Poisson process to assess the performance of the proposed test in finite samples. Two real-life data sets, one from skin cancer chemo prevention trial (Sun and Zhao (2013)) and the other from automobile warranty claims (Somboonsavatdee and Sen (2015)) were analysed to demonstrate the practical utility of the procedures.

The nature of dependence between time to failure and cause of failure is important for modelling competing risks data. Even though the problem is studied under right censoring, it is unexplored for panel count data. We can use either cause specific mean functions or cause specific rate functions to tackle this problem. Works in this direction will be done separately. Regression analysis of panel count data with multiple modes of recurrence using rate functions is also under investigation.

Acknowledgements

The first author would like to thank Science Engineering and Research Board, DST, Government of India and the third author acknowledges the gratitude to Kerala State Council for Science Technology and Environment for the financial support provided to carry out this research work.

References

Aly, E.-E. A., Kochar, S. C., and McKeague, I. W. (1994). Some tests for comparing cumulative incidence functions and cause-specific hazard rates. *Journal of the American Statistical Association*, **89**, 994–999.

Gray, R. J. (1988). A class of k-sample tests for comparing the cumulative incidence of a competing risk. *The Annals of Statistics*, **16**, 1141–1154.

Jiang, H., Su, W., and Zhao, X. (2020). Robust estimation for panel count data with informative observation times and censoring times. *Lifetime Data Analysis*, **26**, 65–84.

Kalbfleisch, J. D. and Lawless, J. F. (1985). The analysis of panel data under a markov assumption. *Journal of the American Statistical Association*, **80**, 863–871.

Lawless, J. F. and Zhan, M. (1998). Analysis of interval-grouped recurrent-event data using piecewise constant rate functions. *Canadian Journal of Statistics*, **26**, 549–565.

Pepe, M. S. and Mori, M. (1993). Kaplan-Meier, marginal or conditional probability curves in summarizing competing risks failure time data? *Statistics in Medicine*, **12**, 737–751.

Sankaran, P. G., Ashlin Mathew, P. M., and Sreedevi, E. P. (2020). Cause specific rate functions for panel count data with multiple modes of recurrence. *Journal of the Indian Statistical Association*, **58**, 193–211.

Sankaran, P. G., Nair, N. U., and Sreedevi, E. P. (2010). A quantile based test for comparing cumulative incidence functions of competing risks models. *Statistics & Probability Letters*, **80**, 886–891.

Somboonsavatdee, A. and Sen, A. (2015). Parametric inference for multiple repairable systems under dependent competing risks. *Applied Stochastic Models in Business and Industry*, **31**, 706–720.

Sreedevi, E. P. and Sankaran, P. G. (2021). Nonparametric inference for panel count data with competing risks. *Journal of Applied Statistics*, **48**, 3102–3115.

Sreedevi, E. P., Sankaran, P. G., and Dewan, I. (2019). Comparison of cumulative incidence functions of current status competing risks data with discrete observation times. *Communications in Statistics-Theory and Methods*, **48**, 5766–5776.

Sreedevi, E. P., Sankaran, P. G., and Dhanavanthan, P. (2012). A nonparametric test for independence of time to failure and cause of failure of current status competing risks data. *Calcutta Statistical Association Bulletin*, **64**, 167–180.

Sreedevi, E. P., Sankaran, P. G., and Dhanavanthan, P. (2014). A nonparametric test for comparing cumulative incidence functions of current status competing risks data. *Journal of Statistical Theory and Practice*, **8**, 743–759.

Sun, J. and Kalbfleisch, J. (1995). Estimation of the mean function of point processes based on panel count data. *Statistica Sinica*, **5**, 279–289.

Sun, J. and Zhao (2013). *Statistical Analysis of Panel Count Data*. Springer.

Sun, L. and Tong, X. (2009). Analyzing longitudinal data with informative observation times under biased sampling. *Statistics & Probability Letters*, **79**, 1162–1168.

Thall, P. F. and Lachin, J. M. (1988). Analysis of recurrent events: Nonparametric methods for random-interval count data. *Journal of the American Statistical Association*, **83**, 339–347.

Wang, J. and Lin, X. (2020). A Bayesian approach for semiparametric regression analysis of panel count data. *Lifetime Data Analysis*, **26**, 402–420.

Wang, W., Wu, X., Zhao, X., and Zhou, X. (2019). Quantile estimation of partially varying coefficient model for panel count data with informative observation times. *Journal of Nonparametric Statistics*, **31**, 932–951.

Wellner, J. A. and Zhang, Y. (2000). Two estimators of the mean of a counting process with panel count data. *The Annals of Statistics*, **28**, 779–814.

Xu, D., Zhao, H., and Sun, J. (2018). Joint analysis of interval-censored failure time data and panel count data. *Lifetime Data Analysis*, **24**, 94–109.

Zhao, X., Balakrishnan, N., and Sun, J. (2011). Nonparametric inference based on panel count data. *Test*, **20**, 1–42.

Zhou, J., Zhang, H., Sun, L., and Sun, J. (2017). Joint analysis of panel count data with an informative observation process and a dependent terminal event. *Lifetime Data Analysis*, **23**, 560–584.

ANNEXURE

Table 4: Automobile warranty data

ID	MIL	FM1	FM2	FM3	TOTAL	ID	MIL	FM1	FM2	FM3	TOTAL
1	1000	1	1	0	2	**45**	1000	2	0	0	2
1	3000	1	0	0	1	**46**	1000	0	1	0	1
2	1000	0	0	2	2	**47**	1000	1	0	0	1
3	3000	0	0	1	1	**47**	3000	0	1	0	1
4	2000	0	0	1	1	**48**	1000	1	1	0	2
5	1000	1	1	1	3	**49**	1000	0	1	0	1
5	2000	1	0	0	1	**50**	1000	0	0	1	1
6	1000	0	0	1	1	**51**	3000	0	0	1	1
7	1000	1	0	0	1	**52**	1000	0	0	1	1
8	1000	1	0	0	1	**53**	2000	1	0	0	1
9	1000	0	1	0	1	**54**	1000	0	1	0	1
10	2000	0	0	2	2	**55**	1000	1	0	0	1
11	1000	1	0	0	1	**56**	1000	0	1	0	1
12	1000	1	0	0	1	**57**	1000	0	2	0	2
13	3000	0	0	1	1	**57**	2000	1	0	1	2
14	1000	0	1	1	2	**58**	1000	0	0	1	1
15	1000	0	1	0	1	**59**	1000	0	1	0	1
15	2000	0	1	0	1	**60**	1000	0	1	0	1
16	2000	0	1	1	2	**61**	2000	1	0	0	1
16	3000	0	1	0	1	**62**	1000	0	1	0	1
17	1000	1	2	1	4	**63**	2000	0	0	1	1
17	2000	1	0	0	1	**64**	1000	0	0	1	1
18	3000	0	0	1	1	**65**	1000	1	0	0	1
19	1000	0	1	0	1	**66**	1000	2	1	0	3
20	1000	1	0	0	1	**67**	1000	1	0	0	1
21	1000	0	1	0	1	**67**	3000	0	0	1	1
22	3000	0	1	0	1	**68**	1000	0	1	0	1
23	1000	1	0	0	1	**69**	2000	0	1	0	1
24	3000	1	0	0	1	**70**	1000	1	0	0	1
25	1000	0	1	0	1	**71**	1000	1	0	0	1
26	1000	1	0	1	2	**72**	2000	0	0	2	2
26	2000	1	2	0	3	**73**	1000	1	0	0	1
26	3000	0	2	0	2	**73**	2000	0	0	1	1
27	3000	0	1	0	1	**74**	1000	1	0	1	2
28	2000	0	0	1	1	**74**	2000	0	0	1	1
29	1000	1	0	1	2	**75**	1000	1	0	0	1
30	3000	0	2	0	2	**76**	1000	0	0	1	1
31	2000	0	1	0	1	**77**	1000	0	1	1	2
32	2000	0	1	0	1	**78**	1000	0	1	0	1
33	3000	0	0	1	1	**79**	3000	0	0	1	1
34	1000	0	1	0	1	**80**	1000	1	0	0	1
35	1000	0	0	1	1	**81**	1000	0	0	1	1
35	2000	1	0	0	1	**82**	1000	1	0	0	1
36	1000	0	1	0	1	**83**	1000	0	0	1	1
37	1000	1	0	0	1	**84**	2000	0	0	1	1
37	2000	0	0	1	1	**85**	1000	0	2	0	2
38	1000	1	1	0	2	**86**	1000	0	0	1	1
39	1000	0	2	0	2	**86**	2000	0	2	0	2
40	1000	0	2	0	2	**87**	1000	1	0	0	1
41	3000	0	0	1	1	**88**	2000	0	0	1	1
42	1000	0	0	1	1	**88**	3000	0	0	1	1
43	1000	0	0	1	1	**89**	3000	1	0	0	1
44	3000	0	1	0	1	**90**	1000	0	0	2	2

ID	MIL	FM1	FM2	FM3	TOTAL	ID	MIL	FM1	FM2	FM3	TOTAL
90	3000	0	0	1	1	**135**	1000	0	0	1	1
91	1000	0	1	0	1	**136**	1000	0	0	1	1
92	1000	0	1	0	1	**137**	1000	0	0	1	1
93	1000	0	0	1	1	**138**	1000	0	0	1	1
94	1000	1	1	0	2	**138**	3000	1	0	0	1
95	1000	1	0	0	1	**139**	1000	1	0	0	1
96	2000	0	0	1	1	**140**	1000	1	0	0	1
97	2000	0	0	1	1	**141**	3000	0	1	0	1
98	1000	0	1	0	1	**142**	1000	0	1	1	2
98	2000	1	1	1	3	**143**	1000	1	0	0	1
99	1000	1	0	0	1	**143**	3000	0	0	1	1
100	1000	1	0	1	2	**144**	1000	0	1	0	1
101	1000	0	0	1	1	**144**	2000	0	0	2	2
102	1000	1	0	0	1	**145**	1000	0	1	0	1
103	1000	1	0	0	1	**146**	1000	1	0	1	2
104	2000	0	0	1	1	**146**	3000	0	0	1	1
105	1000	1	0	0	1	**147**	1000	0	1	0	1
106	1000	0	0	2	2	**148**	3000	0	0	1	1
107	3000	0	1	0	1	**149**	1000	1	0	0	1
108	1000	1	0	0	1	**150**	1000	1	0	0	1
108	3000	0	0	1	1	**151**	1000	0	0	1	1
109	2000	0	0	1	1	**152**	1000	1	0	0	1
109	3000	0	1	0	1	**153**	1000	0	1	0	1
110	1000	1	0	1	2	**154**	3000	1	0	0	1
111	1000	1	0	0	1	**155**	1000	0	1	0	1
112	1000	0	1	0	1	**156**	3000	0	1	0	1
113	1000	0	0	1	1	**157**	2000	0	0	1	1
114	1000	1	0	0	1	**158**	3000	0	0	1	1
115	1000	0	1	1	2	**159**	1000	0	0	1	1
116	1000	1	0	0	1	**160**	3000	0	0	1	1
117	2000	0	1	1	2	**161**	1000	0	1	2	3
118	2000	0	0	1	1	**161**	2000	0	1	2	3
119	2000	1	0	0	1	**161**	3000	1	0	2	3
120	1000	1	0	1	2	**162**	2000	1	0	0	1
121	1000	0	0	1	1	**163**	1000	0	1	0	1
121	3000	0	0	1	1	**164**	3000	0	0	1	1
122	1000	1	0	1	2	**165**	1000	0	2	2	4
123	2000	0	1	0	1	**165**	2000	0	1	1	2
124	1000	1	0	0	1	**165**	3000	0	1	1	2
125	2000	0	0	1	1	**166**	1000	1	0	1	2
126	1000	2	0	1	3	**167**	1000	0	1	0	1
126	3000	0	0	2	2	**167**	3000	0	1	0	1
127	2000	0	0	1	1	**168**	1000	0	1	0	1
128	2000	0	1	0	1	**169**	1000	1	0	0	1
129	1000	2	3	1	6	**169**	2000	0	0	4	4
129	2000	0	0	1	1	**169**	3000	0	1	0	1
130	1000	0	1	0	1	**170**	1000	0	1	0	1
131	1000	1	0	0	1	**170**	2000	0	0	1	1
132	3000	0	0	1	1	**171**	1000	0	0	1	1
133	2000	1	0	1	2	**172**	2000	0	0	1	1
134	2000	0	1	1	2						

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 22, No. 2, 2024 (New Series), pp 63–90
http://www.ssca.org.in/journal

Price Forecasting of TOP (Tomato, Onion and Potato) Commodities using Hidden Markov-based Deep Learning Approach

G. Avinash[1], Ramasubramanian V.[2], Ranjit Kumar Paul[3], Mrinmoy Ray[3], Shashi Dahiya[3], Mir Asif Iquebal[3], Samarth Godara[3] and B. Manjunatha[1]
[1]*The Graduate School, ICAR-Indian Agricultural Research Institute, New Delhi*
[2]*ICAR-National Academy of Agricultural Research Management, Hyderabad*
[3]*ICAR-Indian Agricultural Statistics Research Institute, New Delhi*

Received: 08 October 2023; Revised: 24 November 2023; Accepted: 28 November 2023

Abstract

Accurate prediction of agricultural prices is crucial due to their complex and nonlinear nature. Due to the perishable nature of TOP (Tomato, Onion and Potato) vegetable produce, price fluctuates based on supply and demand. It is necessary to forecast harvest period TOP prices, so growers can make informed production decisions and also farmers can plan their market situation to enhance their profits. This research introduces novel Deep Learning (DL) models based on hidden states to enhance the precision of TOP price forecasting. The Hidden Markov Model (HMM) is employed to identify hidden states and uncover underlying patterns in TOP price data. The hidden states identified by HMM serve as a feature extraction technique and are utilized in four DL models, *viz.*, Multilayer Perceptron (MLP), Recurrent Neural Networks (RNN), Gated Recurrent Units (GRUs) and Long Short-Term Memory (LSTM). The integration of HMM with DL aims to improve forecasting accuracy compared to HMM and traditional DL models. The models are evaluated using a real dataset from Azadpur Mandi in Delhi, providing practical insights into forecasting accuracy. The performance of the models is evaluated using standard metrics such as Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE) and Mean Absolute Error (MAE). Additionally, the Diebold-Mariano (DM) test has been conducted to compare the accuracy of the proposed approach with baseline DL models. The findings demonstrate that the hybrid approach of Hidden Markov (HM) combined with DL models yields superior forecasting performance compared to existing models.

Key words: Diebold mariano; Gated recurrent units; Long short-term memory; Multilayer perceptron; Recurrent neural networks; Vegetable price.

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

1. Introduction

Forecasting of prices for any commodity or product needs hardly be emphasized. Effective planning and strategic decision-making are facilitated by precise and timely price information, coupled with accurate forecasting. However, analyzing agricultural commodity prices presents unique challenges when compared to non-farm goods and services due to their vulnerability owing to unforeseen events like droughts, floods, disease and pest outbreaks, as well as factors such as seasonality, demand fluctuations, climate variability, market imperfections, globalization and speculative trading, see Yin *et al.* (2020) and Manogna and Mishra (2021). Moreover, the nonlinear and nonstationary characteristics of price data further complicate the process of price forecasting, see Xiong (2018).

Vegetable-growing farmers in India are not in a comfortable situation despite the significant increase in the production of tomatoes, onions, and potatoes, collectively known as TOP vegetables. While India holds the position of the world's second-largest producer of overall vegetables, with a total production of 137.99 million metric tonnes (MMT) compared to China's 600.01 MMT, recent statistics show that Tomato, Onion and Potato production reached 21.18, 26.64 and 56.17 MMT respectively in 2021–22 (source; `https://www.statista.com`). Unfortunately, the farmers are currently facing various challenges due to overproduction which resulted in distress sales, crop burning, and the unfortunate practice of discarding their produce on the roads, especially during periods of bearish market conditions Guresen *et al.* (2011)Consequently, it becomes essential to address these issues and stabilize prices by providing storage facilities for farmers during bearish times, offering guidance on selling vegetables during inflationary periods, and imparting knowledge on the supply value chain for increasing the value of vegetables for the betterment of the farmers income. In this study, a real dataset from Azadpur mandi in Delhi has been utilized to shed light on these aspects.

This work has been undertaken on the premise that hybridizing Hidden Markov Models (HMMs) with DL models may offer many advantages over classical DL models. HMMs excel at modeling sequential data, capturing temporal dependencies and leveraging limited labelled data while providing interpretability and handling noisy or incomplete data. Incorporating HMM-based hidden states in DL models provides intermediate representations within these states, facilitating training by providing forecasts within each state and at the same time proceeding with forecasting successively by using the information from previous states. Thus, this explicit modeling of sequential dependencies by HMMs offers a structured framework for DL model training which enhances the predictive capabilities. In addition, while DL approaches offer advantages such as manual feature extraction and resource availability, their effectiveness heavily relies on large datasets. This distinguishes DL techniques from traditional machine learning methods. However, there is still uncertainty regarding the specialization and generalization capabilities of DL models compared to conventional methodologies as the former are computationally intensive, demanding speed and high-end computing resources. DL models are often regarded as black-box models, lacking interpretability and transparency in their decision-making processes. Furthermore, DL models are prone to overfitting, especially when dealing with noisy datasets Singh *et al.* (2023) Hence DL models can be challenging to train and fine-tune, requiring expertise in hyperparameter optimization and architectural design. To address this, our proposed approach combines the strengths of both methodologies by combining an HMM with DL to analyze underlying

patterns in the data series with the aim of overcoming challenges such as overfitting and circumventing local minima traps. By combining the strengths of HMMs and DL models, improved data efficiency, better sequential modeling, interpretability and robustness can be achieved.

The rest of the paper is structured as follows. In Section 2, a review of the literature has been studied. In Section 3, related work along with the background knowledge is discussed. In Section 4, an empirical study has been conducted using a real dataset focusing on the TOP price series with results and discussion. Finally, Section 5 concludes the paper by summarizing the findings with suitable remarks and discussing future prospects following a list of references.

2. Review of literature

Extensive research has focused on improving price forecasting through the utilization of diverse and advanced time series models, see Wang *et al.* (2020). These modeling approaches, designed to enhance price forecasting, can be broadly categorized into two main groups: statistical models and artificial intelligence (AI) based models, see Yu *et al.* (2017).

The ARIMA model, introduced by Box and Jenkins in 1970, is widely used in time series analysis, particularly in forecasting financial data, see [Kocak (2017); Adebiyi *et al.* (2014); Ariyo *et al.* (2014); Jarque and Bera (2011); Avinash *et al.* (2022)]. However, its capabilities are limited when it comes to modeling nonlinear data. To overcome this, alternative nonlinear time series models have emerged, including regime-switching models like SETAR model (Mehdizadeh *et al.* (2019)),STAR model [Athanasopoulos and De Silva (2012)] and GARCH model [Lin (2018)].These models capture nonlinearity but often require specific relationships in the data and lack generalization ability, as highlighted by Weron (2014).

To address the challenges posed by complex dependencies and nonlinear relationships in time series data, HMMs were developed on the basis of pioneering work by Baum and colleagues [Baum and Petrie (1966); Baum and Sell (1968); Baum (1972)]and its first application in the formulation of a statistical method of representing speech was made by Rabiner (1989). HMMs assume that the observed data is generated by a Markov process with hidden states, enabling them to capture nonlinearity and temporal dependencies in the data. By modeling these hidden states, HMMs can effectively uncover latent variables and extract essential features such as trend, seasonality, and volatility. However, it is important to note that the applicability of HMMs may vary depending on the characteristics of the time series data. Chaotic patterns with long-range dependencies may not align well with the assumptions of HMMs [Awad *et al.* (2015); Abdollahi and Ebrahimi (2020)]. Additionally, training an HMM model requires a substantial amount of data, which can be challenging in the context of price forecasting due to noisy data and external factors that may impact the model's performance and also they assume Markovian behavior, which may not hold in all scenarios, limiting their ability to capture complex dependencies. HMMs may struggle to model nonlinear relationships and can be sensitive to initial parameter values, affecting their performance. Determining the appropriate number of hidden states is challenging and HMMs lack transparency and interpretability. Additionally, handling continuous or high-dimensional data can be difficult for HMMs, requiring discretization or dimensionality reduction.

To overcome these challenges, Machine Learning (ML) models have gained prominence in financial time series forecasting due to their ability to learn from data, interpretability and lack of assumptions explained by Makridakis *et al.* (2018).Various ML models, including Artificial Neural Networks (ANNs)/ Multilayer Perceptron (MLP) [Haykin (2009)],Support Vector Regression (SVR) [Henrique *et al.* (2018)], Random Forest (RF) [Nti *et al.* (2019)],eXtreme Gradient Boosting (XGBoost) [Basak *et al.* (2019)] and ensemble models such as stacking [Jiang *et al.* (2020)]and bagging [Wang *et al.* (2009)] have been utilized in financial time series forecasting. ML models, being data-driven and adaptable, offer advantages over traditional model-based approaches. However, ANN has certain limitations such as slow convergence to the optimal solution and the risk of overfitting [Wang *et al.* (2016)]. In the absence of domain knowledge, DL excels at feature extraction, outperforming other methods, except for a few feature engineering techniques like the requirement of a substantial amount of labelled training data to achieve optimal performance.

Deep architectures, achieved by adding additional layers, leverage multiple levels of nonlinear processes by increasing model complexity. Deep Neural Networks (DNNs) with more layers can effectively handle complex functions using fewer parameters. These models include the Recurrent Neural Network (RNN), Gated Recurrent Units (GRUs) [Althelaya *et al.* (2018)] and Long Short-Term Memory (LSTM) [Nelson *et al.* (2017); Jaiswal *et al.* (2022); Zaheer *et al.* (2023); Heidarpanah *et al.* (2023); Latif *et al.* (2023)]

In recent research, several notable studies have explored the integration of Hidden Markov Models (HMMs) with various ML/ DL techniques to enhance the accuracy of time series forecasting across different domains. For instance, Chen *et al.* (2019) proposed a novel approach combining a Generative Adversarial Network (GAN) with an Iteratively Refined HMM for completely unsupervised speech recognition. Hassan (2009)combined hidden Markov and fuzzy model for stock market forecasting. Similarly, Hashish *et al.* (2019) developed a hybrid model that leveraged HMMs and optimized LSTM networks to predict Bitcoin prices. Yao and Cao (2020)introduced a neural network-enhanced HMM based structural time series model tailored explicitly for tourism demand forecasting. Building upon these advancements, Peng *et al.* (2021) devised an HMM-LSTM model for proactive traffic prediction in 6G wireless networks. Additionally, Khan *et al.* (2022) investigated the potential of an HM-BiLSTM-based system for event detection and classification, focusing specifically on food intake recognition. These studies collectively highlight the effectiveness of integrating HMMs with deep learning techniques to tackle complex time series forecasting challenges in diverse domains.

This highlights the need for further research in the area of the agriculture domain. In this study, an attempt has been made on the TOP price series from Azadpur Mandi (Delhi) by using HMM to extract relevant features that can be fed separately to MLP, RNN, GRU, and LSTM to improve the accuracy of forecasting. This approach can be beneficial when the underlying system is complex and difficult to model using traditional methods.

3. Material and methods

In this study, five baseline models *viz.*, HMM, MLP, RNN, GRU and LSTM models and the proposed HMM hybridized with the DL models *viz.* HM-MLP, HM-RNN, HM-GRU and HM-LSTM have been fitted. A brief description of the baseline models considered are

given subsequently followed by the proposed methodology of the hybrid models.

3.1. Hidden markov model (HMM)

HMMs are probabilistic models that generate a series of observations (Y) based on a series of underlying hidden states (S). HMMs are commonly employed to model time-dependent data and have found practical use in diverse fields including speech recognition, molecular biology and computer vision, see Ghahramani (2001).

HMMs are built upon two fundamental assumptions. Firstly, HMM assumes that an observation at a particular time t, denoted as Y_t, is generated by an underlying process where the corresponding state, S_t, remains hidden from the observer. Secondly, it assumes that this hidden state S_t follows a first-order Markov property, meaning that the current state S_t, given the previous state S_{t-1}, is independent of all states prior to $t-1$. Likewise, the output of an HMM also adheres to the Markov property. Consequently, the joint distribution of a sequence of hidden states and observations can be factorized as presented by equation (1).

$$P(S_{1:T}, Y_{1:T}) = P(S_1)P(Y_1|S_1)\prod_{t=2}^{T} P(S_t|S_{t-1})P(Y_t|S_t) \tag{1}$$

where $P(S_{1:T}, Y_{1:T})$ represents the joint distribution of the sequence of hidden states ($S_{1:T}$) and observations ($Y_{1:T}$). $P(S_1)$ is the initial probability distribution of the first hidden state S_1. $P(Y_1|S_1)$ is the probability of observing Y_1 given the state S_1. $P(S_t|S_{t-1})$ is the transition probability from state S_{t-1} to state S_t. $P(Y_t|S_t)$ is the probability of observing Y_t given the state S_t. Overall, the equation describes how the joint distribution of hidden states and observations in an HMM can be factorized based on the initial state probability, observation probabilities given the states, and transition probabilities between states.

HMM is defined by three key components: A, B, and π, while implicitly determined by the number of observations (N) and the number of hidden states (M). Where A represents the state transition probability $M \times M$ matrix, B represents the probability of the observations $M \times N$ matrix, and π is the initial state distribution. Thus, HMM can be defined as $\lambda = (A, B, \pi)$.

Hidden Markov Models (HMMs) are utilized to address three fundamental problems, which can be summarized as follows:

1. Problem 1: Given the model $\lambda = (A, B, \pi)$, along with a sequence of observations Y, determine the likelihood of the observed data with respect to the given model through the forward-backward or Expectation-Maximization algorithm.
2. Problem 2: Given the model $\lambda = (A, B, \pi)$, along with a sequence of observations Y, determine the optimal sequence of hidden states that underlie the Markov process through the Viterbi algorithm.
3. Problem 3: Given a sequence of observations Y, estimate the parameters of the model, namely A, B, and π, through the Baum-Welch algorithm.

In this study, our approach involves constructing an HMM based on a given sequence of observations. Subsequently, by calculating the likelihood of the data and determining

the optimal sequence of hidden states through the Viterbi algorithm, following the standard methodology, which can be found in, Giudici and Abu Hashish (2020).

3.2. Multilayer perceptron (MLP)

ANN is a mathematical model inspired by the human brain's information processing and analysis capabilities, used to solve a wide range of nonlinear problems. ANN offers advantages such as parallel processing, learning from experience (dataset) and the ability to approximate various functions with high accuracy. It finds applications in forecasting and classification tasks, with the Multilayer Perceptron (MLP) being the most well-known ANN model. MLP is particularly popular for time series forecasting [Aizenberg *et al.* (2016)].Typically, an MLP consists of an input layer, hidden layer(s), and an output layer, with neurons connected by weighted links. The mathematical equations describing the neural network are represented by equation (2).

$$\hat{y} = \sum_{j=0}^{s_0} W_{jk} \cdot \xi \left(\sum_{i=0}^{s_h} W_{ij} \cdot x \right) \tag{2}$$

$$\xi(\alpha, x) = \begin{cases} \alpha \cdot (e^x - 1), & \text{if } x < 0 \\ x, & \text{otherwise} \end{cases} \tag{3}$$

where x and $\hat{y}$ are the input and output of the network, respectively. s_0 and s_h are the sizes of the output layer and hidden layers. W_{ij} are the weights of the connections between the input and hidden layers, and W_{jk} are the weights of the connections between the hidden and output layers. ξ is the Exponential Linear Unit (ELU) presented in equation (3) in its general form when $\alpha = 1$. It becomes the Rectified Linear Unit (ReLU) when $\alpha = 0$.

3.3. Recurrent neural networks (RNNs)

RNNs are a type of neural network that is well-suited for modeling time series data. RNNs use a series of interconnected neurons to model the functional relationship between input features in the recent past and a target variable in the future. By repeatedly learning from a training set of historical data, RNNs can capture the transitions of an internal (hidden) state over time and make more accurate predictions about future events as shown in Figure (1).

However, RNNs have a major limitation: they can suffer from the gradient vanishing problem, where the gradient becomes too small over time and the network is unable to retain information from long-term inputs. This can limit the accuracy of RNNs, particularly when modeling time series data with long-term dependencies. To overcome this problem, other variants of RNNs were developed, including the LSTM and the GRU network.

3.4. Long short term memory (LSTM)

Hochreiter and Schmidhuber (1997)recognized that traditional RNNs were unable to retain important historical information for extended periods of time. To address this issue, they developed the LSTM model, which introduced gate mechanisms to the RNN framework.

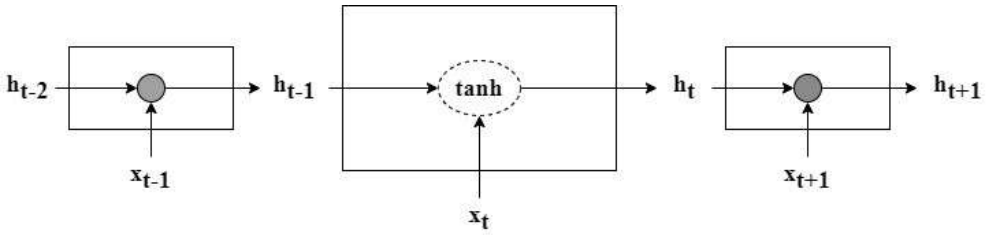


Figure 1: The architecture of RNN model

LSTM is an advanced form of RNN developed specifically for handling sequential data like texts and sentences [Alom *et al.* (2019)]. While a basic RNN is designed to retain and transfer information from one step to the next, it encounters the issue of a vanishing gradient, where long-term information cannot be effectively utilized. Consequently, significant amounts of previous information cannot be stored adequately, resulting in less accurate forecasting. The mathematical formulation of LSTM is represented by equations 4-9

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \tag{4}$$
$$i_t = \sigma(W_i x_t + U_i h_{t-1} + b_i) \tag{5}$$
$$\tilde{c}_t = \gamma(W_c x_t + U_c h_{t-1} + b_c) \tag{6}$$
$$c_t = f_t \times c_{t-1} + i_t \times \tilde{c}_t \tag{7}$$
$$o_t = \sigma(W_0 x_t + U_0 h_{t-1} + b_o) \tag{8}$$
$$h_t = o_t \times \gamma(c_t) \tag{9}$$

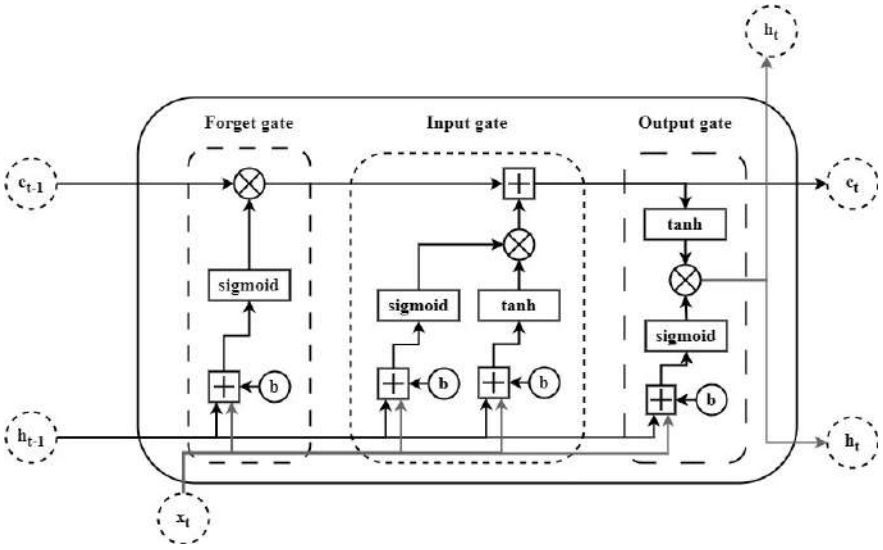


Figure 2: The architecture of LSTM cell

The LSTM mechanism is centred around a cell state, denoted as c_t, which serves as a storage unit for information. This information is regulated through three gates: the forget gate (f_t), the input gate (i_t), and the output gate (o_t). These gates determine whether incoming sequential data should be retained to preserve relevant information for subsequent stages. The forget gate, as indicated by equation 4,decides whether information should be added or omitted. If f_t is close to one (or zero), the information from the input and hidden

state will be preserved (or removed) accordingly. The input gate computes an update to the cell state, evaluating the importance of the input for the subsequent cell. Additionally, the output gate generates the output for the hidden states based on equation 9.Notably, the activation functions used in LSTM are the hyperbolic tangent function (tanh) and the sigmoid function (σ), employed respectively as the activation function as shown in Figure 2.

3.5. Gated recurrent units (GRUs)

GRUs share similarities with LSTM units as they possess a comparable design and can yield similar outcomes in certain cases. However, GRUs differ from LSTM units with the absence of an output gate. Instead, they employ an update gate and a reset gate to control the flow of information into and out of memory as shown in Figure 4. This gating mechanism allows the network to effectively retain information from long-term inputs, enabling more precise predictions in detail explanations by equations 10-13. GRUs offer a potent solution for addressing the vanishing gradient problem in RNNs and find extensive use in various applications including polyphonic music modeling, speech signal processing, handwriting recognition and time series data forecasting. They are particularly beneficial when working with smaller datasets compare to LSTM.

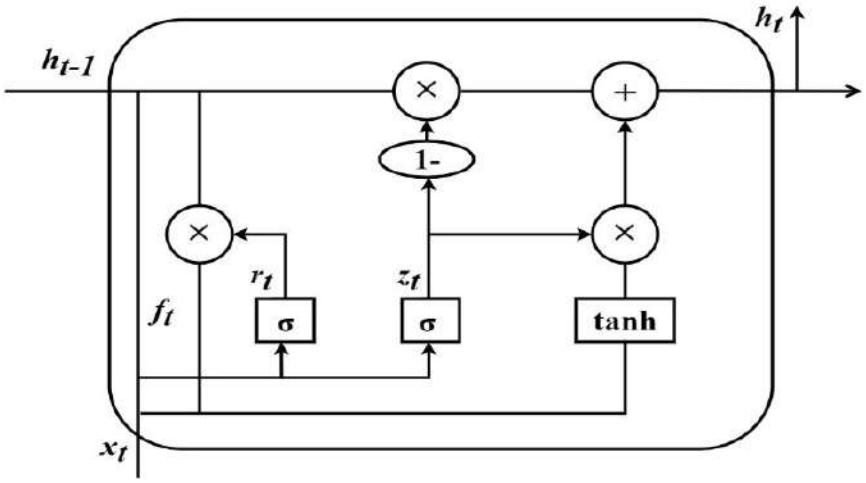


Figure 3: The architecture of GRU cell

The process can be described as:

$$Z_t = \sigma(x_t w^z + h_{t-1} U^z + b_z) \tag{10}$$

$$r_t = \sigma(x_t w^r + h_{t-1} U^r + b_r) \tag{11}$$

$$\tilde{h}_t = \tanh(r_t \cdot h_{t-1} U + x_t W + b) \tag{12}$$

$$h_t = (1 - Z_t) \cdot \tilde{h}_t + Z_t \cdot h_{t-1} \tag{13}$$

where w^z, w^r, and W denote the weight matrices for the corresponding connected input vectors. U^z, U^r, and U represent the weight matrices of the previous time step and b_z, b_r, and b are biases. The σ denotes the sigmoid function, r_t denotes the reset gate, z_t denotes the update gate and $\tilde{h}_t$ denotes the candidate hidden layer.

3.6. Proposed hidden markov based deep learning modeling

The entire analysis was conducted using Python software, employing the "GaussianHMM" and "TensorFlow" libraries (see Appendix). These provided a user-friendly interface for constructing and training DL models. The experiments were conducted on a system equipped with an AMD Ryzen 7 5700U processor and 8 GB of RAM, which proved sufficient for training and evaluating each DL model. The processing time for each model ranged from 25 to 30 minutes while employing the grid search validation technique.

The proposed methodology is represented schematically in Figure 4. To start with, pre-processing is done on the time series data. For this, normalization is employed to rescale the values of the series between 0 and 1 while preserving their shape for the modal price series. The normalization equation 14 used as follows:

$$Y_t = \frac{X_t - X_{\min}}{X_{\max} - X_{\min}} \tag{14}$$

where $X_{\min}$, $X_{\max}$, and X_t are the minimum, maximum and observation at time t, respectively and Y_t is the rescaled value. In Python, the 'Min-Max Scaler' function of the "Scikit-learn" package is used for this purpose. Thereafter, the data is split into, say, 90% training and 10% testing data subsets. The training dataset is then used for training classical Hidden Markov Models (HMM) and baseline Deep Learning (DL) models, such as Multilayer Perceptron (MLP), Recurrent Neural Network (RNN), Gated Recurrent Unit (GRU), and Long Short-Term Memory (LSTM), with optimized hyperparameters obtained through grid search. In addition, the training data is used to fit an HMM and extract hidden states using the Viterbi algorithm, employing grid search cross-validation. These hidden states are then utilized to train the proposed hybrid models, namely HM-MLP, HM-GRU, HM-RNN, and HM-LSTM. Finally, the performance of different models on the time series is evaluated using metrics such as Root Mean Squared Error (RMSE), Mean Absolute Percentage Error (MAPE) and Mean Absolute Error (MAE). Additionally, the Diebold-Mariano (DM) test is conducted to compare the accuracy of the proposed approaches vis-à-vis baseline DL models and also among themselves.

4. Results and discussion

In the present study, the weekly TOP (Tomato, Onion, and Potato) prices (in Rs. /Quintal) from 01 Jan 2006 to 16 June 2023 (obtained from the Agmarknet; `https://agmarknet.gov.in`) of Azadpur market, Delhi were used, whose time plots are depicted in Figure (5, 6 and 7) for TOP commodity price series. This market situated within the Indo-Gangetic plains is characterized by the latitude and longitude coordinates of approximately 28.7078° N and 77.1676° E. This market holds immense significance as one of Asia's largest wholesale fruit and vegetable markets, serving as a crucial link in the agricultural supply chain. Commodities from various regions across the Indo-Gangetic plains converge at Azadpur Mandi, further highlighting its importance as a major hub for agricultural trade. Its strategic location, extensive infrastructure, and role as a price benchmark contribute significantly to its economic importance.

The summary statistics of the datasets are presented in Table 1. Additionally, the Jarque-Bera test [Jarque and Bera (1987)] and Shapiro-Wilk's test [Shapiro and Wilk (1965)]

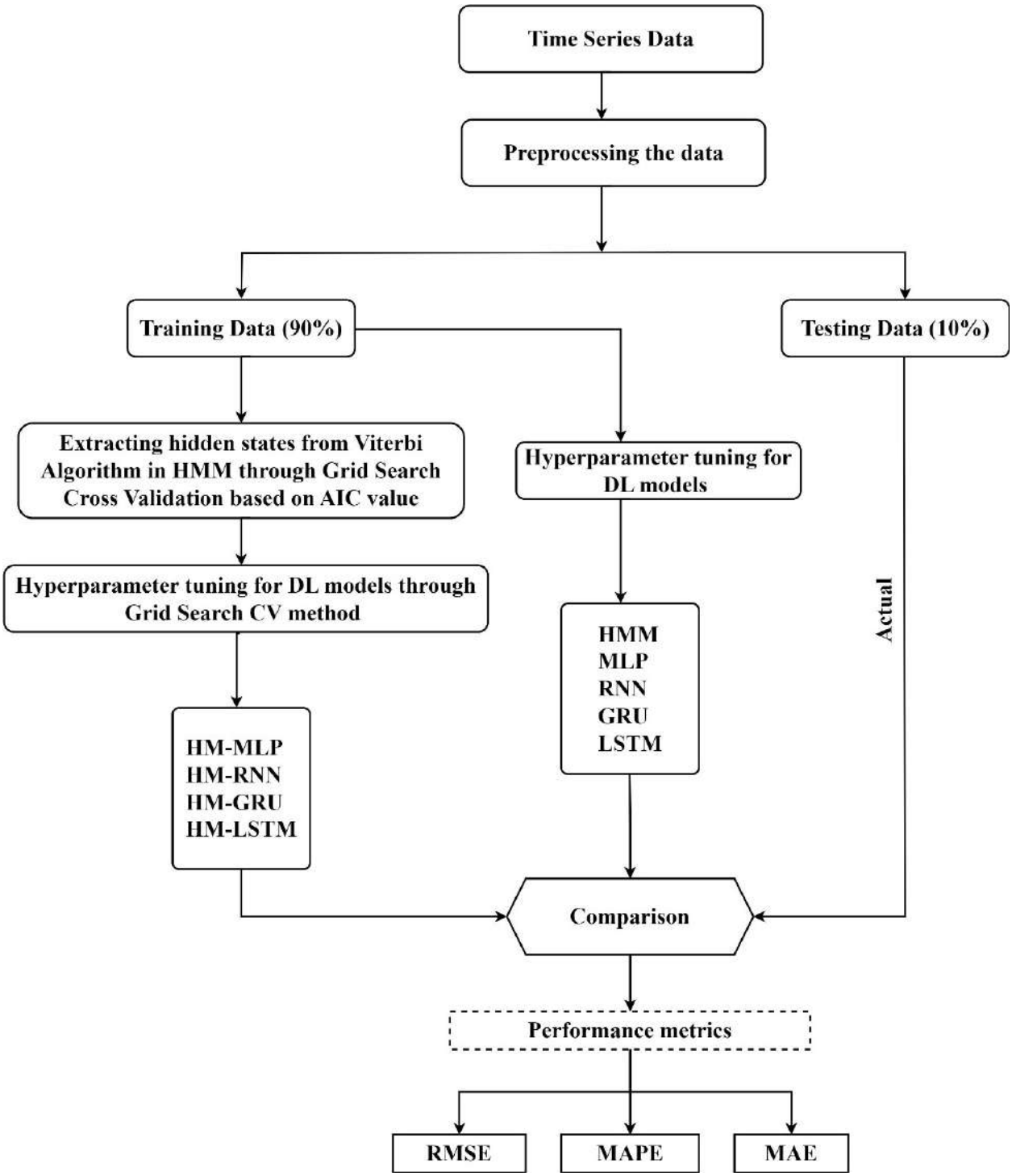


Figure 4: Proposed Hidden Markov (HM) based Deep Learning (DL) modeling

were used to assess the normality of the TOP price series. The tests were significant indicating that all the series are non-normal. Furthermore, the datasets displayed positive skewness but mesokurtic for tomato and potato, while exhibiting leptokurtosis in the case of the Onion price series.

In addition, tests were conducted for the presence of stationarity. The results of the tests (Table 2) reveal weak stationarity under the Augmented Dickey-Fuller (ADF) and Phillips Perron (PP) tests. Subsequently, the nonlinearity of the data series was assessed using the Brock- Dechert-Scheinkman (BDS) test (Table 3). The results highlighted that the weekly TOP price series of all three commodities considered exhibited nonlinear patterns.

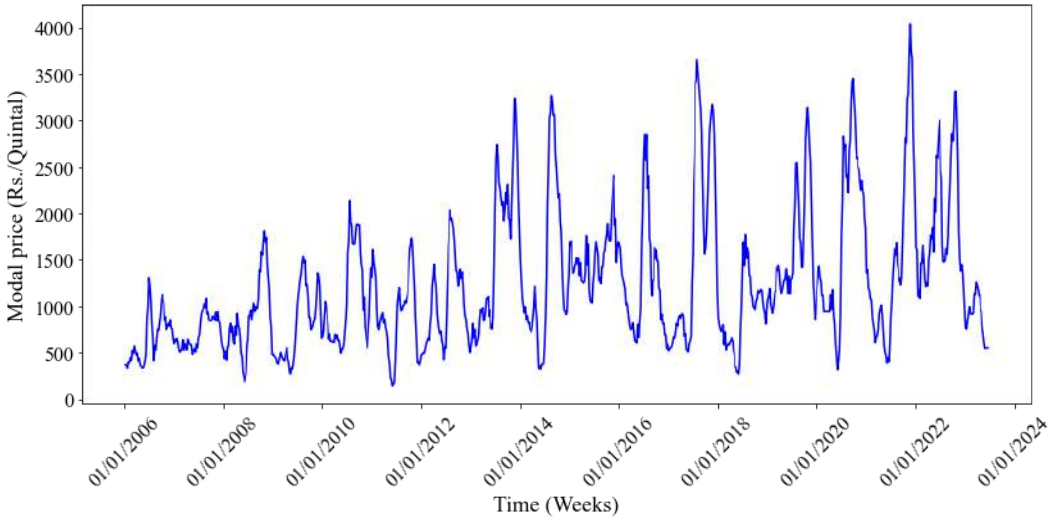


Figure 5: Time plot of weekly Tomato price of Azadpur market

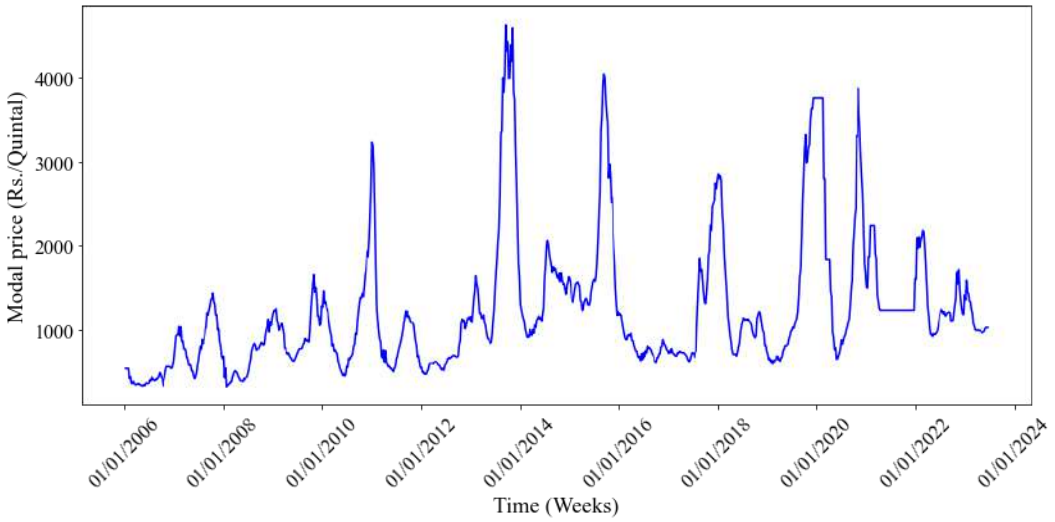


Figure 6: Time plot of weekly Onion price of Azadpur market

The TOP price series comprised 911 observations, which were split into training (90%; 822 data points) and subsequent data points as testing (10%; 89 data points) sets. There were a few missing values in the data series; hence, imputation was done by taking the average of preceding and succeeding observations in the weekly data series. Initially, HMMs were fitted by employing the Viterbi algorithm with grid search cross-validation (2-12 hidden states) to determine the optimal number of hidden states. Results revealed that six hidden states were found for Tomato, and eight hidden states for both Onion and Potato price series,

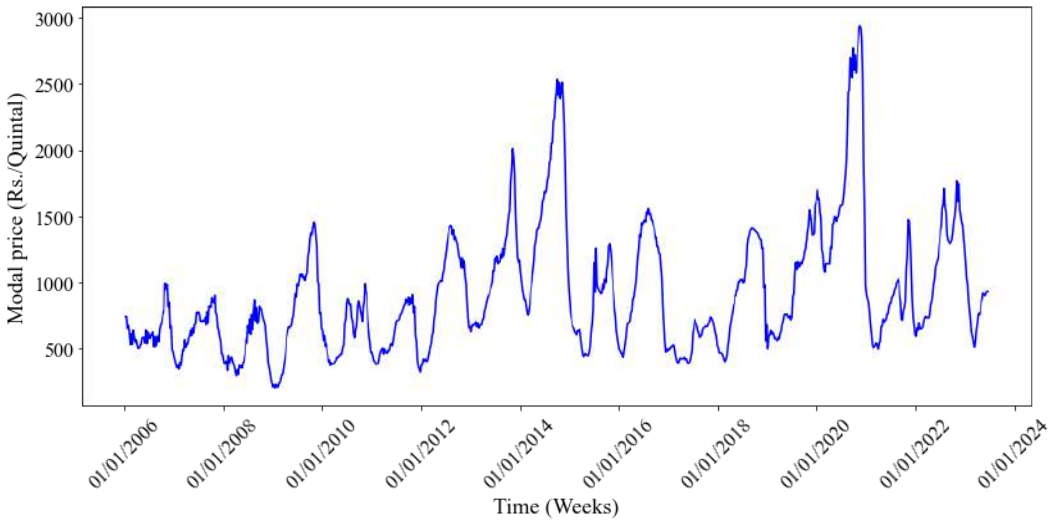


Figure 7: Time plot of weekly Potato price of Azadpur market

Table 1: Descriptive statistics and normality tests of the weekly data series of TOP commodities

Descriptive Statistics	Tomato price series	Onion price series	Potato price series
Mean (Rs. /Quintal)	1244.95	1236.51	916.14
Median (Rs. /Quintal)	1035.06	1031.00	775.92
Maximum (Rs. /Quintal)	4049.75	4638.50	2946.33
Minimum (Rs. /Quintal)	150.08	329.33	207.81
Std. Dev. (Rs. /Quintal)	733.60	799.19	480.28
CV (%)	58.89	64.59	52.39
Skewness	1.19	1.89	1.52
Kurtosis	0.99	3.61	2.91
Jarque-Bera test statistic	253.12^{**}	1039.41^{**}	673.43^{**}
Shapiro-Wilk test statistic	0.89^{**}	0.79^{**}	0.87^{**}

Table 2: Stationarity test of the weekly price series for TOP commodities

Commodities	ADF test		PP test		Conclusion
	Test Statistic	p value	Test Statistic	p value	
Tomato	5.48	< 0.001	4.95	< 0.001	Stationary
Onion	4.56	< 0.001	4.57	< 0.001	Stationary
Potato	4.30	< 0.001	4.37	< 0.001	Stationary

enabling the capturing of complex dynamics and trends to enhance feature engineering in subsequent DL models to be trained (as shown in Figures (8, 9 and 10).

Following the confirmation of stationarity, nonlinearity, feature extraction from HMM, and normalization of the modal price data for TOP, Classical HMM was fitted based on the hidden states obtained by the Viterbi algorithm, and the forecasts were obtained for testing data sets and are shown in Figures (11, 12 and 13). Thereafter, the DL models *viz.*, MLP,

Table 3: Nonlinearity BDS test results with different embedding dimensions for TOP commodities at 0.5, 1, 1.5 and 2 σ respectively

Commodities	Embedding dimension 2		Embedding dimension 3		Conclusion
	Statistics	p value	Statistics	p value	
Tomato	1442.11	<0.001	613663.83	<0.001	Nonlinearity
	1249.30	<0.001	303035.76	<0.001	
	527.43	<0.001	70621.75	<0.001	
	486.41	<0.001	35310.50	<0.001	
Onion	440.29	<0.001	43244.29	<0.001	Nonlinearity
	453.90	<0.001	41327.66	<0.001	
	465.56	<0.001	39211.17	<0.001	
	479.76	<0.001	36843.34	<0.001	
Potato	1248.17	<0.001	326417.51	<0.001	Nonlinearity
	1008.86	<0.001	72434.55	<0.001	
	689.48	<0.001	22456.71	<0.001	
	568.79	<0.001	9869.33	<0.001	

Table 4: Optimal hyperparameters for the various DL models

Model	Batch Size			No. of Epochs			No. of HL			No. of units / HL		
	T	O	P	T	O	P	T	O	P	T	O	P
MLP	64	32	32	57	68	89	2	1	1	32,64	64	128
RNN	128	64	32	76	72	45	1	1	1	32	32	64
GRU	32	64	64	78	56	73	1	1	1	64	64	32
LSTM	64	32	32	87	52	85	1	1	1	128	128	64
HM-MLP	64	16	64	176	147	67	1	1	1	32	32	32
HM-RNN	32	64	32	168	128	112	1	1	1	32	32	8
HM-GRU	16	64	32	52	64	64	1	1	1	64	32	16
HM-LSTM	32	32	32	100	106	65	1	1	1	32	64	16

RNN, LSTM and GRU were also trained. The primary objective of this study is to assess the performance of Hidden Markov hybridized DL (HM-DL) models in forecasting price series.

For training the DL and HM-DL models, hyperparameters play a crucial role as they significantly impact the performance of forecast accuracy to overcome the local minima trap. The batch size in DL models determines how many samples are processed before updating the model's weights. Larger batch sizes can provide more stable gradients but may require more computational resources. The number of epochs specifies how many times the model is trained on the entire dataset. Increasing the number of epochs can potentially improve model performance, but it also increases the risk of overfitting. To mitigate overfitting, early stopping criteria based on mean square error have been applied to select the best weights during training. The number of input units in the model determines the number of variables the model takes as inputs. Having a larger number of input units allows the model to capture more complex relationships in the data but may increase computational costs. A range of hyperparameter values were used in grid search cross-validation on DL models viz., MLP, RNN, GRU, LSTM, and their hybrid HMM cum DL versions, *i.e.*, HM-MLP, HM-RNN,

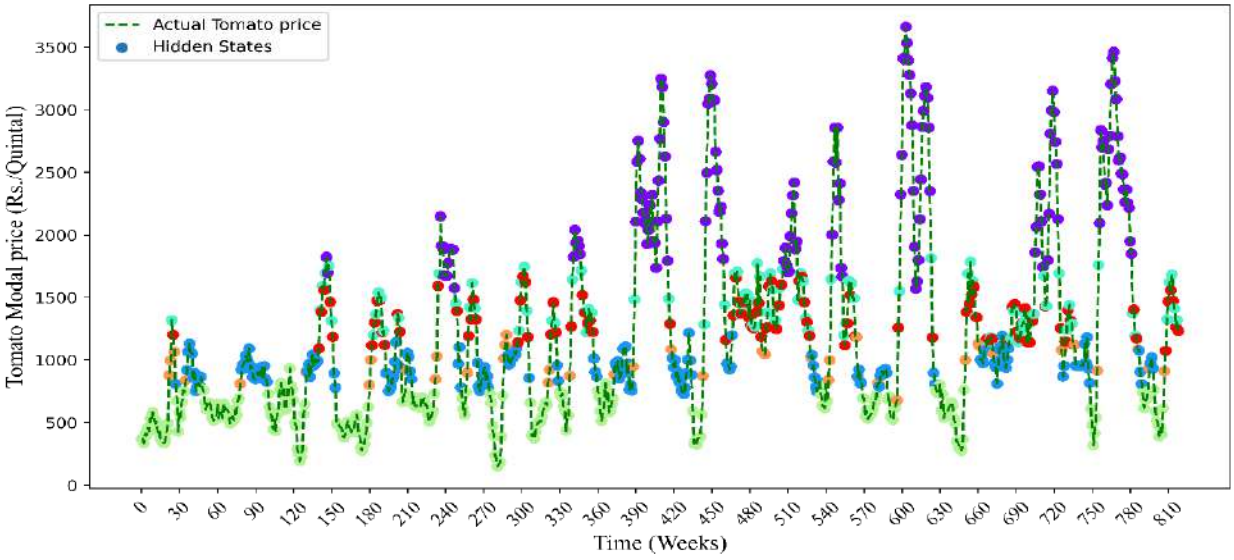


Figure 8: Hidden states obtained from HMM on Tomato price series

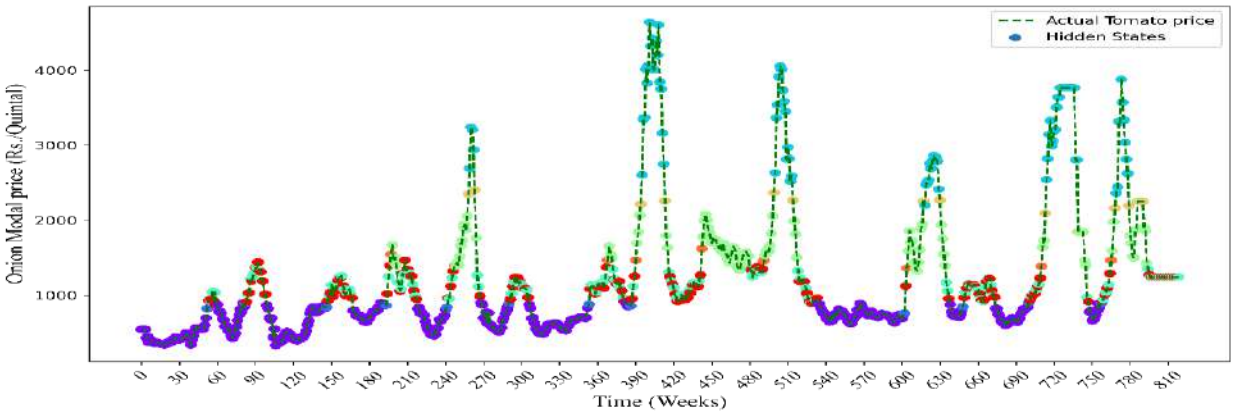


Figure 9: Hidden states obtained from HMM on Onion price series

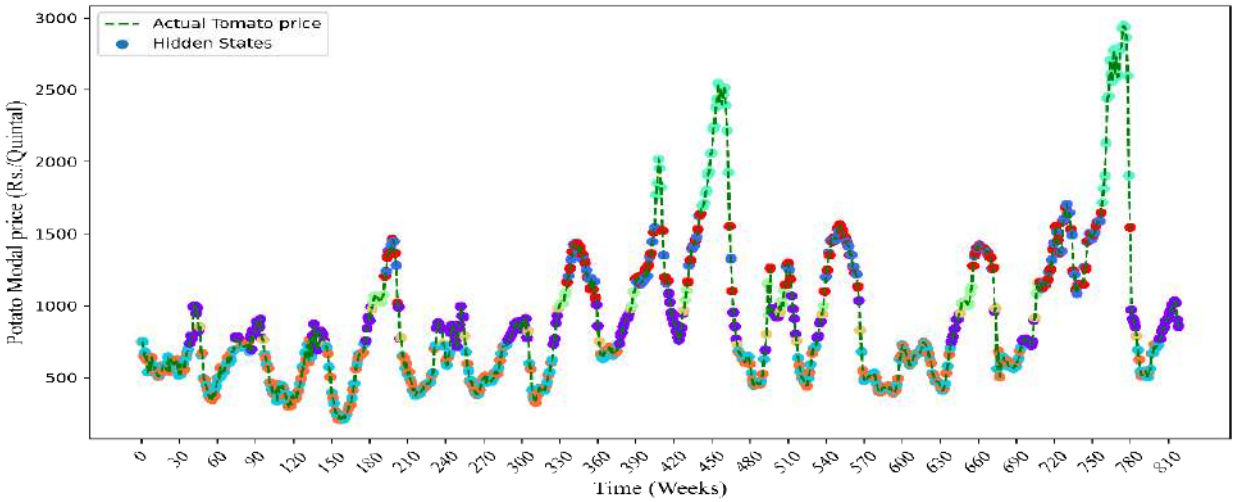


Figure 10: Hidden states obtained from HMM on Potato price series

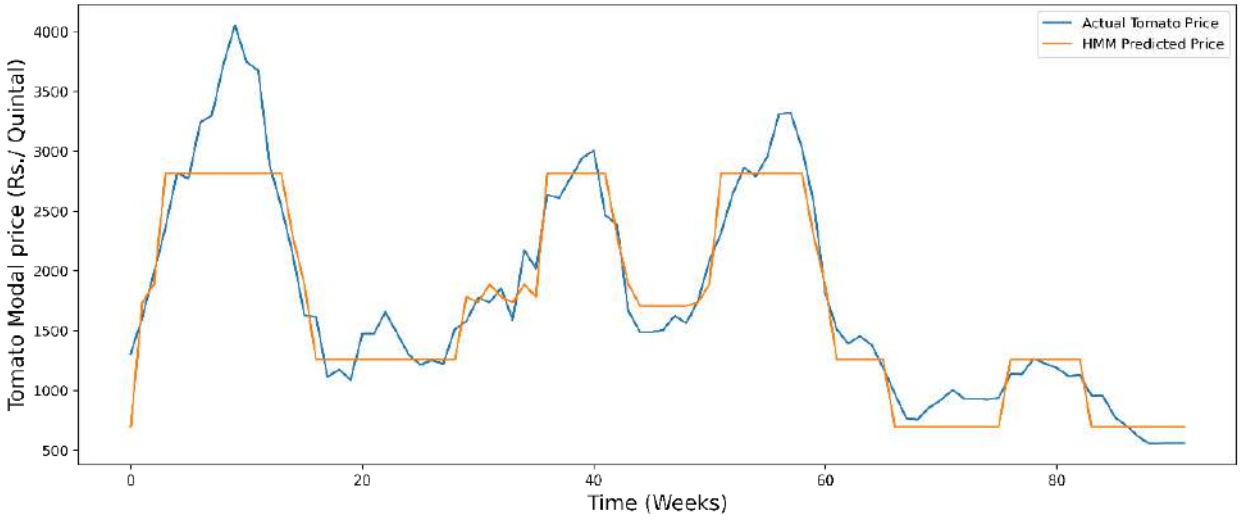


Figure 11: Predicted price by HMM on Tomato price series

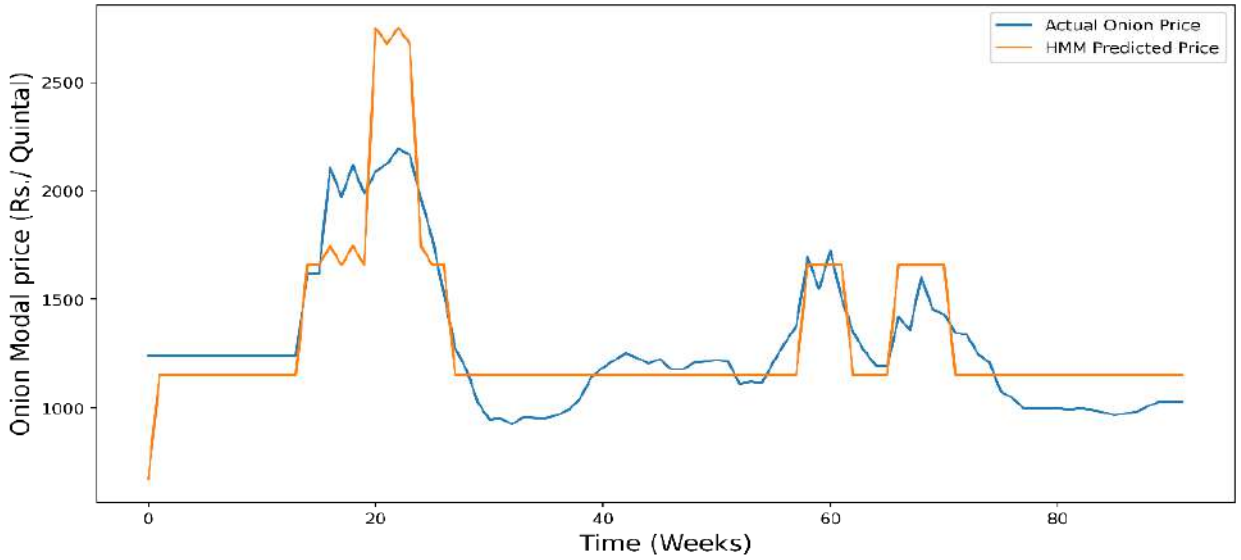


Figure 12: Predicted price by HMM on Onion price series

HM-LSTM, and HM-GRU as follows: the number of lags - fixed as 24 weekly data points for the TOP crop price series; batch size - 8, 16, 32, 64, 128, 256; the number of epochs - 200 with early stopping criteria; the number of hidden layers (HL) - 1, 2, 3; and the number of hidden units - 8, 16, 32, 64, 128, 256, 512, which led to 126 combinations of candidate models for each DL model. For training, each DL model took around 25-30 minutes on average as computing time. The optimal combination of hyperparameters determined is shown in Table (4).

Using these optimal hyper-parameters, DL models trained were utilized for forecasting prices for the test data period. The performance of the models based on RMSE, MAPE and MAE revealed that hybridized HM-DL models perform well as compared to all other models for forecasting of TOP price are shown in Table 5 and Figures (14, 15 and 16).

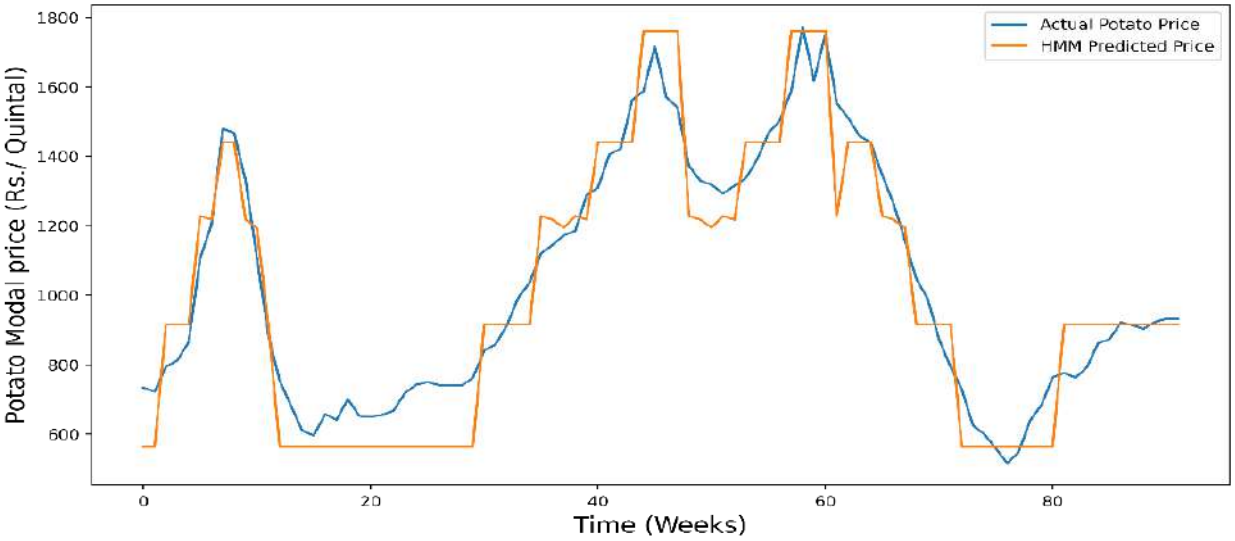


Figure 13: Predicted price by HMM on Potato price series

Table 5 revealed that HMM seems to perform better for forecasting the Potato price series compared to the other two series when considering the results of the testing datasets. Comparison of MAPE values for the baseline DL models with that of the HM-DL models clearly showed that the HM-DL models perform better. By and large, the RMSE of the forecasts for the proposed HM-DL models, namely HM-MLP, HM-RNN, HM-GRU, HM-LSTM, were lower than their corresponding baseline models, namely MLP, RNN, GRU, LSTM, by 9.77–17.50%, 15.02–44.39%, and 7.94–32.60% respectively for the tomato, onion, and potato prices, except for two cases where the baseline DL models seem to be better.

Table 5: Performance of various models on TOP price series data of Azadpur mandi, Delhi

Price series		Tomato			Onion			Potato		
Evaluation measures		RMSE	MAPE	MAE	RMSE	MAPE	MAE	RMSE	MAPE	MAE
Training set (90%)	HMM	221.19	20.29	180.67	274.55	13.75	159.24	165.38	17.31	127.01
	MLP	190.49	10.72	147.63	215.25	16.40	175.39	77.90	7.44	57.21
	RNN	150.18	10.72	111.24	211.20	12.01	144.80	99.66	8.84	70.43
	GRU	154.46	12.48	118.03	135.32	6.86	85.77	91.29	8.22	63.52
	LSTM	152.32	12.73	117.73	133.73	7.16	87.13	69.89	6.51	50.05
	HM-MLP	201.14	17.10	156.20	146.06	9.62	103.23	101.84	10.00	73.85
	HM-RNN	**128.16**	**9.82**	**95.92**	**106.64**	**5.10**	**63.95**	**64.87**	**5.70**	**44.99**
	HM-GRU	147.05	12.64	116.31	116.98	7.23	79.96	73.95	8.00	58.35
	HM-LSTM	129.85	10.06	97.92	115.57	6.20	73.53	71.56	6.48	49.73
Testing set (10%)	HMM	263.76	11.63	189.88	214.58	13.79	169.36	146.32	10.03	101.63
	MLP	294.16	12.90	220.85	168.42	9.44	123.58	111.66	9.04	86.15
	RNN	216.24	9.00	162.60	159.50	9.94	128.64	110.07	8.52	83.85
	GRU	220.87	9.40	163.11	116.69	6.16	84.78	122.18	9.03	90.95
	LSTM	226.01	10.38	176.56	121.50	6.26	85.58	96.66	7.02	68.33
	HM-MLP	265.43	12.66	207.15	122.24	6.33	86.61	113.09	8.44	83.45
	HM-RNN	240.59	12.55	176.23	**88.69**	**4.62**	**63.21**	**79.58**	**5.80**	**58.86**
	HM-GRU	193.94	8.41	135.98	99.16	5.91	78.68	82.35	6.56	64.22
	HM-LSTM	**186.45**	**8.02**	**133.40**	95.57	4.91	68.04	88.99	**5.79**	60.79

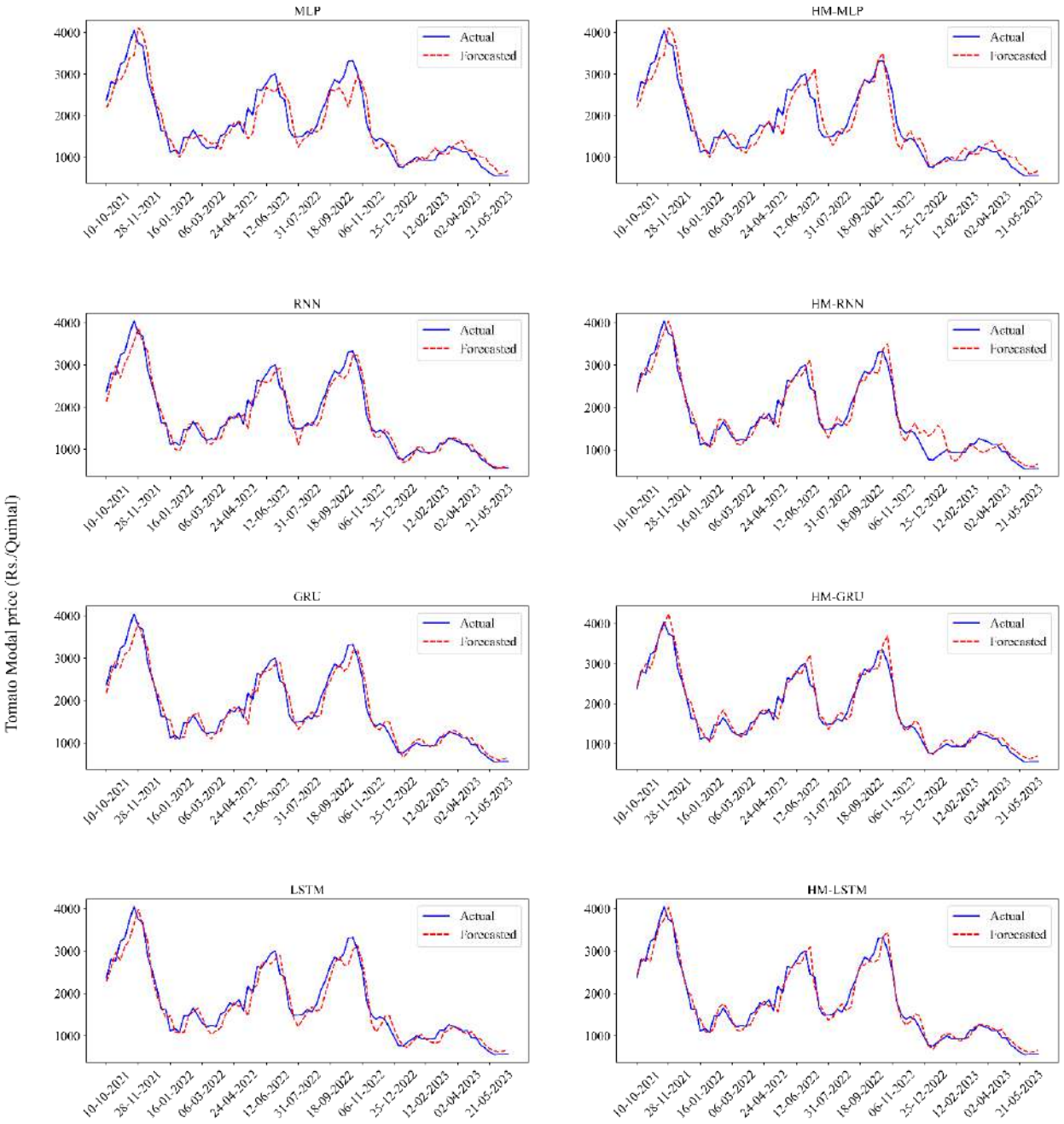


Figure 14: Forecasted Tomato price series obtained from DL and HM-DL models

The proposed HM-RNN model is the best among the HM-DL models proposed. The MAPE of the forecasts for the proposed HM-DL models, namely HM-MLP, HM-RNN, HM-GRU, HM-LSTM, were lower than their corresponding baseline models, namely MLP, RNN, GRU, LSTM, by 0.24–3.55%, 0.25–5.32%, and 0.60–2.72% respectively for the Tomtao, Onion and Potato prices. This reduction is more pronounced for the proposed HM-RNN model with a reduction as high as 5.32%.

It is also emphasized here that the proposed HM-DL models took almost the same computational time while training them when compared to the baseline DL models. The

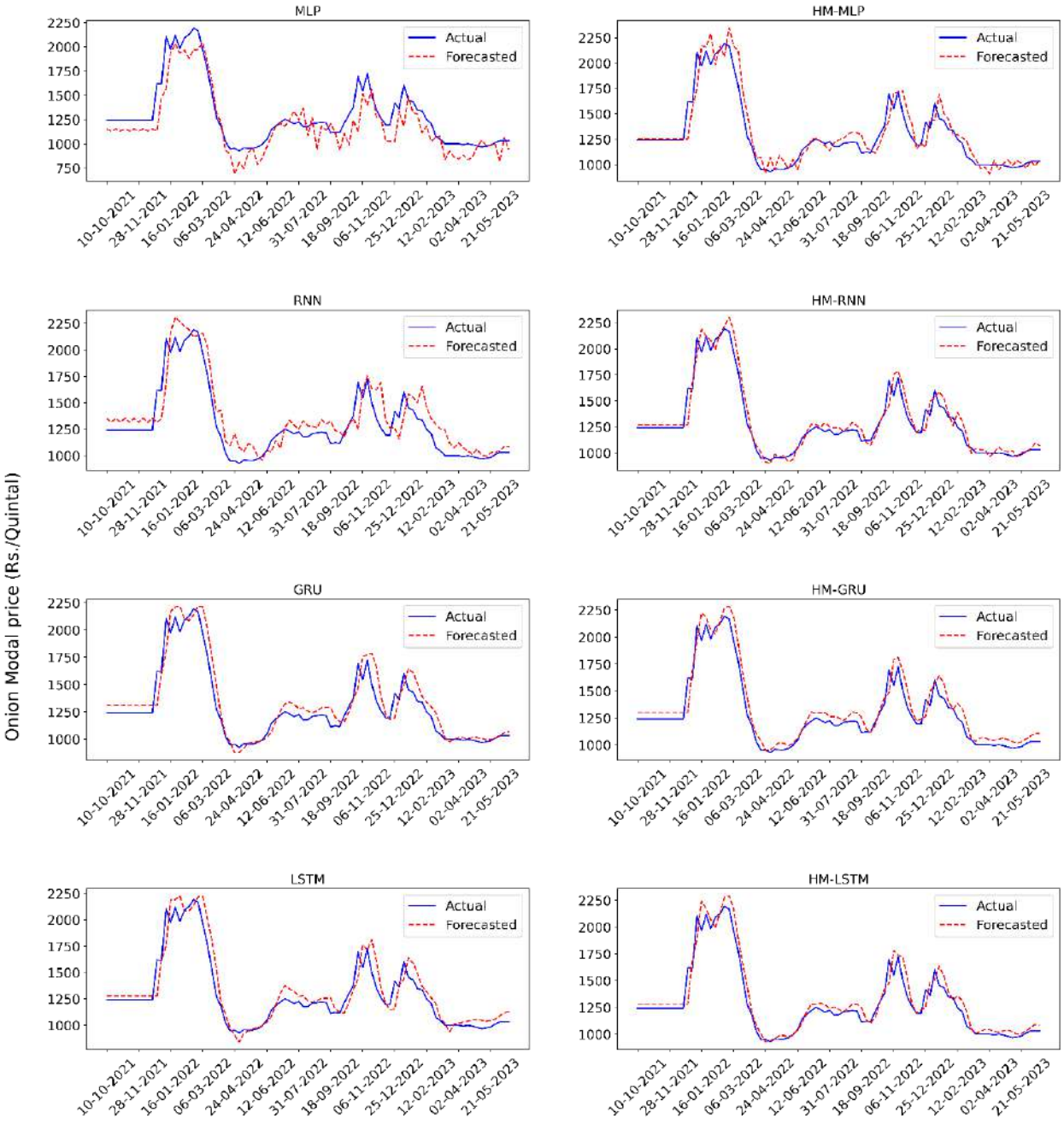


Figure 15: Forecasted Onion price series obtained from DL and HM-DL models

HMM models, when fitted in isolation on the three data series considered, seem to perform inferior as compared to other models, and hence HMM does not capture well the peaks and chaotic patterns present in the price series, even though it could capture the overall trend and latent structure of the data. Moreover, HMM cannot be used for long-term predictions, as the Markovian property assumes a simple conditional dependence of the present on the recent past.

For all the training datasets of the three TOP commodities, consistently HMM-RNN model has been found to be best fitted. On the testing datasets, while for the Onion and

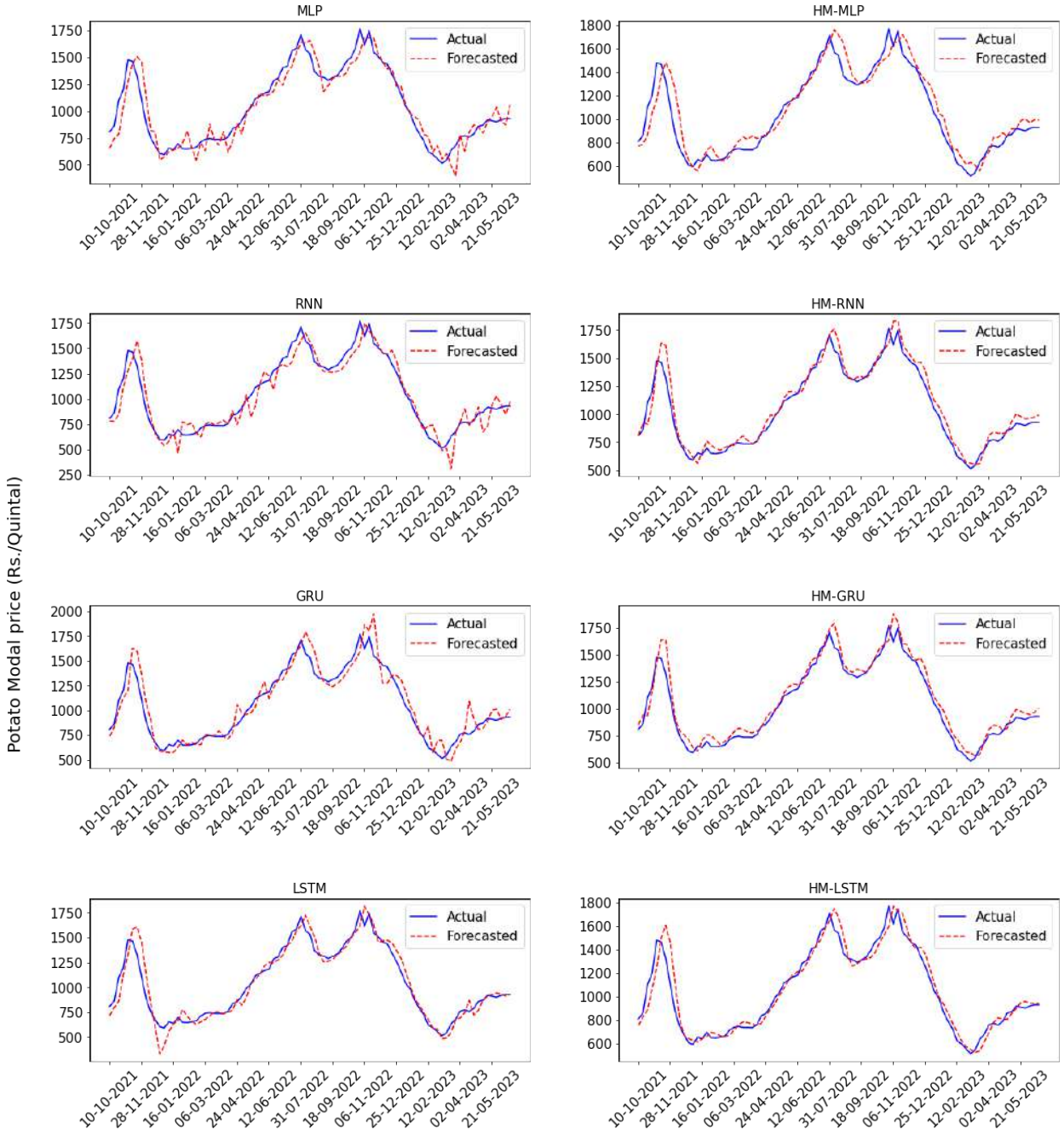


Figure 16: Forecasted Potato price series obtained from DL and HM-DL models

Potato prices, HMM-RNN model performed well as compared to the hybrid and conventional DL models, for the Tomato dataset, HMM-LSTM performed well as compared to HMM-RNN and other models. On further inspection, it has been found that, in the Tomato test dataset, three significant spikes were found with HMM-LSTM also capturing the long memory of the data quite well (as shown in figure 14)while in the other two test datasets (Onion and Potato), only one moderate spike each was present as shown in Figure (15 and 16).

Overall, from the Diebold-Mariano (DM) tests in Figures (17, 18 and 19), it can be

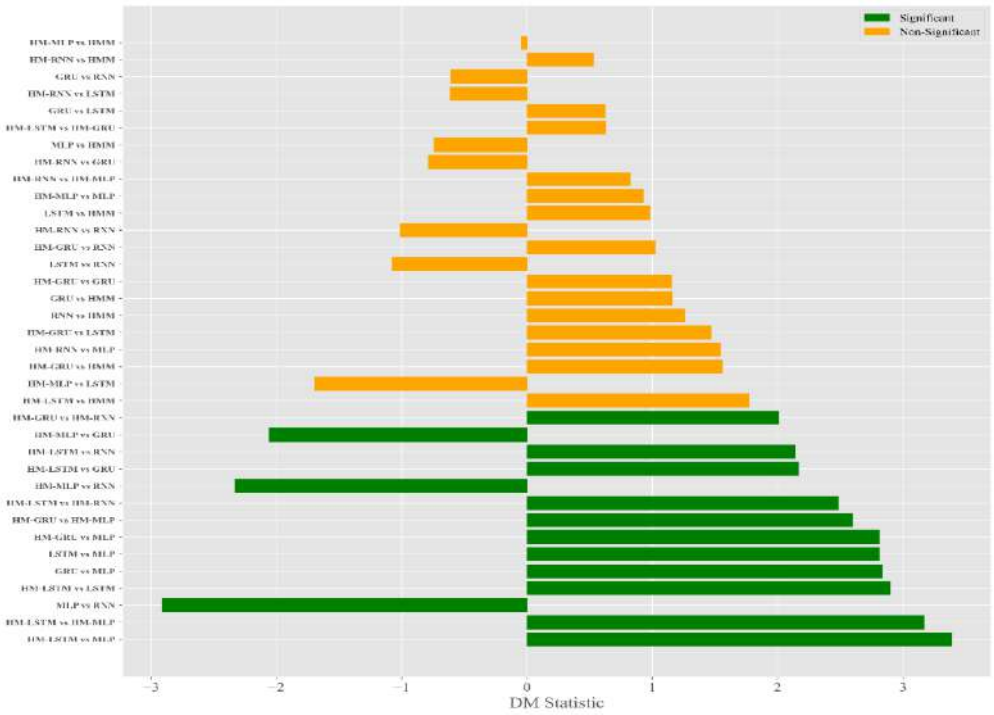


Figure 17: Various forecasting model comparison using DM test on Tomato price series

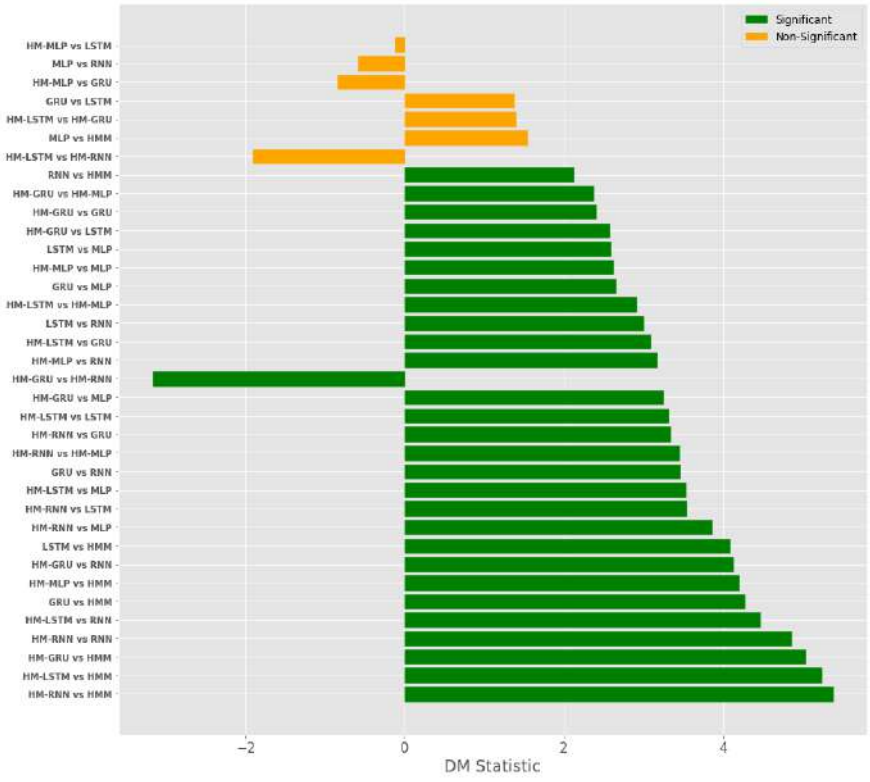


Figure 18: Various forecasting model comparison using DM test on Onion price series

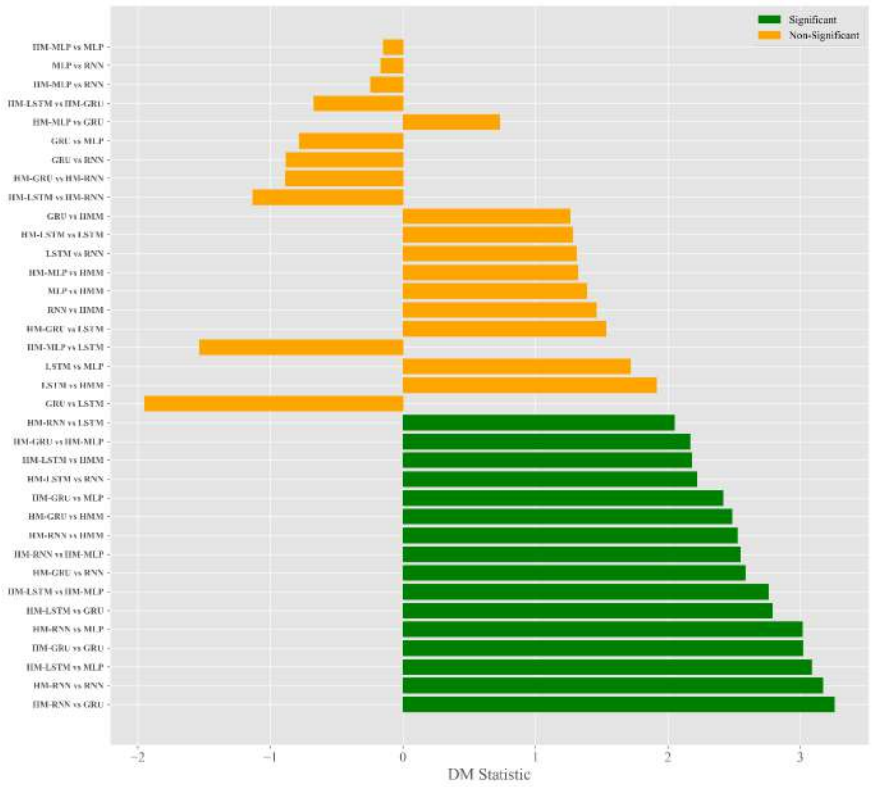


Figure 19: Various forecasting model comparison using DM test on Potato price series

inferred that the RMSEs of the HM-DL models *viz.*, HM-MLP, HM-RNN, HM-GRU, HM-LSTM were all significantly different (read lower) as compared to their DL counterparts for the Onion price series. For the Tomato series, HM-LSTM and for the Potato series, both HM-RNN and HM-GRU were found to be statistically significantly different.

HM based DL models have the advantage of adjusting to the pattern of the data within each of the hidden states found and hence the effect of non-stationarity of the data will be minimal. Meanwhile, HMM-DL models are able to handle data volatility, non-stationarity and non-normality better. However, in situations with datasets which are of relatively lesser size as compared to very large data sets, the application of HMM-DL models might underperform due to overfitting. In this study, careful hyperparameter tuning has been made to ensure model performance that avoided the overfitting issues.

To sum up, it can be concluded that Hidden Markov-Deep Learning (HMM-DL) approaches are more effective in forecasting TOP prices than traditional methods like HMM and baseline DL models. Thus, combining HMM with DL techniques seems to improve prediction accuracy even further, especially for long-term predictions like those required for agricultural commodities pricing.

This study demonstrates the effectiveness of HM-DL models for the accurate prediction of TOP vegetable prices. The proposed models provide farmers, traders, and Mandis with enhanced capabilities for reliable price forecasting and informed decision-making. Furthermore, the analysis enables farmers to optimize storage capacity by identifying periods of

low prices for storing vegetables and selling them during periods of higher prices, minimizing losses, as these can be readily seen by the stakeholders in the plots of forecasts. Overall, the hybrid HM-DL models offer a comprehensive understanding of market dynamics and provide valuable insights for optimizing decision-making in the agricultural sector.

5. Concluding remarks

This work proposed a novel Deep Learning (DL) approach based on hidden states to enhance the precision of TOP price forecasting. The hidden states identified by HMM serve as a feature extraction technique and were utilized in four DL models. The integration of HMM with DL and HMM models improved the forecasting accuracy compared to HMM and traditional DL models. The Diebold-Mariano (DM) tests, by and large, revealed that the RMSEs of the proposed HM-DL models were all significantly different (read lower) as compared to their DL counterparts for the onion price series. It is also emphasized here that the proposed HM-DL models took almost the same computational time while training them when compared to the baseline DL models. The findings demonstrate that the hybrid approach of Hidden Markov (HM) combined with DL models yields superior forecasting performance compared to existing models. Future research directions can extend the current study's univariate analysis of vegetable price series by incorporating multiple related variables, including weather conditions, market demand and economic indicators into the hybrid models. Moreover, other sectors such as finance, energy, and healthcare which involve complex time series data, can benefit from integrating HMM and DL models to improve forecasting accuracy and facilitate informed decision-making.

Acknowledgements

The facilities provided by ICAR-Indian Agricultural Statistics Research Institute (IASRI), New Delhi and the funding granted to the first author by Indian Council of Agricultural Research in the form of an IASRI-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-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.

References

Abdollahi, H. and Ebrahimi, S. B. (2020). A new hybrid model for forecasting brent crude oil price. *Energy*, **200**, 117520.

Adebiyi, A. A., Adewumi, A. O., and Ayo, C. K. (2014). Comparison of arima and artificial neural networks models for stock price prediction. *Journal of Applied Mathematics*, **2014**, 1–7.

Aizenberg, L., Sheremetov, L., Villa-Vargas, J., and Martinez-Munoz (2016). Multilayer neural network with multi-valued neurons in time series forecasting of oil production. *Neurocomputing*, **175**, 980–989.

Alom, M. Z., Taha, T. M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M. S., Hasan, M., Essen, B. C., Awwal, A. A. S., and Asari, V. K. (2019). A state-of-the-art survey on deep learning theory and architectures. *Electronics*, **8**, 292.

Althelaya, K. A., El-Alfy, E. S. M., and Mohammed, S. (2018). Stock market forecast using multivariate analysis with bidirectional and stacked LSTM and GRU. In *21st Saudi Computer Society National Computer Conference*, pages 1–7.

Ariyo, A. A., Adewumi, A. O., and Ayo, C. K. (2014). Stock price prediction using the ARIMA model. In *16th International Conference on Computer Modelling and Simulation*, pages 106–112.

Athanasopoulos, G. and De Silva, A. (2012). Multivariate exponential smoothing for forecasting tourist arrivals. *Journal of Travel Research*, **51**, 640–652.

Avinash, G., Ramasubramanian, V., and Gopalakrishnan, B. N. (2022). Heterogeneous autoregressive modeling based realised volatility forecasting. *Statistics and Applications*, **21**, 121–140.

Awad, M., Khanna, R., Awad, M., and Khanna, R. (2015). Hidden markov model. In *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*, pages 81–104.

Basak, S., Kar, S., Saha, S., Khaidem, L., and Dey, S. R. (2019). Predicting the direction of stock market prices using tree-based classifiers. *The North American Journal of Economics and Finance*, **47**, 552–567.

Baum, L. E. (1972). An inequality and associated maximization technique in statistical estimation for probabilistic functions of markov processes. *Inequalities*, **3**, 1–8.

Baum, L. E. and Petrie, T. (1966). Statistical inference for probabilistic functions of finite state markov chains. *The Annals of Mathematical Statistics*, **37**, 1554–1563.

Baum, L. E. and Sell, G. (1968). Growth transformations for functions on manifolds. *Pacific Journal of Mathematics*, **27**, 211–227.

Ghahramani, Z. (2001). An introduction to hidden markov models and bayesian networks. *International Journal of Pattern Recognition and Artificial Intelligence*, **15**, 9–42.

Giudici, P. and Abu Hashish, I. (2020). A hidden markov model to detect regime changes in cryptoasset markets. *Quality and Reliability Engineering International*, **36**, 2057–2065.

Guresen, E., Kayakutlu, G., and Daim, T. U. (2011). Using artificial neural network models in stock market index prediction. *Expert Systems with Applications*, **38**, 10389–10397.

Hashish, I. A., Forni, F., Andreotti, G., Facchinetti, T., and Darjani, S. (2019). A hybrid model for bitcoin price prediction using hidden markov models and optimized LSTM networks. In *2019 24th International Conference on Emerging Technologies and Factory Automation*, pages 721–728.

Hassan, M. R. (2009). A combination of hidden markov model and fuzzy model for stock market forecasting. *Neurocomputing*, **72**, 3439–3446.

Haykin, S. (2009). *Neural Networks and Learning Machines*. Pearson Education India.

Heidarpanah, M., Hooshyaripor, F., and Fazeli, M. (2023). Daily electricity price forecasting using artificial intelligence models in the iranian electricity market. *Energy*, **263**, 126011.

Henrique, B. M., Sobreiro, V. A., and Kimura, H. (2018). Stock price prediction using support vector regression on daily and up-to-the-minute prices. *The Journal of Finance and Data Science*, **4**, 183–201.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, **9**, 1735–1780.

Jaiswal, R., Jha, G. K., Kumar, R. R., and Choudhary, K. (2022). Deep long short-term memory-based model for agricultural price forecasting. *Neural Computing and Applications*, **34**, 4661–4676.

Jarque, C. M. and Bera, A. K. (1987). A test for normality of observations and regression residuals. *International Statistical Review/Revue Internationale de Statistique*, **55**, 163–172.

Jarque, C. M. and Bera, A. K. (2011). Arima modeling with intervention to forecast and analyze chinese stock prices. *International Journal of Engineering Business Management*, **3**, 53–58.

Jiang, M., Liu, J., Zhang, L., and Liu, C. (2020). An improved stacking framework for stock index prediction by leveraging tree-based ensemble models and deep learning algorithms. *Physica A: Statistical Mechanics and its Applications*, **541**, 122272.

Khan, M. I., Acharya, B., and Chaurasiya, R. K. (2022). Hybrid biLSTM-HMM based event detection and classification system for food intake recognition. In *2022 First International Conference on Electrical, Electronics, Information and Communication Technologies*, pages 1–5.

Kocak, C. (2017). Arma (p, q) type high order fuzzy time series forecast method based on fuzzy logic relations. *Applied Soft Computing*, **58**, 92–103.

Latif, N., Selvam, J. D., Kapse, M., Sharma, V., and Mahajan, V. (2023). Comparative performance of LSTM and ARIMA for the short-term prediction of bitcoin prices. *Australasian Accounting, Business and Finance Journal*, **17**, 256–276.

Lin, Z. (2018). Modelling and forecasting the stock market volatility of sse composite index using garch models. *Future Generation Computer Systems*, **79**, 960–972.

Makridakis, S., Spiliotis, E., and Assimakopoulos, V. (2018). Statistical and machine learning forecasting methods: Concerns and ways forward. *PloS one*, **13**, e0194889.

Manogna, R. and Mishra, A. K. (2021). Forecasting spot prices of agricultural commodities in india: Application of deep-learning models. *Intelligent Systems in Accounting, Finance and Management*, **28**, 72–83.

Mehdizadeh, S., Fathian, F., and Adamowski, J. F. (2019). Hybrid artificial intelligence-time series models for monthly streamflow modeling. *Applied Soft Computing*, **80**, 873–887.

Nelson, D. M., Pereira, A. C., and De Oliveira, R. A. (2017). Stock market's price movement prediction with LSTM neural networks. In *2017 International Joint Conference on Neural Networks*, pages 1419–1426.

Nti, K. O., Adekoya, A., and Weyori, B. (2019). Random forest-based feature selection of macroeconomic variables for stock market prediction. *American Journal of Applied Sciences*, **16**, 200–212.

Peng, Y., Feng, T., Yang, C., Leng, C., Jiao, L., Zhu, X., and Li, R. (2021). HMM-LSTM for proactive traffic prediction in 6g wireless networks. In *21st International Conference on Communication Technology (ICCT)*, pages 544–548.

Shapiro, S. S. and Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, **52**, 591–611.

Singh, K. N., Sharma, K., Avinash, G., Kumar, R. R., Ray, M., Ramasubramanian, V., Ray, M., Lama, A., and Lal, S. B. (2023). LSTM based stacked autoencoder approach for time series forecasting. *Journal of the Indian Society of Agricultural Statistics*, **77**, 71–78.

Wang, C., Yang, H., Bartz, C., and Meinel, C. (2016). Image captioning with deep bidirectional LSTMs. In *Proceedings of the 24th ACM International Conference on Multimedia*, pages 988–997.

Wang, H., Jiang, Y., and Wang, H. (2009). Stock return prediction based on bagging-decision tree. In *International Conference on Grey Systems and Intelligent Services*, pages 1575–1580.

Wang, L., Feng, J., Sui, X., Chu, X., and Mu, W. (2020). Agricultural product price forecasting methods: research advances and trend. *British Food Journal*, **122**, 2121–2138.

Weron, R. (2014). Electricity price forecasting: A review of the state-of-the-art with a look into the future. *International Journal of Forecasting*, **30**, 1030–1081.

Xiong, T. L. (2018). Seasonal forecasting of agricultural commodity price using a hybrid stl and elm method: Evidence from the vegetable market in china. *Neurocomputing*, **275**, 2831–2844.

Yao, Y. and Cao, Y. (2020). A neural network enhanced hidden markov model for tourism demand forecasting. *Applied Soft Computing*, **94**, 106465.

Yin, H., Jin, D., Gu, Y. H., Park, C. J., Han, S. K., and Yoo, S. J. (2020). STL-ATTLSTM: vegetable price forecasting using STL and attention mechanism-based LSTM. *Agriculture*, **10**, 612–619.

Yu, L., Zhao, Y., and Tang, L. (2017). Ensemble forecasting for complex time series using sparse representation and neural networks. *Journal of Forecasting*, **36**, 122–138.

Zaheer, S., Anjum, N., Hussain, S., Algarni, A. D., Iqbal, J., Bourouis, S., and Ullah, S. S. (2023). A multi-parameter forecasting for stock time series data using LSTM and deep learning model. *Mathematics*, **11**, 590.

APPENDIX

The following Python code implements the Hidden Markov Deep Learning (HM-DL) models for time series prediction.

```
#HM-DL models implementation by Python software

import pandas as pd
from sklearn.model_selection import train_test_split

# Load your dataset
my_data = pd.read_csv('path_to_your_data.csv')

# Convert 'Date' column to datetime if necessary
my_data['Date'] = pd.to_datetime(my_data['Date'])

# Ensure the data is sorted by date
my_data.sort_values('Date', inplace=True)

# Split data into train and test sets
train_size = int(len(my_data) * 0.9)
train_data, test_data = my_data[:train_size], my_data[train_size:]

from hmmlearn import hmm

# Initialize Gaussian HMM
# This assumes you've decided on the number of components based on
   your data (AIC/BIC)
model = hmm.GaussianHMM(n_components=number_of_states)

# Train HMM on the prices from the training data
model.fit(train_data[['Price']].values)

# Find the Viterbi path
hidden_states = model.predict(train_data[['Price']].values)

# Append Viterbi path to the training data
train_data['ViterbiPath'] = hidden_states

import tensorflow as tf
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import SimpleRNN, GRU, LSTM, Dense
from tensorflow.keras.optimizers import Adam
from sklearn.metrics import mean_squared_error

import tensorflow as tf
from tensorflow.keras.models import Sequential
from tensorflow.keras.layers import SimpleRNN, GRU, LSTM, Dense
from tensorflow.keras.optimizers import Adam
from sklearn.metrics import mean_squared_error
```

```
# Define a function to create a deep learning model of a specified
   type
def create_dl_model(input_shape, num_hidden_layers, num_hidden_units
   , dl_type):
    model = Sequential()

    if dl_type == 'HM-MLP':
        for i in range(num_hidden_layers):
            model.add(Dense(num_hidden_units, activation='relu',
               input_shape=input_shape if i == 0 else (
               num_hidden_units,)))
    elif dl_type == 'HM-GRU':
        for i in range(num_hidden_layers):
            return_sequences = i < (num_hidden_layers - 1)
            model.add(GRU(num_hidden_units, return_sequences=
               return_sequences, input_shape=input_shape if i == 0
               else (num_hidden_units,)))
    elif dl_type == 'HM-LSTM':
        for i in range(num_hidden_layers):
            return_sequences = i < (num_hidden_layers - 1)
            model.add(LSTM(num_hidden_units, return_sequences=
               return_sequences, input_shape=input_shape if i == 0
               else (num_hidden_units,)))
    elif dl_type == 'HM-RNN':
        for i in range(num_hidden_layers):
            return_sequences = i < (num_hidden_layers - 1)
            model.add(SimpleRNN(num_hidden_units, return_sequences=
               return_sequences, input_shape=input_shape if i == 0
               else (num_hidden_units,)))
    else:
        raise ValueError("Unsupported␣deep␣learning␣type")

    model.add(Dense(1))
    model.compile(loss='mean_squared_error', optimizer=Adam())
    return model

# Grid search over hyperparameters
best_rmse = float('inf')
best_model = None
best_params = {}

batch_sizes = [8,16,32,64,128,256]
num_hidden_layers = [1, 2,3]
num_hidden_units = [8,16,32,64,128,256,512]
num_epochs = 200  # Use early stopping criteria if needed.

for dl_type in ['HM-MLP', 'HM-GRU', 'HM-LSTM', 'HM-RNN']:
    for batch_size in batch_sizes:
        for num_layers in num_hidden_layers:
```

```
            for num_units in num_hidden_units:
                # Create model
                dl_model = create_dl_model(input_shape=(lags, 2),
                   num_hidden_layers=num_layers, num_hidden_units=
                   num_units, dl_type=dl_type)

                # Convert the series to a supervised learning
                   problem, split into input and output
                # X_train, y_train = ...

                # Train model
                dl_model.fit(X_train, y_train, epochs=num_epochs,
                   batch_size=batch_size, verbose=0)

                # Predict on training set and calculate RMSE
                train_predictions = dl_model.predict(X_train)
                train_rmse = mean_squared_error(y_train,
                   train_predictions, squared=False)

                # Update best model if RMSE improves
                if train_rmse < best_rmse:
                    best_rmse = train_rmse
                    best_model = dl_model
                    best_params = {'batch_size': batch_size, '
                       num_layers': num_layers, 'num_units':
                       num_units, 'dl_type': dl_type}
```

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 22, No. 2, 2024 (New Series), pp 91–108
http://www.ssca.org.in/journal

A New Lifetime Distribution: Statistical Inference and it's Applications

Arvind Pandey, Pawan Kumar Singh and Mahendra Saha
Department of Statistics, Central University of Rajasthan, India

Received: 14 March 2022; Revised: 16 October 2023; Accepted: 30 November 2023

Abstract

We have proposed a single parameter lifetime distribution function that has increasing and decreasing hazard rates. The proposed distribution can be used as a heavy-tailed alternative to the exponential, Weibull and gamma distributions. We discuss the different statistical properties, survival characteristics and stress-strength reliability. Different estimation procedures are used to estimate the parameter by using different methods which are given as estimation based on percentiles, least squares estimators, weighted least squares estimators, maximum likelihood estimators, Cramer-von-mises method of estimation and maximum product of the spacing method of estimation. We have compared the performance of these estimators with the help of simulation study. Also for different values of parameter we have found the estimates of hazard rate function and survival function. The proposed distribution is fitted to real data sets and it is observed that the proposed distribution is fitting quite well to the data.

Key words: Exponential distribution; Least squares estimation; Lifetime distribution; Maximum likelihood estimation; Maximum product spacing; Weighted least squares estimation

AMS Subject Classifications: 62K05, 05B05

1. Introduction

Lifetime distributions are common statistical tools used for the modeling and analysis of lifetime phenomena for different characteristics of lifetime data sets. The statistical literature contains very sophisticated multi-parameter distributions to analyze different kinds of data sets. Johnson *et al.* (1995) and Mann *et al.* (1974) discuss the importance of exponential distribution which is a single parameter distribution. The hazard rate of the exponential distribution is constant, which restricts its use in lifetime data analysis.

We are proposing a new single parameter distribution. The proposed distribution is obtained as a survival of a series system, which consists of two components, where one component follows inverse exponential and another follows Lomax distribution. Also, for the proposed distribution, we have studied the distributional properties of the series system.

Corresponding Author: Arvind Pandey
Email: arvindmzu@gmail.com

The proposed distribution has increasing hazard rate, decreasing hazard rate and first increasing then decreasing hazard rate. Also, it can be used as a heavy-tailed alternative to the exponential, Weibull and gamma distributions. This motivates us to introduce a new distribution and study some of its statistical properties. The proposed distribution can be used as a heavy-tailed alternative to the exponential, Weibull and gamma distributions.

The cumulative distribution function (CDF) of the proposed distribution is obtained by multiplying Lomax distribution and inverse exponential

$$F(x \mid \theta) = \begin{cases} \left(1 - \frac{\theta}{x+\theta}\right)(e^{-\theta/x}) & ; x \geq 0, \theta > 0 \\ 0 & ; \text{otherwise} \end{cases} \tag{1}$$

the corresponding probability density function (PDF) is given as,

$$f(x \mid \theta) = \begin{cases} \frac{\theta(2x+\theta)}{x(x+\theta)^2} e^{-\theta/x} & ; x \geq 0, \theta > 0 \\ 0 & ; \text{otherwise} \end{cases} \tag{2}$$

the hazard rate function (HRF) of the distribution for given t and it is denoted as h(t)

$$h(t) = \frac{f(t)}{S(t)} = \frac{\theta(2t+\theta)}{t(t+\theta)(te^{\theta/t} + \theta e^{\theta/t} - t)} \tag{3}$$

and survival function (SF) and it is denoted as S(t)

$$S(t) = 1 - \frac{t}{t+\theta} e^{-\theta/t} \tag{4}$$

the corresponding reversed hazard rate and it is denoted as r(t),

$$r(t) = \frac{f(t)}{F(t)} = \frac{\theta(2t+\theta)}{t^2(t+\theta)}$$

similarly, the cumulative hazard function and it is denoted as H(t)

$$H(t) = -log[S(t)] = -log\left[1 - \frac{t}{(t+\theta)} e^{-\theta/t}\right]$$

The plots of different characteristics of the proposed distribution are given in Figure (1) distribution function, Figure (2) probability density function, Figure (3) survival function and Figure (4) is hazard function. Figure (4) shows that the hazard rate first increases and then decreases for a given θ. This distribution can found its use in wide applications such as the analysis of the business failure lifetime data, income and wealth inequality, size of cities, actuarial science, medical and biological sciences, engineering, lifetime and reliability modeling.

In this article, we have proposed a new single parameter probability distribution and discuss its statistical properties. In section 2, we find the r^{th} moments and moments exist only for $r < 1$ and discuss the quantile function, skewness and kurtosis and also discuss the entropy of distribution.

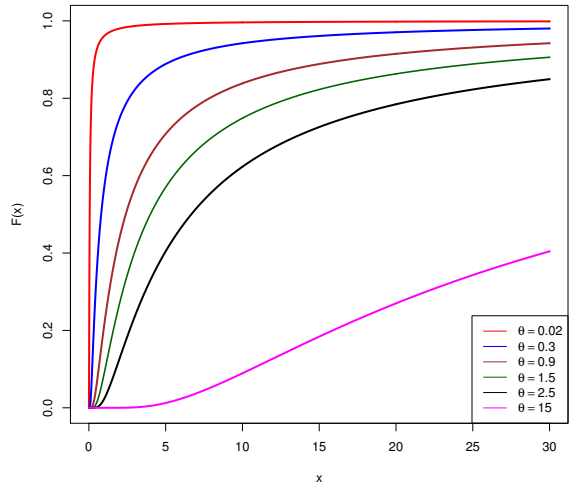


Figure 1: Various CDF forms of proposed distribution

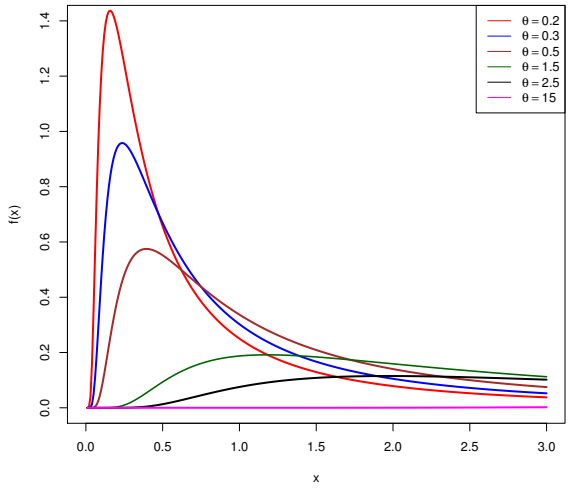


Figure 2: Various PDF forms of proposed distribution

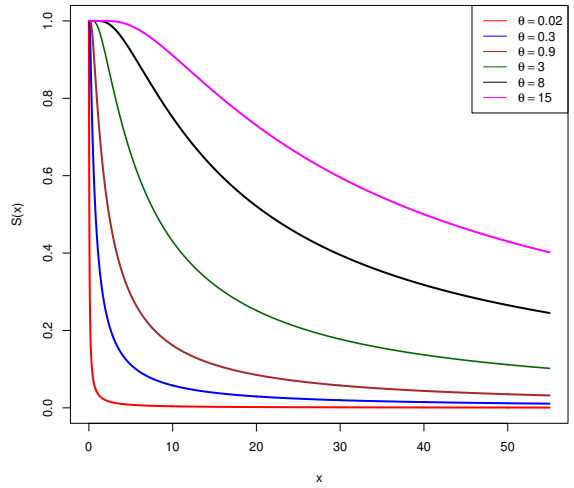


Figure 3: Various SF forms of proposed distribution

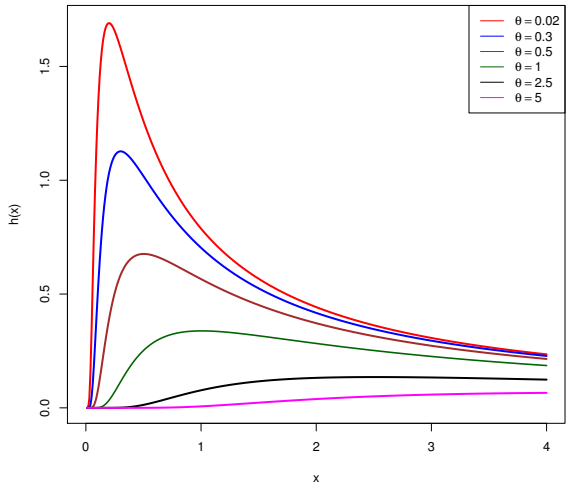


Figure 4: Various HRF forms of proposed distribution

In section (3) and section (4) we discuss the order statistics and stress strength reliability respectively. Similarly, in section (5) we discuss the different methods of estimation to estimate the parameter. In section (6), we mainly compare the different methods of estimations such as maximum likelihood estimators (MLE), estimators based on percentiles (PCE), least squares estimators (LSE), Weighted least squares estimators (WLSE), Cramer-von-Mises method of estimation (CME) and the maximum product of spacings method of estimations (MPSE), by mean squared errors (MSE) using extensive simulation techniques, Similarly, we estimate the HRF and SF. Real-life data applications are presented and discussed in section (7) and Concluding remarks can be found in section (8)

2. Some statistical properties

In this section, we discuss the different statistical properties, viz, moments, quantile function, skewness, kurtosis and entropy.

2.1. Moments

The r^{th} moments about origin is

$$E[X^r] = \int_0^\infty x^r \frac{\theta(2x+\theta)}{x(x+\theta)^2} e^{-\theta/x}\, dx \tag{5}$$

the moments exits only for $r < 1$ and $r \geq 0$ does not exits (see Appendix[1])

$$E[X^r] = \theta^r \left[\sum_{k=0}^{r} \binom{r}{k} (-1)^{(r-k)} \Gamma(k,1) + \sum_{k=1}^{r-1} \binom{r-1}{k} (-1)^{(r-k)} \Gamma(k,1) \right]$$

where, Γ(k,α) is the upper incomplete gamma function defined by

$$\Gamma(k,\alpha) = \int_\alpha^\infty e^{-x} x^{k-1} dx \quad \alpha \geq 0$$

An integral in equation (5) is evaluated by mathematically (see Fisher and Kılıcman (2012)) and also computational method using the Monte Carlo method. First, we draw a sample from the proposed distribution with parameter $\theta = 1.5$ with sample size n. We calculate the sample mean $E[x^r] = \frac{1}{n}\sum_{i=1}^{n} x_i^r$. The sample mean values are plotted with respect to different sample sizes. For $r = 0.2$, it is convergent which is shown in the Figure (5).

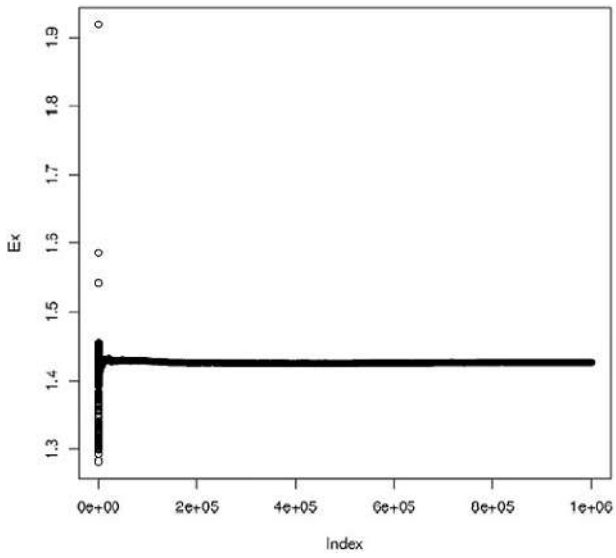


Figure 5: Moments at r=0.2

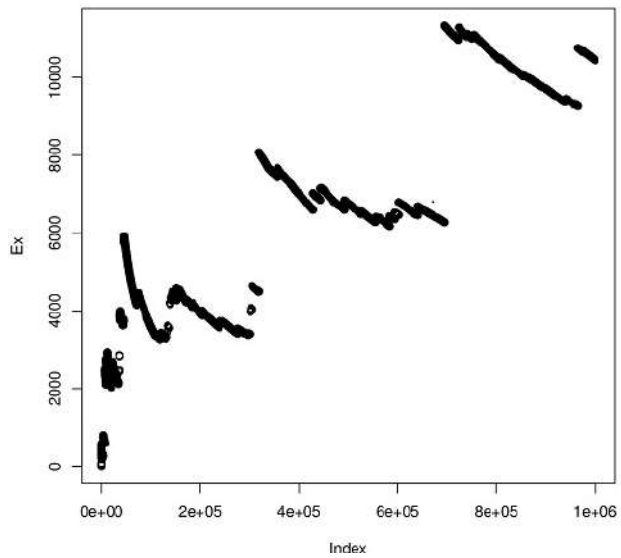


Figure 6: Moments at r=1.58

The value of the integration for $\theta = 1.5$ is 1.44. Which can be seen in Figure (5). The above integral equation (5) is not convergent for $r \geq 1$. This integral is evaluated by using the Monte Carlo method for $\theta = 1.5$ and $r = 1.5$. Observing the plot in Figure (6) the integration is not convergent, as can be seen from Figure (6).

2.2. Quantile function

The quantile function $Q(p)$ can be obtained by using the equation (1)

$$p = \frac{Q(p)}{Q(p)+\theta} e^{-\theta/Q(p)} \tag{6}$$

where Q(p) is the quantile of order p and $0 < p < 1$. If we put p=$\frac{1}{2}$ in equation (6) then we get the value of the median. Since moments of the proposed distribution does not exist, we can use the Moor measures of kurtosis (see Moors (1988)) and the Bowley measures of skewness(see Bowley (1920)) based on quantile and corresponding expressions are given in equation (7) and (8), respectively

$$K = \frac{Q(\frac{7}{8}) - Q(\frac{5}{8}) + Q(\frac{3}{8}) - Q(\frac{1}{8})}{Q(\frac{6}{8}) - Q(\frac{2}{3})} \tag{7}$$

$$S = \frac{Q(\frac{3}{4}) - 2Q(\frac{1}{2}) + Q(\frac{1}{4})}{Q(\frac{3}{4}) - Q(\frac{1}{4})} \tag{8}$$

2.3. Entropy

The concept of information entropy was introduced by Shannon (1948). Entropy measures the expected amount of information or "uncertainty" inherent in the possible outcomes of the variable. If the entropy is high then it indicates higher uncertainty.

2.4. Shannon entropy

Shannon's Entropy is simply the amount of information contained in a variable. It is defined as $H(f) = E\left[-log f(x)\right]$,

$$H(f) = E\left[-log\left[\frac{\theta(2x+\theta)}{x(x+\theta)^2} e^{-\theta/x}\right]\right] \tag{9}$$

To calculate the Shannon entropy by equation (9), solving by Monte Carlo integration method. The generate the x_i from proposed distribution and calculating,
$\frac{-1}{n}\sum_{i=1}^{n} log\left[\frac{\theta(2x_i+\theta)}{x_i(x_i+\theta)^2} e^{-\theta/x_i}\right]$, we see that above Figure (7) value converges at 3.24 for given value of $\theta = 1.5$.

2.5. Renyi entropy

In the information theory, if X is a random variable with density function f(x), the Renyi entropy is a measure of uncertainty of the random variable defined and it is denoted as $H(\gamma)$

$$H(\gamma) = \frac{1}{1-\gamma} log\left(\int_0^\infty \left[f(x)\right]^\gamma \, dx\right)$$

$$H(\gamma) = \frac{1}{1-\gamma} log\left(\int_0^\infty \left[\frac{\theta(2x+\theta)}{x(x+\theta)^2} e^{-\theta/x}\right]^\gamma \, dx\right) \tag{10}$$

Similarly, we can calculate the integral of equation (10) by the same method we see that the integral does not converge because in the Figure (8) plot does not converge at any point and mathematically when moments does not exits for $r \geq 1$ then Renyi entropy does not exits.

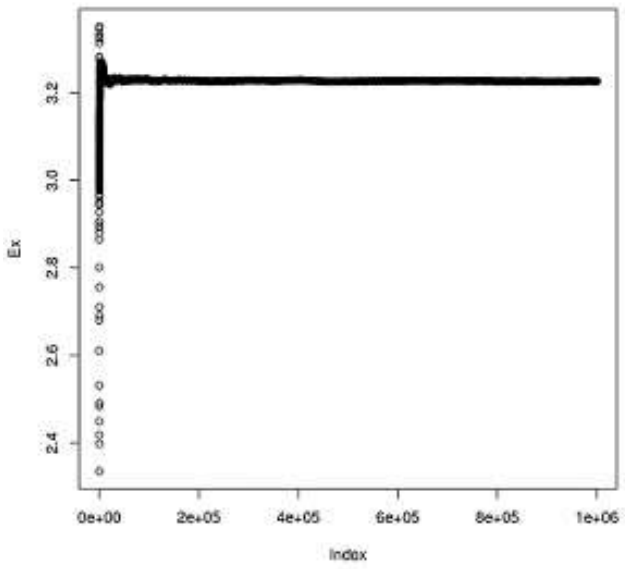


Figure 7: Shannon Entropy

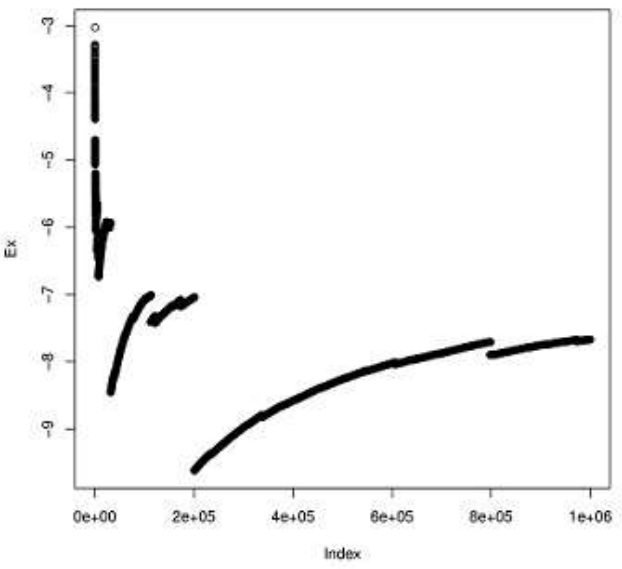


Figure 8: Renyi Entropy

2.6. Regularly varying tail behavior of the proposed distribution

A distribution function F with survival function (SF), $S(y) = 1 - F(y)$, is said to be heavy-tailed if for every t ≥ 0, $\lim_{y\to\infty} \frac{1-F(y)}{e^{-ty}} = \infty$ (see section 2.4 of Rolski *et al.* (2009)). A distribution function F is said to belong to the regularly varying class if

$$\lim_{y\to\infty} \frac{1 - \frac{y}{y+\theta}e^{-\theta/y}}{e^{-ty}}$$

$$\lim_{y\to\infty} \frac{e^{tx}}{(y+\theta)^2} \lim_{y\to\infty} \frac{\theta(2+\frac{\theta}{y})e^{\frac{-\theta}{y}}}{2t\theta}$$

$$\lim_{y\to\infty} \theta t e^{ty} = \infty \qquad \forall \;\; t \geq 0$$

3. Distributions of order statistics and ordering property

First, we know the PDF, CDF and the moment of the i^{th} order statistics $x_{(i)}$. Let $X_1, X_2, ..., X_n$ are random samples of size n from the proposed continuous distribution, then the PDF of order statistics is given by the following formula

$$f_{(i,n)}(x \mid \theta) = \frac{n!}{(n-i)!(i-1)!} \left[\frac{x}{x+\theta}e^{-\theta/x}\right]^{i-1} \left[1 - \frac{x}{x+\theta}e^{-\theta/x}\right]^{n-i} \frac{\theta(2x+\theta)}{x(x+\theta)^2}e^{-\theta/x}$$

The PDF of maximum of $X_1, X_2, ..., X_n$ is

$$f(x_{(n)} \mid \theta) = n\left[F(x)\right]^{n-1} f(x) = n\left[\frac{x}{x+\theta}e^{-\theta/x}\right]^{n-1} \frac{\theta(2x+\theta)}{x(x+\theta)^2}e^{-\theta/x}$$

The PDF of minimum of $X_1, X_2, ..., X_n$ is

$$f(x_{(1)} \mid \theta) = n\left[1-F(x)\right]^{n-1} f(x) = n\left[1-\frac{x}{x+\theta}e^{-\theta/x}\right]^{n-1} \frac{\theta(2x+\theta)}{x(x+\theta)^2}e^{-\theta/x}$$

The joint CDF of minimum and maximum distribution of $X_1, X_2, ..., X_n$ is

$$F_{x_{(1)},x_{(n)}}(x,y) = P[x_{(1)} \leq x, x_{(n)} \leq y] = [F(y)]^n - [F(y)-F(x)]^n$$

$$F_{x_{(1)},x_{(n)}}(x,y) = \left[\frac{y}{y+\theta}e^{-\theta/y}\right]^n - \left[\frac{y}{y+\theta}e^{-\theta/y} - \frac{x}{x+\theta}e^{-\theta/x}\right]^n$$

The joint PDF of minimum and maximum of $X_1, X_2, ..., X_n$ is

$$f_{x_{(1)},x_{(n)}}(x,y) = n(n-1)[F(y)-F(x)]^{n-2} f(x) f(y)$$

$$f_{x_{(1)},x_{(n)}}(x,y) = n(n-1)\left[\frac{y}{y+\theta}e^{-\theta/y} - \frac{x}{x+\theta}e^{-\theta/x}\right]^{n-2} \frac{\theta^2(2x+\theta)(2y+\theta)}{xy(x+\theta)^2(y+\theta)^2}e^{-\theta(\frac{x+y}{xy})}$$

4. Stress-strength reliability

Stress-strength reliability (SSR) describe the life of a component having a random strength X that is subject to a random stress Y. The component fails at an instant, when the stress applied to it exceeds the strength and the component function satisfactorily whenever $X > Y$. Let X and Y follows the proposed distribution with parameters θ_1 and θ_2 then,

$$R = P[X > Y] = \int_0^\infty P[X > Y | Y = y] f_Y(y) dy$$

$$P[X > Y] = \int_0^\infty \frac{\theta_2(2y+\theta_2)}{(y+\theta_1)(y+\theta_2)^2} e^{-(\frac{\theta_1+\theta_2}{y})} dy$$

Table 1: True value of SSR of given value of θ_1 and θ_2

θ_2 \ θ_1	0.2	0.8	1.0	1.5	2.0	2.5	5.0
0.2	0.5000	0.2156	0.1820	0.1313	0.1027	0.0845	0.0005
0.8	0.7844	0.5000	0.4486	0.3584	0.2994	0.2575	0.1524
1.0	0.8179	0.5514	0.5000	0.4073	0.3447	0.2994	0.1820
1.5	0.8687	0.6415	0.5926	0.5000	0.4338	0.3839	0.2461
2.0	0.8972	0.7005	0.6552	0.5661	0.5000	0.4485	0.2994
2.5	0.9155	0.7424	0.7005	0.6160	0.5514	0.5000	0.3447
5.0	0.9552	0.8475	0.8179	0.7539	0.7005	0.6552	0.5000

5. Statistical inference

In this section, we discuss six classical methods of estimation, viz. method of MLE, method of LSE, method of WLSE, method of CME, method of PCE and methods of MPSE.

5.1. Estimation based on percentiles

Among the most easily obtained estimators of the parameters of the Weibull distribution are the graphical approximation to the best linear unbiased estimators. It can be obtained by fitting a straight line to the theoretical points obtained from the distribution function and the sample percentile points. This method was originally explored by Kao (1959), see also Mann *et al.* (1974) and Johnson *et al.* (1995). It is possible for the Weibull case because of the nature of its distribution function. In the case of a proposed distribution also it is possible to use the same concept to obtain the estimator of θ based on the percentiles, because of the structure of its distribution function.
Now

$$F(x \mid \theta) = \frac{x}{x+\theta} e^{-\theta/x}$$

If p_i denotes some estimate of $F(x \mid \theta)$ then the estimate of θ can be obtained by minimizing

$$\sum_{i=1}^{n} [p_i(x_i + \theta) - x_i e^{-\theta/x_i}]^2 \tag{11}$$

Partially differentiate equation (11) with respect to θ, we get

$$\sum_{i=1}^{n} [(p_i((x_i + \theta) - x_i e^{-\theta/x_i})(p_i + e^{-\theta/x_i})] = 0 \tag{12}$$

Solving equation (12) using non-linear method and find the value of θ for different value of p_i, for example $p_i = (i/(n+1))$ is the most used estimator of $F(x_{(i)})$, as $(i/(n+1))$ is the expected value of $F(x_{(i)})$. We have also used this p_i here. Some of the other choices of p_i's are $p_i = ((i - 3/8)/(n + 1/4))$ or $p_i = ((i - 1/2)/n)$ (see Mann *et al.* (1974)) although they have not pursued here. We get the value of θ is know as $\hat{\theta}_{PCE}$ substituting the $\hat{\theta}_{PCE}$ in equation (3) and (4), we can get the estimators of HRF estimate h(x) SF S(x) given as

$$\hat{h}(x)_{PCE} = \frac{\hat{\theta}_{PCE}(2x + \hat{\theta}_{PCE})}{x(x + \hat{\theta}_{PCE})(xe^{\hat{\theta}_{PCE}/x} + \hat{\theta}_{PCE} e^{\hat{\theta}_{PCE}/x} - x)} \tag{13}$$

and

$$\hat{S}(x)_{PCE} = 1 - \frac{x}{x + \hat{\theta}_{PCE}} e^{-\hat{\theta}_{PCE}/x} \tag{14}$$

5.2. Least squares estimators

In this method we provide the regression based method estimators of the unknown parameters, which was originally suggested by Swain *et al.* (1988) to estimate the parameters of Beta distributions. It can be using some other cases also. Suppose $X_1, ..., X_n$ is a random sample of size n from distribution function $G(.)$ and suppose $x_{(i)}; i = 1, ..., n$ denotes the

ordered sample. The proposed method uses the distribution of $G(X_{(i)})$. For a sample of size n, we have

$$E(G(X_{(i)})) = \frac{i}{n+1}$$
$$V(G(X_{(i)})) = \frac{i(n-i+1)}{(n+1)^2(n+2)}$$

See Johnson *et al.* (1995). Using the expectations and the variances, two variants of the least squares methods can be used. The least square estimate can be obtained by minimizing the,

$$LS(\theta) = \sum_{i=1}^{n} \left[\frac{x_i}{x_i+\theta} e^{-\theta/x_i} - \frac{i}{n+1} \right]^2 \tag{15}$$

When differentiated equation (15) with respect to θ and equated to 0. Then we get the value of θ which is called $\hat{\theta}_{LSE}$. Substituting these $\hat{\theta}_{LSE}$ in equation (3) and (4), we can get the estimators of HRF estimate h(x), SF S(x) given as,

$$\hat{h}(x)_{LSE} = \frac{\hat{\theta}_{LSE}(2x+\hat{\theta}_{LSE})}{x(x+\hat{\theta}_{LSE})(xe^{\hat{\theta}_{LSE}/x} + \hat{\theta}_{LSE}e^{\hat{\theta}_{LSE}/x} - x)} \tag{16}$$

and

$$\hat{S}(x)_{LSE} = 1 - \frac{x}{x+\hat{\theta}_{LSE}} e^{-\hat{\theta}_{LSE}/x} \tag{17}$$

5.3. Weighted least squares estimators

The weighted least squares estimation minimizes the equation given below,

$$\sum_{i=1}^{n} W_i \left(G(X_i) - \frac{i}{n+1} \right)^2$$

with respect to the unknown parameters, where

$$W_i = \frac{1}{V(G(X_{(i)}))} = \frac{(n+1)^2(n+2)}{i(n-i+1)}$$

Therefore, in case of distribution the weighted least squares of θ can be obtained by minimizing

$$W(\theta) = \sum_{i=1}^{n} \frac{(n+1)^2(n+2)}{i(n-i+1)} \left[F(x_i \mid \theta) - \frac{i}{n+1} \right]^2$$
$$W(\theta) = \sum_{i=1}^{n} \frac{(n+1)^2(n+2)}{i(n-i+1)} \left[\frac{x_i}{x_i+\theta} e^{-\theta/x_i} - \frac{i}{n+1} \right]^2 \tag{18}$$

Differentiating the above equation (18) with respect to θ and equating to 0 we get the estimate of θ and the estimated value of θ is denoted as $\hat{\theta}_{WLSE}$. Substituting the $\hat{\theta}_{WLSE}$, in equation (3) and (4) , we can get the estimate $\hat{h}(x)$ and $\hat{S}(x)$ given below,

$$\hat{h}(x)_{WLSE} = \frac{\hat{\theta}_{WLSE}(2x+\hat{\theta}_{WLSE})}{x(x+\hat{\theta}_{WLSE})(xe^{\hat{\theta}_{WLSE}/x} + \hat{\theta}_{WLSE}e^{\hat{\theta}_{WLSE}/x} - x)} \tag{19}$$

and

$$\hat{S}(x)_{WLSE} = 1 - \frac{x}{x+\hat{\theta}_{WLSE}} e^{-\hat{\theta}_{WLSE}/x} \tag{20}$$

5.4. Maximum likelihood estimator

In this section, the MLE of the distribution function is considered. If $X_1, X_2..., X_n$ is a random sample from proposed distribution with parameter θ, then the likelihood function, L(θ), is

$$L = \prod_{i=1}^{n} \frac{\theta(2x_i+\theta)}{x_i(x_i+\theta)^2} e^{-\theta/x_i}$$

$$log(L(\theta)) = nlog(\theta) + \sum_{i=1}^{n} log(2x_i+\theta) - \sum_{i=1}^{n} log(x_i) - \sum_{i=1}^{n} 2log(x_i+\theta) - \sum_{i=1}^{n} \frac{\theta}{x_i} \tag{21}$$

Differentiating the above equation (21) with respect to θ and $\frac{\partial logL(\theta)}{\partial \theta} = 0$ we get

$$\frac{n}{\theta} + \sum_{i=1}^{n} \frac{1}{(2x_i+\theta)} - \sum_{i=1}^{n} \frac{2}{(x_i+\theta)} - \sum_{i=1}^{n} \frac{1}{x_i} = 0 \tag{22}$$

The above equation is a non-linear equation which can be solved by a simple iterative procedure can be used to find a solution and we can estimate the value of θ and the estimated value is called as $\hat{\theta}_{MLE}$ substituting the $\hat{\theta}_{MLE}$, in equation (3) and (4), we can get the estimators of HRF estimate h(x), SF S(x) given as,

$$\hat{h}(x)_{MLE} = \frac{\hat{\theta}_{MLE}(2x+\hat{\theta}_{MLE})}{x(x+\hat{\theta}_{MLE})(xe^{\hat{\theta}_{MLE}/x} + \hat{\theta}_{MLE}e^{\hat{\theta}_{MLE}/x} - x)} \tag{23}$$

and

$$\hat{S}(x)_{MLE} = 1 - \frac{x}{x+\hat{\theta}_{MLE}} e^{-\hat{\theta}_{MLE}/x} \tag{24}$$

5.5. Cramer-von-Mises method of estimation

To motivate our choice of CME type minimum distance estimators, Macdonald (1971) provided empirical evidence that the bias of the estimator is smaller than the other minimum distance estimators. Thus, the proposed estimators are based on the Cramer-von Mises statistics given by,

$$C(\theta) = \frac{1}{12n} + \sum_{i=1}^{n} \left[F(x,\theta) - \frac{2i-1}{2n} \right]^2$$

$$C(\theta) = \frac{1}{12n} + \sum_{i=1}^{n} \left[\frac{x_i}{x_i+\theta} e^{-\theta/x_i} - \frac{2i-1}{2n} \right]^2 \tag{25}$$

Then the Cramer-von-Mises estimator are obtained by minimizing the above equation (25) with respect to θ, $\frac{\partial C(\theta)}{\partial \theta} = 0$. This estimator can also be obtained by solving the following non-linear equation. We get the estimated value of θ and the estimated value is called as $\hat{\theta}_{CME}$

$$\sum_{i=1}^{n} \left[\left(\frac{x_i}{x_i+\theta} e^{-\theta/x_i} - \frac{2i-1}{n} \right) \left(-\frac{2x_i+\theta}{(x_i+\theta)^2} e^{-\theta/x_i} \right) \right] = 0 \tag{26}$$

Substituting the $\hat{\theta}_{CME}$, in equation (3) and (4), we can get the estimators of HRF estimate h(x), SF S(x) given as,

$$\hat{h}(x)_{CME} = \frac{\hat{\theta}_{CME}(2x+\hat{\theta}_{CME})}{x(x+\hat{\theta}_{CME})(xe^{\hat{\theta}_{CME}/x} + \hat{\theta}_{CME}e^{\hat{\theta}_{CME}/x} - x)} \tag{27}$$

and

$$\hat{S}(x)_{CME} = 1 - \frac{x}{x+\hat{\theta}_{CME}} e^{-\hat{\theta}_{CME}/x} \tag{28}$$

5.6. Maximum product of spacing method of estimation

The Maximum Product of Spacing method of estimation (MPSE) is an alternative to MLE for the estimation of the unknown parameters of continuous uni-variate distributions. This method is used for estimating parameter, In this method estimation of the parameter θ is obtained by maximizing the geometric mean of the spacing with respect to θ

$$G(\theta) = \left[\prod_{i=1}^{n+1} \left[F(x_{(i,n)}|\theta) - F(x_{(i-1,n)}|\theta)\right]\right]^{\frac{1}{n+1}}$$

Taking log on both sides the above equation we get,

$$logG(\theta) = \frac{1}{n+1} \sum_{1}^{n+1} log\left[F(x_{(i,n)}|\theta) - F(x_{(i-1,n)}|\theta)\right] \tag{29}$$

To maximize the above equation (29) we differentiate with respect to θ, $\frac{\partial logG(\theta)}{\partial \theta} = 0$, we get the estimated value of θ and the estimated value is called as $\hat{\theta}_{MPSE}$ shown in equation (30), substituting the $\hat{\theta}_{MPSE}$, in equation (3) and (4), we can get the estimators of HRF estimate h(x), SF S(x) given as,

$$\frac{1}{n+1} \frac{\sum_{i=1}^{n} \left(\frac{(2x_i+\theta)}{(x_i+\theta)^2} e^{-\theta/x_i} - \frac{(2x_{i-1}+\theta)}{(x_{i-1}+\theta)^2} e^{-\theta/x_{i-1}}\right)}{\sum_{i=1}^{n} \left(\frac{x_i}{x_i+\theta} e^{-\theta/x_i} - \frac{x_{i-1}}{x_{i-1}+\theta} e^{-\theta/x_{i-1}}\right)} = 0 \tag{30}$$

$$\hat{h}(x)_{MPSE} = \frac{\hat{\theta}_{MPSE}(2x+\hat{\theta}_{MPSE})}{x(x+\hat{\theta}_{MPSE})(xe^{\hat{\theta}_{MPSE}/x} + \hat{\theta}_{MPSE}e^{\hat{\theta}_{MPSE}/x} - x)} \tag{31}$$

and

$$\hat{S}(x)_{MPSE} = 1 - \frac{x}{x+\hat{\theta}_{MPSE}} e^{-\hat{\theta}_{MPSE}/x} \tag{32}$$

6. Simulation study

In this section, we generate random sample from the proposed distribution by using the inverse transformation method. We used the Monto Carlo simulation study to assess the performance of the proposed estimators (PSE, LSE, WLSE, MLE, CME, MPSE) of the parameter θ for the proposed distribution. We used the particular values of θ= 0.2, 1.05, 2.5, 5 and the corresponding sample size is n= 5, 15, 30, 50, 100, 500 for each design we draw the sample of size n from the original sample and it is replicated 10,000 times. We calculate the mean of parameters θ using PSE, LSE, WLSE, MLE, CME and MPSE, and their corresponding MSEs, the results are reported in Table (6). It is observed that all the methods follow the same pattern, by increasing sample size the corresponding MSEs are decreasing in all the methods of estimation. On basis of MSEs we find the Maximum Product of Spacing method of estimation is the best method of estimation among the method used.

Similarly, the simulation for hazard estimations and survival estimations for same sample sizes and for a given value of t= 0.75, 1, 1.5 and given $\hat{\theta}_{MLE}$ we calculate the hazard and survival estimations and corresponding calculate MSEs for 10,000 times. The calculated average estimate of hazard and survival estimation with corresponding MSE and the results are reported in Table (7) and Table (8). It is observed that the HRF estimation and SF estimation also follow the same pattern as with, the increase in the sample size corresponding MSEs decreases.

7. Real data applications

In this Section, we use two data sets and comparing the existing model. For more details we see in Shukla (2019) and Shanker *et al.* (2015).

7.1. Data set 1

We use the survival times of a group of patients suffering from head and neck cancer disease and treated using a combination of radiotherapy and chemotherapy which is reported by Shukla (2019) and Shanker *et al.* (2015).
432.00 140.00 119.00 47.38 58.36 195.00 155.00 339.00 209.00 112.00 194.00 519.00 68.46 25.87 179.00 78.26 159.00 84.00 31.98 110.00 1776.00 725.00 173.00 41.35 94.00 74.47 633.00 319.00 127.00 146.00 281.00 23.56 92.00 249.00 37.00 23.74 133.00 130.00 12.20 63.47 81.43 55.46 817.00 469.00

Table 2: Values of the estimate of parameter for given real data set 1 and corresponding AIC, AICc, BIC, KS and p-value

Model	$\hat{\theta}$	-2ln L	AIC	AICc	BIC	KS	p-value
Lindley	0.00891	579.16	581.16	581.26	582.95	0.219	0.0243
Exponential	0.00447	564.02	566.02	566.11	567.80	0.145	0.2838
PD_{MLE}	45.33775	558.77	560.77	560.86	562.55	0.089	0.8509
PD_{MPSE}	41.95271	558.98	560.98	561.07	562.76	0.074	0.9562

Table 3: Value of hazard estimate and survival estimate for real data set 1 for given value of time t

t	θ_{MPSE}	$\hat{h(t)}_{MPSE}$	$\hat{S(t)}_{MPSE}$	θ_{MLE}	$\hat{h(t)}_{MLE}$	$\hat{S(t)}_{MLE}$
t=223.48(mean)	41.9527	0.00357	0.30216	45.33775	0.00351	0.32131
t=128.50(median)		0.00531	0.45611		0.00516	0.48057

7.2. Data set 2

We use the times between successive failures of air conditioning equipment in a Boeing 720 airplane data set which is reported by Shukla (2019) and Shanker *et al.* (2015) come from data set (13). 386 70 57 12 59 29 74 27 153 48 326 21 26 29 502.

In Table 2 and Table 4 estimated values of parameter, -2lnL, AIC, AICc, BIC, KS statistics and p value are given and the basis of these value our proposed distribution is better

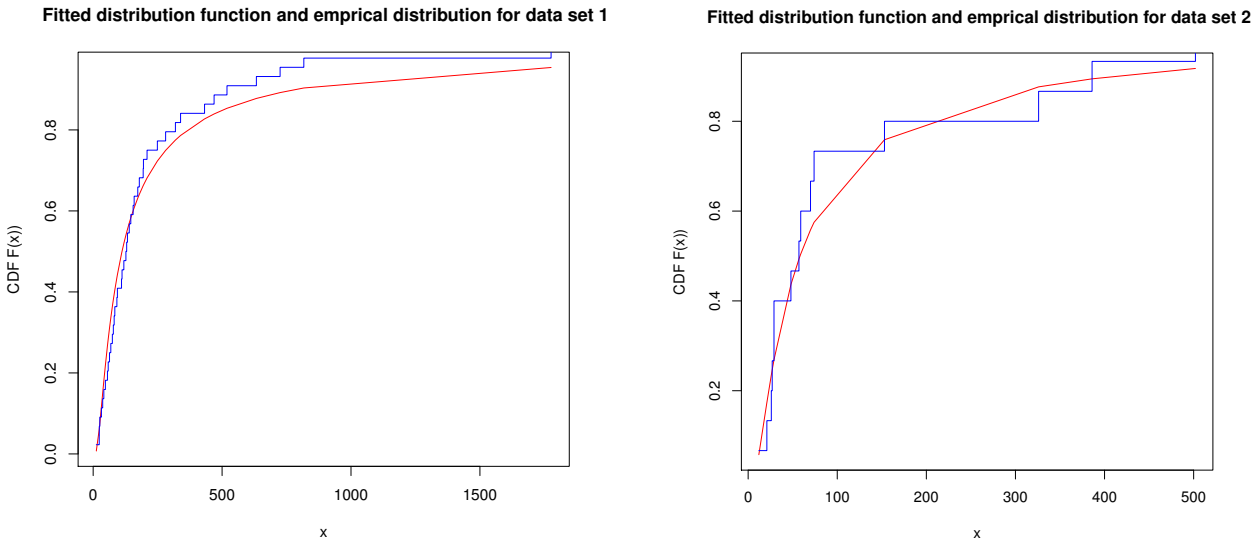
Table 4: Values of the estimate of parameter for given real data set 2 and corresponding AIC, AICc, BIC, KS and p-value

Model	Estimate	-2ln L	AIC	AICc	BIC	KS	p-value
Lindley	0.01636	181.34	183.34	183.65	184.05	0.386	0.0159
Exponential	0.00824	173.94	175.94	176.25	176.65	0.277	0.1662
PD_{MLE}	23.12604	169.25	171.25	171.56	171.96	0.176	0.6789
PD_{MPSE}	21.82070	169.29	171.29	171.60	172.00	0.158	0.7923

Table 5: Value of hazard estimate and survival estimate for real data set 2 for given value of time t

t	θ_{MPSE}	$\hat{h(t)}_{MPSE}$	$\hat{S(t)}_{MPSE}$	θ_{MLE}	$\hat{h(t)}_{MLE}$	$\hat{S(t)}_{MLE}$
t=121.2667(mean)	21.8207	0.00664	0.29206	23.12604	0.00656	0.30598
t=57(median)		0.01126	0.50685		0.01098	0.52587

than exponential distribution and one parameter Lindley distribution. It is also observed that estimate by using MPSE value provide better fit for both data sets 1 and 2 on basis of KS statistics and p value but there is no difference of -2lnL, AIC, AICc, BIC. It is same line supports the results for simulation study for our proposed distribution MPSE is better results than other methods.



(a) data set-1 **(b) data set-2**

8. Conclusions

In this article, we have proposed a new probability distribution, which is having an increasing hazard rate and decreasing hazard rate, heavy-tailed properties hold. Moments exist only for $r < 1$ and moments for $r \geq 1$ is divergent. Oder statistics and stress strength reliability properties are obtained and also find the quantile function, skewness and kurtosis. Simulation is conducted for different estimation methods (MLE, PSE, LSE, WLSE, CME,

and MPSE) are used to estimate used to obtain the estimate of $\theta, h(t)$ and $S(t)$. Discuss the different statistical properties of the proposed distribution. In the simulation, it found that MPSE performed best among the different methods of estimation. We used two data sets to find the suitability proposed distribution and use the MPSE for this purpose. It is observed that our proposed probability distribution is fitted well to the data sets. The new proposed probability distribution performed well as compared to competitor models like exponential, Lindley, etc in terms of various model selection criteria like AIC, AICc, BIC, KS, and p value. The above comparison is tabulated in Tables 2, 3, 4, and 5. Several inferential aspects of the proposed model are yet to study, which we may be adjust in further another publication.

References

Bowley, A. L. (1920). *Elements of Statistics, 4th Edn.* Charles Scribner, New York.

Fisher, B. and Kılıcman, A. (2012). Some results on the gamma function for negative integers. *Applied Mathematics and Information Sciences*, **6**, 173–176.

Johnson, N. L., Kotz, S., and Balakrishnan, N. (1995). *Continuous Univariate Distributions.* John Wiley and Sons.

Kao, J. H. (1959). A graphical estimation of mixed weibull parameters in life-testing of electron tubes. *Technometrics*, **1**, 389–407.

Macdonald, P. (1971). Comments and queries comment on "an estimation procedure for mixtures of distributions" by choi and bulgren. *Journal of the Royal Statistical Society: Series B (Methodological)*, **33**, 326–329.

Mann, N. R., Schafer, R. E., and Singpurwalla, N. D. (1974). *Methods for Statistical Analysis of Reliability and Life Data.* John Wiley and Sons, Inc.

Moors, J. (1988). A quantile alternative for kurtosis. *Journal of the Royal Statistical Society: Series D (The Statistician)*, **37**, 25–32.

Rolski, T., Schmidli, H., Schmidt, V., and Teugels, J. L. (2009). *Stochastic Processes for Insurance and Finance.* John Wiley and Sons.

Shanker, R., Hagos, F., and Sujatha, S. (2015). On modeling of lifetimes data using exponential and lindley distributions. *Biometrics and Biostatistics International Journal*, **2**, 1–9.

Shannon, C. E. (1948). Claude elwood shannon. *Bell System Technical Journal*, **27**, 379–423.

Shukla, K. (2019). A comparative study of one parameter lifetime distributions. *Biometrics and Biostatistics International Journal*, **8**, 111–123.

Swain, J. J., Venkatraman, S., and Wilson, J. R. (1988). Least-squares estimation of distribution functions in johnson's translation system. *Journal of Statistical Computation and Simulation*, **29**, 271–297.

Appendix 1

Table 6: True values of θ and estimates (MLE, LSE, WLSE, PCE, CME, MPSE) and corresponding MSEs.

	Methods	$n=5$	$n=15$	$n=30$	$n=50$	$n=100$	$n=500$
$\theta=0.2$	MLE	0.25281	0.21462	0.20761	0.20496	0.20222	0.20047
		(0.02638)	(0.00446)	(0.00198)	(0.00115)	(0.00053)	(0.00010)
	LSE	0.47376	0.21083	0.20579	0.20364	0.20134	0.20042
		(24.1709)	(0.0056)	(0.0025)	(0.0014)	(0.00065)	(0.00013)
	WLSE	0.49936	0.20992	0.20543	0.20346	0.20141	0.20042
		(24.7952)	(0.0052)	(0.0023)	(0.0013)	(0.00059)	(0.00012)
	PCE	0.12575	0.17549	0.18882	0.19144	0.19647	0.19927
		(0.0211)	(0.0241)	(0.0097)	(0.0032)	(0.0018)	(0.00031)
	CME	0.34324	0.21302	0.20692	0.20433	0.20169	0.20049
		(5.8857)	(0.0057)	(0.0025)	(0.0014)	(0.00065)	(0.00013)
	MPSE	0.16493	0.18012	0.18796	0.19204	0.19501	0.19866
		(0.0114)	(0.0034)	(0.0017)	(0.0010)	(0.00052)	(0.00009)
$\theta=1.05$	MLE	1.28750	1.13026	1.08756	1.07410	1.06000	1.05201
		(0.6111)	(0.1270)	(0.0539)	(0.0304)	(0.0145)	(0.0028)
	LSE	1.33088	1.11065	1.07748	1.06844	1.05637	1.05169
		(1.5951)	(0.1558)	(0.0672)	(0.0375)	(0.0179)	(0.0035)
	WLSE	1.31587	1.10524	1.07561	1.06750	1.05628	1.05153
		(1.5474)	(0.1450)	(0.0624)	(0.0344)	(0.0164)	(0.0032)
	PCE	0.67763	0.91507	0.99103	1.01397	1.03069	1.04593
		(0.8187)	(0.4546)	(0.2164)	(0.1123)	(0.1210)	(0.00884)
	CME	1.34438	1.12206	1.08345	1.07204	1.05818	1.05205
		(1.3668)	(0.1588)	(0.0681)	(0.0378)	(0.0181)	(0.00356)
	MPSE	0.86779	0.948460	0.98447	1.00631	1.02222	1.04252
		(0.3252)	(0.0954)	(0.0474)	(0.0282)	(0.0142)	(0.00278)
$\theta=2.5$	MLE	3.00848	2.69749	2.59059	2.55393	2.53204	2.50549
		(3.1287)	(0.7112)	(0.3106)	(0.1735)	(0.0841)	(0.0161)
	LSE	3.16709	2.64741	2.56482	2.54427	2.52676	2.50419
		(9.0752)	(0.8941)	(0.3808)	(0.2230)	(0.1057)	(0.0199)
	WLSE	3.13132	2.63481	2.55994	2.54115	2.525	2.50432
		(8.7689)	(0.8299)	(0.3514)	(0.2038)	(0.0965)	(0.0182)
	PCE	1.64480	2.18839	2.33701	2.41562	2.46184	2.4732
		(7.2999)	(2.0666)	(1.1109)	(0.6250)	(0.2943)	(0.0505)
	CME	3.21388	2.67504	2.57896	2.55280	2.53108	2.50505
		(8.5114)	(0.9103)	(0.3854)	(0.2248)	(0.1062)	(0.0199)
	MPSE	2.07919	2.6378	2.34567	2.39264	2.44134	2.48291
		(1.9694)	(0.5314)	(0.2727)	(0.1613)	(0.0802)	(0.0161)
$\theta=5$	MLE	5.57307	5.17582	5.19181	5.10321	5.05499	5.01262
		(10.2351)	(10.1722)	(1.2233)	(0.7032)	(0.3364)	(0.0643)
	LSE	6.18395	5.29498	5.15390	5.07877	5.03796	5.01025
		(30.3224)	(3.6853)	(1.5322)	(0.8862)	(0.4194)	(0.0808)
	WLSE	6.09790	5.26990	5.14333	5.07328	5.03828	5.01056
		(28.0626)	(3.42865)	(1.4120)	(0.8149)	(0.3841)	(0.0736)
	PCE	3.54518	4.37968	4.70454	4.80893	4.91191	4.98667
		(1324.831)	(8.5011)	(5.1263)	(2.8744)	(1.1243)	(0.2073)
	CME	6.29332	5.34993	5.18218	5.09584	5.04661	5.01199
		(30.64088)	(3.7582)	(1.5522)	(0.8934)	(0.4212)	(0.0809)
	MPSE	4.11762	4.52001	4.70706	4.78218	4.87486	4.96747
		(7.5738)	(2.2041)	(1.0637)	(0.6579)	(0.3261)	(0.0641)

Table 7: True value of $h(t)$ and estimate(MLE) & corresponding MSEs for $t = 0.75$, $t = 1$, & $t = 1.5$

θ & t	Method	$n = 5$	$n = 15$	$n = 30$	$n = 50$	$n = 100$	$n = 500$
	$\hat{\theta}$	0.25496	0.21564	0.20654	0.20479	0.20201	0.20018
$\theta = 0.2$	h(t)	0.89696	0.95072	0.96372	0.96625	0.97028	0.97295
$t = 0.75$	$\hat{h(t)}$	0.92154	0.95615	0.96617	0.96768	0.97099	0.97309
		(0.07605)	(0.01336)	(0.00475)	(0.00274)	(0.00118)	(0.00021)
	$\hat{\theta}$	1.32072	1.13019	1.08827	1.07346	1.06291	1.05226
$\theta = 1.05$	h(t)	0.22596	0.29048	0.30731	0.31433	0.31943	0.32204
$t = 0.75$	$\hat{h(t)}$	0.29245	0.29895	0.30068	0.30124	0.30149	0.30264
		(0.04675)	(0.01539)	(0.00744)	(0.00446)	(0.00221)	(0.00042)
	$\hat{\theta}$	3.15398	2.69375	2.59982	2.56007	2.52611	2.50674
$\theta = 2.5$	h(t)	0.01920	0.03520	0.03982	0.04196	0.04387	0.04500
$t = 0.75$	$\hat{h(t)}$	0.06857	0.05317	0.04951	0.04783	0.46915	0.45614
		(0.00997)	(0.00227)	(0.00107)	(0.00061)	(0.00029)	(0.00005)
	$\hat{\theta}$	6.16170	5.39945	5.20226	5.11701	5.06865	5.01182
$\theta = 5$	h(t)	0.00035	0.00098	0.00127	0.00142	0.00152	0.001642
$t = 0.75$	$\hat{h(t)}$	0.01562	0.00395	0.00281	0.00236	0.00198	0.00173
		(0.00936)	(0.00005)	(0.00001)	(0.000006)	(0.000002)	(0.0000003)
	$\hat{\theta}$	0.25237	0.21597	0.20773	0.20481	0.20250	0.20052
$\theta = 0.2$	h(t)	0.74177	0.77309	0.78042	0.78303	0.78511	0.78690
$t = 1$	$\hat{h(t)}$	0.75393	0.77579	0.78165	0.78374	0.78546	0.78697
		(0.03941)	(0.00668)	(0.00238)	(0.00122)	(0.00052)	(0.00008)
	$\hat{\theta}$	1.34594	1.13448	1.09333	1.07245	1.06215	1.05246
$\theta = 1.05$	h(t)	0.23958	0.29552	0.30791	0.31441	0.31767	0.32076
$t = 1$	$\hat{h(t)}$	0.29889	0.31263	0.31598	0.31926	0.32005	0.32123
		(0.03540)	(0.01036)	(0.00493)	(0.00293)	(0.00145)	(0.00028)
	$\hat{\theta}$	3.17655	2.69800	2.59951	2.55627	2.52797	2.50549
$\theta = 2.5$	h(t)	0.13842	0.18462	0.19426	0.19889	0.20248	0.20504
$t = 1$	$\hat{h(t)}$	0.09068	0.08216	0.07953	0.07877	0.07798	0.07737
		(0.01150)	(0.00324)	(0.00157)	(0.00092)	(0.00045)	(0.00009)
.	$\hat{\theta}$	6.08392	5.39642	5.20642	5.11668	5.05185	5.01264
$\theta = 5$	h(t)	0.002234	0.00442	0.00534	0.00584	0.00622	0.00647
$t = 1$	$\hat{h(t)}$	0.02941	0.01040	0.00853	0.00779	0.00724	0.00667
		(0.01348)	(0.00020)	(0.00007)	(0.00003)	(0.00001)	(0.000002)
	$\hat{\theta}$	0.25543	0.21631	0.20744	0.20483	0.20254	0.200619
$\theta = 0.2$	h(t)	0.54324	0.56002	0.56392	0.56507	0.56609	0.56694
$t = 1.5$	$\hat{h(t)}$	0.54795	0.56097	0.56434	0.56532	0.56621	0.56697
		(0.01839)	(0.00264)	(0.00078)	(0.00039)	(0.00015)	(0.00002)
	$\hat{\theta}$	1.34439	1.13237	1.09132	1.07585	1.06092	1.05250
$\theta = 1.05$	h(t)	0.25025	0.28903	0.29727	0.30045	0.30355	0.30531
$t = 1.5$	$\hat{h(t)}$	0.28143	0.29741	0.30125	0.30273	0.30467	0.30554
		(0.01869)	(0.00488)	(0.00226)	(0.00129)	(0.00062)	(0.00012)
	$\hat{\theta}$	3.15152	2.69424	2.59652	2.55265	2.52997	2.50530
$\theta = 2.5$	h(t)	0.07607	0.10256	0.10932	0.11251	0.11419	0.11605
$t = 1.5$	$\hat{h(t)}$	0.11653	0.11604	0.11597	0.11643	0.11614	0.11644
		(0.00998)	(0.00313)	(0.00150)	(0.00089)	(0.00044)	(0.00008)
	$\hat{\theta}$	6.04436	5.41528	5.18005	5.1028	5.0607	5.01264
$\theta = 5$	h(t)	0.01142	0.01728	0.02017	0.02122	0.02181	0.02251
$t = 1.5$	$\hat{h(t)}$	0.040103	0.026824	0.025033	0.024219	0.02329	0.02281
		(0.00668)	(0.00062)	(0.00027)	(0.00015)	(0.00007)	(0.00001)

Table 8: True value of $S(t)$ and estimate(MLE) & corresponding MSEs for $t = 0.75$, $t = 1$, & $t = 1.5$

θ & t	Method	$n = 5$	$n = 15$	$n = 30$	$n = 50$	$n = 100$	$n = 500$
	$\hat{\theta}$	0.25496	0.21564	0.20654	0.20479	0.20201	0.20018
$\theta = 0.2$	S(t)	0.46877	0.41739	0.40466	0.40218	0.39821	0.39557
$t = 0.75$	$\hat{S(t)}$	0.43577	0.40951	0.40106	0.40007	0.39716	0.39537
		(0.02359)	(0.00719)	(0.00349)	(0.00207)	(0.001044)	(0.00021)
	$\hat{\theta}$	1.32072	1.13019	1.08827	1.07346	1.06291	1.05226
$\theta = 1.05$	S(t)	0.93211	0.90319	0.89505	0.89159	0.88904	0.88773
$t = 0.75$	$\hat{S(t)}$	0.88859	0.89339	0.89531	0.89619	0.89697	0.89715
		(0.01056)	(0.00358)	(0.00174)	(0.00104)	(0.00052)	(0.00010)
	$\hat{\theta}$	3.15398	2.69375	2.59982	2.56007	2.52611	2.50674
$\theta = 2.5$	S(t)	0.99713	0.99399	0.99300	0.99253	0.99211	0.99185
$t = 0.75$	$\hat{S(t)}$	0.98202	0.98870	0.99015	0.99080	0.99121	0.99167
		(0.00133)	(0.00016)	(0.00006)	(0.00003)	(0.00001)	(0.000003)
	$\hat{\theta}$	6.16170	5.39945	5.20226	5.11701	5.06865	5.01182
$\theta = 5$	S(t)	0.99997	0.99990	0.99987	0.99986	0.99985	0.99983
$t = 0.75$	$\hat{S(t)}$	0.99339	0.99946	0.99966	0.99973	0.99978	0.99982
		(0.00502)	(0.000001)	(0.0000003)	(0.0000001)	(0.00000003)	(0.000000004)
	$\hat{\theta}$	0.25237	0.21597	0.20773	0.20481	0.20250	0.20052
$\theta = 0.2$	S(t)	0.37961	0.33735	0.32731	0.32370	0.32084	0.31837
$t = 1$	$\hat{S(t)}$	0.35642	0.33190	0.32479	0.32225	0.32011	0.31823
		(0.01929)	(0.00562)	(0.00274)	(0.00161)	(0.00081)	(0.00015)
	$\hat{\theta}$	1.34594	1.13448	1.09333	1.07245	1.06215	1.05246
$\theta = 1.05$	S(t)	0.88904	0.84933	0.83992	0.83489	0.83235	0.82992
$t = 1$	$\hat{S(t)}$	0.83001	0.82987	0.83040	0.82908	0.82946	0.82935
		(0.01692)	(0.00569)	(0.00283)	(0.00175)	(0.00088)	(0.00017)
	$\hat{\theta}$	3.17655	2.69800	2.59951	2.55627	2.52797	2.50549
$\theta = 2.5$	S(t)	0.94869	0.92345	0.91775	0.91496	0.91278	0.91120
$t = 1$	$\hat{S(t)}$	0.96381	0.97194	0.97422	0.97506	0.97579	0.97638
		(0.00329)	(0.00063)	(0.00028)	(0.00016)	(0.00007)	(0.00001)
	$\hat{\theta}$	6.08392	5.39642	5.20642	5.11668	5.05185	5.01264
$\theta = 5$	S(t)	0.99967	0.99929	0.99911	0.99901	0.99894	0.99889
$t = 1$	$\hat{S(t)}$	0.98283	0.99773	0.99831	0.99853	0.99869	0.99884
		(0.01260)	(0.000014)	(0.000004)	(0.000002)	(0.0000008)	(0.0000001)
	$\hat{\theta}$	0.25543	0.21631	0.20744	0.20483	0.20254	0.200619
$\theta = 0.2$	S(t)	0.27930	0.24339	0.23495	0.23245	0.23024	0.22838
$t = 1.5$	$\hat{S(t)}$	0.26506	0.24039	0.23361	0.23166	0.22985	0.22831
		(0.01376)	(0.00352)	(0.00166)	(0.00097)	(0.00048)	(0.00009)
	$\hat{\theta}$	1.34439	1.13237	1.09132	1.07585	1.06092	1.05250
$\theta = 1.05$	S(t)	0.78479	0.73215	0.72035	0.71576	0.71124	0.70866
$t = 1.5$	$\hat{S(t)}$	0.72525	0.71401	0.71150	0.71062	0.70868	0.70815
		(0.02417)	(0.00839)	(0.00430)	(0.00254)	(0.00129)	(0.00026)
	$\hat{\theta}$	3.15152	2.69424	2.59652	2.55265	2.52997	2.50530
$\theta = 2.5$	S(t)	0.96055	0.94065	0.93515	0.93250	0.93108	0.92951
$t = 1.5$	$\hat{S(t)}$	0.91684	0.92465	0.92706	0.92764	0.92865	0.92902
		(0.00834)	(0.00238)	(0.00111)	(0.00066)	(0.00032)	(0.00006)
	$\hat{\theta}$	6.04436	5.41528	5.18005	5.1028	5.0607	5.01264
$\theta = 5$	S(t)	0.99646	0.99413	0.99289	0.99243	0.99216	0.99185
$t = 1.5$	$\hat{S(t)}$	0.97250	0.98850	0.99002	0.99065	0.99129	0.99167
		(0.01145)	(0.00018)	(0.00006)	(0.00003)	(0.00001)	(0.000002)

Appendix 2

1: Distribution Function

$$F(x \mid \theta) = \begin{cases} \left(1 - \frac{\theta}{x+\theta}\right)(e^{-\theta/x}) \; ; x \geq 0, \theta > 0 \\ 0 \qquad\qquad ; \text{otherwise} \end{cases} \tag{33}$$

Proof

- $F(-\infty) = 0$
- $F(+\infty) = 1$
- If $x \leq y$ then $F(x|\theta) \leq F(y|\theta)$
- $F(x|\theta)$ is right continuous.

Hence $F(x|\theta)$ is a distribution function.

2: r^{th} Moment

The r^{th} moments about origin exits only for $r < 1$

$$E[X^r] = \int_0^\infty x^r \frac{\theta(2x+\theta)}{x(x+\theta)^2} e^{-\theta/x} \, dx \tag{34}$$

when we put the $\theta/x = t$ and simplify it then we get

$$E[X^r] = \theta^r \left[\int_0^\infty \frac{1}{t^r(1+t)} e^{-t} \, dt + \int_0^\infty \frac{1}{t^r(1+t)^2} e^{-t} \, dt \right]$$

First term of integral we say I_1 and second term of integral we say that I_2
Now, put r=1 we get I_1 is

$$I_1 = \int_0^\infty \left[\frac{1}{t} - \frac{1}{1+t} \right] e^{-t} \, dt$$

First term of above integral we say I_3 and second term of integral we say that I_4
Now, The I_3 is negative integer for gamma function which is divergent (see Fisher and Kılıcman (2012)) and I_4 is incomplete gamma function which is finite, so equation (34) is divergent

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 22, No. 2, 2024 (New Series), pp 109–119
http://www.ssca.org.in/journal

LDPC Codes Based on New Combinatorial Designs

Shyam Saurabh[1] **and Kishore Sinha**[2]
[1] *Tata College, Kolhan University, Chaibasa, India*
[2] *Formerly at Birsa Agricultural University, Ranchi, India and*
Presently at 201D, Oceanus Tranquil, Dr. B. R. Ambedkar Road, Bangalore–560036

Received: 14 July 2023; Revised: 28 November 2023; Accepted: 02 December 2023

Abstract

Earlier binary (k, r)-regular LDPC codes have been constructed using balanced incomplete block designs, mutually orthogonal Latin rectangles, partial geometries, group divisible designs, resolvable group divisible designs and finite geometries. Here we have constructed LDPC codes from certain triangular and L_2-type designs which are free of 4–cycles.

Key words: LDPC Codes; Association schemes; Partially balanced incomplete block designs; Triangular designs; L_2-type designs

AMS Subject Classifications: 62K10, 05B05

1. Introduction

1.1. LDPC codes

A binary (k, r)-regular low–density parity–check *(LDPC) code* is the null space of a $s \times t$ sparse parity-check matrix H (*i.e.* the majority of entries must be zero) over a Galois field $GF(2)$ of order 2 such that each row has r nonzero elements and each column has k nonzero elements where $r \ll t$ and $k \ll s$. The minimum distance of a code is equal to the minimum number of nonzero columns in the parity-check matrix such that a nontrivial linear combination of these columns sums to zero over $GF(2)$ [see Wicker (1995), p. 84 and Johnson and Weller (2003), p. 1416].

Parity-check matrices (or LDPC codes) may be represented as *Tanner bipartite graphs* with vertex set $V \cup W$ where V is comprised of code bits and W is comprised of parity-check equations. There exists an edge $\{v, w\}, v \in V$ and $w \in W$, in this bipartite graph if and only if v is a term in the check equation w. A *cycle* in a graph is a sequence of connected vertices which start and end at the same vertex in the graph and no other vertices occur more than once. The length of the cycle is the number of edges it contains and the *girth* of a graph is the length of its smallest cycle. Since the Tanner graph is bipartite, the length of a cycle must be even and at least 4.

Corresponding Author: Shyam Saurabh
Email: shyamsaurabh785@gmail.com

An LDPC code performs well with iterative decoding provided the corresponding Tanner graph have a reasonably large girth, *i.e.*, the graph should be free of short cycles. The cycles that affect the performance the most are the cycles of length four. These short cycles severely limit the performance of iterative decoding. For codes with these short cycles, iterative decoding becomes correlated after two iterations. Therefore, cycles of length four must be avoided in code construction [see Bonello *et al.* (2011) and Xu *et al.* (2005)].

An LDPC code is free of 4-cycles if no two distinct columns (or two distinct rows) of H have more than one nonzero component in common or the inner product of any two distinct rows or any two distinct columns of the parity-check matrix H is less than or equal to 1. This constraint on H is known as the row-column constraint (or RC constraint), see Diao *et al.* (2013). The RC-constraint confirms that the girth of the LDPC codes generated by such H is at least six.

1.2. Balanced incomplete block design

A *balanced incomplete block design* (BIBD) or a $2-(v,k,\lambda)$ design is an arrangement of v elements into $b = (\lambda(v^2 - v))/((k^2 - k))$ blocks, each of size $k(< v)$ such that every element is replicated r times and any two distinct elements occur together in λ blocks.

A BIBD is *resolvable* if the b blocks each of size k can be partitioned into r resolution classes such that

(i) Each resolution class contains b/r blocks;
(ii) Every element is replicated exactly once in each resolution class.

A BIBD with $k = 3$ and $\lambda = 1$ is usually known as Steiner's triple system (STS) or Steiner 2–design and a resolvable Steiner's triple system is known as Kirkman triple system (KTS), see Raghavarao (1971), Johnson and Weller (2001) and Ray–Chaudhuri and Wilson (1971).

Example 1: Consider a resolvable BIBD with parameters: $v = 9, b = 12, r = 4, k = 3, \lambda = 1$ whose resolution classes are:
RI: [(1 2 3) (4 5 6) (7 8 9)]; RII: [(1 4 7) (2 5 8) (3 6 9)]; RIII: [(1 5 9) (2 6 7) (3 4 8)]; RIV: [(1 6 8) (2 4 9) (3 5 7)].

1.3. Association scheme

A relationship defined on a set of v elements is called an *association scheme* with two associate classes if it satisfies the following conditions:

(a) Any two distinct elements are either 1st or 2nd associates of each other and any element is the 0-th associate of itself,
(b) Each element has n_j; j-th associates ($j = 0, 1, 2$) and
(c) For every pair of elements which are j-th associates of each other, there are $p^j_{u,w}$ elements that are u-th associates of one and w-th associates of the other ($j, u, w = 0, 1, 2$).

1.4. Partially balanced incomplete block (PBIB) design

Given an association scheme with two associate classes on a set of v elements, a *PBIB design* based on this association scheme is a block design with v elements and b blocks satisfying the following conditions:

(i) Each element appears at most once in a block,
(ii) Each block has a fixed number of elements, say k,
(iii) Each element appears in a fixed number of blocks, say r, and
(iv) Every pair of elements which are j-th ($j = 1, 2$) associates of each other appear together in λ_j blocks ($\lambda_1 \neq \lambda_2$).

Some special classes of PBIB designs known as group divisible, triangular and L_2-type Latin square designs are described below:

1.5. Group divisible design

Let $v = mn (m, n \geq 2)$ elements be arranged in an $m \times n$ array, say M. A *group divisible (GD) association scheme* on these $v = mn$ elements is defined as follows: two elements are first associates if they occur in the same row of M and second associates, otherwise.

A PBIB design based on GD association scheme is said to be GD design. The integers: $v = mn, b, r, k, \lambda_1$ and λ_2 are known as parameters of the GD design and they satisfy the relations: $bk = vr; (n-1)\lambda_1 + n(m-1)\lambda_2 = r(k-1)$. Furthermore, if $r - \lambda_1 = 0$ then the GD design is singular (S); if $r - \lambda_1 > 0$ and $rk - v\lambda_2 = 0$ then it is semi–regular (SR) and if $r - \lambda_1 > 0$ and $rk - v\lambda_2 > 0$ then the design is regular (R).

Example 2: Consider the following resolvable solution of an SRGD design SR9 with parameters: $v = 8, b = 16, r = 4, k = 2, \lambda_1 = 0, \lambda_2 = 1, m = 2, n = 4$ as given in Clatworthy (1973):
RI: [(1 5) (2 6) (3 7) (4 8)]; RII: [(2 7) (1 8) (4 5) (3 6)]; RIII: [(4 6) (3 5) (2 8) (1 7)]; RIV: [(3 8) (4 7) (1 6) (2 5)].
The arrangement of $v = 8$ elements in 2×4 array is given as: $\begin{matrix} 1 & 2 & 3 & 4 \\ 5 & 6 & 7 & 8 \end{matrix}$.

1.6. Triangular design

A *triangular association scheme* is an arrangement of $v = (s(s-1))/2$ elements in an $s \times s$ array such that the positions on the principal diagonal are left blank, the $(s(s-1))/2$ positions above and below the principal diagonal are filled with the v elements in such a way that the resultant arrangement is symmetric about the principal diagonal. Then any two elements which occur in the same row or same column are first associates; otherwise they are second associates. A PBIB design based on triangular association scheme is called a triangular design. The integers $v = (s(s-1))/2, b, r, k, \lambda_1$ and λ_2 are known as parameters of the triangular design and they satisfy the relations: $bk = vr$; $2(s-2)\lambda_1 + ((s-2)(s-3))/2\lambda_2 = r(k-1)$.

Example 3: Consider a triangular design T9 given in Clatworthy (1973) with parameters: $v = b = 10, r = k = 3, \lambda_1 = 1, \lambda_2 = 0$ whose blocks are given as: (1 2 5); (8 9 10); (2 3 8); (5 7 9); (2 4 9); (5 6 8); (3 4 10); (6 7 10); (1 4 7); (1 3 6)

The arrangement of 10 elements in 5×5 array is given as:

$$\begin{matrix} * & 1 & 2 & 3 & 4 \\ 1 & * & 5 & 6 & 7 \\ 2 & 5 & * & 8 & 9 \\ 3 & 6 & 8 & * & 10 \\ 4 & 7 & 9 & 10 & * \end{matrix}.$$

1.7. L_2-type design

An L_2-*association scheme* is an arrangement of $v = s^2$ elements into an $s \times s$ array such that any two elements in the same row or in the same column of the array are 1st associates; otherwise they are 2nd associates. A PBIB design based on L_2-association scheme is known as an L_2- type design. The integers $v = s^2, b, r, k, \lambda_1$ and λ_2 are known as parameters of the L_2-type design and they satisfy the relations: $bk = vr$; $2(s-1)\lambda_1 + (s-1)^2\lambda_2 = r(k-1)$.

Example 4: Consider an L_2-type design as given in Clatworthy (1973) with parameters: LS26: $v = b = 9, r = k = 4, n_1 = n_2 = 4, \lambda_1 = 1, \lambda_2 = 2$ whose blocks are given as:
(1 2 6 9); (2 4 6 8); (1 4 8 9); (2 5 7 9); (2 3 4 7); (3 4 5 9); (1 5 6 7); (3 6 7 8); (1 3 5 8)

The arrangement of $v = 9$ elements in 3×3 array is given as:

$$\begin{matrix} 1 & 4 & 7 \\ 2 & 5 & 8 \\ 3 & 6 & 9 \end{matrix}.$$

SRX, TX and LSX numbers are from Clatworthy (1973). For details on BIB, GD, triangular and L_2-type designs, we refer to Dey (2010), Raghavarao (1971), Raghavarao and Padgett (2005).

2. Earlier constructions

Low-density parity-check (LDPC) codes were introduced by Gallager (1962). LDPC codes can be divided into two types: random codes and structured codes. Random LDPC codes are constructed by computer search while structured LDPC codes are constructed by algebraic and combinatorial methods. Earlier constructions of regular LDPC codes from combinatorial designs may be summarized below in Table 1:

A recent survey on algebraic constructions of LDPC codes may also be found in Saurabh and Sinha (2023). The purpose of this paper is to construct binary regular LDPC codes based on triangular and L_2- type designs. The incidence matrix of such block design is used as parity-check matrix of the code which satisfies row-column constraint which ensures that the girth of the proposed code is at least six and the corresponding LDPC code (or Tanner graph) is free of 4-cycles. We are describing below the method to obtain LDPC codes from BIB and GD designs:

2.1. LDPC codes from BIB and GD designs

The following Lemmas [see Saurabh and Sinha (2023)] describe the constructions of LDPC codes from BIB and two associate class PBIB designs:

Lemma 1: The existence of two associate classes PBIB design with parameters: v, b, r, k, λ_1, $\lambda_2 \in 0, 1$ implies the existence of a (k, r)-regular LDPC codes free of four cycles with code length b and code rate about $1 - k/r(k < r)$.

Table 1: Combinatorial structures and corresponding LDPC codes

No.	Combinatorial Structure	Reference
1	BIB designs	Ammar *et al.* (2004), Lan *et al.* (2008)
2	Resolvable BIB designs	Johnson and Weller (2001, 2003)
3	Group divisible designs	Shan and Li (2013)
4	Resolvable group divisible designs	Xu *et al.* (2015)
5	$\alpha(>1)$-resolvable group divisible designs	Saurabh and Sinha (2023)
6	Semipartial geometries	Li *et al.* (2008)
7	Partial geometries	Johnson and Weller (2004), Diao *et al.* (2016), Xu *et al.* (2019)
8	Finite geometries	Kou *et al.* (2001)
9	Mutually orthogonal Latin rectangles	Vasic *et al.* (2002)
10	Euclidean geometries and partial BIBDs	Mahadevan and Morris (2002)
11	Oval designs	Weller and Johnson (2003)
12	Cyclic 2-(v,3,1) designs	Vasic and Milenkovic (2004)
13	Mutually orthogonal Latin squares	Zhang *et al.* (2010)
14	Difference covering arrays	Donovan *et al.* (2022)
15	Cubic semi–symmetric graphs	Crnkovic *et al.* (2022)

As a special case of Lemma 1, we can obtain the following result:

Lemma 2: The existence of a BIB design with parameters: $v, b, r, k, \lambda = 1$ implies the existence of a (k, r)-regular LDPC codes free of four cycles with code length b and code rate about $1 - k/r(k < r)$.

2.2. LDPC codes from resolvable BIB and GD designs

2.2.1. LDPC codes from resolvable BIB designs

Johnson and Weller (2003) used following series of Kirkman triple systems (KTSs) in the construction of LDPC codes:
Series 1: $v = 3(4t + 1), b = (4t + 1)(6t + 1), r = 6t + 1, k = 3, \lambda = 1$.
Series 2: $v = 3(6t + 1), b = (6t + 1)(9t + 1), r = 9t + 1, k = 3, \lambda = 1$;
where $s = 6t + 1$ is a prime or prime power.

Since the series I and II of KTSs are resolvable BIB designs, their incidence matrices N may be partitioned in to 'r' submatrices as $N = (N_1|N_2|N_3|N_4|\ldots|N_r)$ where each N_i is $v \times (v/k)$ matrix such that each row sum of N_i $(1 \leq i \leq r)$ is one. Further juxtaposing set of any $p(4 \leq p \leq r)$ submatrices of N we obtain series of LDPC codes with length $vp/3$ and code rate about $1 - 3/p$ [see Saurabh and Sinha (2023)]. This method may also be used to obtain LDPC codes from a resolvable BIB design with $\lambda = 1$ other than above Series (1 and 2) of resolvable BIB designs.

2.2.2. LDPC codes from resolvable GD designs

Xu *et al.* (2015) considered submatrices of the incidence matrix of a resolvable GD design with parameters: $v = mn, b, r, k, \lambda_1 = 0, \lambda_2 = 1$ as the parity-check matrix to con-

struct series of regular LDPC codes as follows:
Consider a resolvable GD design with parameters: $v = mn, b, r, k, \lambda_1 = 0, \lambda_2 = 1$. Then its incidence matrix N may be partitioned in to 'r' submatrices as $N = (N_1|N_2|N_3|N_4|\ldots|N_r)$ where each N_i is $v \times (v/k)$ matrix such that each row sum of $N_i (1 \leq i \leq t)$ is one. Further juxtaposing set of any $p (1 \leq p \leq r)$ submatrices of N we obtain series of LDPC codes with length vp/k and code rate about $1 - k/p$.

Xu *et al.* (2015) used the results of Assaf and Hartman (1989), Greig (1999) and Sun and Ge (2009) on the existence of resolvable GD designs for the construction of LDPC codes. Some of the series obtained as special cases of their results are given below:

Series 3 [Assaf and Hartman (1989)]: There exists a resolvable GD design with parameters: $v = 12s, b = 16s^2, r = 4s, k = 3, \lambda_1 = 0, \lambda_2 = 1, m = 3, n = 4s$.

Series 4 [Greig (1999)]: There exists a resolvable GD design with parameters: $v = 32s, b = 64s^2, r = 8s, k = 4, \lambda_1 = 0, \lambda_2 = 1, m = 4, n = 8s$.

Series 5 [Sun and Ge (2009)]: There exist resolvable GD designs with parameters: $v = s(s^2-1), b = (s^2-1)^2, r = s^2-1, k = s, \lambda_1 = 0, \lambda_2 = 1, m = s, n = s^2-1$ where s is a prime or prime power.

The following series of resolvable designs obtained as a special case of Theorem 10 of Saurabh and Sinha (2023) may also be used in LDPC codes:

Series 6: There exists a resolvable SRGD design with parameters: $v = q(q-t), b = q^2, r = q, k = q-t, \lambda_1 = 0, \lambda_2 = 1, m = q-t, n = q (1 \leq t \leq q-1)$ where q is a prime or prime power.

3. LDPC codes from triangular and L_2-type designs

3.1. Some series of BIB designs

The following series of BIB designs may be found in Raghavarao (1971, pp. 77–78) for s being a prime or prime power:

Series 7: $v' = b' = s^2 + s + 1, r' = k' = s + 1, \lambda' = 1$.
Series 8: $v' = (s+1)(s^2+1), b' = (s^2+1)(s^2+s+1), r' = s^2+s+1, k' = s+1, \lambda' = 1$.
Series 9: $v' = s^2, b' = s(s+1), r' = s+1, k' = s, \lambda' = 1$.
Series 10: $v' = s^3, b' = s^2(s^2+s+1), r' = s^2+s+1, k' = s, \lambda' = 1$.

3.2. Some series of triangular designs

The series (10–12) of triangular designs given below may be found in Raghavarao (1971) and Dey (2010):

Series 11: For $v = (s(s-1))/2; s \geq 5$ and block size $k = 2$, there exist triangular designs with parameters:
(i) $b = (s(s-1)(s-2))/8, r = 2(s-2), \lambda_1 = 1, \lambda_2 = 0$.
(ii) $b = (s(s-1)(s-2)(s-3))/8, r = ((s-2)(s-3))/8, \lambda_1 = 0, \lambda_2 = 1$.

Series 12: The existence of a triangular design with parameters $v = (2s-1)s, b = (2s -$

$1)(2s-3), r = 2s-3, k = s, \lambda_1 = 0, \lambda_2 = 1$ implies the existence of another triangular design with parameters $v = (2s-1)(s-1), b = (2s-1)(2s-3), r = 2s-3, k = s-1, \lambda_1 = 0, \lambda_2 = 1$.

Series 13: The existence of a BIB design: $v' = s-1, b', r', k', \lambda = 1$ implies the existence of a triangular design with parameters: $v = (s(s-1))/2, b = sb', r = 2r', k = k', \lambda_1 = 1, \lambda_2 = 0$.

Then utilizing the series (7–10) of BIB designs respectively in Series 13, we obtain the following series (14–17) of triangular designs respectively:

Series 14: $v = ((s^2+s+1)(s^2+s+2))/2, b = 2v, r = 2(s+1), k = s+1, \lambda_1 = 1, \lambda_2 = 0$.

Series 15: $v = \{(s+1)(s^2+1)+1\}\{(s+1)(s^2+1)\}/2, b = \{(s+1)(s^2+1)+1\}\{(s^2+1)(s^2+s+1)\}, r = 2(s^2+s+1), k = s+1, \lambda_1 = 1, \lambda_2 = 0$.

Series 16: $v = (s^2(s^2+1))/2, b = s(s^2+1)(s+1), r = 2(s+1), k = s, \lambda_1 = 1, \lambda_2 = 0$.

Series 17: $v = (s^3(s^3+1))/2, b = s^2(s^3+1)(s^2+s+1), r = 2(s^2+s+1), k = s, \lambda_1 = 1, \lambda_2 = 0$.

3.3. Some series of L_2-type designs

Consider the following series of L_2-type design as given in Raghavarao (1971) and Dey (2010):

Series 18: The existence of a BIB design with parameters: $v' = s$, b', r', k', $\lambda = 1$ implies the existence of an L_2-type design with parameters: $v = s^2, b = 2sb'$, $r = 2r'$, $k = k'$, $\lambda_1 = 1$, $\lambda_2 = 0$.

Further applying the series (7–10) of BIB designs respectively in series 18, we obtain the following series (19–22) of L_2-type designs respectively:

Series 19: $v = (s^2+s+1)^2, b = 2v, r = 2(s+1), k = s+1, \lambda_1 = 1, \lambda_2 = 0$.

Series 20: $v = (s+1)^2(s^2+1)^2, b = 2(s+1)(s^2+1)^2(s^2+s+1), r = 2(s^2+s+1), k = s+1, \lambda_1 = 1, \lambda_2 = 0$.

Series 21: $v = s^4, b = 2s^3(s+1), r = 2(s+1), k = s, \lambda_1 = 1, \lambda_2 = 0$.

Series 22: $v = s^6, b = 2s^5(s^2+s+1), r = 2(s^2+s+1), k = s, \lambda_1 = 1, \lambda_2 = 0$.

Example 5: Using BIB design: $v = 4, b = 6, r = 3, k = 2, \lambda = 1$ in Series 13, we obtain a triangular design $T1 : v = 10, b = 30, r = 6, k = 2, \lambda_1 = 1, \lambda_2 = 0$ whose incidence matrix N is:

$$\begin{bmatrix}
1&0&0&0&0&1&0&0&0&0&1&0&0&0&0&1&0&0&0&0&1&0&0&0&0&1&0&0&0&0\\
1&0&0&0&0&0&1&0&0&0&0&1&0&0&0&0&1&0&0&0&0&1&0&0&0&0&1&0&0&0\\
0&1&0&0&0&1&0&0&0&0&0&1&0&0&0&0&0&1&0&0&0&0&1&0&0&0&0&1&0&0\\
0&1&0&0&0&0&0&1&0&0&1&0&0&0&0&0&1&0&0&0&0&0&0&1&0&0&0&0&1&0\\
0&0&1&0&0&0&0&0&1&0&0&0&1&0&0&1&0&0&0&0&0&1&0&0&0&0&0&0&0&1\\
0&0&0&1&0&0&0&0&0&1&0&0&0&1&0&0&0&1&0&0&1&0&0&0&0&0&0&0&0&1\\
0&0&0&1&0&0&0&1&0&0&0&0&1&0&0&0&0&0&0&1&0&0&0&0&1&1&0&0&0&0\\
0&0&0&0&1&0&0&0&1&0&0&0&0&1&0&0&0&0&1&0&0&0&1&0&0&0&1&0&0&0\\
0&0&1&0&0&0&1&0&0&0&0&0&0&0&1&0&0&0&1&0&0&0&0&0&1&0&0&0&1&0\\
0&0&0&0&1&0&0&0&0&1&0&0&0&0&1&0&0&0&0&1&0&0&0&1&0&0&0&1&0&0
\end{bmatrix}$$

Since any two distinct columns (or rows) of N intersect in at most one element, N satisfies RC-constraint. Hence we obtain a (2,6)-regular LDPC code free of 4-cycles with code length 30 and code rate about 0.66.

Further using BIB design: $v = 4, b = 6, r = 3, k = 2, \lambda = 1$ in Series 18, we obtain an L_2-type design $LS3 : v = 16, b = 48, r = 6, k = 2, \lambda_1 = 1, \lambda_2 = 0$. This design may be used to obtain (2,6)-regular LDPC code free of 4-cycles with code length 48 and code rate about 0.66.

A correspondence between resolvable BIB/GD designs, triangular/ L_2-type designs and LDPC codes is given in Table 2 where p denotes the number of resolution classes:

Table 2: Resolvable BIB/GD designs, Triangular/ L_2-type designs and LDPC codes

No.	Resolvable BIB designs	Code length	Code rate
1	Series 1	$p(4m+1)$	$1-(3/p); 4 \le p \le 6m+1$
2	Series 2	$p(6m+1)$	$1-(3/p); 4 \le p \le 9m+1$
3	Series 3	$4sp$	$1-(3/p)(4 \le p \le 4s)$
4	Series 4	$8sp$	$1-(4/p)(5 \le p \le 8s)$
5	Series 5	$p(s^2-1)$	$1-(s/p); (s+1 \le p \le s^2-1)$
6	Series 6	pq^2	$1-((q-t))/p$; q is a prime or prime power
7	Series 11 (i)	$\frac{s(s-1)(s-2)}{8}$	$\frac{(s-3)}{(s-2)}$
8	Series 11 (ii)	$\frac{s(s-1)(s-2)(s-3)}{8}$	$\frac{(s^2-5s-10)}{(s-2)(s-3)}$
9	Series 14	$(s^2+s+1)(s^2+s+2)$	0.5
10	Series 15	$\{(s+1)(s^2+1)+1\}\times (s^2+1)(s^2+s+1)$	$\frac{(2s^2+s+1)}{2(s^2+s+1)}$
11	Series 16	$s(s^2+1)(s+1)$	$\frac{(s+2)}{2(s+1)}$
12	Series 17	$s^2(s^3+1)(s^2+s+1)$	$\frac{(2s^2+s+2)}{2(s^2+s+1)}$
13	Series 19	$2(s^2+s+1)^2$	0.5
14	Series 20	$2(s+1)(s^2+1)^2 \times (s^2+s+1)$	$\frac{(2s^2+s+1)}{2(s^2+s+1)}$
15	Series 21	$2s^3(s+1)$	$\frac{(s+2)}{2(s+1)}$
16	Series 22	$2s^5(s^2+s+1)$	$\frac{(2s^2+s+2)}{2(s^2+s+1)}$

4. Discussion and conclusion

It is observed that LDPC codes obtained from series 1 of resolvable BIB designs have shorter code length than those obtained from series 2 for the same code rate. Hence the codes obtained from series1 are better than those obtained from series 2. Further by putting $p = 4s$ in series 3 we obtain LDPC codes with code length $16s^2$ and code rate $R_1 = 1-(3/4s)$ whereas by putting $p = 2s$ in series 4 we obtain LDPC codes with same code length $16s^2$ but different code rate $R_2 = 1-(2/s) = 1-(8/4s) < R_1$. Thus the codes obtained from series 3 are better than those obtained from series 4.

LDPC codes obtained from series 11(i) have shorter code length in comparison to the codes obtained from series 11 (ii) with better code rates if $s < 19$ as their differences is

$\frac{(s-3)}{(s-2)} - \frac{(s^2-5s-10)}{(s-2)(s-3)} = \frac{-(s-19)}{(s-2)(s-3)}$.

Further LPDC codes obtained from series (14-17) of triangular designs have shorter code length in comparison to the codes obtained from series (19-22) of L_2-type designs w.r.t. fixed code rate. Hence the codes obtained from series (14-17) of triangular designs are better than those obtained from series (19-22) of L_2-type designs respectively. For example consider LDPC codes obtained from series 14 and 19. The two codes have same code rate 0.5 but the differences of their code lengths is $2(s^2+s+1)^2-(s^2+s+1)(s^2+s+2) = s(s+1)(s^2+s+1) > 0$.

Similarly differences of code lengths between series 15 and 20 is $s(s^2+s+1) > 0$, series 16 and 21 is $s(s+1)(s^2-1) > 0$ and series 17 and 22 is $(s^2+s+1)(s^3-1) > 0$. Also the differences of code lengths obtained from series 15 and 17 is $\{(s+1)(s^2+1)+1\}(s^2+1)(s^2+s+1) - s^2(s^3+1)(s^2+s+1) > 0$ with almost same code. Hence series 17 yields better LDPC codes in comparison to series 15.

Apart from the above discussed designs, PBIB designs based on partial geometry may also be used in LDPC codes. For example the PBIB designs: PG2, PG5, PG6a from Clatworthy (1973) yield LDPC codes free of 4–cycles and positive code rates. Some LDPC codes with shorter code lengths and higher code rates from Table 2 are given below in Table 3:

Table 3: LPDC codes with shorter code length and higher code rate from Table 1

No.	Designs	Code length	Code rate
1	Series 1	$p(4m+1)$	$1-(3/p); 4 \le p \le 6m+1$
2	Series 3	$16s^2$	$1-(3/4s)$
3	Series 5	$p(s^2-1)$	$1-(s/p); (s+1 \le p \le s^2-1)$
4	Series 6	pq^2	$1-((q-t))/p$; q is a prime or prime power
5	Series 11 (i)	$\frac{s(s-1)(s-2)}{8}$	$\frac{(s-3)}{(s-2)}; s < 19$
6	Series 14	$(s^2+s+1)(s^2+s+2)$	0.5
7	Series 16	$s(s^2+1)(s+1)$	$\frac{(s+2)}{2(s+1)}$
8	Series 17	$s^2(s^3+1)(s^2+s+1)$	$\frac{(2s^2+s+2)}{2(s^2+s+1)}$

Acknowledgement

The authors are thankful to anonymous referees and Editor-in-Chief for their nice comments.

References

Ammar, B., Honary, B., and Kou, Y. (2004). Construction of low-density parity-check codes based on balanced incomplete block designs. *IEEE Transactions on Information Theory*, **50(6)**, 1257–1268.

Assaf, A. M. and Hartman, A. (1989). Resolvable group divisible designs with block size 3. *Discrete Mathematics*, **77**, 5–20.

Bonello, N., Chen, S., and Hanzo, L. (2011). Low-density parity-check codes and their rateless relatives. *IEEE Communications Surveys & Tutorials*, **13**, 1–24.

Clatworthy, W. H. (1973). *Tables of Two-associate-class Partially Balanced Designs*. National Bureau of Standards (U.S.), *Applied Mathematics*, Series 63.

Crnkovic, D., Rukavina, S., and Simac, M. (2022). LDPC codes from cubic semisymmetric graphs. *ARS Mathematica Contemporanea*, doi.org/10.26493/1855–3974.2501.4c4.

Dey, A. (2010). *Incomplete Block Designs*. Hindustan Book Agency, New Delhi.

Diao, Q., Tai, Y., Lin, S., and Abdel-Ghaffar, K. (2013). LDPC codes on partial geometries: construction, trapping set structure, and puncturing. *IEEE Transactions on Information Theory*, **59** (12), 7898-7914.

Donovan, D., Price, A., Rao, A., Uskuplu, E., and Syazici, E. (2022). High rate LDPC codes from partially balanced incomplete block designs. *Journal of Algebraic Combinatorics*, **55**, 259-275.

Gallager, R. G. (1962). Low-density parity-check codes. *IEEE Transactions on Information Theory*, **8**, 21–28.

Greig, M. (1999). Designs from projective planes and PBD bases. *Journal of Combinatorial designs*, 7, 341–374.

Johnson, S. J. and Weller, S. R. (2001).Construction of low-density parity-check codes from Kirkman triple systems, *In Proc. IEEE Globecom Conf.*, San Antonio, TX, vol. 2, pp. 970-974.

Johnson, S. J. and Weller, S. R. (2003). Resolvable 2-design for regular low-density parity-check codes. *IEEE Transactions on Communications*, **51(9)**, 1413 – 1419.

Johnson, S. J. and Weller, S. R. (2004). Codes for iterative decoding from partial geometries. *IEEE Transactions on Communications*, **52**, 236–243.

Kou, Y., Lin, S., and Fossorier, M. P. C. (2001). Low-density parity-check codes based on finite geometries: a rediscovery and new results. *IEEE Transactions on Information Theory*, **47(7)**, 2711–2736.

Lan, L., Tai, Y., Lin, S., Memari, B., and Honary, B. (2008). New constructions of quasi-cyclic LDPC codes based on special classes of BIBDs for the AWGN and binary erasure channels. *IEEE Transactions on Communications*, **56(1)**, 39-48.

Li, Xiuli , Zhang, X. C., and Shen, J. (2008). Regular LDPC codes from semipartial geometries. *Acta Applicandae Mathematicae*, **102(1)**, 25–35.

Mahadevan, A. and Morris, J. M. (2002). *On RCD SPC codes as LDPC codes based on arrays and their equivalence to some codes constructed from Euclidean geometries and partial BIBDs*. Technical Reports No.: CSPLTR: 2002–1, Communication Signal Process Labview, CS and EE Department, University of Maryland, USA.

Mathon, R. and Rosa, A. (2007). 2-(v, k, λ) designs of small order. *The CRC Handbook of Combinatorial Designs*, (Eds.): C. J. Colbourn and J. H. Dinitz, CRC Press, 35–58.

Raghavarao, D. (1971). *Constructions and Combinatorial Problems in Design of Experiments*. John Wiley, New York.

Raghavarao, D. and Padgett, Lakshmi V. (2005). *Block designs. Analysis, Combinatorics and Applications*. Series on Applied Mathematics, 17, World Scientific, Singapore.

Ray Chaudhuri, D. K. and Wilson, R. M. (1971). Solution of Kirkman's schoolgirl problem. *Proceedings of Symposia in Pure Mathematics*, **19**, 187–203.

Saurabh, S. and Sinha, K. (2023). LDPC codes based on α-resolvable designs. *SN Computer Science*, 4, 617, https:doi.org/10.1007/s42979-023-02088-2.

Shan, X. and Li, T. (2013). A construction of low-density parity-check codes. *Journal of Mathematical Research with Applications*, **33(3)**, 330-336.

Sun, X. and Ge, G. (2009). Resolvable group divisible designs with block size four and general index. *Discrete Mathematics*, **309**, 2982–2989.

Vasic, B. V., Kurtas, E. M., and Kuznetsov, A. V. (2002). LDPC codes based on mutually orthogonal Latin rectangles and their applications in perpendicular magnetic recording. *IEEE Transactions on Magnetics*, **38**, 2346–2348.

Vasic, B. V. and Milenkovic, O. (2004). Combinatorial constructions of low-density parity-check codes for iterative decoding. *IEEE Transactions on Information Theory*, **50(6)**, 1156–1176.

Weller, S. R. and Johnson, S. J. (2003). Regular low-density parity-check codes from oval designs. *European Transactions on Telecommunications*, **14**, 399–409.

Wicker, S. B. (1995). *Error Control Systems for Digital Communication and Storage*. Prentice-Hall, Upper Saddle River, NJ 07458.

Xu, J., Chen, L., Zeng, L., and Lin, S. (2005). Construction of low-density parity-check codes superposition. *IEEE Transactions on Communications*, **53(2)**, 243–251.

Xu, H., Feng, D., Sun, C., and Bai, B. (2015). Construction of LDPC codes based on resolvable Group divisible designs. *2015 International Workshop on High Mobility Wireless Communications (HMWC)*, DOI: 10.1109/HMWC.2015.7354346, 111–115.

Xu, H., Yu, Z., Feng, D., and Zhu, H. (2019). New construction of partial geometries based on group divisible designs and their associated LDPC codes. *Physical Communication*, https://doi.org/10.1016/j.phycom.2019.100970.

Zhang, L., Huang, Q., Lin, S., Abdel-Ghaffar, A., and Blake, I. F. (2010). Quasi-cyclic LDPC codes: an algebraic construction, rank analysis and codes on Latin squares. *IEEE Transactions on Communications*, **58(11)**, 3126–3139.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 22, No. 2, 2024 (New Series), pp 121–132
http://www.ssca.org.in/journal

Calibration Approach for Estimating Mean of a Stratified Population in the Presence of Non-response

Manoj K. Chaudhary, Anil Prajapati and Basant K. Ray
Department of Statistics, Institute of Science, Banaras Hindu University, Varanasi, India

Received: 31 December 2022; Revised: 26 November 2023; Accepted: 03 December 2023

Abstract

Calibration approach is a systematic way of including the auxiliary information in order to increase the precision of the estimates of a population parameter. In this paper, we have suggested some calibration estimators for estimating the mean of a stratified population under non-response. An efficient use of suitable auxiliary information has been elaborated to obtain a better estimate of the population mean under certain conditions. We have obtained new stratum weights for which the variance of the suggested calibration estimators would achieve its minimum. An empirical study has also been carried out to verify the theoretical outcomes.

Key words: Calibration approach; Auxiliary information; Stratified random sampling; Population mean; Non-response.

AMS Subject Classifications: 62K05, 05B05

1. Introduction

Calibration offers a methodical approach to incorporate the auxiliary information in increasing the precision of the estimates. The concept behind the calibration is to find out the new calibrated weights in such a way that the mean square error of the estimators would be minimized. To construct the new calibrated weights, the chi-square distance measure and some calibration constraints based on auxiliary information can be utilized. Deming and Stephan (1940) were the first to pick up the idea of calibration in a sample survey. Deville and Särndal (1992) adopted the idea of calibration approach in estimating the population parameters. Särndal (2007) provides a deep study of the calibration approach, including methods for avoiding extreme weights, estimation of complex parameters and estimation under a complex sampling design. Kim and Park (2010) prove that an instrumental variable calibration estimator and a functional-form calibration estimator are asymptotically equivalent.

Consider a sample 's' of size n which is drawn from a population of Nunits by simple random sampling without replacement (SRSWOR) scheme. The study variable Y is observed for each unit in the sample hence the observation y_i is known for all $i \in s$ since

Corresponding Author: Manoj K. Chaudhary
Email: mk15@bhu.ac.in

the values $y_1, y_2, ..., y_N$are known for the entire population. To estimate the population total $Y^* = \sum_{i=1}^{N} y_i$, Deville and Särndal (1992) have suggested the calibration estimator, which is constructed as $\widehat{Y}= \sum_{i\in S} p_i y_i$, where the calibration weights p_i's are chosen to minimize their average distance from the basic design weights $d_i = 1/\pi_i$ that are used in the Horvitz and Thompson (1952) estimator given by

$$\widehat{Y}_{HT}= \sum_{i\in S} d_i y_i \tag{1}$$

subject to the constraint

$$\sum_{i\in S} p_i x_i = X^* \tag{2}$$

where $X^*$is the known population total for the auxiliary variable Xwhich is observed for each unit in the sample hence the observation x_i is known for all $i \in s$. The most common distance measure is given as

$$\phi = \sum_{i\in S} \frac{(d_i - p_i)^2}{d_i q_i} \tag{3}$$

where q_i's are known positive weights uncorrelated withd_i. Then the resulting calibration estimator is given as follows:

$$\widehat{Y}= \sum_{i\in S} p_i y_i =\widehat{Y}_{HT} + \widehat{B}\left(X^*- \widehat{X}_{HT}\right) \tag{4}$$

where$\widehat{B}= \left[\sum_{i\in S} d_i q_i x_i^2\right]^{-1} \left[\sum_{i\in S} d_i q_i x_i y_i\right]$ and $\widehat{X}_{HT}= \sum_{i\in S} d_i x_i$. The definition of $\widehat{Y}$is equivalent to a generalized estimator with the choice of q_i.

The authors such as Rao (1994), Estevao and Särndal (2009), Sud *et al.* (2014), Han (2018), Gautam *et al.* (2020), Jaiswal *et al.* (2023), Singh *et al.* (2023) and others have contributed a lot to the survey sampling in estimating the population parameters with a view to justify the concept of calibration approach.

2. Literature reviews under stratified random sampling

Consider a finite population $U = (U_1, U_2, ..., U_N)$ of size $N$and it is divided into k homogeneous groups (called strata). Let the size of i^{th} stratum be $N_i\,(i = 1, 2, ..., k)$ and hence $\sum_{i=1}^{k} N_i = N$. Let Y and X be the study and auxiliary variables with respective population means $\overline{Y}$and $\overline{X}$. A sample of size n_i is drawn by SRSWOR scheme from the i^{th} stratum such that $\sum_{i=1}^{k} n_i = n$. Let (y_{ij},x_{ij}) be the observed values of (Y, X) on the j^{th} unit in the i^{th} stratum $(j = 1, 2, ..., N_i)$. The classical unbiased estimator of the population mean $\overline{Y}$ is given by

$$\overline{y}_{st} = \sum_{i=1}^{k} w_i \overline{y}_i \tag{5}$$

where $\overline{y}_i$ is the mean based on n_i units for the study variable and $w_i = \frac{N_i}{N}$.

In the availability of auxiliary information, Singh *et al.* (1998) suggested a new calibration estimator of the population mean $\overline{Y}$ as

$$\overline{y}_{c,st} = \sum_{i=1}^{k} w_i^* \overline{y}_i \tag{6}$$

where w_i^* is a new calibrated weight such that it minimizes the chi- square distance function

$$\varphi = \sum_{i=1}^{k} \frac{\left(w_i^* - w_i\right)^2}{w_i q_i} \tag{7}$$

subject to the calibration constraint

$$\sum_{i=1}^{k} w_i^* \overline{x}_i = \overline{X} \tag{8}$$

where q_i is the tuning parameter for the i^{th} stratum and $\overline{x}_i$ is the mean based on n_i units for the auxiliary variable.

The calibration constraint given in equation (8) is similar as used by Dupont (1995) and Hidiroglou and Särndal (1998) for two-phase sampling design. Minimization of chi-square distance function given in equation (7) subject to the calibration constraint (8) leads to the calibrated weights

$$w_i^* = w_i + \frac{w_i q_i \overline{x}_i}{\sum_{i=1}^{k} w_i q_i \overline{x}_i^2} \left[\overline{X} - \sum_{i=1}^{k} w_i \overline{x}_i\right] \tag{9}$$

Substituting the value of w_i^* from equation (9) into equation (6), one can get combined regression-type estimator given by

$$\overline{y}_{c,st} = \sum_{i=1}^{k} w_i \overline{y}_i + \frac{\sum_{i=1}^{k} w_i q_i \overline{x}_i \overline{y}_i}{\sum_{i=1}^{k} w_i q_i \overline{x}_i^2} \left[\overline{X} - \sum_{i=1}^{k} w_i \overline{x}_i\right] \tag{10}$$

An estimator of the variance of the calibration estimator $\overline{y}_{c,st}$ is represented as

$$\widehat{V}\left(\overline{y}_{c,st}\right) = \sum_{i=1}^{k} \frac{w_i^2 \left(1 - f_i\right) s_{ei}^2}{n_i} \tag{11}$$

where $s_{ei}^2 = \frac{1}{n_i - 1} \sum_j^{n_i} e_{ij}^2$, $f_i = \frac{n_i}{N_i}$, $e_{ij} = \left(y_{ij} - \overline{y}_i\right) - b\left(x_{ij} - \overline{x}_i\right)$and $b = \frac{\sum_{i=1}^{k} w_i q_i \overline{x}_i \overline{y}_i}{\sum_{i=1}^{k} w_i q_i \overline{x}_i^2}$.

Moreover, there are several authors who have implemented the notion of calibration approach in estimating the parameters of a stratified population. Tracy *et al.* (2003), Kim *et al.* (2007), Koyuncu and Kadilar (2013, 2014), Clement and Enang (2015), Nidhi *et al.* (2017), Rao *et al.* (2017), Ozgul (2019) and others have proposed a number of calibration estimators in stratified random sampling.

The occurrence of non-response is inherent in sample surveys. Rubin (1976) delineated three key concepts, *viz.*, (i) Missing at Random (MAR), (ii) Missing Completely at Random (MCAR) and (iii) Observed at Random (OAR). MAR method addresses non-response scenarios by assuming that missing data occur randomly and depend only on observed information. Utilizing Multiple Imputation (MI), this technique generates multiple plausible imputations, which reflect the uncertainty associated with missing values. MCAR

is a category of missing data mechanism in which the likelihood of a data point being missing has no connection to either observed or unobserved data. In the context of OAR, the data adhere to this pattern if, for every conceivable missing data value, the probability of the observed missing pattern, given both observed and unobserved data, is independent of the specific values within the observed data. It is to be noted that the non-response error is not so important if the characteristics of the non-responding units are similar to those of the responding units. But, such similarity of characteristics between the responding and non-responding units is not always attained in custom. In such a situation, it is much difficult to get the précised estimates of the parameters. To deal with the problem of non-response, Hansen and Hurwitz (1946) suggested a technique of sub-sampling of non-respondents. Later on, Khare (1987), Chaudhary *et al.* (2012, 2018) have discussed the problem of non-response in estimating the parameters of a stratified population.

It is to be mentioned that there are two types of non-response; (i) unit non-response and (ii) item non-response. In the subsequent sections, we have tried to propose an efficient calibration method of estimation of the population mean $\overline{Y}$ in stratified random sampling utilizing the information on an auxiliary variable X under unit non-response. The calibration estimators have been pioneered out under the situation in which the knowledge about the population mean of the auxiliary variable is available in advance. It is further assumed that the study variable is suffering from the non-response, whereas the auxiliary variable does not suffer from the non-response. The theoretical facts have been demonstrated through an empirical study.

3. Proposed calibration estimators

In the presence of non-response, the sampling strategy given in section-2 has been extended to the further process. It is noted that out of n_i units, n_{i1} units respond and n_{i2}units do not respond on the study variable Y. Adopting Hansen and Hurwitz (1946) technique of sub-sampling of non-respondents, a sub-sample of $h_{i2}\left(=\frac{n_{i2}}{g_i}; g_i > 1\right)$ units is selected from the n_{i2} non-responding units using SRSWOR scheme and information is collected from all the h_{i2} units. The usual estimator of the population mean $\overline{Y}$ under non-response (without using auxiliary information) is given by

$$\overline{y}^*_{st} = \sum_{i=1}^{k} w_i \overline{y}^*_i \tag{12}$$

where $\overline{y}^*_i = \frac{n_{i1}\overline{y}_{ni1}+n_{i2}\overline{y}_{hi2}}{n_i}$. $\overline{y}_{ni1}$ and $\overline{y}_{hi2}$ are respectively the means based on n_{i1} responding units and h_{i2} non-responding units for study variable in the i^{th} stratum.

The estimate of the variance of the estimator $\overline{y}^*_{st}$ is given as

$$V\left[\overline{y}^*_{st}\right] = \sum_{i=1}^{k} \frac{w_i^2\left(1-f_i\right)}{n_i} s^{*2}_{yi} + \sum_{i=1}^{k} \frac{w_i^2\left(g_i-1\right)W_{i2}}{n_i} s^2_{yi(2)} \tag{13}$$

where $s^{*2}_{yi} = \frac{1}{n^*_i-1}\sum_j^{n^*_i}\left(y_{ij}-\overline{y}^*_i\right)^2$, $s^2_{yi(2)} = \frac{1}{h_{i2}-1}\sum_j^{h_{i2}}\left(y_{ij}-\overline{y}_{hi2}\right)^2$, $n^*_i = n_{i1}+h_{i2}$ and $W_{i2}\left(=\frac{N_{i2}}{N_i}\right)$ is the non-response rate in the population for the i^{th} stratum.

Here, we have considered the situation in which the non-response occurs on the study variable, whereas the auxiliary variable is free from the non-response. In this situation, we

have suggested some calibration estimators of the population mean $\overline{Y}$ when the information about the population mean $\overline{X}$ of the auxiliary variable is known in advance. Following Singh *et al.* (1998), we now propose a calibration estimator of the population mean $\overline{Y}$ in the presence of non-response as

$$\overline{y}^*_{st(C)} = \sum_{i=1}^{k} \delta_i^* \overline{y}_i{}^* \tag{14}$$

where δ_i^* is an adjusted calibrated weight for the i^{th} stratum.

In order to get the optimum value of calibrated weight δ_i^*, we now minimize the chi-square distance function

$$\varphi^* = \sum_{i=1}^{k} \frac{(\delta_i^* - w_i)^2}{w_i q_i} \tag{15}$$

subject to the calibration constraint

$$\sum_{i=1}^{k} \delta_i^* \overline{x}_i = \overline{X} \tag{16}$$

Let us define the Lagrange function

$$L = \sum_{i=1}^{k} \frac{(\delta_i^* - w_i)^2}{w_i q_i} - 2\lambda \left(\sum_{i=1}^{k} \delta_i^* \overline{x}_i - \overline{X} \right) \tag{17}$$

where λ is the Lagrange multiplier.

Differentiating the equation (17) with respect to δ_i^* and equating the derivative to zero, we get

$$\begin{aligned} \frac{\partial L}{\partial \delta_i^*} &= 2\frac{(\delta_i^* - w_i)}{w_i q_i} - 2\lambda \overline{x}_i = 0 \\ \Rightarrow \delta_i^* &= w_i + \lambda w_i q_i \overline{x}_i \end{aligned} \tag{18}$$

Putting the value δ_i^* from equation (18) into the equation (16), we have

$$\lambda = \frac{\overline{X} - \sum_{i=1}^{k} w_i \overline{x}_i}{\sum_{i=1}^{k} w_i q_i \overline{x}_i^2} \tag{19}$$

Substituting the value of λ form equation (19) into equation (18), we get the optimum calibrated weights as

$$\delta_i^* = w_i + \frac{w_i q_i \overline{x}_i}{\sum_{i=1}^{k} w_i q_i \overline{x}_i^2} \left[\overline{X} - \sum_{i=1}^{k} w_i \overline{x}_i \right] \tag{20}$$

Putting the value of δ_i^* from equation (20) into the equation (14), the proposed calibration estimator becomes

$$\overline{y}^*_{st(C)} = \sum_{i=1}^{k} w_i \overline{y}_i{}^* + \frac{\sum_{i=1}^{k} w_i q_i \overline{x}_i \overline{y}_i{}^*}{\sum_{i=1}^{k} w_i q_i \overline{x}_i^2} \left[\overline{X} - \sum_{i=1}^{k} w_i \overline{x}_i \right] \tag{21}$$

The estimators of the bias and variance of the calibration estimator $\overline{y}^*_{st(C)}$ are respectively given by

$$\widehat{B}\left(\overline{y}^*_{st(C)}\right) = \sum_{i=1}^{k} w_i b^* \overline{x}_i \left[\frac{N_i\left(N_i - n_i\right)}{\left(N_i - 1\right)\left(N_i - 2\right)} \cdot \frac{1}{n_i \overline{x}_i} \left\{ \frac{\widehat{\mu}_{30i}}{s^2_{xi}} - \frac{\widehat{\mu}_{21i}}{s^*_{xyi}} \right\} + \frac{W_{i2}\left(g_i - 1\right)}{n_i \overline{x}_i} \left\{ \frac{\widehat{\mu}_{30i(2)}}{s^2_{xi}} - \frac{\widehat{\mu}_{21i(2)}}{s^*_{xyi}} \right\} \right] \tag{22}$$

$$\widehat{V}\left(\overline{y}^*_{st(C)}\right) = \sum_{i=1}^{k} \frac{w_i^2\left(1 - f_i\right)}{n_i} \left(s^{*2}_{yi} + b^{*2} s^2_{xi} - 2b^* s^*_{xyi}\right) + \sum_{i=1}^{k} \frac{w_i^2\left(g_i - 1\right) W_{i2}}{n_i} s^2_{yi(2)} \tag{23}$$

where $b^* = \frac{\sum_{i=1}^{k} w_i q_i \overline{x}_i \overline{y}_i{}^*}{\sum_{i=1}^{k} w_i q_i \overline{x}_i^2}$, $s^2_{xi} = \frac{1}{n_i - 1}\sum_j^{n_i}\left(x_{ij} - \overline{x}_i\right)^2$, $s^*_{xyi} = \frac{1}{n_i^* - 1}\sum_j^{n_i^*}\left(x_{ij} - \overline{x}_i^*\right)\left(y_{ij} - \overline{y}_i^*\right)$, $\overline{x}_i^* = \frac{n_{i1}\overline{x}_{ni1} + n_{i2}\overline{x}_{hi2}}{n_i}$
and $n_i^* = n_{i1} + h_{i2}$. $\overline{x}_{ni1}$ and $\overline{x}_{hi2}$ are respectively the means based on n_{i1} responding units and h_{i2} non-responding units for auxiliary variable in the i^{th} stratum.

$\widehat{\mu}_{30i} = \frac{1}{n_i - 1}\sum_j^{n_i}\left(x_{ij} - \overline{x}_i\right)^3$, $\widehat{\mu}_{21i} = \frac{1}{n_i^* - 1}\sum_j^{n_i^*}\left(x_{ij} - \overline{x}_i^*\right)^2\left(y_{ij} - \overline{y}_i^*\right)$,
$\widehat{\mu}_{30i(2)} = \frac{1}{h_{i2} - 1}\sum_j^{h_{i2}}\left(x_{ij} - \overline{x}_{hi2}\right)^3$ and $\widehat{\mu}_{21i(2)} = \frac{1}{h_{i2} - 1}\sum_j^{h_{i2}}\left(x_{ij} - \overline{x}_{hi2}\right)^2\left(y_{ij} - \overline{y}_{hi2}\right)$.

Particular cases:

(i) For instance, if $q_i = \frac{1}{\overline{x}_i}$, then the equation (21) reduces to the well-known combined ratio-type estimator of the population mean $\overline{Y}$ under non-response, *i.e.*, $\overline{y}^*_{st(C)R} = \frac{\sum_{i=1}^{k} w_i \overline{y}_i^*}{\sum_{i=1}^{k} w_i \overline{x}_i}\overline{X}$.

(ii) Putting $q_i = 1$ into the equation (21), it reduces to the combined regression-type estimator of the population mean $\overline{Y}$ under non-response, *i.e.*, $\overline{y}^*_{st(C)Reg} = \sum_{i=1}^{k} w_i \overline{y}_i{}^* + \frac{\sum_{i=1}^{k} w_i \overline{x}_i \overline{y}_i{}^*}{\sum_{i=1}^{k} w_i \overline{x}_i^2}\left[\overline{X} - \sum_{i=1}^{k} w_i \overline{x}_i\right]$.

We now propose an improved calibration estimator of the population mean $\overline{Y}$ under non-response as follows:

$$\overline{y}^{**}_{st(C)} = \sum_{i=1}^{k} \delta_i^{**} \overline{y}_i{}^* \tag{24}$$

where δ_i^{**} is the new calibrated weight for the i^{th}stratum.

The new calibrated weight δ_i^{**} is chosen such that the chi-square type distance

$$\varphi^{**} = \sum_{i=1}^{k} \frac{\left(\delta_i^{**} - w_i\right)^2}{w_i q_i} \tag{25}$$

is minimum, subject to the constraints

$$\sum_{i=1}^{k} \delta_i^{**} \overline{x}_i = \overline{X} \tag{26}$$

$$\sum_{i=1}^{k} \delta_i^{**} = 1 \tag{27}$$

Let us consider the Lagrange function

$$\Delta = \sum_{i=1}^{k} \frac{\left(\delta_i^{**} - w_i\right)^2}{w_i q_i} - 2\phi_1 \left(\sum_{i=1}^{k} \delta_i^{**} \overline{x}_i - \overline{X} \right) - 2\phi_2 \left(\sum_{i=1}^{k} \delta_i^{**} - 1 \right) \tag{28}$$

where ϕ_1 and ϕ_2 are the Lagrange multipliers.

Differentiating the equation (28) with respect to δ_i^{**} and equating the derivative to zero, we get

$$\frac{\partial \Delta}{\partial \delta_i^{**}} = 2\frac{(\delta_i^{**} - w_i)}{w_i q_i} - 2\phi_1 \overline{x}_i - 2\phi_2 = 0$$

$$\left\{ Since \quad \frac{\partial}{\partial x}\left(F_1 \pm F_2 \pm \cdots \pm F_n\right) = \frac{\partial}{\partial x}F_1 \pm \frac{\partial}{\partial x}F_2 \pm \cdots \pm \frac{\partial}{\partial x}F_n \right\}$$

$$\Rightarrow \delta_i^{**} = w_i + w_i q_i \left(\phi_1 \overline{x}_i + \phi_2\right) \tag{29}$$

Let us put the value of δ_i^{**} from equation (29) into the equation (26). The resulting equation is given as

$$\phi_1 \sum_{i=1}^{k} w_i q_i \overline{x}_i^2 + \phi_2 \sum_{i=1}^{k} w_i q_i \overline{x}_i = \overline{X} - \sum_{i=1}^{k} w_i \overline{x}_i \tag{30}$$

Let us now substitute the value of δ_i^{**} from equation (29) into the equation (27). The resulting equation becomes

$$\phi_1 \sum_{i=1}^{k} w_i q_i \overline{x}_i + \phi_2 \sum_{i=1}^{k} w_i q_i = 0 \tag{31}$$

The equations (30) and (31) can be written in the following matrix form:

$$A\phi = B \tag{32}$$

where $A = \begin{bmatrix} \sum_{i=1}^{k} w_i q_i \overline{x}_i^2 & \sum_{i=1}^{k} w_i q_i \overline{x}_i \\ \sum_{i=1}^{k} w_i q_i \overline{x}_i & \sum_{i=1}^{k} w_i q_i \end{bmatrix}$, $\phi = \begin{bmatrix} \phi_1 \\ \phi_2 \end{bmatrix}$ and $B = \begin{bmatrix} \overline{X} - \sum_{h=1}^{L} w_h \overline{x}_h \\ 0 \end{bmatrix}$.

The inverse of the matrix A is given as

$$A^{-1} = \frac{1}{|A|} \begin{bmatrix} \sum_{i=1}^{k} w_i q_i & -\sum_{i=1}^{k} w_i q_i \overline{x}_i \\ -\sum_{i=1}^{k} w_i q_i \overline{x}_i & \sum_{i=1}^{k} w_i q_i \overline{x}_i^2 \end{bmatrix}$$

where $|A| = \sum_{i=1}^{k} w_i q_i \overline{x}_i^2 \sum_{i=1}^{k} w_i q_i - \left(\sum_{i=1}^{k} w_i q_i \overline{x}_i\right)^2$.

The solution of the system of equation (32) is given by

$$\begin{bmatrix} \phi_1 \\ \phi_2 \end{bmatrix} = \frac{1}{|A|} \begin{bmatrix} \sum_{i=1}^{k} w_i q_i \left(\overline{X} - \sum_{i=1}^{k} w_i \overline{x}_i\right) \\ -\sum_{i=1}^{k} w_i q_i \overline{x}_i \left(\overline{X} - \sum_{i=1}^{k} w_i \overline{x}_i\right) \end{bmatrix} \tag{33}$$

From equation (33), we have

$$\left.\begin{aligned}\phi_1 &= \frac{\sum_{i=1}^{k} w_i q_i \left(\overline{X} - \sum_{i=1}^{k} w_i \overline{x}_i\right)}{|A|} \\ \phi_2 &= \frac{-\sum_{i=1}^{k} w_i q_i \overline{x}_i \left(\overline{X} - \sum_{i=1}^{k} w_i \overline{x}_i\right)}{|A|}\end{aligned}\right\} \tag{34}$$

Thus, the optimum weight δ_i^{**} becomes

$$\delta_i^{**} = w_i + \frac{w_i q_i \overline{x}_i \sum_{i=1}^{k} w_i q_i - w_i q_i \sum_{i=1}^{k} w_i q_i \overline{x}_i}{\sum_{i=1}^{k} w_i q_i \overline{x}_i^2 \sum_{i=1}^{k} w_i q_i - \left(\sum_{i=1}^{k} w_i q_i \overline{x}_i\right)^2} \left(\overline{X} - \sum_{i=1}^{k} w_i \overline{x}_i\right) \tag{35}$$

Substituting the value of δ_i^{**} from equation (35) into the equation (24), the proposed calibration estimator becomes

$$\overline{y}_{st(C)}^{**} = \sum_{i=1}^{k} w_i \overline{y}_i^{*} + \hat{\beta} \left(\overline{X} - \sum_{i=1}^{k} w_i \overline{x}_i\right) \tag{36}$$

where $\hat{\beta} = \dfrac{\left(\sum_{i=1}^{k} w_i q_i \overline{x}_i \overline{y}_i^{*}\right)\left(\sum_{i=1}^{k} w_i q_i\right) - \left(\sum_{i=1}^{k} w_i q_i \overline{y}_i^{*}\right)\left(\sum_{i=1}^{k} w_i q_i \overline{x}_i\right)}{\left(\sum_{i=1}^{k} w_i q_i \overline{x}_i^2\right)\left(\sum_{i=1}^{k} w_i q_i\right) - \left(\sum_{i=1}^{k} w_i q_i \overline{x}_i\right)^2}$.

Now, the estimators of the bias and variance of the proposed calibration estimator $\overline{y}_{st(C)}^{**}$ are respectively represented as

$$\begin{aligned}\widehat{B}\left(\overline{y}_{st(C)}^{**}\right) = \sum_{i=1}^{k} w_i \hat{\beta} \overline{x}_i \Bigg[&\frac{N_i \left(N_i - n_i\right)}{\left(N_i - 1\right)\left(N_i - 2\right)} \cdot \frac{1}{n_i \overline{x}_i} \left\{ \frac{\widehat{\mu}_{30i}}{s_{xi}^2} - \frac{\widehat{\mu}_{21i}}{s_{xyi}^{*}} \right\} \\ &+ \frac{W_{i2}\left(g_i - 1\right)}{n_i \overline{x}_i} \left\{ \frac{\widehat{\mu}_{30i(2)}}{s_{xi}^2} - \frac{\widehat{\mu}_{21i(2)}}{s_{xyi}^{*}} \right\} \Bigg]\end{aligned} \tag{37}$$

$$\widehat{V}\left(\overline{y}_{st(C)}^{**}\right) = \sum_{i=1}^{k} \frac{w_i^2 \left(1 - f_i\right)}{n_i} \left(s_{yi}^{*2} + \hat{\beta}^2 s_{xi}^2 - 2\hat{\beta} s_{xyi}^{*}\right) + \sum_{i=1}^{k} \frac{w_i^2 \left(g_i - 1\right) W_{i2}}{n_i} s_{yi(2)}^2 \tag{38}$$

Note: The equation (37) can provide the non-response versions of a number of combined-type estimators of the population mean $\overline{Y}$ for the suitable choices of q_i.

4. Simulation study

In this section, a simulation study has been carried out with a view to verify the performance of the proposed calibration estimators. We have considered a hypothetical data set which is generated using R software under the condition of normal distribution. Here, we first define the two random variables Y^{*} and X^{*} *i.e.*, $Y^{*} \sim N\left(0, 1\right)$ and $X^{*} \sim N\left(0, 1\right)$. Now, we generate a set of correlated random variables with correlation coefficient ρ using the transformations $Y^{**} = Y^{*}$ and $X^{**} = \rho Y^{*} + \sqrt{1 - \rho^2} X^{*}$ [See Reddy *et al.* (2010)]. Finally, we define the random variables Y and X using the transformations $Y = \mu_Y + \sigma_Y Y^{**}$ and $X = \mu_x + \sigma_X X^{**}$. The above transformations constitute the random variables which

are normally distributed with some meansμ_Y, μ_X and variancesσ_Y^2, σ_X^2. In this data set, a population of 15000 units has been shaped out. The population is divided into four strata with respective sizes 6000, 3000, 1500 and 4500. The sample size has been fixed as 3000. The sample size for each stratum has been determined under proportional allocation. To carry out the simulation analysis, the number of runs has been considered as 1000. The summary of the data set is given in Table 1.

Table 1: Distribution of population

Stratum No.(i)	Stratum size(N_i)	Sample size(n_i)	Distribution of of Y *i.e.* $Y \sim N(\mu_Y, \sigma_Y)$	Distribution of X *i.e.* $X \sim N(\mu_X, \sigma_X)$	Correlation coefficient between Y and X
1	6000	1200	$N(200, 20)$	$N(100, 10)$	0.78
2	3000	600	$N(230, 17)$	$N(120, 15)$	0.82
3	1500	300	$N(240, 22)$	$N(145, 22)$	0.8
4	4500	900	$N(235, 23)$	$N(135, 19)$	0.75

Table 2 depicts the estimate of the variance of the usual estimator $\overline{y}^*_{st}$ and proposed calibration estimators $\overline{y}^*_{st(C)}$ and $\overline{y}^{**}_{st(C)}$ at the different levels of non-response rate W_{i2} and inverse sampling rate g_i. The percentage relative efficiency (PRE) of the proposed calibration estimators $\overline{y}^*_{st(C)}$ and $\overline{y}^{**}_{st(C)}$ with respect to the usual estimator $\overline{y}^*_{st}$ has also been computed.

Table 2: Estimate of variance and PRE of the estimators $\overline{y}^*_{st}$, $\overline{y}^*_{st(C)}$, and $\overline{y}^{**}_{st(C)}$

$W_{i2}\forall i$	$g_i\forall i$	Estimate of Variance			PRE		
		$\overline{y}^*_{st}$	$\overline{y}^*_{st(C)}$	$\overline{y}^{**}_{st(C)}$	$\overline{y}^*_{st}$	$\overline{y}^*_{st(C)}$	$\overline{y}^{**}_{st(C)}$
0.1	1	0.04522	0.03329	0.02214	100	135.843	204.295
	2	0.04794	0.03592	0.02482	100	133.47	193.16
	2.5	0.05055	0.03857	0.02745	100	131.076	184.169
	3	0.05325	0.04123	0.03013	100	129.164	176.754
0.2	1.5	0.04789	0.03591	0.02479	100	133.376	193.223
	2	0.05325	0.04128	0.03015	100	128.988	176.627
	2.5	0.05861	0.04667	0.03551	100	125.572	165.059
	3	0.06385	0.05186	0.04075	100	123.124	156.708
0.3	1.5	0.05057	0.03858	0.02746	100	131.069	184.145
	2	0.05838	0.04648	0.03534	100	125.61	165.184
	2.5	0.06642	0.05437	0.04329	100	122.161	153.414
	3	0.07448	0.06241	0.05133	100	119.352	145.1
0.4	1.5	0.05224	0.04027	0.02915	100	129.712	179.227
	2	0.06175	0.04984	0.03869	100	123.902	159.61
	2.5	0.07157	0.05951	0.04843	100	120.266	147.784
	3	0.08125	0.06921	0.0581	100	117.403	139.848

From the Table 2, it is revealed that the estimates of the variance of the proposed

calibration estimators $\overline{y}^{*}_{st(C)}$ and $\overline{y}^{**}_{st(C)}$ are much smaller than the usual estimator $\overline{y}^{*}_{st}$ and hence the PRE of the proposed calibration estimators $\overline{y}^{*}_{st(C)}$ and $\overline{y}^{**}_{st(C)}$ is much higher as compared to the usual estimator$\overline{y}^{*}_{st}$. It is further revealed that the estimates of the variance of the proposed calibration estimators $\overline{y}^{*}_{st(C)}$ and $\overline{y}^{**}_{st(C)}$ increase with the increase in non-response rate W_{i2} and inverse sampling rate g_i as well. Such kind of outcomes is intuitively anticipated. Table 3 represents the estimate of the bias of the proposed calibration estimators $\overline{y}^{*}_{st(C)}$ and $\overline{y}^{**}_{st(C)}$ at the different levels of non-response rate W_{i2} and inverse sampling rate g_i.

Table 3: Estimate of bias of estimators $\overline{y}^{*}_{st(C)}$and $\overline{y}^{**}_{st(C)}$

$W_{i2}\forall i$	$g_i\forall i$	Estimate of Bias	
		$\overline{y}^{*}_{st(C)}$	$\overline{y}^{**}_{st(C)}$
0.1	1	-0.00119	-0.0005
	2	-0.00128	-0.0006
	2.5	-0.00134	-0.0006
	3	-0.00139	-0.0006
0.2	1	-0.00115	-0.0005
	2	-0.00137	-0.0006
	2.5	-0.00163	-0.0007
	3	-0.00168	-0.0008
0.3	1	-0.00116	-0.0005
	2	-0.00153	-0.0007
	2.5	-0.0017	-0.0008
	3	-0.00186	-0.0008
0.4	1	-0.00115	-0.0005
	2	-0.00171	-0.0008
	2.5	-0.00192	-0.0009
	3	-0.00209	-0.0009

The Table 3 reveals that both calibration estimators $\overline{y}^{*}_{st(C)}$ and $\overline{y}^{**}_{st(C)}$ provide negative bias of very less magnitude. A negative bias in the estimator of the finite population mean suggests that on an average, the estimator leads to underestimate the true mean of the population. Alternatively, if one has to take multiple samples from the population and compute the mean using the estimator, the average of these computations would be below the true population mean.

5. Concluding remarks

We have suggested some calibration estimators for estimating the mean of a stratified population in the presence of non-response. The information on a single auxiliary variable has been utilized to develop the calibration estimators. The chi-square distance measure has been used in obtaining the new stratum weights under the given constraints. The calibration estimators have been proposed under the situation in which the non-response occurs on study variable, whereas the auxiliary variable is free from the non-response. The basic properties of the proposed calibration estimators have been discussed in detail. The expressions for

the estimators of the bias and variance of the proposed calibration estimators have been derived. To examine the behavior of the proposed calibration estimators, a simulation study has been carried out by generating an artificial data set. The Table 2 shows that the proposed calibration estimators $\overline{y}^{*}_{st(C)}$ and $\overline{y}^{**}_{st(C)}$ perform very well as compared to the usual estimator$\overline{y}^{*}_{st}$. From Table 3, it is also revealed that both calibration estimators $\overline{y}^{*}_{st(C)}$ and $\overline{y}^{**}_{st(C)}$ confer bias of very less extent.

References

Chaudhary, M. K., Kumar, A., and Prajapati, A. (2018). Estimating the mean of a stratified population using two auxiliary variables under double sampling scheme in the presence of non-response. *International Journal of Statistics and Economics*, **19**, 123-134.

Chaudhary, M. K., Singh, V. K., and Shukla, R. K. (2012). Combined-type family of estimators of population mean in stratified random sampling under non-response. *Journal of Reliability and Statistical Studies (JRSS)*, **5**, 133-142.

Clement, E. P. and Enang, E. I. (2015). Calibration approach alternative ratio estimator for population mean in stratified sampling. *International Journal of Statistics and Economics*, **16**, 83-93.

Deming, W. E. and Stephan, F. F. (1940). On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics*, **11**, 427-444.

Deville, J. C. and Särndal, C. E. (1992). Calibration estimators in survey sampling. *Journal of the American statistical Association*, **87**, 376-382.

Dupont, F. (1995). Alternative adjustments where there are several levels of auxiliary information. *Survey Methodology*, **21**, 125–135.

Estevao, V. M. and Särndal, C. E. (2009). A new face on two-phase sampling with calibration estimators. *Survey Methodology*, **35**, 3-14.

Gautam, A. K., Sharma, M. K., and Sisodia, B. V. S. (2020). Development of calibration estimator of population mean under non-response. *International Journal of Chemical Studies*, **8**, 1360-1371.

Han, P. (2018). Calibration and multiple robustness when data are missing not at random. *Statistica Sinica*, **28**, 1725-1740.

Hansen, M. H. and Hurwitz, W. N. (1946). The problem of non-response in sample surveys. *Journal of the American Statistical Association*, **41**, 517-529.

Hidiroglou, M. A. and Särndal, C. E. (1998). Use of auxiliary information for two-phase sampling. *Survey Methodology*, **24**, 11-20.

Horvitz, D. G. and Thompsom, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American statistical Association*, **47**, 663-685.

Jaiswal, A. K., Usman, M., and Singh, G. N. (2023). New calibration estimation procedure in the presence of unit non response. *Ain Shams Engineering Journal*, **14**, 101910.

Khare, B. B. (1987). Allocation in stratified sampling in presence of non-response. *Metron*, **45**, 213-221.

Kim, J. K. and Park, M. (2010). Calibration estimation in survey sampling. *International Statistical Review*, **78**, 21-39.

Kim, J.-M., Sungur, E. A., and Heo, T. Y. (2007). Calibration approach estimators in stratified sampling. *Statistics & Probability Letters*, **77**, 99-103.

Koyuncu, N. and Kadilar, C. (2013). Calibration estimator using different distance measures in stratified random sampling. *International Journal of Modern Engineering Research*, **3**, 415-419.

Koyuncu, N. and Kadilar, C. (2014). A new calibration estimator in stratified double sampling. *Hacettepe Journal of Mathematics and Statistics*, **43**, 337-346.

Nidhi, Sisodia, B. V. S., Singh, S., and Singh, S. K. (2017). Calibration approach estimation of the mean in stratified sampling and stratified double sampling. *Communications in Statistics-Theory and Methods*, **46**, 4932-4942.

Ozgul, N. (2019). New calibration estimator based on two auxiliary variables in stratified sampling. *Communications in Statistics-Theory and Methods*, **48**, 1481-1492.

Rao, D. K., Khan, M. G. M., and Singh, G. K. (2017). On calibrated weights in stratified sampling. *ANZIAM Journal*, **59**, C190-C204.

Rao, J. N. K. (1994). Estimating totals and distribution functions using auxiliary information at the estimation stage. *Journal of official statistics*, **10**, 153.

Reddy, M. K., Rao, K. R., and Boiroju, N. K. (2010). Comparison of ratio estimators using Monte Carlo simulation. *International Journal of Agriculture and Statistical Sciences*, **6**, 517-527.

Rubin, B. D. (1976). Inference and missing data. *Biometrika*, **63**, 581-592.

Särndal, C. E. (2007). The calibration approach in survey theory and practice. *Survey Methodology*, **33**, 99-119.

Singh, G. N., Bhattacharyya, D., and Bandyopadhyay, A. (2023). A general class of calibration estimators under stratified random sampling in presence of various kinds of non-sampling errors. *Communications in Statistics-Simulation and Computation*, **52**, 320-333.

Singh, S., Horn, S., and Yu, F. (1998). Estimation of variance of general regression estimator: Higher level calibration approach. *Survey Methodology*, **24**, 41-50.

Sud, U.C., Chandra, H., and Gupta, V.K. (2014). Calibration-based product estimator in single- and two phase sampling. *Journal of Statistical theory and Practice*, **8**, 1-11.

Tracy, D., Singh, S., and Arnab, R. (2003). Note on calibration in stratified and double sampling. *Survey Methodology*, **29**, 99–104.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 22, No. 2, 2024 (New Series), pp 133–143
http://www.ssca.org.in/journal

Reliability Assessment of Two-Component Series System Shock Model

S. B. Munoli and Abhijeet Jadhav
Department of Statistics, Karnatak University, Dharwad - 580003

Received: 24 January 2023; Revised: 21 July 2023; Accepted: 05 December 2023

Abstract

Configuration of sub-assemblies in series is recommended in some environments. Reliability assessment of two-component series system receiving shocks from a single source is studied. Shocks are of two types: damage shocks and fatal shocks. The component fails either due to exceedance of damage to its threshold or when it experiences a fatal shock. The two cases of fixed and random thresholds are considered. Computation and comparison of estimators of two models is done through simulation.

Key words: Series system; Damage shock; Fatal shock; Threshold; Reliability Assessment.

AMS Subject Classifications: 62N02, 62N05

1. Introduction

Configuration of sub-assemblies of a system in different ways is attempted (explored) to meet certain requirements. The two fundamental configurations of subassemblies are series and parallel. These have been studied extensively by reliability engineers, economists, life science and social science researchers. One can quote several examples, wherein series configurations of sub-assemblies is considered such as water heaters, lamps in a circuit, water pumps, freezers and refrigerators, solar panels, etc. Series configuration is needed when the same current must flow through all the sub-assemblies, easy overheating of components is to be avoided, voltage is to be increased to meet the minimum operating requirements of the inverter in solar appliances. In this paper, an attempt is made to study the reliability of two component series system receiving shocks from single source wherein the shocks are of two types: damage shocks and fatal shocks. The pioneering work on shock models is by Esary and Marshall (1973). A-hameed and Proschan (1973),A-Hameed and Proschan (1975) have considered non-stationary shock models and shock models with underlying birth process. Ross (1981) has studied generalized Poisson shock models. Survival under the pure birth shock model was studied by Klefsjö (1981). Shanthikumar and Sumita (1983) have discussed on general shock models with correlated renewal sequences. Semi-Markov shock models with additive damages is studied by Posner and Zuckerman (1986). Anderson (1987) has proposed limit theorems for general shock models. Some multivariate distributions were

Corresponding Author: Abhijeet Jadhav
Email: abhijadhav10292@gmail.com

derived from non-fatal shock models by Savits (1988). Gut (1990) and Skoulakis (2000) have contributed to the literature on general shock models for reliability system and cumulative shock models respectively. Mallor and Santos (2003) have dealt with classification of shock models in system reliability. Applications of Poisson shock models in insurance and credit risk is studied by Lindskog and McNeil (2003). Inference for reliability of shock models is studied by Chikkagoudar and Palaniappan (1981), Kunchur and Munoli (1993), Munoli and Suranagi (2007),Munoli and Suranagi (2009) and Munoli and Bhat (2011).

Several researchers have contributed to the fields of modelling system reliability, its optimization, bounds on system reliability and inference for system reliability. Here are few references of contributions to these fields: Rutemiller (1966), Zacks and Even (1966), Chung (1995), Råde (1976), Nakagawa and Rosenfeld (1979), Weier (1981), Necsulescu and Krausz (1986), Fujii and Sandoh (1984), Wani and Kabe (1971), Hanagal (1996), Munoli and Mutkekar (2011a), Munoli and Mutkekar (2011b). In the present study, a two-component series system is subjected to a sequence of shocks occurring randomly in time as events of Poisson process. Shocks are occurring with intensity $\lambda,\ \lambda > 0$. Shocks are of two types; damage shocks and fatal shocks. Any shock will be a damage shock with probability 'p' and fatal shock with probability $(1-p)$. Every damage shock causes some amounts of damage to both components. Damages are non-accumulating. The component fails whenever the damage exceeds threshold (u) of the component. If not, the component functions as good as new one. On the other hand, the component may also fail when it experiences a fatal shock. The two components function independently. The system fails when either of the two components fail (series system). Let X and Y denote respectively the amount of damages to first failing component of the system and surviving component of the system. X and Y are assumed to be exponential random variables *(r.v.'s)* with parameter $\theta,\ \theta > 0$. The system reliability at time 't' is given by

$$S_1(t) = \sum_{k=0}^{\infty} \frac{e^{-p\lambda t}(p\lambda t)^k}{k!}\ e^{-(1-p)\lambda t}\ \overline{P}_k \tag{1}$$

The above expression represents the following:

The first term $\frac{e^{-p\lambda t}(p\lambda t)^k}{k!}$ is the expression for the probability that system has experienced 'k' number of damage shocks during $(0,\ t)$, $e^{-(1-p)\lambda t}$ represents the probability that the system did not experience a fatal shock during $(0,\ t)$. $\overline{P}_k$ is the probability that the system survives with k number of shocks that it has experienced during $(0,\ t)$. The system may experience during $(0, t)$ no shock or one shock or two shocks,...; hence the summation with $k = 0, 1, 2, \ldots$. $\overline{P}_k$ is given by

$$\begin{aligned}
\overline{P}_k &= P(\text{Both components survive with k number of damage shocks}) \\
&= P\left(X_1 < u,\ \ldots,\ X_k < u\right).\ P\left(Y_1 < u, \ldots, Y_k < u\right) \\
&= \left(1 - e^{-u\theta}\right)^k\ .\ \left(1 - e^{-u\theta}\right)^k \\
&= \left(1 - e^{-u\theta}\right)^{2k}
\end{aligned} \tag{2}$$

Now, substituting the value of $\overline{P}_k$ from expression (2) in the expression for $S_1(t)$ (expression

(1)) and simplifying, we get the expression for reliability of series system as

$$S_1(t) = e^{-\lambda t[1-p\left(1-e^{-u\theta}\right)^2]} \tag{3}$$

The real-life examples from health science and finance for a shock model with damage shocks and fatal shocks are:

Example 1: Heart disease is the leading cause of death worldwide. The common heart diseases are heart attack and cardiac arrest. Heart attacks occur when blood flow to the heart muscle is temporarily blocked, starving the muscle tissue of oxygen which causes scarring and damage to heart muscle (damage shock with amount of damage tolerable and the person survives with this heart attack). For a heart attack to lead to death the damage to the heart needs to be large enough resulting in irregular heart beat and stop eventually (failure due to damage exceeding threshold). Cardiac arrest is an abrupt loss of heart function, breathing and consciousness. It results from an electric disturbance in the heart that disrupts its pumping action, stopping blood flow to the different organs and can lead to death (fatal shock).

Example 2: While lending loans to customers, financial institutes choose customers who fetch the institute high profit. If a loanee defaults (fatal shock), it will be a loss to financial institute. On the other hand, the loanee may do some partial repayments, close the loan account by paying off the loan early. In this case the lender will lose a proportion of the interest. Here partial repayments are damages due to shocks and due to closing the loan account early is failure due to damage exceeding the shock.

These examples are discussed in detail in Munoli and Suhas (2019).

The rest of the paper is organized as follows: Life testing experiment is explained in Section 2. MLE's of the parameters of the model and their asymptotic distribution are obtained in Section 3. Computation of estimators is dealt with in Section 4. Section 5 deals with the case of thresholds of the components being *r.v's*. Comparison of two cases of fixed and random thresholds is made in Section 6, conclusions are also outlined in the same section.

2. Life testing experiment

Suppose, 'r' two-component systems having life distribution $(1 - S_1(t))$ are subjected to a life testing experiment, and the experiment continues until all the systems fail. For the i^{th} system, let the first failure (of two components) occur at the m_i^{th} shock, $i = 1, 2, \ldots, r$, which coincides with system failure (also known as a series system). Out of 'r' number of first failing components, 'r_1' components fail due to damage exceeding threshold 'u', and $r_2(= r - r_1)$ components fail due to experiencing a fatal shock. Let X_{ij} and Y_{ij}, $i = 1, 2, \ldots, r;\ j = 1, 2, \ldots, m_i$, be the random variables representing damages due to the j^{th} damage shock to failing and surviving components of the i^{th} system. The $X_{ij}\prime s$ and $Y_{ij}\prime s$ are assumed to be independent exponential random variables with parameter θ $(\theta > 0)$. Let t_{ij} be the time epoch at which the i^{th} system experienced the j^{th} shock $(j = 1, 2, \ldots, m_i;\ i = 1, 2, \ldots, r)$. The inter-arrival times $(t_{i,j} - t_{i,j-1})$ are exponential random variables with parameter $p\lambda$.

For 'r_1' systems, first failure (out of two components) has occurred due to damage exceeding the threshold, the joint pdf of the *r.v's* $m_i,\ t_{i1},\ t_{i2}, \ldots, t_{im_i},\ X_{i1}, \ldots, X_{im_i-1}, Y_{i1}, \ldots, Y_{im_i}$ is

$$\prod_{i=1}^{r_1} (p\lambda)^{m_i} e^{-p\lambda t_{m_i}}\ \theta^{m_i-1} e^{-\theta \sum_{j=1}^{m_i-1} x_{ij}}\ e^{-u\theta}\ \theta^{m_i}\ e^{-\theta \sum_{j=1}^{m_i} y_{ij}} \tag{4}$$

It is assumed that the amount of damage due to a shock at which component has failed (due to damage exceeding its threshold) is not observable but is known to exceeds its threshold.

For 'r_2' systems, the first failure (out of 2 components) has occurred due to fatal shock and the joint pdf of *r.v's* $m_i,\ t_{i1}, \ldots, t_{im_i},\ \ X_{i1}, \ldots, X_{im_i-1}, Y_{i1}, \ldots,\ Y_{im_i-1}$ is given by

$$\prod_{i=1}^{r_2} (p\lambda)^{m_i-1} e^{-p\lambda t_{m_i-1}} \theta^{m_i-1} e^{-\theta \sum_{j=1}^{m_i-1} x_{ij}} .(1-p)\lambda\ e^{-(1-p)\lambda(t_{m_i}-t_{m_i-1})} \theta^{m_i-1} e^{-\theta \sum_{j=1}^{m_i-1} y_{ij}} \tag{5}$$

In this case, as the system failure has occurred due to experiencing a fatal shock at m_i^{th} shock, the damages due to fatal shock for both surviving and failing components are not observed. Combining (4) and (5), the joint pdf L_1 of all random variables of the experiment is given by

$$L_1 = p^{m-r_2} \lambda^m e^{-p\lambda t_{..}}\ \theta^{2m-r_1-2r_2}\ e^{-\theta\left(x_{..}+y_{..}+r_1 u\right)} (1-p)^{r_2}\, e^{-\lambda t'} \tag{6}$$

where,

$t_{..} \ = \ \sum_{i=1}^{r_1} t_{m_i} + 2\sum_{i=1}^{r_2} t_{m_i-1} - \sum_{i=1}^{r_2} t_{m_i}$

$t' = \sum_{i=1}^{r-r_1} (t_{m_i} - t_{m_i-1})$

$x_{..} = \sum_{i=1}^{r} \sum_{j=1}^{ri-1} x_{ij}$

$y_{..} = \sum_{i=1}^{r_1} \sum_{j=1}^{m_i} y_{ij} + \sum_{i=1}^{r_2} \sum_{j=1}^{m_i-1} y_{ij}$

$m = \sum_{i=1}^{r} m_i$

3. MLE's of parameters

Treating L_1 as function of parameters, the MLE's of parameters are obtained as

$$\hat{p} = \frac{[m-(r-r_1)]\, t'}{mt' + (r-r_1)\, t_{..}} \tag{7}$$

$$\hat{\lambda} = \frac{mt' + (r-r_1)\, t_{..}}{t'(t_{..} + t')} \tag{8}$$

$$\hat{\theta} = \frac{2m - 2r + r_1}{x_{..} + y_{..} + r_1 u} \tag{9}$$

Using invariance property of MLE's, the MLE $\widehat{S_1}(t)\ of\ S_1(t)$ is obtained by substituting $\widehat{p}$, $\widehat{\lambda}$ and $\widehat{\theta}$ for $p,\ \lambda\ \ and\ \theta$ respectively in expression (3).

In order to obtain asymptotic distribution of $\widehat{p}$, $\widehat{\lambda}$ and $\widehat{\theta}$, Fisher information matrix $I(p,\ \lambda,\ \theta)$ is obtained as

$$I(p,\ \lambda,\ \theta) = \begin{bmatrix} \frac{2m-2r+r_1}{\theta^2} & 0 & 0 \\ 0 & \frac{m-(r-r_1)}{p^2} + \frac{r-r_1}{(1-p^2)} & E(t..) \\ 0 & E(t..) & \frac{m}{\lambda^2} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{2m-2r+r_1}{\theta^2} & 0 & 0 \\ 0 & b & E(t..) \\ 0 & E(t..) & \frac{m}{\lambda^2} \end{bmatrix} \tag{10}$$

where,

$$b = \frac{m-(r-r_1)}{p^2} + \frac{r-r_1}{(1-p^2)}, \qquad E(t..) = \sum_{i=1}^{r_1} \frac{m_i}{p\lambda} + 2\sum_{i}^{r-r_1} \frac{m_i-1}{p\lambda} - \sum_{i=1}^{r-r_1} [\frac{m_i-1}{p\lambda} + \frac{1}{(1-p)\lambda}]$$

Using multivariate central limit theorem and asymptotic properties of MLE (under regularity conditions), we have

$$((p-\widehat{p}), \left(\lambda - \widehat{\lambda}\right), (\theta - \widehat{\theta})) \to N_3(0,\ I^{-1})$$

and I^{-1} is given by

$$I^{-1} = \begin{bmatrix} \frac{\theta^2}{2m-2r+r_1} & 0 & 0 \\ 0 & \frac{m}{mb-\lambda^2(E(t..))^2} & -\frac{\lambda^2 E(t..)}{mb-\lambda^2[E(t..)]^2} \\ 0 & -\frac{\lambda^2[E(t..)]^2}{mb-\lambda^2[E(t..)]^2} & \frac{b\lambda^2}{mb-\lambda^2[E(t..)]^2} \end{bmatrix} \tag{11}$$

4. Simulation study

Validation of the model and computation of estimators is made through Monte-Carlo simulation. The random variables of the model are generated for different values of r, t and parameter combinations as below:

For the i^{th} series system of two components, the *r.v's* $t_{i1},\ t_{i2}, \ldots, t_{im_i}, X_{i1}, \ldots, X_{im_i-1}$, $Y_{i1}, Y_{i2}, \ldots, Y_{im_i-1}, Y_{im_i}$ in the case of system failure due to damage shock are generated as follows:

Step 1: Let $u = u_0$ and a random number w_i is generated from $U(0,1)$. If $0 < w_i < p(= p_0)$, then the system failure is considered as failure due to damage shock.

Step 2: Initialize $m_i = 0$, for $\theta = \theta_0$ the r.v. X_{i1}, Y_{i1} following exponential distribution with parameter θ_0 are generated. With this m_i is incremented by 1. X_{i1} *and* Y_{i1} are compared with u_0. If both $X_{i1} < u_0$ and $Y_{i1} < u_0$, the process of exponential r.v's generation with parameter θ_0 and comparing with u_0 is repeated with incrementation of m_i by 1 with every repeatation. The iteration at which either of X_{ij} *or* Y_{ij} exceeds u_0, m_i is noted.

Step 3: m_i number of inter-arrival times having exponential distribution with parameter $p_0\lambda_0$ are generated. Addition of these inter-arrival times results in t_{im_i}.

Step 4: If $w_i > p_0$, then system failure is considered as failure due to fatal shock. Step 2 and 3 are repeated with the difference that $(m_i - 1)$ interarrival times having exponential

distribution with parameter $(p_0\lambda_0)$ are generated, adding all these, t_{i,m_i-1} is obtained. One inter-arrival time $(t_{i,m_i} - t_{i,m_i-1})$ is generated having $exp((1-p_0)\,\lambda_0)$ distribution.

Steps 1 to 4 are repeated for $r = 25, 30, 40, 50$.

The statistics $x..,\ y..,\ r_1,\ m,\ t..,\ t'$ are computed, using which MLE $\widehat{S}_1(t)$ *of* $S_1(t)$ are obtained at given mission time. Also using the considered set of parameter combinations $S_1(t)$ is also obtained for same mission times. The discrepancy between theoretical $S_1(t)$ and estimated $\widehat{S}_1(t)$ is studied through bias for three sets of parameter combinations and are presented in Table 1, Table 2 and Table 3.

Table 1: Survival probabilities and bias for $\mathbf{p = 0.7},\ \lambda = 0.3,\ \mathbf{u = 0.95},\ \theta = 0.8$

		Absolute Bias			
t	$S_1(t)$	r=25	r=30	r=40	r=50
0.5	0.886703	0.024142	0.009437	0.008541	0.002917
0.75	0.834963	0.034296	0.013352	0.01208	0.004119
1	0.786242	0.043307	0.016791	0.015188	0.005171
1.25	0.740364	0.051269	0.019796	0.017901	0.006085
1.5	0.697163	0.058268	0.022405	0.020256	0.006874
1.75	0.656483	0.064384	0.024654	0.022283	0.00755
2	0.618176	0.06969	0.026574	0.024013	0.008124

Table 2: Survival probabilities and bias for $\mathbf{p = 0.4},\ \lambda = 0.4,\ \mathbf{u = 1.75},\ \theta = 0.7$

		Absolute Bias			
t	$S_1(t)$	r=25	r=30	r=40	r=50
0.5	0.849827	0.047447	0.042041	0.005374	0.003764
0.75	0.783422	0.066517	0.058848	0.007443	0.005211
1	0.722206	0.082895	0.073223	0.009163	0.006412
1.25	0.665773	0.096855	0.085421	0.010575	0.007397
1.5	0.61375	0.108646	0.095668	0.011717	0.008192
1.75	0.565792	0.118495	0.104174	0.012622	0.00882
2	0.521581	0.126606	0.111126	0.013319	0.009303
2	0.52709	0.091367	0.047135	0.04284	0.007813

Table 3: Survival probabilities and bias for $\mathbf{p = 0.65},\ \lambda = 0.5,\ \mathbf{u = 1.2},\ \theta = 1.1$

		Absolute Bias			
t	$S_1(t)$	r=25	r=30	r=40	r=50
0.5	0.849827	0.047447	0.042041	0.005374	0.003764
0.75	0.783422	0.066517	0.058848	0.007443	0.005211
1	0.722206	0.082895	0.073223	0.009163	0.006412
1.25	0.665773	0.096855	0.085421	0.010575	0.007397
1.5	0.61375	0.108646	0.095668	0.011717	0.008192
1.75	0.565792	0.118495	0.104174	0.012622	0.00882
2	0.521581	0.126606	0.111126	0.013319	0.009303

5. A random threshold case

Assuming that the thresholds of two components of the series system are independent *r.v's* having exponential distribution with parameter $\sigma,\ \sigma > 0$; and with other modelling features same as in Section 1, the reliability of the system at mission time 't' is given by

$$S_2(t) = e^{-\lambda t\left[1-p\left(\frac{\theta}{\theta+\sigma}\right)\right]^2} \tag{12}$$

In order to assess $S_2(t)$, considering the life testing experiment of 'r' systems with life distribution $(1 - S_2(t))$ and following on the lines of Section 2, the joint distribution of the random variables $m_i, t_{i1},\ t_{i2}, \ldots, t_{im_i}, X_{i1}, \ldots, X_{im_i-1}, Y_{i1}, Y_{i2}, \ldots, Y_{im_i-1}, Y_{im_i},\ u_{i,1},\ u_{i,2}$ for all 'r' systems is given by

$$L_2 = p^{m-r_2}\lambda^m e^{-p\lambda t..}\theta^{2m-r_1-2r_2}e^{-\theta(x..+y..)}(1-p)^{r_2}e^{-\lambda t'}\left(\frac{\sigma}{\sigma+\theta}\right)^{r_1}\sigma^{2r}e^{-\sigma u} \tag{13}$$

where, $t..,\ t',\ x..,\ y..,\ m$ are as defined in (6) with $u. = \sum_{i=1}^{r_1} u_{i,1} + \sum_{i=1}^{r_2} u_{i,2}$ Using L_2, the MLE's of $p,\ \lambda,\ \theta,\ \sigma$ are obtained as

$$\hat{p} = \frac{[m-(r-r_1)]\,t'}{mt' + (r-r_1)\,t..} \tag{14}$$

$$\hat{\lambda} = \frac{mt' + (r-r_1)\,t..}{t'(t.. + t')} \tag{15}$$

and $\hat{\sigma}$ and $\hat{\theta}$ are obtained numerically using Newton-Raphson method by solving the equations given below

$$(\sigma+\theta)\,(x.. + y..) + r_1 = 0 \tag{16}$$

$$\frac{1}{\sigma}\left(2r_1 + 2r_2 + \frac{r_1\theta}{\sigma+\theta}\right) - u. = 0 \tag{17}$$

Using invariance property of MLE, MLE $\widehat{S}_2(t)\ of\ S_2(t)$ is obtained as

$$\widehat{S}_2(t) = e^{-\hat{\lambda}t\left[1-\hat{p}\left(\frac{\hat{\theta}}{\hat{\theta}+\hat{\sigma}}\right)\right]^2} \tag{18}$$

$\widehat{S}_2(t)$ is computed using Monte-Carlo simulation procedure. For the i^{th} system, for generation of random variables m_i,t_{i1},t_{i2},$\ldots$,t_{im_i},X_{i1},$\ldots$,X_{im_i-1},Y_{i1},Y_{i2},$\ldots$,Y_{im_i} and computation of $\widehat{S}_2(t)$, Section 4 is referred. The random thresholds U_{i1}, U_{i2} are generated from exponential distribution with parameter $\sigma = \sigma_0$ and results are presented in Table 4, Table 5 and Table 6.

From above tables, it is evident that for both the sets of parameter combinations under the two cases of fixed and random thresholds of components, bias of estimators decreases as the number of systems on test (r) increases.

Table 4: Survival probabilities and bias for $\mathbf{p = 0.7},\ \lambda = 0.3,\ \sigma = 0.7,\ \theta = 0.8$

		Absolute Bias			
t	$S_2(t)$	$r = 25$	$r = 30$	$r = 40$	$r = 50$
0.5	0.886802	0.072797	0.069117	0.06289	0.061482
0.75	0.835103	0.104912	0.099509	0.090392	0.088335
1	0.786418	0.134412	0.127362	0.115497	0.112826
1.25	0.740571	0.161466	0.152842	0.138364	0.135112
1.5	0.697397	0.186231	0.176103	0.159144	0.155342
1.75	0.65674	0.208854	0.19729	0.177978	0.173656
2	0.618454	0.229475	0.216541	0.194998	0.190186

Table 5: Survival probabilities and bias for $\mathbf{p = 0.4},\ \lambda = 0.4,\ \sigma = 0.3,\ \theta = 0.7$

		Absolute Bias			
t	$S_2(t)$	$r = 25$	$r = 30$	$r = 40$	$r = 50$
0.5	0.851462	0.054733	0.045812	0.040406	0.035388
0.75	0.785685	0.076962	0.064254	0.056585	0.049486
1	0.724988	0.096202	0.080112	0.070441	0.061515
1.25	0.668981	0.112745	0.093647	0.082213	0.07169
1.5	0.6173	0.126859	0.105096	0.092118	0.08021
1.75	0.569612	0.138785	0.114674	0.100354	0.087252
2	0.525608	0.148745	0.122579	0.107099	0.092979

Table 6: Survival probabilities and bias for $\mathbf{p = 0.65},\ \lambda = 0.5,\ \sigma = 0.4,\ \theta = 1.1$

		Absolute Bias			
t	$S_2(t)$	$r = 25$	$r = 30$	$r = 40$	$r = 50$
0.5	0.849922	0.097285	0.087879	0.069802	0.029125
0.75	0.783553	0.138312	0.124615	0.098482	0.04062
1	0.722367	0.174835	0.157104	0.123524	0.050357
1.25	0.665959	0.207239	0.185722	0.14527	0.058528
1.5	0.613955	0.23588	0.210813	0.164031	0.065305
1.75	0.566013	0.261086	0.232694	0.180093	0.070845
2	0.521814	0.283157	0.251655	0.193718	0.075288

6. Comparison, results analysis and conclusion

The estimators for two models of series systems with components having fixed threshold and random threshold are compared by computing the mean square errors of $\widehat{S}_i(t),\ \ i = 1, 2$ using

$$MSE_i\left(\hat{S}_i(t)\right) = \tfrac{1}{M}\sum_{j=1}^{M}\left(S_i(t) - \hat{S}_{ij}(t)\right)^2;\quad i = 1, 2 \quad \text{for} \quad m = 10000$$

The relative efficiencies of $\widehat{S}_2(t)$ as compared $\widehat{S}_1(t)$ are obtained as the ratio of MSE $(\widehat{S}_1(t))$ to MSE $(\widehat{S}_2(t))$ and are presented in Table 7.

From Table 7, it is clear that the estimators of the series system with fixed threshold are more efficient as compared to estimators of series system with random threshold. Hence,

Table 7: Relative efficiency of $\widehat{\mathbf{S}}_{\mathbf{2}}(\mathbf{t})$ as compared to $\widehat{\mathbf{S}}_{\mathbf{1}}(\mathbf{t})$

	P=0.7, λ=0.3, u=0.95, σ=0.7, θ=0.8			P=0.4, λ=0.4, u=1.75, σ=0.3, θ=0.7			P=0.65, λ=0.5, u=1.2, σ=0.4, θ=1.1		
t	$S_1(t)$	$S_2(t)$	Efficiency	$S_1(t)$	$S_2(t)$	Efficiency	$S_1(t)$	$S_2(t)$	Efficiency
0.5	0.8848	0.8868	0.0423	0.8521	0.8515	0.06055	0.84985	0.84995	0.1773
0.75	0.8323	0.8351	0.0411	0.7865	0.7857	0.0593	0.7834	0.7836	0.1732
1	0.7830	0.7864	0.0400	0.7260	0.7250	0.0581	0.7222	0.7224	0.1692
1.25	0.7365	0.7406	0.0389	0.6702	0.6690	0.0570	0.6658	0.6660	0.1652
1.5	0.6928	0.6974	0.0378	0.6186	0.6173	0.0558	0.6138	0.6140	0.1612
1.75	0.6517	0.6567	0.0368	0.5710	0.5696	0.0547	0.5658	0.5660	0.1573
2	0.6130	0.6185	0.0357	0.5271	0.5256	0.0536	0.5216	0.5218	0.1534

the study is suggestive of series system with components having fixed threshold, which results in gain in reliability of series system. This is because when the thresholds are *r.v.'s* and if one of the component's thresholds turns out to be too small, then system will be less reliable. Instead, maintaining the threshold of weakest component at certain level (optimum) would be the wise criteria to enhance system reliability.

Acknowledgements

The second author is thankful to Karnatak University, Dharwad for financial support.

References

A-hameed, M. and Proschan, F. (1973). Nonstationary shock models. *Stochastic Processes and Their Applications*, **1**, 383–404.

A-Hameed, M. and Proschan, F. (1975). Shock models with underlying birth process. *Journal of Applied Probability*, **12**, 18–28.

Anderson, K. K. (1987). Limit theorems for general shock models with infinite mean inter-shock times. *Journal of Applied Probability*, **24**, 449–456.

Chikkagoudar, M. and Palaniappan, K. (1981). Uniformly minimum variance unbiased estimation of reliability in shock models. *Journal of the Indian Statistical Association*, **19**, 9–13.

Chung, W. K. (1995). Reliability of imperfect switching of cold stanby systems with multiple non-critical and critical errors. *Microelectronics Reliability*, **35**, 1479–1482.

Esary, J. and Marshall, A. (1973). Shock models and wear processes. *The Annals of Probability*, **1**, 627–649.

Fujii, S. and Sandoh, H. (1984). Bayes reliability assessment of a 2-unit hot-standby redundant system. *IEEE Transactions on Reliability*, **33**, 297–300.

Gut, A. (1990). Cumulative shock models. *Advances in Applied Probability*, **22**, 504–507.

Hanagal, D. D. (1996). Estimation of system reliability from stress-strength relationship. *Communications in Statistics-Theory and Methods*, **25**, 1783–1797.

Klefsjö, B. (1981). Survival under the pure birth shock model. *Journal of Applied Probability*, **18**, 554–560.

Kunchur, S. and Munoli, S. B. (1993). Estimation of reliability in a shock model for a two component system. *Statistics & Probability Letters*, **17**, 35–38.

Lindskog, F. and McNeil, A. J. (2003). Common poisson shock models: applications to insurance and credit risk modelling. *ASTIN Bulletin: The Journal of the IAA*, **33**, 209–238.

Mallor, F. and Santos, J. (2003). Classification of shock models in system reliability. *Monografıas del Semin. Matem. Garcıa de Galdeano*, **27**, 405–412.

Munoli, S. and Bhat, S. V. (2011). Reliability estimation in shock model when successive shocks cause greater damage (non-accumulating damages). *Research Journal of Mathematics and Statistics*, **3**, 61–66.

Munoli, S. and Mutkekar, R. R. (2011a). Estimation of reliability for a two component survival stress-strength model. *Journal of Quality and Reliability Engineering*, **2011**, 1–8.

Munoli, S. and Mutkekar, R. R. (2011b). Estimation of reliability in a non-accumulating damage shock model for a two-out-of-three component system. *Proceedings of International congress in Productivity, Quality, Reliability, Optimization and Modelling*, **01**, 57–64.

Munoli, S. and Suranagi, M. (2007). Estimation of reliability for (k+ 1) component lightly loaded standby system with single repair. *Communications in Statistics-Theory and Methods*, **36**, 193–202.

Munoli, S. and Suranagi, M. (2009). Estimation of reliability in non-accumulating damage shock model for two-component (non iid) parallel system. *Research Journal of Mathematics and Statistics*, **1**, 23–26.

Munoli, S. B. and Suhas (2019). Modelling and assessment of survival probability of shock model with two kinds of shocks. *Open Journal of Statistics*, **9**, 484–493.

Nakagawa, Y. and Rosenfeld, A. (1979). Some experiments on variable thresholding. *Pattern recognition*, **11**, 191–204.

Necsulescu, D. S. and Krausz, A. S. (1986). A multi-step stress-strength model of a parallel system. *IEEE Transactions on Reliability*, **35**, 119–123.

Posner, M. and Zuckerman, D. (1986). Semi-markov shock models with additive damage. *Advances in Applied Probability*, **18**, 772–790.

Råde, L. (1976). Reliability systems in random environment. *Journal of Applied Probability*, **13**, 407–410.

Ross, S. M. (1981). Generalized poisson shock models. *The Annals of Probability*, **9**, 896–898.

Rutemiller, H. C. (1966). Point estimation of reliability of a system comprised of k elements from the same exponential distribution. *Journal of the American Statistical Association*, **61**, 1029–1032.

Savits, T. H. (1988). Some multivariate distributions derived from a non-fatal shock model. *Journal of Applied Probability*, **25**, 383–390.

Shanthikumar, J. G. and Sumita, U. (1983). General shock models associated with correlated renewal sequences. *Journal of Applied Probability*, **20**, 600–614.

Skoulakis, G. (2000). A general shock model for a reliability system. *Journal of Applied Probability*, **37**, 925–935.

Wani, J. and Kabe, D. (1971). Point estimation of reliability of a system comprised of k elements from the same gamma model. *Technometrics*, **13**, 859–864.

Weier, D. (1981). Bayes estimation for a bivariate survival model based on exponential distributions. *Communications in Statistics-Theory and Methods*, **10**, 1415–1427.

Zacks, S. and Even, M. (1966). Minimum variance unbiased and maximum likelihood estimators of reliability functions for systems in series and in parallel. *Journal of the American Statistical Association*, **61**, 1052–1062.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 22, No. 2, 2024 (New Series), pp 145–167
http://www.ssca.org.in/journal

A Modified Measurement Error Model for Replicated Method Comparison Data with Skewness and Heavy Tails

Jeevana Duwarahan[1,2] **and Lakshika S. Nawarathna**[3]
[1]*Postgraduate Institute of Science, University of Peradeniya, Peradeniya, Sri Lanka*
[2]*Department of Mathematics and Statistics, University of Jaffna, Jaffna, Sri Lanka*
[3]*Department of Statistics and Computer Science*
Faculty of Science, University of Peradeniya, Peradeniya, Sri Lanka

Received: 24 January 2023; Revised: 21 July 2023; Accepted: 06 December 2023

Abstract

Measurement error models (MEMs) provide a flexible framework to model the method comparison data by incorporating measurement errors. However, these models often rely on normality assumptions, which are frequently violated in practice due to skewness and heavy tails. Furthermore, repeated data with measurement errors (MEs) are often observed in medical research, epidemiological studies, economics, and the environment. Thus, this research aims to assess the extent of similarity and agreement between the two methods using the replicated measurement error model (RMEM) under asymmetric and heavy-tailed distributions with a matching degree for true covariate and errors. The expectation-maximization (EM) approach is applied to fit the model. A simulation study is used to test the proposed methodology, demonstrated by evaluating subcutaneous fat data. The Total Deviation Index (TDI) and Concordance Correlation Coefficient (CCC) were used to further assess the agreement between the methods. Our suggested model works well for analyzing replicated method comparison data with measurement errors, skewness, and heavy tails.

Key words: Agreement; Heavy-tailed distributions; Replicated measurement error model; EM algorithm; Concordance correlation; Total deviation index.

1. Introduction

Method comparison study refers to comparing two or more methods that analyze the outcome for understanding the agreement between the methods. Generally, a comparison is made between the already established methods and the new methods to see whether there is enough agreement between them. If the method comparison study of the continuous variables is agreeable to each other or similar, it reflects that both methods can be interchangeably used. With the vast development in the field of health and sciences, method comparisons play a vital role in determining the better of the existing practices and new innovative methods that are put into use. The methods include an assay, equipment, medical device,

Corresponding Author: Lakshika S. Nawarathna
Email: lakshikas@sci.pdn.ac.lk

observation, measurement techniques, and variables of interest such as blood pressure, pulse rate, level of cholesterol, the level of concentration of the chemical used, *etc.* Currently, we have new techniques evolving in the health sector as a result of advancement, and these new techniques might be much more effective, less invasive, economical, faster, and easy to handle. However, the medical practitioner needs analysis of these more recent methods that are to be compared with the already existing methods or standards to understand the outcome.

Numerous method comparison studies are being conducted in the field of health and science to evaluate the techniques used. Comparing the measurements of continuous variables helps us determine the better of the prevailing methods or if they can be used interchangeably. In method comparison studies, every subject has at least one measurement from each method. Our focus in this research is where measurements are replicated. The first step in the methodology is to model the method comparison data where the mixed-effects model is commonly used. It is to be noted that the model assumes independent normal distribution for both random effects and errors, and the measurement variability is constant over the whole measurement range (Bland and Altman, 1999, 2007; Carrasco and Jover, 2003; Carstensen *et al.*, 2008; Carrasco *et al.*, 2009; Hedayat *et al.*, 2009; Choudhary, 2008). Secondly, the evaluation of agreement between the methods is conducted using inference on one or more measures of agreements that quantify how much they agree well. When the difference in measurements is small, it reflects a good agreement between the two methods. There are numerous agreement measures, including the CCC (Lin, 1989; Barnhart *et al.*, 2007; Nawarathna and Choudhary, 2013, 2015) and the TDI (Lin, 1989; Nawarathna and Choudhary, 2013, 2015; Choudhary, 2009; Choudhary and Yin, 2010) have attracted the greatest attention in the statistical literature.

In many real-world situations, accurately measuring the true value of a variable is challenging. Instead, we can only observe it with some degree of error. This discrepancy between the observed and true values is known as "Measurement Error (ME)". Imagine trying to hit a target with a bow and arrow. The true bullseye represents the actual value we aim to measure, while the observed values are scattered around it due to measurement error. These errors can arise from various factors, such as different measurement methods, instruments, human error, or external influences. Ignoring these errors can lead to biased estimates and increased variability in statistical inferences. Therefore, it is essential to consider measurement errors to ensure accurate and reliable statistical analysis.

MEMs, which have been discussed in Nawarathna and Choudhary (2015), Dunn and Roberts (1999), Alanen (2010), typically assume normality for both the true covariate and error terms. However, in practice, the method comparison data often reflects skewness and heavy tails, indicating departures from normality. This is exemplified by the subcutaneous fat data discussed in Carstensen *et al.* (2020), demonstrating these characteristics. While data transformations can be used to achieve normality, limiting transformations to (natural) logarithmic transformations in method comparison studies is generally advised, as Bland and Altman (1999) recommend. However, the log transformation may not always be successful. In such cases, alternative approaches should be considered. Nonparametric methods, as suggested by King and Chinchilli (2001), King *et al.* (2007), and Choudhary (2010), do not rely on distributional assumptions. Generalized Estimating Equations (GEE), discussed by Barnhart *et al.* (2002, 2005), and Lin *et al.* (2007), offer a semiparametric approach

by directly modeling the moments of the data without assuming a specific distribution. Additionally, parametric models can be utilized based on distributions other than the normal distribution, as explored by Sengupta *et al.* (2015). These alternative approaches provide flexibility in modeling method comparison data, accounting for its specific characteristics beyond the assumptions of normality.

The parametric mixed-effects model approach developed by Sengupta *et al.* (2015) offers a methodology for analyzing method comparison data with skewness and heavy tails. In the context of MEM, Duwarahan and Nawarathna (2022) modified the STcT-MEM (Tomaya and de Castro, 2018) specifically for unreplicated method comparison data with known error variances. However, no MEM model is designed for replicated method comparison data. Inspired by this gap, we aim to modify a model within the MEM framework to analyze replicated method comparison data with skewed and heavy-tailed features. In our approach, building upon the work of Cao *et al.* (2017), we consider MEMs for replicated data under scale mixtures of skew-normal (SMSN) distributions for the true covariate and scale mixtures of normal (SMN) distributions for the error terms with the matching degree. Specifically, we use the skew-t (ST) distribution for the true covariate and the t distribution for the error term. We also consider the skew-normal (SN) and normal (N) distributions for comparative purposes. The primary objective of this paper is to modify the above model to analyze method comparison data, assess the agreement between the two methods, and determine if they can be used interchangeably.

Additionally, our proposed methodology provides a unified framework that can handle various types of data, including normally distributed, skewed, and heavy-tailed data. It encompasses the N-RMEM (normal-distributed replicated measurement error model) and SN-RMEM (skew-normal-distributed replicated measurement error model) as special cases. Specifically, when the degrees of freedom reach infinity, the SN-RMEM turns into a special case of the ST-RMEM (skew-t-distributed replicated measurement error model). Similarly, when the degrees of freedom tends to infinity and the skewness parameter is zero, the N-RMEM becomes a special case of the ST-RMEM. This flexibility allows for comprehensive analysis and comparison of different types of method comparison data.

The remainder of the paper is organized as follows. Section 2 introduces the ST-RMEM for method comparison data. Section 3 discusses the proposed methodology for evaluating similarity and agreement under ST-RMEM. Section 4 investigates the proposed model's performance using simulated studies. Section 5 illustrates our method using subcutaneous fat data, and the concluding section summarizes the findings and conclusions. The statistical program R (R Core Team, 2021) was used to perform all of the computations given in this research.

2. Framework for method comparison data

This section describes a framework for analyzing research that compares two methods when taking several measurements on each subject. Let $Y_{ijk}, k = 1, 2, \ldots, n_j, j = 1, 2, i = 1, 2, \ldots, m$ denote the k^{th} replicate measurement of the j^{th} method on the i^{th} subject. Here m is the number of subjects in the study, and n_j is the number of measurements on method j. It is to be noted that $n_j \geq 2$. Here Method 1 is the reference method, while Method 2 is the test method. Let $n = n_1 + n_2$ represent the total number of measurements taken on the

subject and $N = nm$ represent the total number of measurements in the dataset.

If multiple measurements are found on each subject, it is referred to as 'repeated measurements data' and categorized as unlinked, linked, and longitudinal data. These categories are essential as it influences the way the data are modeled. In this research, we focus on unlinked data. Unlinked data refers to the measurements obtained from the two methods separately, and multiple measurements on a subject taken by a method are independent replications of the same underlying measurement. In this case, it is not mandatory for the methods to have the same number of replications on a subject. As always, measurements from various subjects are presumed to be independent.

Let $N_p(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $SN_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda})$, $t_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \nu)$, and $ST_p(\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{\lambda}, \nu)$ indicate the p dimensional N, SN, t, and ST distributions, respectively. Here, $\boldsymbol{\mu} \in \mathbb{R}^p$ is a location vector, $\boldsymbol{\Sigma}$ is a $p \times p$ positive definite scale matrix, $\boldsymbol{\lambda} \in \mathbb{R}^p$ is a vector of skewness parameters, and $\nu(> 0)$ is degrees of freedom. Let $G(\alpha, \beta)$ represent the gamma distribution with parameters $\alpha(> 0)$ and $\beta(> 0)$, and $HN(0, \sigma^2)$ represent the half-normal $(0, \sigma^2)$ distribution. Let $\mathbf{I}_p$ denote a $p \times p$ identity matrix. The symbol $\boldsymbol{\Sigma}^{1/2}$ represents a square root of the symmetric and positive definite matrix $\boldsymbol{\Sigma}$. This implies that $\boldsymbol{\Sigma}^{1/2}(\boldsymbol{\Sigma}^{1/2})^T = \boldsymbol{\Sigma}$, where the symbol T denotes transposition. The inverse of $\boldsymbol{\Sigma}$ is denoted as $\boldsymbol{\Sigma}^{-1}$.

2.1. Definition of ST-RMEM

The classical replicated measurement error model is

$$\begin{aligned} X_{ik} &= b_i + \delta_{ik}; \quad k = 1, 2, \ldots, p \text{ and} \\ Y_{il} &= y_i + \epsilon_{ik}; \quad l = 1, 2, \ldots, q \\ y_i &= \alpha + \beta b_i + e_i; \quad i = 1, 2, \ldots, m \end{aligned} \tag{1}$$

where b_i, y_i be the unobserved true covariate and response, and they are observed p and q times, respectively; α is the fixed bias; slope β is its proportional bias; δ_{ik}, ϵ_{ik} are measurement errors of X_{ik} and Y_{il}, respectively; e_i is the equation error, which indicates that the true variables b_i and y_i are not completely connected if other factors other than b_i are also involved in the variation in y_i, and δ_{ik}, ϵ_{ik}, e_i are uncorrelated with each other. Moreover, e_i is known as 'method-subject interaction' in a mixed-effects model. When a measurement error model is used, it may be noted that they are frequently incorporated in the testing method but not in the standard method. However, when a mixed-effects model is used, they are always included in both methods. A slope β with a non-unit value suggests a difference in the proportionate biases (or scales) of the methods.

Consider a $(p+q)$ dimensional random vector $\boldsymbol{Z}_i = (\boldsymbol{X}_i^T, \boldsymbol{Y}_i^T)^T$, where $\boldsymbol{X}_i = (X_{i1}, X_{i2}, \ldots, X_{ip})^T$ is a p dimensional random vector and $\boldsymbol{Y}_i = (Y_{i1}, Y_{i2}, \ldots, Y_{iq})^T$ is a q dimensional random vector. From (1), the model can be written as

$$\boldsymbol{Z}_i = \boldsymbol{A} + \boldsymbol{B} b_i + \boldsymbol{\psi}_i \tag{2}$$

where $\mathrm{A} = \begin{bmatrix} \mathbf{0}_p \\ \alpha \mathbf{1}_q \end{bmatrix}$, $\mathrm{B} = \begin{bmatrix} \mathbf{1}_p \\ \beta \mathbf{1}_q \end{bmatrix}$, $\boldsymbol{\psi}_i = \begin{bmatrix} \boldsymbol{\delta}_i \\ e_i \mathbf{1}_q + \boldsymbol{\epsilon}_i \end{bmatrix}$ with $\boldsymbol{\delta}_i = (\delta_{i1}, \ldots, \delta_{ip})^T$, $\boldsymbol{\epsilon}_i = (\epsilon_{i1}, \ldots, \epsilon_{iq})^T$. It is standard to assume that b_i, e_i, δ_{ik} and ϵ_{ik} are independent and

$$b_i \sim N_1(\mu_b, \phi_b), \quad e_i \sim N_1(0, \phi_e), \quad \delta_{ik} \sim N_1(0, \phi_\delta), \quad \text{and} \quad \epsilon_{ik} \sim N_1(0, \phi_\epsilon). \tag{3}$$

Normality assumption is sometimes unfeasible due to the skewness, heavy-tailed ness, and outliers. To overcome this problem, Cao *et al.* (2017) developed the ST-RMEM by considering ST for true covariate and t distribution for error terms with the same degrees of freedom. It follows

$$b_i \sim ST_1(\mu_b, \phi_b, \lambda_b, \nu), \quad e_i \sim t_1(0, \phi_e, \nu), \quad \delta_{ik} \sim t_1(0, \phi_\delta, \nu), \quad \text{and} \quad \epsilon_{ik} \sim t_1(0, \phi_\epsilon, \nu). \tag{4}$$

It can be hierarchically represented as

$$\begin{aligned} \boldsymbol{Z}_i \mid b_i, U_i = u_i &\sim N_n(\boldsymbol{A} + \boldsymbol{B}b_i, \boldsymbol{\Sigma}_1/U_i), \\ b_i \mid U_i = u_i, V_i = v_i &\sim N_1(\mu_b + \gamma_b v_i, \tau_b/U_i), \\ V_i \mid U_i = u_i &\sim HN(0, 1/U_i), \\ U_i &\sim G(\frac{\nu}{2}, \frac{\nu}{2}). \end{aligned} \tag{5}$$

where $n = p + q$, $\gamma_b = \phi_b^{1/2}\delta_b$, $\delta_b = \frac{\lambda_b}{\sqrt{1+\lambda_b^2}}$, $\tau_b = \phi_b(1-\delta_b^2)$, $\boldsymbol{\Sigma}_1 = \begin{bmatrix} \phi_\delta \mathbf{I}_p & \mathbf{0}_{p\times q} \\ \mathbf{0}_{q\times p} & \phi_e \mathbf{1}_q \mathbf{1}_q^T + \phi_\varepsilon \mathbf{I}_q \end{bmatrix}$.

The mean vector and variance matrix of $\boldsymbol{Z}_i$ are as follows.

$$\begin{aligned} E(\boldsymbol{Z}_i) &= \boldsymbol{A} + \boldsymbol{B}E(b_i), \quad \nu > 1 \quad \text{and} \\ Var(\boldsymbol{Z}_i) &= \frac{\nu}{\nu-2}\phi_b \boldsymbol{B}\boldsymbol{B}^T - \zeta^2 \boldsymbol{B}\gamma_b\gamma_b^T\boldsymbol{B}^T + \frac{\nu}{\nu-2}\boldsymbol{\Sigma}_1, \quad \nu > 2 \end{aligned} \tag{6}$$

where $E(b_i) = \mu_b + \zeta\gamma_b$, with $\zeta = \sqrt{\frac{\nu}{\pi}}\frac{\Gamma(\frac{\nu-1}{2})}{\Gamma(\frac{\nu}{2})}$ and $\Gamma(\cdot)$ denotes the gamma function, and its conditional distribution is expressed as

$$\boldsymbol{Z}_i \mid U_i \sim SN_n(\boldsymbol{A} + \boldsymbol{B}\mu_b, \boldsymbol{\Sigma}/U_i, \boldsymbol{\lambda}), \tag{7}$$

where $\boldsymbol{\Sigma} = \phi_b \boldsymbol{B}\boldsymbol{B}^T + \boldsymbol{\Sigma}_1$ and $\boldsymbol{\lambda} = \frac{\lambda_b \phi_b \boldsymbol{\Sigma}^{-1/2}\boldsymbol{B}}{\sqrt{\phi_b + \lambda_b^2 \Lambda_b}}$ with $\Lambda_b = \frac{\phi_b}{c}$, $c = 1 + \phi_b \boldsymbol{B}^T\boldsymbol{\Sigma}_1^{-1}\boldsymbol{B}$.
Cao *et al.* (2017) used an EM algorithm to estimate the parameters due to the complexity of the likelihood function.

2.2. ST-RMEM for method comparison data

Let $\boldsymbol{Y}_{ij} = (Y_{ij1}, Y_{ij2}, \ldots, Y_{ijn_j})^T$ denote the n_j measurement vector from method $j(= 1, 2)$. The vector $\boldsymbol{Y}_i = (\boldsymbol{Y}_{i1}^T, \boldsymbol{Y}_{i2}^T)^T$ denote all measurements on subject i. Let $\tilde{\boldsymbol{Y}} = (\tilde{Y}_1, \tilde{Y}_2)^T$ represent paired observations from the two methods on a randomly chosen subject from the population. The basic ST-RMEM can now be used flexibly to model replicated method comparison data. This model implies that Method 1 is a well-known method used as a reference method in the comparison. It is of the form

$$\begin{aligned} &Y_{i1k} = b_i + \delta_{i1k}; \\ &Y_{i2k} = \alpha + \beta b_i + e_i + \epsilon_{i2k}; \quad i = 1, 2, \ldots, m \quad k = 1, 2, \ldots, n_j. \end{aligned} \tag{8}$$

where α and β are the fixed bias and proportional bias of method 2, respectively, b_i denotes the true unobservable measurement for the i^{th} subject, e_i is the equation error, and δ_{i1k}, ϵ_{i2k} are random errors. Both fixed and proportional biases result in systematic measuring

mistakes. However, Method 1 does not assume a fixed or proportionate bias for identifiability reasons.

This model can be expressed in the matrix notation of (2) by setting $\boldsymbol{Z}_i = \boldsymbol{Y}_i, \boldsymbol{\psi}_i = \begin{bmatrix} \boldsymbol{\delta}_{i1} \\ e_i \mathbf{1}_{n_2} + \boldsymbol{\epsilon}_{i2} \end{bmatrix}$ where $\boldsymbol{\delta}_{i1} = (\delta_{i11}, \ldots, \delta_{i1n_1})^T$, $\boldsymbol{\epsilon}_{i2} = (\epsilon_{i21}, \ldots, \epsilon_{i2n_2})^T$, and $(p, q) = (n_1, n_2)$. It further assumes that

$$b_i \sim ST_1(\mu_b, \phi_b, \lambda_b, \nu), \quad e_i \sim t_1(0, \phi_e, \nu), \quad \delta_{i1k} \sim t_1(0, \phi_\delta, \nu), \quad \text{and} \quad \epsilon_{i2k} \sim t_1(0, \phi_\epsilon, \nu). \tag{9}$$

where b_i, e_i, δ_{i1k}, and ϵ_{i2k} are mutually independent. This model is a modification of the ST-RMEM, previously mentioned in this section. It can be considered when the data shows skewness and heavy-tailedness in method comparison. It can handle two or more measuring methods, replicated and un-replicated measurements, as well as balanced and unbalanced designs. In the un-replicated case ($i.e., n_j = 1$) there is no need to include the equation error term. The unknown parameter vector of the model (9) is denoted by $\boldsymbol{\theta} = (\alpha, \beta, \mu_b, \phi_b, \lambda_b, \phi_e, \phi_\delta, \phi_\varepsilon)^T$, and we use the EM algorithm to obtain the maximum likelihood estimates (MLEs) of these parameters. The SN-RMEM gets to be a special case of the ST-RMEM (9) when $\nu \to \infty$.

$$b_i \sim SN_1(\mu_b, \phi_b, \lambda_b), \quad e_i \sim N_1(0, \phi_e), \quad \delta_{i1k} \sim N_1(0, \phi_\delta), \quad \text{and} \quad \epsilon_{i2k} \sim N_1(0, \phi_\epsilon) \tag{10}$$

When the skewness parameter $\lambda_b = 0$ and the degrees of freedom parameter $\nu \to \infty$, it is a standard N-RMEM.

$$b_i \sim N_1(\mu_b, \phi_b), \quad e_i \sim N_1(0, \phi_e), \quad \delta_{i1k} \sim N_1(0, \phi_\delta), \quad \text{and} \quad \epsilon_{i2k} \sim N_1(0, \phi_\epsilon) \tag{11}$$

3. Assessment of similarity and agreement

3.1. Similarity measures

A method comparison study aims to assess the similarity of measuring methods and their agreement. This evaluation is performed by drawing conclusions based on similarity and agreement measures, which are functions of the model parameters. Evaluation of similarity is a comparison of characteristics, including biases, precisions, and scales of the methods, to find out how the methods differ. In the case of the model (8), the similarity is assessed by analyzing biases with intercept (α) and slope (β). The scales of the methods are the same if the slope is 1. In addition, the true values of the methods are also the same if the intercept is zero. Method precisions can be determined using the ratio, denoted as $\lambda = \frac{\text{error variance of Method 1}}{\text{error variance of Method 2}}$. If $\lambda = 1$, methods 1 and 2 are equally accurate, but if $\lambda < 1$, Method 1 is more accurate than Method 2, and if $\lambda > 1$, Method 2 is more accurate than Method 1. However, this necessitates that these methods be on the same scale. For example, the precisions of two thermometers measured in Fahrenheit and Celsius cannot be compared until one is converted to the same scale. Hence, the test method's scale can be adjusted to equal that of the reference method by dividing $\tilde{Y}_2$ by the slope β. As a result, the precision ratio is $\beta^2\lambda$ and is referred to as the 'squared sensitivity ratio'.

3.2. Agreement measures

The evaluation of similarity is just a comparison of the methods' marginal distributions. Evaluation of agreement is an analysis of the methods' joint distribution, including

their marginal distributions. Further, the closeness of the two methods' measurements is referred to as agreement. When their measurements are identical, the methods agree perfectly. In this ideal case, the bivariate distribution of $\tilde{Y}_1$ and $\tilde{Y}_2$ is concentrated on the 45° line; as a result, the joint distribution becomes degenerate at zero.

In practice, we use measures of an agreement to quantify the extent of the agreement. In spite of the fact that a number of agreement measures are available (Barnhart *et al.*, 2007), two among them, to be specific, the CCC and the TDI, have received the foremost consideration in the statistical literature. These are explained below.

3.2.1. Concordance correlation coefficient

This measure was introduced by Lin (1989), and it is defined as

$$CCC = \frac{2\text{cov}(\tilde{Y}_1, \tilde{Y}_2)}{[E(\tilde{Y}_1) - E(\tilde{Y}_2)]^2 + \text{Var}(\tilde{Y}_1) + \text{Var}(\tilde{Y}_2)} \tag{12}$$

It lies in $[-1, 1]$, and a high CCC score indicates good agreement. A score of 1 indicates perfect positive agreement, whereas -1 denotes excellent negative agreement. More information on this measure's properties and generalizations to various data types, and models can be found in Barnhart *et al.* (2007) and Lin *et al.* (2012).

3.2.2. Total Deviation Index

Lin (2000) introduced this measure, defined as

$$TDI(p) = 100\, p^{\text{th}}\text{percentile of} |\tilde{D} = \tilde{Y}_1 - \tilde{Y}_2| \text{ for a specified } p. \tag{13}$$

In practice, the value of p is assumed to be between 0.80 and 0.95. It is a non-negative measure, with smaller values indicating higher agreement and zero indicating perfect agreement. The confidence bounds of this measure reflect how significant a measurement difference may be in a given large fraction of the population. As a result, if all of the discrepancies in this interval are acceptable from a practical standpoint, the methods are said to be in satisfactory agreement. Lin *et al.* (2002) and Choudhary (2008) provided extensive information on this measure.

In order to evaluate the agreement between methods, we first fit a model to the method comparison data Y_{ijk}, $k = 1, \ldots, n_j$, $j = 1, 2$, $i = 1, \ldots, m$, using the maximum likelihood (ML) approach, as indicated in section (2.2). Let $\hat{\boldsymbol{\theta}}$ be the ML estimator of the parameter vector of $\boldsymbol{\theta}$. According to asymptotic theory, when n is large, the sampling distribution of $\hat{\boldsymbol{\theta}}$ approximately follows a multivariate normal distribution with mean $\boldsymbol{\theta}$ and variance $\boldsymbol{I}^{-1}$ under specific regularity constraints, where $\boldsymbol{I}$ is the observed information matrix Lehmann (1998).

Consider φ as a scalar measure of agreement between two methods. Its ML estimator $\hat{\varphi}$ is produced by substituting $\boldsymbol{\theta}$ with $\hat{\boldsymbol{\theta}}$. When φ is a differentiable function of $\boldsymbol{\theta}$, the delta method can be used to estimate the sample distribution of $\hat{\varphi}$, expressed as $\hat{\varphi} \sim N(\varphi, \boldsymbol{D}^T \boldsymbol{I}^{-1} \boldsymbol{D})$, where $\boldsymbol{D} = \frac{\partial \varphi}{\partial \boldsymbol{\theta}}$ is the Jacobian matrix evaluated at $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, and they are typically estimated numerically. The $100(1-\alpha)\%$ two-sided confidence bounds for the

agreement measure φ are $\hat{\varphi} \pm z_{1-\alpha}\text{SE}(\hat{\varphi})$, where $z_{1-\alpha}$ is the $(1-\alpha)^{\text{th}}$ percentile of $N_1(0,1)$ and $\text{SE}(\hat{\varphi}) = (\boldsymbol{D}^T\boldsymbol{I}^{-1}\boldsymbol{D})^{\frac{1}{2}}$. In specific, in case small values for φ infer good agreement (*e.g.*, TDI), at that point, require an upper bound. Though in case large values for φ infer good agreement (*e.g.*, CCC), at that point, require a lower bound. After applying a normalizing transformation, the confidence intervals are computed to make stride accuracy for parameters or parameter functions whose range does not span the entire real line. The results are rearranged back to the initial scale. Particularly, TDI is transformed using a log transformation, while CCC is transformed using Fisher's z-transformation. These confidence boundaries are then used to assess if the methods agree sufficiently.

This approach makes sense only if there is no proportionate bias in the test procedure. Thus, the test method needs to be adjusted such that its scale matches to that of the reference method before the agreement can be evaluated (Nawarathna and Choudhary, 2015; Choudhary and Nagaraja, 2017). Therefore, we first transform $\tilde{Y}_2$ as $\tilde{Y}_2^{\,*} = \tilde{Y}_2/\beta$ to make $\tilde{Y}_2$ on the same scale as $\tilde{Y}_1$. The measures of agreement in the transformed case are functions of parameters of the bivariate distribution of $(\tilde{Y}_1, \tilde{Y}_2^{\,*})$, respectively, and the equation of these agreement measures can be determined by inserting the moments from their respective bivariate distributions into their definitions. After this transformation, these measures follow from (12)-(13) that

$$CCC^* = \frac{2\text{cov}(\tilde{Y}_1, \tilde{Y}_2^{\,*})}{[E(\tilde{Y}_1) - E(\tilde{Y}_2^{\,*})]^2 + \text{Var}(\tilde{Y}_1) + \text{Var}(\tilde{Y}_2^{\,*})} \tag{14}$$

$$TDI^* = 100\,p^{\text{th}}\text{percentile of}\,|\tilde{D}^* = \tilde{Y}_1 - \tilde{Y}_2^{\,*}|\ \text{for a specified}\ p. \tag{15}$$

3.3. Agreement evaluation under different models

3.3.1. ST-RMEM

As previously mentioned, $\tilde{Y}_j$ denotes a single measurement using the j^{th} method $(j = 1, 2)$ on a randomly selected subject from the population to derive the expressions for measures of the agreement under the assumed ST-RMEM. Moreover, a companion model for $\tilde{\boldsymbol{Y}} = (\tilde{Y}_1, \tilde{Y}_2)^T$ is generated from the model (8).

$$\tilde{\boldsymbol{Y}} = \boldsymbol{A} + \boldsymbol{B}\tilde{b} + \tilde{\boldsymbol{\psi}} \tag{16}$$

where $\boldsymbol{A} = \begin{bmatrix} 0 \\ \alpha \end{bmatrix}$; $\boldsymbol{B} = \begin{bmatrix} 1 \\ \beta \end{bmatrix}$; $\tilde{\boldsymbol{\psi}} = \begin{bmatrix} \delta_1 \\ e + \epsilon_2 \end{bmatrix}$.

Further, $\tilde{b} \sim \text{ST}_1(\mu_b, \phi_b, \lambda_b, \nu)$ and $\tilde{\boldsymbol{\psi}} \sim t_2(\boldsymbol{0}, \tilde{\boldsymbol{\Sigma}}_{\boldsymbol{1}}, \nu)$ with $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{1}} = \begin{bmatrix} \phi_\delta & 0 \\ 0 & \phi_e + \phi_\varepsilon \end{bmatrix}$.

The mean vector and variance matrix of $\tilde{\boldsymbol{Y}}^* = (\tilde{Y}_1, \tilde{Y}_2^* = \tilde{Y}_2/\beta)$ are as follows from (16) that

$$\begin{aligned} E(\tilde{\boldsymbol{Y}}^*) &= \boldsymbol{A}^* + \boldsymbol{B}^* E(b_i), \quad \nu > 1 \,\text{and} \\ Var(\tilde{\boldsymbol{Y}}^*) &= \frac{\nu}{\nu-2}\phi_b \boldsymbol{B}^* \boldsymbol{B}^{*T} - \zeta^2 \boldsymbol{B}^* \gamma_b \gamma_b^T \boldsymbol{B}^{*T} + \frac{\nu}{\nu-2}\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{1}}^{\,*}, \quad \nu > 2 \end{aligned} \tag{17}$$

where $\boldsymbol{A}^* = \begin{bmatrix} 0 \\ \alpha/\beta \end{bmatrix}$; $\boldsymbol{B}^* = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$; $E(b_i) = \mu_b + \zeta\gamma_b$ with $\zeta = \sqrt{\frac{\nu}{\pi}}\frac{\Gamma(\frac{\nu-1}{2})}{\Gamma(\frac{\nu}{2})}$ and $\tilde{\boldsymbol{\Sigma}}_{\boldsymbol{1}}^{\boldsymbol{*}} = \begin{bmatrix} \phi_\delta & 0 \\ 0 & \frac{1}{\beta^2}(\phi_e + \phi_\varepsilon) \end{bmatrix}$.

Further, we can write using the hierarchical representation (7),

$$\tilde{\boldsymbol{Y}} = \begin{pmatrix} \tilde{Y}_1 \\ \tilde{Y}_2 \end{pmatrix} | U \sim \mathrm{SN}_2(\boldsymbol{A} + \boldsymbol{B}\mu_b, \tilde{\boldsymbol{\Sigma}}/U, \tilde{\boldsymbol{\lambda}}) \tag{18}$$

where $\boldsymbol{A} = \begin{bmatrix} 0 \\ \alpha \end{bmatrix}; \boldsymbol{B} = \begin{bmatrix} 1 \\ \beta \end{bmatrix}; \tilde{\boldsymbol{\Sigma}} = \begin{bmatrix} \phi_b + \phi_\delta & \beta\phi_b \\ \beta\phi_b & \beta^2\phi_b + \phi_e + \phi_\epsilon \end{bmatrix}$ and $\tilde{\boldsymbol{\lambda}} = \frac{\lambda_b \phi_b \tilde{\boldsymbol{\Sigma}}^{-1/2} \boldsymbol{B}}{\sqrt{\phi_b + \lambda_b^2 \Lambda_b}}$ are counterparts of $\boldsymbol{\Sigma}$ and $\boldsymbol{\lambda}$. After the transformation, it becomes

$$\tilde{\boldsymbol{Y}}^* = \begin{pmatrix} \tilde{Y}_1 \\ \tilde{Y}_2^* \end{pmatrix} | U \sim \mathrm{SN}_2(\boldsymbol{A}^* + \boldsymbol{B}^*\mu_b, \tilde{\boldsymbol{\Sigma}}^*/U, \tilde{\boldsymbol{\lambda}}^*) \tag{19}$$

where $\tilde{\boldsymbol{\Sigma}}^* = \begin{bmatrix} \phi_b + \phi_\delta & \phi_b \\ \phi_b & \phi_b + \frac{1}{\beta^2}(\phi_e + \phi_\epsilon) \end{bmatrix}$ and $\tilde{\boldsymbol{\lambda}}^* = \frac{\lambda_b \phi_b \tilde{\boldsymbol{\Sigma}}^{*-1/2} \boldsymbol{B}^*}{\sqrt{\phi_b + \lambda_b^2 \Lambda_b}}$ with $\Lambda_b = \frac{\phi_b}{c}$,
$c = 1 + \phi_b \boldsymbol{B}^{*T} \tilde{\boldsymbol{\Sigma}}_1^{*-1} \boldsymbol{B}^*, \tilde{\boldsymbol{\Sigma}}_1^* = \begin{bmatrix} \phi_\delta & 0 \\ 0 & \frac{1}{\beta^2}(\phi_e + \phi_\epsilon) \end{bmatrix}$.

Furthermore, we know, if $Y \sim \mathrm{SN}_q(\mu_b, \phi_b, \lambda_b)$ and $\delta_b = \frac{\lambda_b}{(1+\lambda_b^2)^{\frac{1}{2}}}$, $\gamma_b = \phi_b^{\frac{1}{2}} \delta_b$, $\tau_b = \phi_b(1 - \delta_b^2)$. Then

$$\boldsymbol{m}^T Y \sim \mathrm{SN}_1 \left(\boldsymbol{m}^T \mu_b, \boldsymbol{m}^T \phi_b \boldsymbol{m}, \boldsymbol{m}^T \phi_b^{\frac{1}{2}} \delta_b / (\boldsymbol{m}^T \tau_b \boldsymbol{m})^{\frac{1}{2}} \right), \tag{20}$$

where $\boldsymbol{m} \in \mathbb{R}^q$ with at least one non-zero element (Sengupta *et al.*, 2015). It follows from (20) that the difference $\tilde{D} = \tilde{Y}_1 - \tilde{Y}_2$ is

$$\tilde{D}|U \sim \mathrm{SN}_1 \left(\boldsymbol{m}^T(\boldsymbol{A} + \boldsymbol{B}\mu_b), \boldsymbol{m}^T \tilde{\boldsymbol{\Sigma}} \boldsymbol{m}/U, \boldsymbol{m}^T \tilde{\boldsymbol{\Sigma}}^{1/2} \tilde{\boldsymbol{\delta}} / (\boldsymbol{m}^T \tilde{\boldsymbol{\Gamma}}\, \boldsymbol{m})^{1/2} \right), \tag{21}$$

where $\boldsymbol{m} = (1, -1)^T$, $\tilde{\boldsymbol{\delta}} = \tilde{\boldsymbol{\lambda}}/(1 + \tilde{\boldsymbol{\lambda}}^T \tilde{\boldsymbol{\lambda}})^{1/2}$ and $\tilde{\boldsymbol{\Gamma}} = \tilde{\boldsymbol{\Sigma}} - \tilde{\boldsymbol{\Sigma}}^{1/2} \tilde{\boldsymbol{\delta}} \tilde{\boldsymbol{\delta}}^T \tilde{\boldsymbol{\Sigma}}^{1/2}$.
When considering transformation, $\tilde{D}^* = \tilde{Y}_1 - \tilde{Y}_2^*$ is

$$\tilde{D}^*|U \sim \mathrm{SN}_1 \left(\boldsymbol{m}^T(\boldsymbol{A}^* + \boldsymbol{B}^*\mu_b), \boldsymbol{m}^T \tilde{\boldsymbol{\Sigma}}^* \boldsymbol{m}/U, \boldsymbol{m}^T \tilde{\boldsymbol{\Sigma}}^{*1/2} \tilde{\boldsymbol{\delta}}^* / (\boldsymbol{m}^T \tilde{\boldsymbol{\Gamma}}^*\, \boldsymbol{m})^{1/2} \right), \tag{22}$$

where $\tilde{\boldsymbol{\delta}}^* = \tilde{\boldsymbol{\lambda}}^*/(1 + \tilde{\boldsymbol{\lambda}}^{*T} \tilde{\boldsymbol{\lambda}}^*)^{1/2}$ and $\tilde{\boldsymbol{\Gamma}}^* = \tilde{\boldsymbol{\Sigma}}^* - \tilde{\boldsymbol{\Sigma}}^{*1/2} \tilde{\boldsymbol{\delta}}^* \tilde{\boldsymbol{\delta}}^{*T} \tilde{\boldsymbol{\Sigma}}^{*1/2}$.

We can now derive the equations for CCC and TDI under ST-RMEM (16) for transformed data. As described in (14), the CCC^* for transformed data can be computed as

$$CCC^* = \frac{2\left[\frac{\nu}{\nu-2}\phi_b - \zeta^2 \gamma_b \gamma_b^T\right]}{\left[\boldsymbol{m}^T(\boldsymbol{A}^* + \boldsymbol{B}^*\mu_b)\right]^2 + \left[\frac{\nu}{\nu-2}\phi_b - \zeta^2 \gamma_b \gamma_b^T + \frac{\nu}{\nu-2}\phi_\delta\right] + \left[\left(\frac{\nu}{\nu-2}\phi_b - \zeta^2 \gamma_b \gamma_b^T\right) + \frac{1}{\beta^2}\left(\frac{\nu}{\nu-2}\phi_\epsilon + \frac{\nu}{\nu-2}\phi_e\right)\right]} \tag{23}$$

Next, as we know from equation (15), the TDI^* is defined as the p^{th} quantile of $\tilde{D}^*$ with a given large probability of $0 < p < 1$, and it can be obtained by solving

$$TDI^* = P(|\tilde{D}^*| \leq t) = \int_0^\infty \left\{F^*(t) - F^*(-t)\right\} f(u|\nu)\, du, \quad t > 0 \tag{24}$$

where F^* is the distribution function of $\tilde{D}^*|U$ and $f(u|\nu)$ is the density of U.

3.3.2. SN-RMEM

For the model (10), the mean vector and variance matrix of $\tilde{\boldsymbol{Y}}^*$ are

$$\begin{aligned} E(\tilde{\boldsymbol{Y}}^*) &= \boldsymbol{A}^* + \boldsymbol{B}^* E(b_i), \quad \text{and} \\ Var(\tilde{\boldsymbol{Y}}^*) &= \phi_b \boldsymbol{B}^* \boldsymbol{B}^{*T} \left(1 - \frac{2\delta_b^T \delta_b}{\pi}\right) + \tilde{\boldsymbol{\Sigma}}_1^*, \end{aligned} \tag{25}$$

where $E(b_i) = \mu_b + \sqrt{\frac{2}{\pi}}\gamma_b$.

The hierarchical representation of $\tilde{\boldsymbol{Y}}$ is $\begin{pmatrix} \tilde{Y}_1 \\ \tilde{Y}_2 \end{pmatrix} \sim \mathrm{SN}_2(\boldsymbol{A} + \boldsymbol{B}\mu_b, \tilde{\boldsymbol{\Sigma}}, \tilde{\boldsymbol{\lambda}})$, and after the transformation, the marginal distribution of $(\tilde{Y}_1, \tilde{Y}_2^*)$ is $\begin{pmatrix} \tilde{Y}_1 \\ \tilde{Y}_2^* \end{pmatrix} \sim \mathrm{SN}_2(\boldsymbol{A}^* + \boldsymbol{B}^*\mu_b, \tilde{\boldsymbol{\Sigma}}^*, \tilde{\boldsymbol{\lambda}}^*)$.

Then, CCC^* can be defined as

$$CCC^* = \frac{2\phi_b \left(1 - \frac{2\delta_b^2}{\pi}\right)}{\left(\frac{\alpha}{\beta}\right)^2 + \left[\phi_b \left(1 - \frac{2\delta_b^2}{\pi}\right) + \phi_\delta\right] + \left[\phi_b \left(1 - \frac{2\delta_b^2}{\pi}\right) + \frac{1}{\beta^2}(\phi_\varepsilon + \phi_e)\right]} \tag{26}$$

Next, $\tilde{D}^* = \tilde{Y}_1 - \tilde{Y}_2^*$ and $\boldsymbol{m} = (1, -1)^T$.

$$\tilde{D}^* \sim \mathrm{SN}_1\left(\alpha/\beta, \boldsymbol{m}^T \tilde{\boldsymbol{\Sigma}}^* \boldsymbol{m}, \boldsymbol{m}^T \tilde{\boldsymbol{\Sigma}}^{*1/2} \tilde{\boldsymbol{\delta}}^* / (\boldsymbol{m}^T \tilde{\boldsymbol{\Gamma}}^* \boldsymbol{m})^{1/2}\right). \tag{27}$$

The TDI* for SN-RMEM is

$$P(|\tilde{D}^*| \le t) = F^*(t) - F^*(-t); \quad t > 0 \tag{28}$$

where F^* is the distribution function of $\tilde{D}^*$.

3.3.3. N-RMEM

The mean vector and variance matrix of $\tilde{\boldsymbol{Y}}^*$ are, according to the standard model (11),

$$\begin{aligned} E(\tilde{\boldsymbol{Y}}^*) &= \boldsymbol{A}^* + \boldsymbol{B}^* \mu_b \quad \text{and} \\ Var(\tilde{\boldsymbol{Y}}^*) &= \phi_b \boldsymbol{B}^* \boldsymbol{B}^{*T} + \tilde{\boldsymbol{\Sigma}}_1^*. \end{aligned} \tag{29}$$

The marginal distribution of $(\tilde{Y}_1, \tilde{Y}_2^*)$ is $\begin{pmatrix} \tilde{Y}_1 \\ \tilde{Y}_2^* \end{pmatrix} \sim \mathrm{N}_2(\boldsymbol{A}^* + \boldsymbol{B}^*\mu_b, \tilde{\boldsymbol{\Sigma}}^*)$.

Next, $\tilde{D}^* = \tilde{Y}_1 - \tilde{Y}_2^*$ can be represented as

$$\tilde{D}^* \sim \mathrm{N}_1\left(\alpha/\beta, \boldsymbol{m}^T \tilde{\boldsymbol{\Sigma}}^* \boldsymbol{m}\right). \tag{30}$$

The N-RMEM adaptation of CCC^* can now be defined as

$$CCC^* = \frac{2\phi_b}{\left(\frac{\alpha}{\beta}\right)^2 + [\phi_b + \phi_\delta] + \left[\phi_b + \frac{1}{\beta^2}(\phi_\varepsilon + \phi_e)\right]} \tag{31}$$

The TDI* under N-RMEM can be determined as

$$P(|\tilde{D}^*| \le t) = \Phi\left(\frac{t - E(\tilde{D}^*)}{\mathrm{sd}(\tilde{D}^*)}\right) - \Phi\left(\frac{-t - E(\tilde{D}^*)}{\mathrm{sd}(\tilde{D}^*)}\right) \tag{32}$$

where Φ denotes the cumulative distribution function (CDF) of a standard N distribution.

4. Simulation study

A simulation study is performed to evaluate the performance of the MLEs under the ST-RMEM, SN-RMEM, and N-RMEM models designed for analyzing method comparison data. We generated the data for two different methods, considering sample sizes of $m = 20$, 50, and 100, using models (5) and (8) that incorporated ST distribution for true covariate and t distribution for the error term ($\nu = 5$). The initial values of the parameters are $\mu_b = 1$, $\alpha = 0.02$, $\beta = 0.96$, $\log(\phi_b) = 0.03$, $\log(\phi_\delta) = -5$, $\log(\phi_\varepsilon) = -6$, $\log(\phi_e) = -6$, and we set $\lambda_b = 5$ and 10 for comparison. These findings were inspired by the ML estimate from the real data set studied in Section 5. Furthermore, we assume that the repeated number of observations per method is three. We then compute the MLEs using the EM algorithm on the sample data using the ST, SN, and N distributions with equation error, respectively. For the assessment of the estimations, we compute the sample bias (BIAS), standard deviation (SD), root mean square error (RMSE), and coverage probability (CP) after 1000 repeats. Table 1 summarizes the findings. The R programming language (R Core Team, 2021) was used to do all calculations.

Table 1 shows BIAS, SD, and RMSE values for the ST distribution are lower in all circumstances. As a result, the performance of the ST distribution is better than that of the SN and N distributions, which may be due to their skewed and heavy-tailed characteristics. Additionally, the estimates become more exact when the sample size rises from 20 to 100. When $m = 100$, all coverage probabilities (CPs) are near the nominal value of 95 percent. For smaller and moderate sample sizes, most of the CPs are also around 95 percent, and some are considerably lower. However, the CPs for all cases rise as the sample size increases. As a result, whether the skewness is moderate or heavy, we may state that the ST-RMEM CPs outperform other models.

Table 2 presents the efficiencies of ST-RMEM-based estimators in relation to the SN-RMEM and N-RMEM models calculated by dividing the MSE under the SN-RMEM and N-RMEM models by the MSE under the ST-RMEM. Notice that the relative efficiencies are greater than one in all situations, meaning that ST-RMEM is more accurate than SN-RMEM and N-RMEM. Furthermore, when n rises, the relative efficiencies improve.

We also compute the Akaike information criterion (AIC) and Bayesian information criterion (BIC) values when the data is produced via ST-RMEM. These values are shown in Table 3, and the findings reveal that ST-RMEM performs better than other models since it has lower values. Furthermore, as the sample size rises, the estimates become more exact. Table 4 presents estimated type I error probabilities for the 5% level Likelihood Ratio (LR) test, where the null hypothesis claims that a smaller model (SN-RMEM or N-RMEM) gives a good fit and the alternative hypothesis states that a larger model (ST-RMEM) provides a good fit. For the small sample size, values are close to 5%, showing the minimal difference between the two models. The values are fewer than 5% for moderate and large sample sizes, indicating that ST-RMEM is preferable. In summary, the ST-RMEM performs much better than the N-RMEM and SN-RMEM in the presence of skewed and heavy-tailedness.

5. Application to fat data

The subcutaneous fat thickness (Carstensen *et al.*, 2020) was measured in centimeters at the Steno Diabetes Center to compare the measurements of two experienced observers, 'KL' (Method 1) and 'SL' (Method 2). The study includes 43 persons (subjects), and the measurements (cm) from each method are repeated three times on each subject. The three replicates are interchangeable within the subject and method, and the repeated measurements are unlinked. The design is balanced with 43×3×2=258 observations, and measurements vary from 0.39 to 4.20 cm. Figure 1

depicts a histogram and normal Q-Q plot for subcutaneous fat, revealing that the data is positively skewed and heavy-tailed. The trellis plot of this data, shown in Figure 2, reveals considerable overlap in the measurements given by the methods. At the same time, it is evident that SL values are lower than KL for the majority of persons. A few cases show quite substantial disparities, implying a skewed distribution of differences. The measures show significant within-subject variation, although it is small when compared to between-subject variation. The dataset is homoscedastic, and there are no obvious outliers.

Figure 3 shows scatterplots and Bland-Altman plots of randomly chosen and averaged over replications. The scatter plots reveal a high correlation between the methods, confirming that KL readings are greater than SL readings since most points are above the line of equality, and the Bland-Altman plots indicate that the scales of the methods may differ. Moreover, it should be noted that the data were obtained on persons from a Diabetes Center, and numerous factors, such as a person's food habits and laboratory conditions, might influence a measurement. As a result, these measures are prone to inaccuracy. Thus, the measurement error model gives a better fit for this data.

The modeling of data is the preliminary step in the analysis. Initially, we fit the data using the modified ST-RMEM (9), where b_i follows ST distribution, measurement errors $(\delta_{ik}, \epsilon_{ik})$, and equation errors (e_i) follows multivariate t distribution. In this case, the degree of freedom (ν) is treated as a known parameter, determined by the Schwarz information criteria (Schwarz, 1978). There are a total of eight parameters in this model. The *numDeriv* package (Gilbert and Varadhan, 2019) in R is used to compute the required numerical derivatives. Secondly, we fit the SN-RMEM (10) where b_i follows SN distribution, measurement errors $(\delta_{ik}, \epsilon_{ik})$, and equation errors (e_i) follow multivariate N distribution. This model also has eight unknown parameters. Next, we fit the N-RMEM (11), which has seven unknown parameters. We then compute the MLEs of parameter $\boldsymbol{\theta}$ using the EM algorithm and their standard errors (SEs) under the above models.

Table 5 provides these parameter estimates, SEs, and 95% confidence limits for the above RMEMs. AIC and BIC values based on the RMEM model under the above distributions are shown in Table 6. The model is better when the AIC value is small, and we find that the AIC value is small for ST-RMEM. Furthermore, the LR test is used to determine if the null hypothesis H_0 : SN-RMEM model is preferred to the alternative hypothesis H_1 : ST-RMEM model is preferable. It is important to test the hypothesis H_0 to see if the inclusion of the degrees of freedom (ν) is meaningful. The p-value for this LR test is < 0.0001. Therefore, the parameter (ν) must be taken into account. Thus, the modified ST-RMEM fits significantly better than N-RMEM and SN-RMEM.

The examination of similarity is the second step in the analysis. The proportionate bias estimate (β) is 0.97 (SE = 0.02), and the 95% confidence interval is [0.93, 0.99]. As a result, there is evidence of a minor downward proportionate bias, but it is marginal. Furthermore, the estimated fixed bias (α) is 0.02 (SE = 0.03), with a 95% confidence range of [-0.04, 0.08]. Although this interval includes 0, it also demonstrates a minor fixed bias. Because there is evidence of small bias, the methods have unequal scales. Thus, their precision is measured using a squared sensitivity ratio, and the value is 1.16 (> 1), indicating that Method 2 (SL) is more precise than Method 1(KL).

The next step is an assessment of the agreement. As previously indicated, due to a little bias in SL measurement, we rescaled its measurement $\tilde{Y_2}$ as $\tilde{Y_2}^* = \tilde{Y_2}/\beta$. The estimated transformation is $\tilde{Y_2}^* = \tilde{Y_2}/0.97$. We compute the estimates and 95% one-sided confidence limits for the agreement measures examined in Section 3.2 for the converted data, which are also shown in Table 6. To obtain these estimates, first, perform Fisher's z-transformation for CCC^* and the log transformation for

TDI^*. The CCC^* estimate for ST-RMEM is 0.990, as defined in (23), and its lower bound is 0.984; both are close to one, suggesting good agreement amongst the methods. Next, we make an inference on the agreement measure TDI^* (with $p = 0.90$), which is given by (24). It has an estimate of 0.034 and an upper bound of 0.050. This upper bound indicates that 90% of discrepancies in measurements from the methods lie within -0.05 to 0.05 with 95% confidence. Since the true value range is around 4, this discrepancy may be regarded as acceptable. As a result, we may conclude that the methods are in good agreement. This obviously suggests that the KL and rescaled SL methods agree sufficiently to be deemed interchangeable.

6. Conclusions

This paper develops the methodology for analyzing replicated method comparison data using the MEM framework with the ST distribution for true covariate and the t distribution for errors. We considered the same degree for true covariates and errors. This methodology is sufficient enough to accommodate normally distributed, skewed, heavy-tailed data and both together. The main advantage of this model is that it can assess similarity and agreement between methods, regardless of whether or not the methods use the same nominal unit of measurement. We concentrated here on a comparison of two methods. However, the model may be expanded to include more than two methods. Simulation experiments and the use of subcutaneous fat data confirmed the efficiency and reliability of findings under the ST-RMEM model. Furthermore, we determined that the ST-RMEM model performs best with skewed and heavy-tailed data. Our proposed model would yield appropriate results for method comparison data with measurement error, skewness, and heavy tails, which are frequent in many fields such as economics, health, and the environment.

Data availability statement

The subcutaneous fat dataset is available in Carstensen *et al.* (2020).

References

Alanen, E. (2010). Everything all right in method comparison studies? *Statistical Methods in Medical Research*, **21**, 297–309.

Barnhart, H. X., Haber, M., and Song, J. (2002). Overall concordance correlation coefficient for evaluating agreement among multiple observers. *Biometrics*, **58**, 1020–1027.

Barnhart, H. X., Lokhnygina, Y., Kosinski, A. S., and Haber, M. (2007). Comparison of concordance correlation coefficient and coefficient of individual agreement in assessing agreement. *Journal of Biopharmaceutical Statistics*, **17**, 721–738.

Barnhart, H. X., Song, J., and Haber, M. J. (2005). Assessing intra, inter and total agreement with replicated readings. *Statistics in Medicine*, **24**, 1371–1384.

Bland, J. M. and Altman, D. G. (1999). Measuring agreement in method comparison studies. *Statistical Methods in Medical Research*, **8**, 135–160.

Bland, J. M. and Altman, D. G. (2007). Agreement between methods of measurement with multiple observations. *Journal of Biopharmaceutical Statistics*, **17**, 571–582.

Cao, C., Wang, Y., Shi, J. Q., and Lin, J. (2017). Measurement error models for replicated data under asymmetric heavy-tailed distributions. *Computational Economics*, **52**, 531–553.

Carrasco, J. L. and Jover, L. (2003). Estimating the generalized concordance correlation coefficient through variance components. *Biometrics*, **59**, 849–858.

Carrasco, J. L., King, T. S., and Chinchilli, V. M. (2009). The concordance correlation coefficient for repeated measures estimated by variance components. *Journal of Biopharmaceutical Statistics*, **19**, 90–105.

Carstensen, B., Gurrin, L., Ekstrøm, C., and Figurski, M. (2020). Methcomp: Analysis of agreement in method comparison studies. *R package version*, **1**.

Carstensen, B., Simpson, J., and Gurrin, L. C. (2008). Statistical models for assessing agreement in method comparison studies with replicate measurements. *The International Journal of Biostatistics*, **4**, 1–26.

Choudhary, P. K. (2008). A tolerance interval approach for assessment of agreement in method comparison studies with repeated measurements. *Journal of Statistical Planning and Inference*, **138**, 1102–1115.

Choudhary, P. K. (2009). *Methods and Applications of Statistics in the Life and Health Sciences*, chapter Interrater agreement. John Wiley Sons, New York.

Choudhary, P. K. (2010). A unified approach for nonparametric evaluation of agreement in method comparison studies. *The International Journal of Biostatistics*, **6**, 1–24.

Choudhary, P. K. and Nagaraja, H. N. (2017). *Measuring Agreement: Models, Methods, and Applications*. J. Wiley amp; Sons.

Choudhary, P. K. and Yin, K. (2010). Bayesian and frequentist methodologies for analyzing method comparison studies with multiple methods. *Statistics in Biopharmaceutical Research*, **2**, 122–132.

Dunn, G. and Roberts, C. (1999). Modeling method comparison data. *Statistical Methods in Medical Research*, **8**, 161–179.

Duwarahan, J. and Nawarathna, L. S. (2022). An improved measurement error model for analyzing unreplicated method comparison data under asymmetric heavy-tailed distributions. *Journal of Probability and Statistics*, **2**, 1–13.

Gilbert, P. and Varadhan, R. (2019). numderiv: Accurate numerical derivatives.

Hedayat, A. S., Lou, C., and Sinha, B. K. (2009). A statistical approach to assessment of agreement involving multiple raters. *Communications in Statistics - Theory and Methods*, **38**, 2899–2922.

King, T. S. and Chinchilli, V. M. (2001). A generalized concordance correlation coefficient for continuous and categorical data. *Statistics in Medicine*, **20**, 2131–2147.

King, T. S., Chinchilli, V. M., Wang, K.-L., and Carrasco, J. L. (2007). A class of repeated measures concordance correlation coefficients. *Journal of Biopharmaceutical Statistics*, **17**, 653–672.

Lehmann, E. L. (1998). *Elements of Large-sample Theory*. Springer.

Lin, L., Hedayat, A., and Wu, W. (2012). *Statistical Tools for Measuring Agreement*. Springer.

Lin, L., Hedayat, A. S., Sinha, B., and Yang, M. (2002). Statistical methods in assessing agreement: Models, issues, and tools. *Journal of the American Statistical Association*, **97**, 257–270.

Lin, L., Hedayat, A. S., and Wu, W. (2007). A unified approach for assessing agreement for continuous and categorical data. *Journal of Biopharmaceutical Statistics*, **17**, 629–652.

Lin, L. I.-K. (1989). A concordance correlation coefficient to evaluate reproducibility. *Biometrics*, **45**, 255–268.

Lin, L. I.-K. (2000). Total deviation index for measuring individual agreement with applications in laboratory performance and bioequivalence. *Statistics in Medicine*, **19**, 255–270.

Nawarathna, L. S. and Choudhary, P. K. (2013). Measuring agreement in method comparison studies with heteroscedastic measurements. *Statistics in Medicine*, **32**, 5156–5171.

Nawarathna, L. S. and Choudhary, P. K. (2015). A heteroscedastic measurement error model for method comparison data with replicate measurements. *Statistics in Medicine*, **34**, 1242–1258.

R Core Team, R. (2021). *R: A Language and Environment for Statistical Computing*.

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, **6**, 461–464.

Sengupta, D., Choudhary, P. K., and Cassey, P. (2015). Modeling and analysis of method comparison data with skewness and heavy tails. In Choudhary, P., Nagaraja, C. H., and Ng, H. K. T., editors, *Ordered Data Analysis, Modeling and Health Research Methods*, pages 169–187, Cham. Springer International Publishing.

Tomaya, L. C. and de Castro, M. (2018). A heteroscedastic measurement error model based on skew and heavy-tailed distributions with known error variances. *Journal of Statistical Computation and Simulation*, **88**, 2185–2200.

ANNEXURE

Table 1: Simulated bias, sample standard deviation (SD), root mean square error (RMSE), and coverage probability (CP) of ML for the ST-RMEM, SN-RMEM, and N-RMEM

λ	m	Parameter	ST-RMEM				SN-RMEM				N-RMEM			
			BIAS	SD	RMSE	CP	BIAS	SD	RMSE	CP	BIAS	SD	RMSE	CP
5	20	α	0.000	0.042	0.042	0.968	-0.002	0.054	0.054	0.932	-0.001	0.053	0.053	0.944
		β	0.000	0.022	0.022	0.974	0.002	0.029	0.029	0.890	0.000	0.028	0.028	0.904
		μ_b	0.015	0.080	0.082	0.794	-0.022	0.099	0.101	1.000	0.003	0.214	0.214	0.902
		$\log(\phi_b)$	-0.044	0.152	0.159	0.876	0.097	0.215	0.236	0.960	-0.244	0.593	0.641	0.682
		λ_b	0.003	0.405	0.405	0.802	0.347	1.367	1.410	1.000	-	-	-	-
		$\log(\phi_\delta)$	-0.015	0.215	0.216	0.982	-0.097	0.357	0.370	0.740	-0.095	0.353	0.366	0.746
		$\log(\phi_\varepsilon)$	0.002	0.258	0.258	0.970	-0.059	0.380	0.385	0.748	-0.059	0.380	0.385	0.748
		$\log(\phi_e)$	-0.259	0.617	0.669	0.996	-0.513	0.906	1.041	0.976	-0.438	0.802	0.914	0.980
	50	α	0.000	0.024	0.024	0.984	-0.003	0.038	0.038	0.908	0.000	0.036	0.036	0.924
		β	0.000	0.013	0.013	0.988	0.001	0.021	0.021	0.868	0.000	0.020	0.020	0.882
		μ_b	0.003	0.039	0.039	0.884	-0.034	0.058	0.067	0.996	-0.004	0.133	0.133	0.900
		$\log(\phi_b)$	-0.011	0.067	0.068	0.908	0.109	0.141	0.178	0.942	-0.126	0.445	0.462	0.612
		λ_b	0.001	0.175	0.175	0.884	0.195	0.699	0.726	0.996	-	-	-	-
		$\log(\phi_\delta)$	0.001	0.114	0.114	0.994	-0.036	0.235	0.238	0.772	-0.036	0.232	0.234	0.784
		$\log(\phi_\varepsilon)$	-0.003	0.140	0.140	0.982	-0.037	0.246	0.248	0.758	-0.037	0.246	0.248	0.758
		$\log(\phi_e)$	-0.065	0.256	0.264	0.998	-0.176	0.560	0.587	0.968	-0.149	0.457	0.481	0.980
	100	α	-0.001	0.015	0.015	0.988	-0.005	0.030	0.031	0.904	-0.002	0.030	0.030	0.914
		β	0.001	0.008	0.008	0.984	0.003	0.016	0.017	0.848	0.001	0.016	0.016	0.856
		μ_b	0.002	0.020	0.021	0.934	-0.033	0.038	0.050	0.996	0.000	0.094	0.094	0.934
		$\log(\phi_b)$	-0.005	0.032	0.033	0.938	0.100	0.088	0.133	0.948	-0.067	0.330	0.337	0.614
		λ_b	-0.003	0.099	0.099	0.934	0.097	0.359	0.372	0.998	-	-	-	-
		$\log(\phi_\delta)$	-0.003	0.069	0.069	1.000	-0.020	0.193	0.194	0.724	-0.022	0.190	0.192	0.732
		$\log(\phi_\varepsilon)$	-0.001	0.093	0.093	0.984	-0.023	0.187	0.189	0.708	-0.023	0.187	0.189	0.708
		$\log(\phi_e)$	-0.027	0.136	0.138	0.998	-0.104	0.363	0.378	0.952	-0.089	0.303	0.316	0.964

Table 1: Continued

λ	m	Parameter	ST-RMEM				SN-RMEM				N-RMEM			
			BIAS	SD	RMSE	CP	BIAS	SD	RMSE	CP	BIAS	SD	RMSE	CP
10	20	α	0.000	0.043	0.043	0.972	-0.002	0.055	0.055	0.934	-0.001	0.055	0.055	0.932
		β	0.000	0.023	0.023	0.972	0.001	0.029	0.029	0.900	0.000	0.029	0.029	0.906
		μ_b	0.006	0.033	0.034	0.752	-0.006	0.041	0.041	0.946	0.004	0.211	0.211	0.890
		$log(\phi_b)$	-0.015	0.057	0.059	0.944	0.026	0.097	0.101	0.954	-0.258	0.608	0.660	0.666
		λ_b	0.026	0.548	0.549	0.758	0.291	2.083	2.103	0.950	-	-	-	-
		$log(\phi_\delta)$	-0.015	0.210	0.211	0.986	-0.095	0.355	0.368	0.746	-0.095	0.353	0.366	0.746
		$log(\phi_\varepsilon)$	0.001	0.256	0.256	0.968	-0.059	0.380	0.385	0.748	-0.059	0.380	0.385	0.748
		$log(\phi_e)$	-0.247	0.597	0.646	0.994	-0.478	0.915	1.032	0.976	-0.437	0.804	0.915	0.976
	50	α	0.000	0.024	0.024	0.982	-0.003	0.038	0.038	0.914	0.000	0.038	0.038	0.920
		β	0.000	0.013	0.013	0.988	0.001	0.021	0.021	0.866	0.000	0.021	0.021	0.878
		μ_b	0.001	0.015	0.015	0.798	-0.008	0.023	0.024	0.994	-0.004	0.130	0.130	0.892
		$\log(\phi_b)$	-0.003	0.024	0.024	0.934	0.027	0.060	0.066	0.978	-0.131	0.456	0.474	0.588
		λ_b	0.007	0.230	0.230	0.796	0.086	0.888	0.892	0.990	-	-	-	-
		$log(\phi_\delta)$	0.003	0.112	0.112	0.992	-0.035	0.233	0.235	0.780	-0.036	0.232	0.235	0.784
		$log(\phi_\varepsilon)$	-0.001	0.138	0.138	0.980	-0.037	0.246	0.248	0.758	-0.037	0.246	0.248	0.760
		$log(\phi_e)$	-0.063	0.247	0.255	0.994	-0.167	0.511	0.538	0.970	-0.150	0.454	0.478	0.980
	100	α	-0.001	0.015	0.015	0.990	-0.005	0.029	0.030	0.918	-0.002	0.029	0.030	0.928
		β	0.001	0.008	0.008	0.986	0.003	0.016	0.016	0.854	0.001	0.016	0.016	0.858
		μ_b	0.001	0.008	0.008	0.844	-0.008	0.013	0.016	0.998	0.000	0.093	0.093	0.934
		$\log(\phi_b)$	-0.001	0.012	0.012	0.946	0.023	0.025	0.034	0.996	-0.073	0.339	0.346	0.596
		λ_b	0.001	0.126	0.126	0.844	0.001	0.381	0.381	1.000	-	-	-	-
		$log(\phi_\delta)$	-0.002	0.067	0.067	1.000	-0.019	0.192	0.193	0.722	-0.022	0.190	0.191	0.732
		$log(\phi_\varepsilon)$	-0.001	0.092	0.092	0.984	-0.023	0.187	0.189	0.708	-0.023	0.187	0.189	0.708
		$log(\phi_e)$	-0.024	0.129	0.131	0.998	-0.091	0.313	0.326	0.958	-0.089	0.309	0.321	0.962

Table 2: Relative efficiencies of ST-RMEM-based estimators relative to the N-RMEM and SN-RMEM

m	Quantity	$\lambda = 5$		$\lambda = 10$	
		MSE_{SN}/MSE_{ST}	MSE_N/MSE_{ST}	MSE_{SN}/MSE_{ST}	MSE_N/MSE_{ST}
20	α	1.669	1.564	1.645	1.615
	β	1.697	1.591	1.689	1.664
	μ_b	1.542	6.880	1.494	39.476
	$log(\phi_b)$	2.214	16.298	2.899	125.044
	λ_b	12.140	-	14.691	-
	$log(\phi_\delta)$	2.941	2.869	3.041	3.004
	$log(\phi_\epsilon)$	2.230	2.230	2.261	2.261
	$log(\phi_e)$	2.420	1.864	2.552	2.008
50	α	2.518	2.353	2.478	2.436
	β	2.599	2.427	2.542	2.479
	μ_b	2.987	11.626	2.784	79.822
	$log(\phi_b)$	6.929	46.536	7.531	384.150
	λ_b	17.174	-	15.026	-
	$log(\phi_\delta)$	4.310	4.195	4.420	4.386
	$log(\phi_\epsilon)$	3.144	3.145	3.244	3.244
	$log(\phi_e)$	4.950	3.322	4.451	3.514
100	α	4.281	4.007	4.079	4.016
	β	4.466	4.172	4.272	4.160
	μ_b	5.886	20.898	4.224	150.331
	$log(\phi_b)$	16.624	106.344	8.354	860.975
	λ_b	14.090	-	9.108	-
	$log(\phi_\delta)$	8.006	7.790	8.370	8.245
	$log(\phi_\epsilon)$	4.114	4.114	4.215	4.215
	$log(\phi_e)$	7.444	5.211	6.147	5.975

Table 3: Results of model selection criteria when the ST-RMEM is the data generating model

Set	m	Criterion	Models		
			ST-RMEM	SN-RMEM	N-RMEM
$\lambda = 5$	20	AIC	-127.966	-114.127	-114.254
		BIC	-120.001	-106.161	-107.284
	50	AIC	-334.621	-291.235	-286.153
		BIC	-319.325	275.939	-272.769
	100	AIC	-681.815	-585.752	-571.630
		BIC	-660.974	-564.911	-553.394
$\lambda = 10$	20	AIC	-130.099	-115.218	-115.177
		BIC	-122.133	-107.252	-108.206
	50	AIC	-341.021	-295.815	-288.066
		BIC	-325.725	-280.519	-274.681
	100	AIC	-695.410	-596.829	-575.334
		BIC	-674.569	-575.987	-557.098

Table 4: Estimated type I error probabilities for 5% level likelihood ratio test

Set	m	H_0 : SN-RMEM model is preferable H_1 : ST-RMEM model is preferable	H_0 : N-RMEM model is preferable H_1 : ST-RMEM model is preferable
$\lambda = 5$	20	0.067	0.044
	50	0.008	0.002
	100	< 0.001	< 0.001
$\lambda = 10$	20	0.062	0.032
	50	0.008	0.001
	100	< 0.001	< 0.001

Table 5: ML estimates, their SEs, and 95% confidence intervals for parameters of subcutaneous fat data

Parameter	ST-RMEM				SN-RMEM				N-RMEM			
	Estimate	SE	95% CI		Estimate	SE	95% CI		Estimate	SE	95% CI	
			LCL	UCL			LCL	UCL			LCL	UCL
α	0.018	0.029	-0.039	0.075	0.032	0.034	-0.035	0.099	0.033	0.035	-0.037	0.102
β	0.967	0.017	0.934	0.999	0.958	0.017	0.925	0.991	0.957	0.017	0.923	0.992
μ_b	0.959	0.202	0.564	1.355	1.237	0.503	0.252	2.223	1.830	0.134	1.567	2.093
$log(\phi_b)$	0.027	0.339	-0.638	0.691	0.105	0.426	-0.729	0.939	-0.258	0.216	-0.682	0.166
λ_b	1.687	0.842	0.036	3.338	0.988	1.288	-1.537	3.513	-	-	-	-
$log(\phi_\delta)$	-5.318	0.180	-5.671	-4.966	-5.134	0.151	-5.430	-4.837	-5.135	0.151	-5.431	-4.840
$log(\phi_\epsilon)$	-5.532	0.186	-5.896	-5.168	-5.251	0.152	-5.549	-4.952	-5.251	0.152	-5.548	-4.953
$log(\phi_e)$	-5.754	0.470	-6.676	-4.832	-5.139	0.354	-5.832	-4.445	-5.030	0.344	-5.705	-4.356

Table 6: Model selection criteria and measures of agreement for transformed subcutaneous fat data. Lower bound for CCC^* and upper bound for TDI^* are presented

Models	AIC	BIC	CCC^*		TDI^*	
			Estimate	95% Bound	Estimate	95% Bound
ST-RMEM	-278.806	-259.171	0.990	0.984	0.034	0.050
SN-RMEM	-261.149	-241.514	0.987	0.977	0.056	0.083
N-RMEM	-264.469	-247.289	0.987	0.980	0.232	0.263

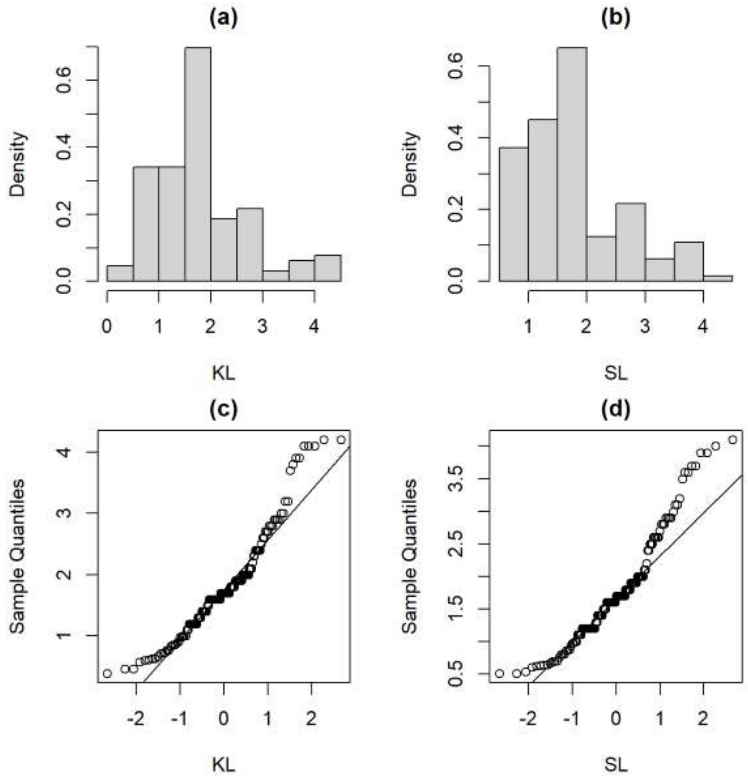


Figure 1: Histogram (a-b) and normal Q-Q plot (c-d) of the subcutaneous fat data. Left panel for 'KL' observer and right panel for 'SL' observer

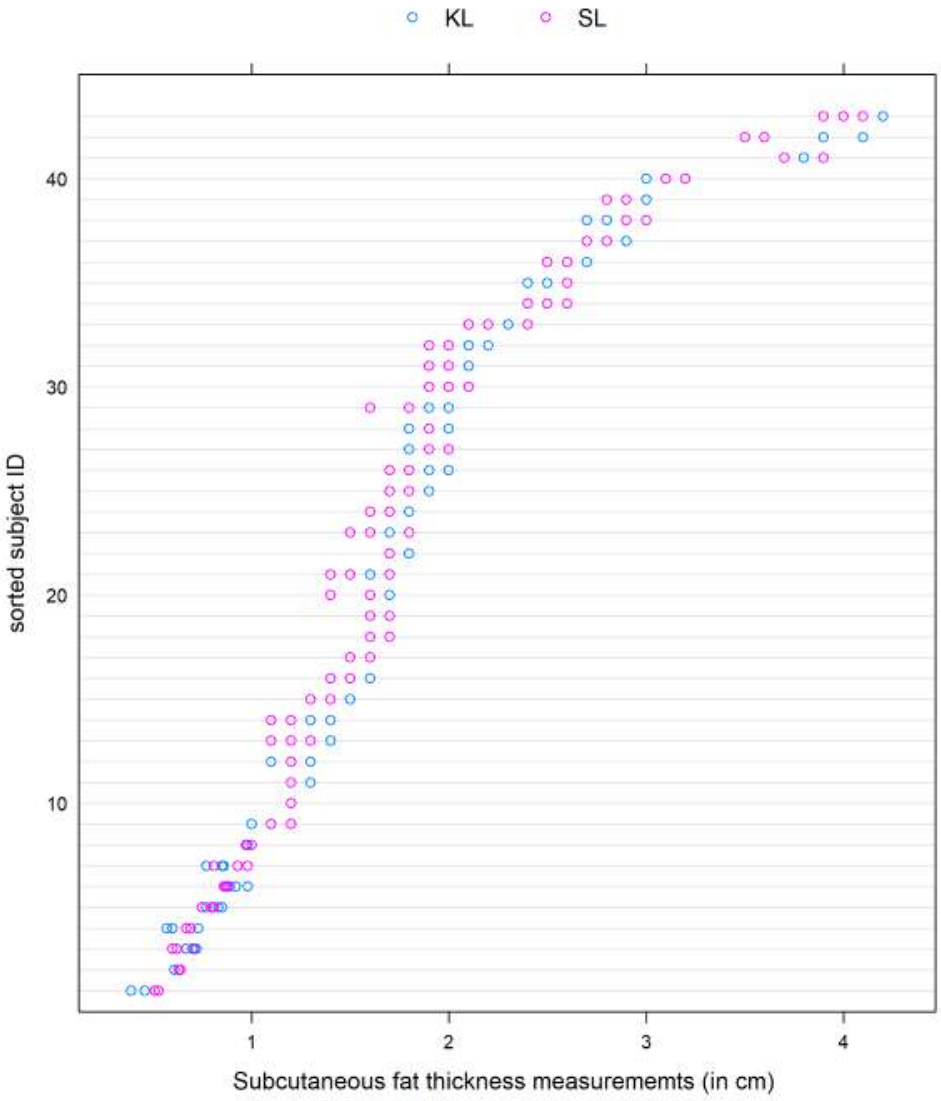


Figure 2: Trellis plot for subcutaneous fat data

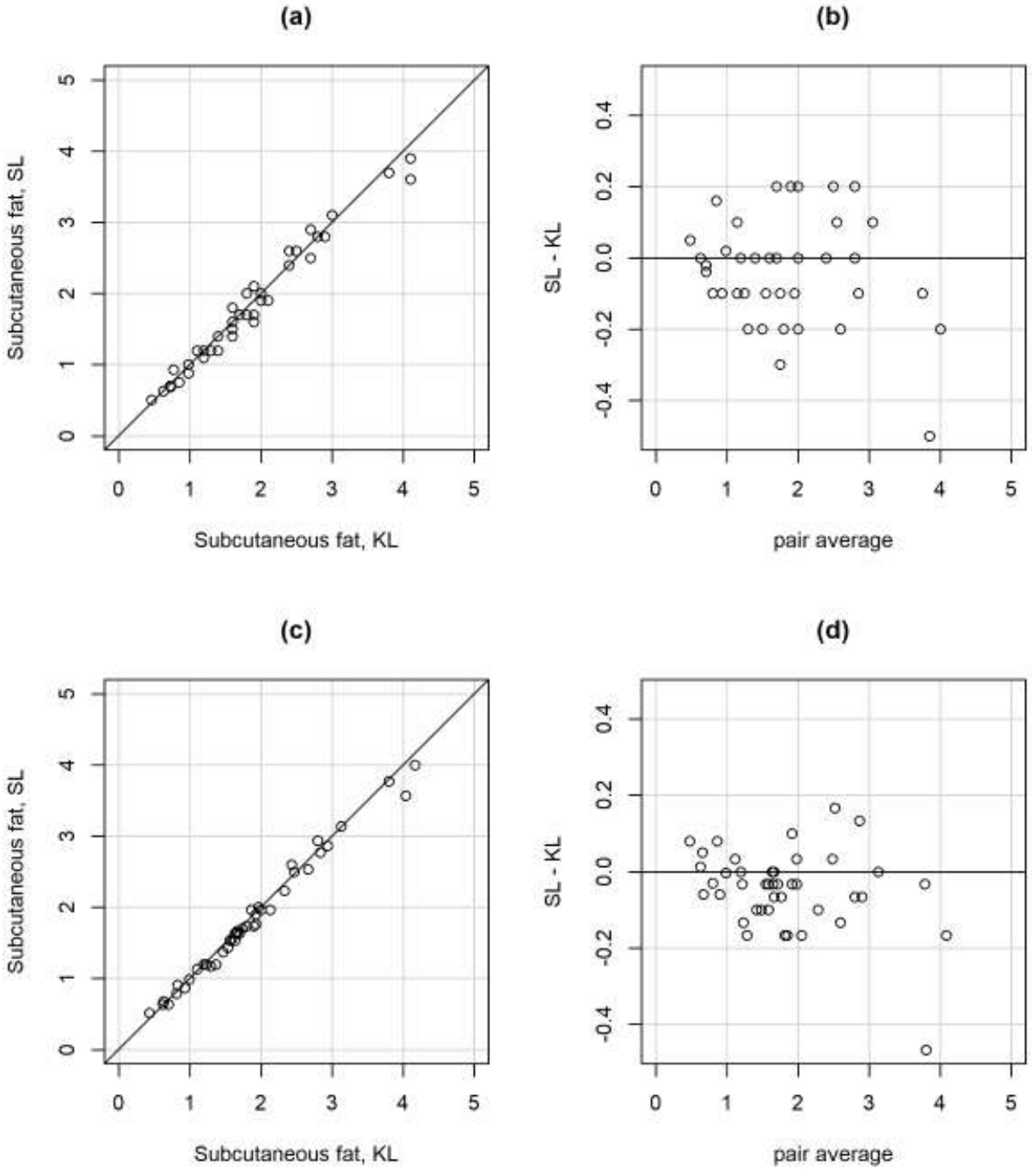


Figure 3: Scatterplot with line of equality (left) and Bland-Altman plots with zero line (right) for subcutaneous fat thickness measurements. One measurement per method from each of the 43 subjects is randomly selected for this plot. Same as the top panel but based on 43 average measurements

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 22, No. 2, 2024 (New Series), pp 169–180
http://www.ssca.org.in/journal

Study of Priority Based Network Nodes Using Quasi Birth and Death Process

Malla Reddy Perati and Abhilash Vollala
Department of Mathematics
Kakatiya University, Warangal, India

Received: 15 December 2022 ; Revised: 16 July 2023; Accepted: 07 December 2023

Abstract

In this paper, network node with self-similar priority based input traffic is modeled into finite buffer single server queuing system, and is analyzed through level dependent quasi birth-death (QBD) process with preemptive priority mechanism. Here, input process follows transient Markovian arrival process (MAP), and service time (packet lengths) follows Phase type (PH) distribution, which is more general than deterministic and exponential distributions. The queuing behavior of the system at arbitrary times through the performance metrics, namely, queue length, mean waiting time, and packet loss probability is investigated. For this, time dependent state probability vector of transition rate matrix is obtained using method of product integrals which in turn gives performance measures, and computational complexity of analysis is presented. This type of analysis is useful in dimensioning the network node to provide Quality of Service (QoS) guarantee.

Key words: Self-similar; Quasi birth and death process; Markovain arrival process; Phase type; Transition rate matrix; Waiting time; Loss probability.

AMS Subject Classifications: 60G18, 60K25, 68M20

1. Introduction

Performance of communication system depends on network nodes. The network nodes namely, switch, router, and multiplexer in B-ISDN (Broadband Integrated Switching Digital Network), play a vital role in communication, and therefore it is essential to analyze the performance of nodes for providing QoS. In general, analysis of network nodes is made by queueing methods, and this queueing based analysis has a long history of success in planning and dimensioning of networks. The fundamental studies of network traffic namely LAN (Leland *et al.*, 1994), WAN (Paxson and Floyd, 1995), and WWW (Crovella and Bestavros, 1997) at AT & T Bell labs disclosed that these traffic are self-similar, and degrade performance of system. It is clear that self-similar nature of traffic is emulated by homogeneous Markovian Modulated Poisson process which was superposition of Interrupted Poisson Process (IPP) or Switched Poisson Process (SPP) over different time scales. In the papers

Corresponding Author: Abhilash Vollala
Email: ashwesa.abhi@gmail.com

(Andersen and Nielsen, 1998; Yoshihara *et al.*, 2001; Shao *et al.*, 2005), performance analysis was made under steady state conditions, as such is not so useful for real time network traffic analysis. Recently, Abhilash and Malla Reddy (2022) proposed a fitting procedure for time dependent Markovian process, namely, MMPP with Sinusoidal arrival rates based on second order statistics, and proved that resultant MMPP emulates self-similar nature of network traffic in prescribed time scales. On the other hand, in B-ISDN, high demand causes congestion, and pertinent issues can be handled using priority queueing mechanism. Prioritization based on the importance is most common feature in all modern internet applications to offer QoS. Priority mechanism is a concept of scheduling of different classes of arrivals to a single server. It has wide range of applications not only in engineering, but in inventory of manufacturing industries and health care systems (Zhao and Alfa, 1995; Brahimi and Worthington, 1991; Cohen *et al.*, 1988). There are different priority disciplines like preemptive, non-preemptive and discretionary priority. Each discipline has a scheduling procedure. In the literature, there are number of supplements based on priority scheduling, the outline of few fundamental priority queueing models in continuous-time was evident in the papers (Miller, 1960; Kleinrock, 1976; Takagi, 1991) and references therein. White and Christie (1958) studied M/M/1 queues with multiclass arrivals using preemptive priorities and analysis is made by generating functions of state probabilities. Later, Marks (1973) proposed an algorithm for computing probabilities of queue length. Sandhu and Posner (1989) analyzed voice/data communication using priority M/G/1 queue. Boxma *et al.* (1999) worked on heavy traffic using M/G/1 queue with priority classes and regularly varying heavy tailed service time distributions. Sharma and Virtamo (2002) consider finite buffer queue with priorities to model the system in the internet and obtain algorithms for workload, waiting time, and packet loss. Takine and Hasegawa (1994) derived LST of waiting time of customers based on MAP/G/1 queue with state dependent service time distributions. Takahashi and Miyazawa (1994) gave relation between queue length and waiting time distribution in a priority queue with batch arrivals. Takada and Miyazawa (2002) obtain moments of buffer contents for a Markov modulated fluid queue with preemptions. Jin and Min (2007) propose a novel analytical model for priority queueing system with heterogeneous LRD input traffic. Tarabia (2007) investigated the impact of catastrophes on single server preemptive priority queue using generating functions. Sampath *et al.* (2013)studied performance of wavelength division multiplexing optical packet switch employing wave length conversion techniques under self-similar input traffic. Zhao *et al.* (2015) analyzed sojourn time of two classes of customers using MAP/PH/1 queue with discretionary priority based on service stages. Ravi Kumar *et al.* (2017) evaluated performance of self-similar traffic input model in terms of high priority and low priority packet loss probabilities using MMPP/PH/c/K queuing system. Also, Malla Reddy and Ravi Kumar (2014, 2016, 2021)explored performance of network routers (synchronous and asynchronous) with self-similar input traffic using various multiserver queueing systems employing priority mechanism in the papers. From above cited papers, one can observe that priority discipline was used in various contexts to analyze systems, but in all the above cases performance analysis was made under steady state with homogeneous arrival and service processes, which are not realistic. As mentioned earlier, in present work, a network node with self-similar input traffic is modeled into transient MAP/PH/1/N queue with preemptive priority mechanism. Time dependent analysis of the system is made by level dependent quasi birth and death process, and arrival process follows MMPP with sinusoidal arrival rates (which is a special case of MAP). Performance measures, namely, queue length, mean waiting time, and packet loss of high priority and low priority

packets are presented numerically.

The paper is organized as follows: Queueing model description is given in section 2. In section 3, performance analysis of system is presented. In section 4, computation complexity of algorithm is presented, and numerical results are illustrated in section 5. Finally, conclusions are given in section 6.

2. Queuing model

It is assumed that the packet arrivals are of high priority (Type I) and low priority (Type II) packets. Assume that Type I packet arrivals follows the MMPP characterized by $(Q^I, \Lambda^I(t))$, where $Q^I, \Lambda^I(t)$ are matrices of order n_I. Where as, Type II packet arrivals follow the MMPP characteriazed by $(Q^{II}, \Lambda^{II}(t))$, where $Q^{II}, \Lambda^{II}(t)$ are matrices of order n_{II}. Here n_I, n_{II} represent number of states of underlying Markov chains of Type I and Type II arrivals, respectively. As in Andersen and Nielsen (1998); Yoshihara *et al.* (2001); Shao *et al.* (2005); Abhilash and Malla Reddy (2022), modeling of self-similar traffic involves superposition of two-state MMPPs (In particular IPPs). The i^{th} IPP of Type I and Type II arrival process are given as follows:

$$\begin{aligned} Q_i^I &= \begin{bmatrix} -c_{1i} & c_{1i} \\ c_{2i} & -c_{2i} \end{bmatrix}, \Lambda_i^I(t) = \begin{bmatrix} \lambda_i^I(t) & 0 \\ 0 & 0 \end{bmatrix} \\ and, Q_i^{II} &= \begin{bmatrix} -d_{1i} & d_{1i} \\ d_{2i} & -d_{2i} \end{bmatrix}, \Lambda_i^{II}(t) = \begin{bmatrix} \lambda_i^{II}(t) & 0 \\ 0 & 0 \end{bmatrix}, 1 \le i \le r \end{aligned} \tag{1}$$

The superposition of r IPPs and a Poisson process of Type I and Type II arrival process are, respectively, given as

$$\begin{aligned} Q^I &= Q_1^I \bigoplus Q_2^I \bigoplus \cdots \bigoplus Q_r^I \\ \Lambda^I &= \Lambda_1^I(t) \bigoplus \Lambda_2^I(t) \bigoplus \cdots \bigoplus \Lambda_r^I(t) \bigoplus \lambda_p^I(t) \\ Q^{II} &= Q_1^{II} \bigoplus Q_2^{II} \bigoplus \cdots \bigoplus Q_r^{II} \\ \Lambda^{II} &= \Lambda_1^{II}(t) \bigoplus \Lambda_2^{II}(t) \bigoplus \cdots \bigoplus \Lambda_r^{II}(t) \bigoplus \lambda_p^{II}(t) \end{aligned} \tag{2}$$

Here, $\oplus, \otimes$ represent Kronecker's sum and product respectively, and $\lambda_p^I(t), \lambda_p^{II}(t)$ are time dependent Poisson arrival rates of Type I and Type II arrivals. The superposition of MMPPs $(Q^I, \Lambda^I(t))$, $(Q^{II}, \Lambda^{II}(t))$ is turned into a MAP with representation of $(D_0, D_1(t), D_2(t))$, where $D_0 = D_0^I \oplus D_0^{II}$ denote transitions of no arrival in both types, $D_1(t) = \Lambda^I(t) \otimes I_{n_{II}}$ and $D_2(t) = I_{n_I} \otimes \Lambda^{II}(t)$ denotes transitions corresponding to Type I and Type II arrivals, respectively, where $D_0^{II} = Q^{II} - \Lambda^{II}(t), D_0^I = Q^I - \Lambda^I(t)$. The mean arrival rate of Type I and Type II packets are given by (Abhilash and Malla Reddy, 2022)

$$\lambda_m^I(t) = \frac{1}{t}\left(\pi \int_0^t D_1(x)\, dx\, e\right), \lambda_m^{II}(t) = \frac{1}{t}\left(\pi \int_0^t D_2(x)\, dx\, e\right) \tag{3}$$

where e represents column vector of 1's with appropriate size, and π is unique vector satisfying $\pi(Q^I + Q^{II}) = 0, \pi e = 1$. The system is modeled into a single server queue with finite buffer capacity. Server provides priority scheduled service for Type I and Type II packets

with preemptive discipline. That is, if Type I packet arrives, when Type II packet is in service, service process is interrupted, and after completion of the service, if there are no Type I packets, it starts processing of left over Type II packet as it is new one. Otherwise, it would go for another Type I packet. Assume that service process of the Type I and Type II packets follows continuous-time PH distributions denoted by (α, T) and (β, S) respectively, with same dimension p , where, α, β are vectors of size $1 \times p$, T, S are $p \times p$ matrices, and $t^0 = -Te, s^0 = -Se$. The mean service time of Type I and Type II packets are obtained by $\mu^I = -\alpha T^{-1}e$, $\mu^{II} = -\beta S^{-1}e$, respectively. Buffer capacity of system is taken to be N . Finally, the resultant queueing system of network node is MAP/PH/1/N queue with preemptive priority. Let$N_{II}(t)(N_I(t))$ be the number of Type II (Type I) packets in the system at time t, including packet in service. The thresholds for Type I and Type II packets are K_1, K_2 respectively, where N equals to $K_1 + K_2$. The arrival phase of superposed MAP at time t is denoted by $A(t)$, and service phase of system is denoted by $B(t)$. Therefore, arrival process of system is characterized by a multi-dimensional continuous time Markov chain $F(t) = \{N^I(t), N^{II}(t), A(t), B(t), t \geq 0\}$, The state space of is given by:

$$
\begin{aligned}
F_1 &= \{(0, 0, a, 0), a = 1, \ldots, n\} \\
F_2 &= \{(m_I, 0, a, b), m_{II} > 0; a = 1, \ldots, n; b = 1, \ldots, p\} \\
F_3 &= \{(0, m_{II}, a, b), m_{II} > 0; a = 1, \ldots, n; b = 1, \ldots, p\} \\
F_4 &= \{(m_I, m_{II}, a, b), m_I > 0, m_{II} > 0; a = 1, \ldots, n; b = 1, \ldots, p\}
\end{aligned}
$$

Here, F_1 represents idle state of server with arrival at phase at a. F_2 represent $m_I(> 0)$ Type I packets and no Type II packets in queue. F_3 represent $m_{II}(> 0)$ Type II packets and no Type I packet in queue. F_4 represent there are $m_I(> 0)$, $m_{II}(> 0)$ of Type I and Type II packets are in queue. In three cases, arrival is in phase a, and service is in phase b. If stages of $F(t)$ are arranged in lexicographical order. The level dependent block tridiagonal generator matrix of system occupancy at time t is given by

$$
Q(t) = \begin{bmatrix}
\overline{A_0(t)} & \overline{A_1(t)} & 0 & 0 & \ldots & 0 & 0 & 0 \\
\overline{B_2(t)} & A_0(t) & A_1(t) & 0 & \ldots & 0 & 0 & 0 \\
0 & A_2(t) & A_0(t) & A_1(t) & \ldots & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & 0 & \ldots & A_2(t) & A_0(t) & A_1(t) \\
0 & 0 & 0 & 0 & \ldots & 0 & A_2(t) & A_0(t) + A_1(t)
\end{bmatrix}
$$

where all block matrices in $Q(t)$ are square matrices of finite order, and are defined as follows,

$$
\overline{A_0(t)} = \begin{bmatrix}
D_0 & D_2(t) \otimes \beta & 0 & 0 & \ldots & 0 & 0 & 0 \\
I \otimes s^0 & D_0 \oplus S & D_2(t) \otimes I & \ldots & 0 & 0 & 0 & \\
0 & I \otimes s^0\beta & D_0 \oplus S & D_2(t) \otimes I & \ldots & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
0 & 0 & 0 & 0 & \ldots & I \otimes s^0\beta & D_0 \oplus S & D_2(t) \otimes I \\
0 & 0 & 0 & 0 & \ldots & 0 & I \otimes s^0\beta & M
\end{bmatrix}
$$

$$
A_0(t) = \begin{bmatrix}
D_0 \oplus T & D_1(t) \otimes I & 0 & 0 & 0 & 0 \\
0 & D_0 \oplus T & D_1(t) \otimes I & 0 & 0 & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\
0 & 0 & 0 & 0 & D_0 \oplus T & D_1(t) \otimes I \\
0 & 0 & 0 & 0 & 0 & D_0 \oplus T + D_1(t) \otimes I
\end{bmatrix}_{(K_2+1)\times(K_2+1)}
$$

$$\overline{A_1(t)} = \begin{bmatrix} D_1(t) \bigotimes \alpha & 0 & 0 & 0 \\ 0 & D_1(t) \bigotimes I & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & D_1(t) \bigotimes I \end{bmatrix}_{(K_2+1)\times(K_2+1)}$$

$$\overline{B_2(t)} = \begin{bmatrix} I \bigotimes t^0 & 0 & 0 & 0 \\ 0 & I \bigotimes t^0\alpha & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & I \bigotimes t^0\alpha \end{bmatrix}_{(K_2+1)\times(K_2+1)}$$

$$A_1(t) = diag[D_1(t) \bigotimes I, D_1(t) \bigotimes I, \ldots, D_1(t) \bigotimes I]_{K_2+1},$$

$$A_2(t) = diag[I \bigotimes t^0\alpha, I \bigotimes t^0\alpha, \ldots, I \bigotimes t^0\alpha]_{K_2+1},$$

where $M = D_0 \bigoplus S + D_2(t) \bigotimes I$, I is an identity matrix of an appropriate order and $\overline{A_0(t)}$ is of order $(K_2 + 1) \times (K_2 + 1)$.

3. Performance analysis

Let $\pi(t) = (\pi_0(t), \pi_1(t), \ldots, \pi_{K_1}(t))$ be transient state probability vector of $Q(t)$. That is, $\pi(t)$ satisfies (Stewart, 1994)

$$\frac{d}{dt}\pi(t) = Q(t)\pi(t) \tag{4}$$

$$\implies \pi(t) = \pi(0)\ exp\left(\int_0^t Q(x)\,dx\right) \tag{5}$$

By using Theorem. 2.4.3 in (Slavík, 2007), one can get

$$\pi(t) = \pi(0) \prod_{k=0}^{n} (I + Q(t_k)h) \tag{6}$$

where $h = t_k - t_{k-1}$, n is number of partitions of the interval $(0, t]$, and $\pi(0)$ is state probability vector at time $t = 0$. Here, each $\pi_j(t)$ is vector corresponding to the set of states with j Type I packets, and is in the form of $\pi_j(t) = (\pi_{j0}(t), \pi_{j1}(t), \ldots, \pi_{jK_2}(t)$. Each $\pi_{jk}(t)$ represents the probability that there exist j Type I packets, and k Type II packets are in the system. The performance measures are given as follows:
The Mean waiting time of Type I packets (Zhao *et al.*, 2015)

$$MWT_{TypeI} = \frac{E[N^I(t)]}{\lambda_m^I(t)} = \frac{1}{\lambda_m^I(t)} \sum_{j=1}^{K_1} j\pi_j(t)e \tag{7}$$

Let assume $Y = \sum_{i=1}^{K_1} \pi_j(t)$, and $Y = \{Y_{:0}(t), Y_{:1}(t), \ldots, Y_{:K_2}(t)\}$, where each $Y_{:m_{II}}(t)$ is a row vector corresponding to m_{II} Type II customers in the system. The Mean waiting time of Type II packets is

$$MWT_{TypeII} = \frac{E[N^{II}(t)]}{\lambda_m^{II}(t)} = \frac{1}{\lambda_m^{II}(t)} \left(\sum_{k=1}^{K_2} k\left(Y_{:k}(t) + \pi_{0k}(t)\right)e\right) \tag{8}$$

Since, buffer capacity is finite, If Type I (Type II) packet arrives, and finds that there are $K_1\,(K_2)$ packets in system, then the packet is lost. The loss probability of Type I and Type II packets (Zhao *et al.*, 2015) in small time duration Δ are, respectively

$$P_{loss}^{I} = \frac{1}{\lambda_m^I(t+\Delta)}\left[\pi_{K_1}(t)\left(\int_t^{t+\Delta}\left(D_1(x)\bigotimes I\right)\,dx\right)e\right] \tag{9}$$

$$P_{loss}^{II} = \frac{1}{\lambda_m^{II}(t+\Delta)}\left[(Y_{:K_2}(t)+\pi_{0K_2}(t))\left(\int_t^{t+\Delta}\left(D_2(x)\bigotimes I\right)\,dx\right)e\right] \tag{10}$$

4. Computational complexity

In this section, one can present computational complexity of performance measures (Malla Reddy and Ravi Kumar, 2016; Chen *et al.*, 2007; Wang *et al.*, 2000), namely, mean waiting time and packet loss probability of Type I and Type II packets, which are given in Eqs.(7-10). The complexity of $MWT_{TypeI}, MWT_{TypeII}$ is of the order $\mathcal{O}\ ((K_2+1)nm)$, $\mathcal{O}\ ((K_1+1)nm)$ respectively, due to it involves product of several row and column vectors. The complexity of $P_{loss}^{I}, P_{loss}^{II}$ is of the order $\mathcal{O}\ ((K_2+1)^2n^2m^2)$, $\mathcal{O}\ ((K_1+1)^2n^2m^2)$ respectively. But, the Eqs. (7-10) involves transient state probability vector of generator matrix $Q(t)$ (with dimensions$((K_1+1)(K_2+1)nm)$, which is obtained by using method of Product integrals, and it is given in Eq. 6. Since, the problem of finding state probability vector involves addition and product of matrix $Q(t)$ several times. The computational complexity of finding product $\prod_{i=0}^{n}(I+Q(t_i)h)$ is of the order $\mathcal{O}\left(((K_1+1)(K_2+1)nm)^{2.37}\right)$ (using Coppersmith-Winograd algorithm), and complexity of addition is of the order $\mathcal{O}((K_1+1)^2(K_2+1)^2n^2m^2)$. Therefore, the overall computation complexity of the algorithm according to Big-O analysis is of order $\mathcal{O}\left(((K_1+1)(K_2+1)nm)^{2.37}\right)$.

5. Numerical results

In this section, performance measures of the system are presented numerically. For arrival process of Type I and Type II packets, the numerical values given in Table 1 and 2 are used. The number superposed MMPPs are taken to be 2, and sinusoidal arrival rates are taken in the form of $a + b_j sint$, where a is whole arrival rate and b_j varies in between $(0, 1)$. For Type I packet arrivals transition rates are given in Table 1 and arrival rates are same for three samples (based on traffic parameters) of Type I packets, these are $\lambda_1^I(t) = 1 + 0.3 \times \sin t$, $\lambda_2^I(t) = 1 + 0.7 \times \sin t$. For Type II packet arrivals transition rates are given in Table 2, and arrival rates are same for three samples (based on traffic parameters), these are $\lambda_1^{II}(t) = 1 + 0.4 \times \sin t$, $\lambda_2^{II}(t) = 1 + 0.8 \times \sin t$. Assume that service distribution follows two phase distribution, i.e, Erlang distribution(E_2) with varying service rates. Figs. 1-6 show that waiting time and packet loss for Type I and Type II packets increases as traffic intensity increases at every time instant, and also represent that waiting time and packet loss increase as Hurst parameter (H) increases. From Figures 7 and 8, one can observe that mean waiting time increases, and packet loss decreases as threshold of Type I increases at every particular instant of time for $H = 0.9$.

Table 1: Values of traffic parameters and fitting parameters of Type I arrival rates

Sample	Parameters of Self-similar Input Traffic	r=2	
		c_{11}	c_{21}
Sample 1	$H = 0.7, \lambda_w(t) = 1, and\, \sigma^2 = 0.6$	0.196	0.001
Sample 2	$H = 0.8, \lambda_w(t) = 1, and\, \sigma^2 = 0.6$	0.0102	0.000188
Sample 3	$H = 0.9, \lambda_w(t) = 1, and\, \sigma^2 = 0.6$	0.005198	0.0005

Table 2: Values of traffic parameters and fitting parameters of Type II arrival rates

Sample	Parameters of Self-similar Input Traffic	r=2	
		d_{11}	d_{21}
Sample 1	$H = 0.7, \lambda_w(t) = 1, and\, \sigma^2 = 0.6$	0.23	0.0013
Sample 2	$H = 0.8, \lambda_w(t) = 1, and\, \sigma^2 = 0.6$	0.05	0.00092
Sample 3	$H = 0.9, \lambda_w(t) = 1, and\, \sigma^2 = 0.6$	0.003	0.000272

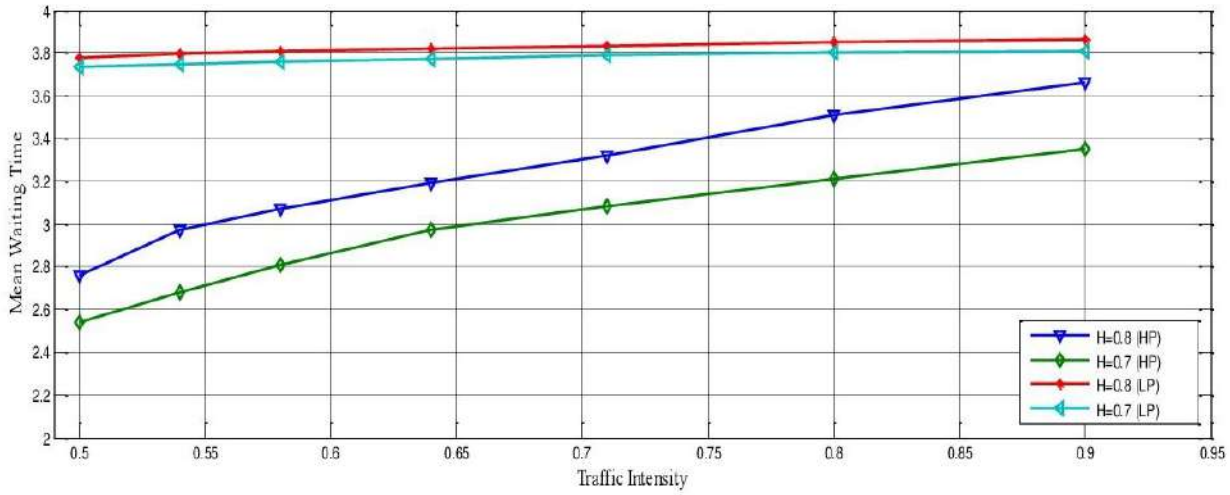


Figure 1: Traffic intensity vs mean waiting time with $N = 10, t = 1$

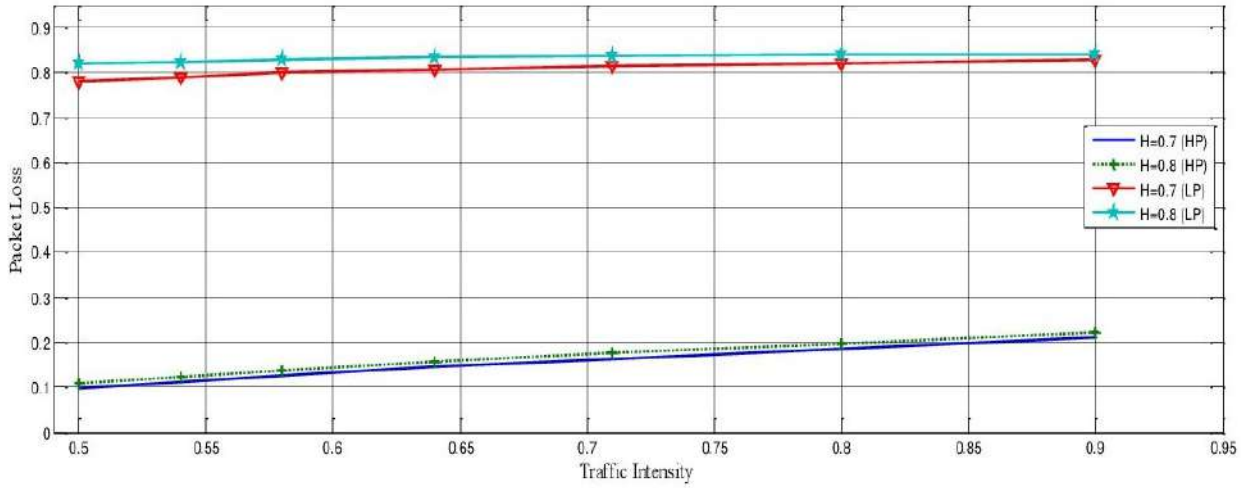


Figure 4: Traffic intensity vs packet loss with $N = 10, t = 3, \Delta = 0.5$

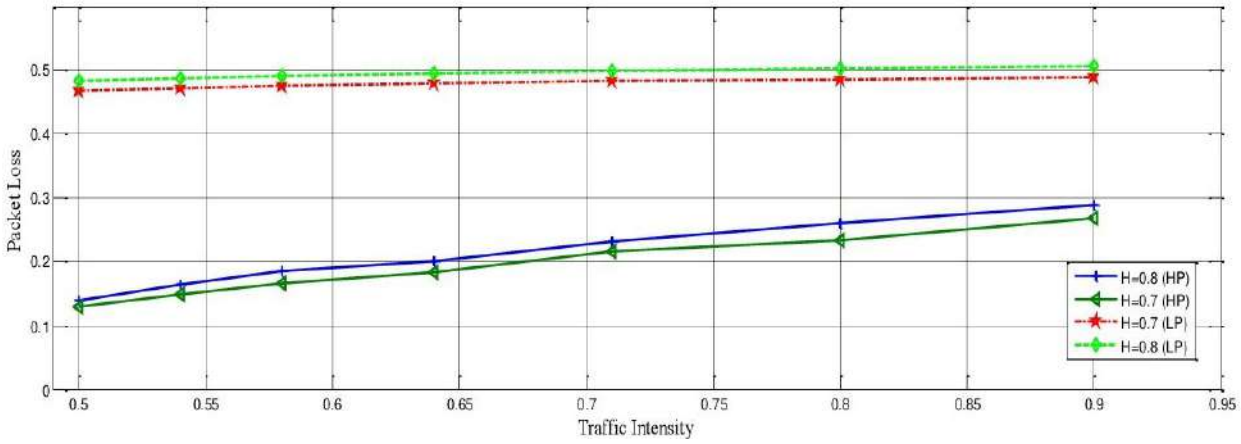


Figure 2: Traffic intensity vs packet loss with $N = 10, t = 1, \Delta = 0.5$

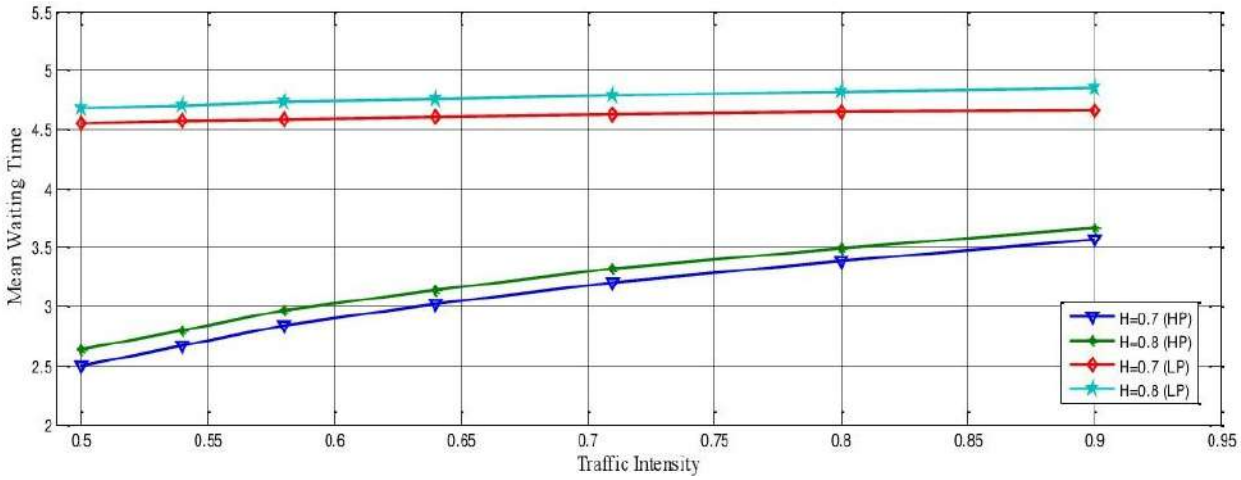


Figure 3: Traffic intensity vs mean waiting time with $N = 10, t = 3$

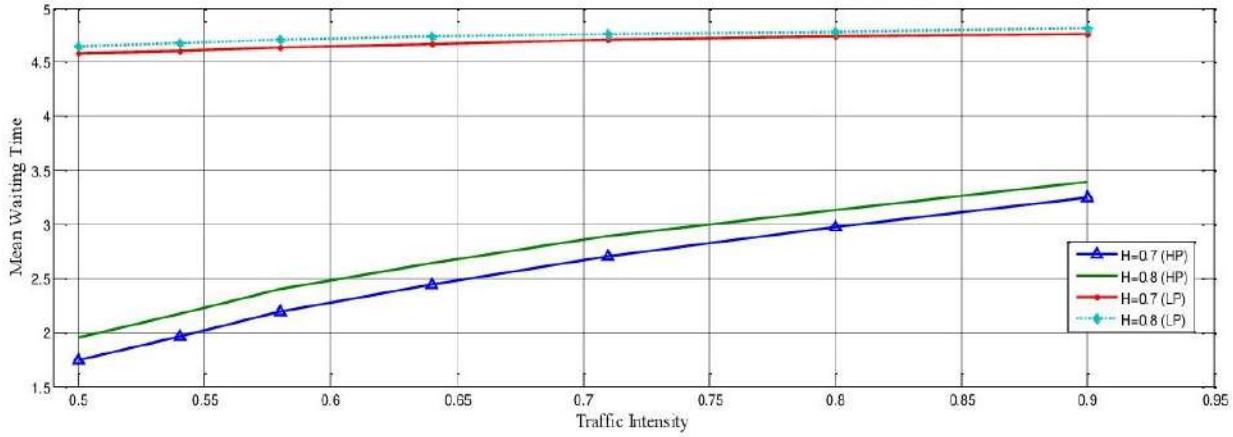


Figure 5: Traffic intensity vs mean waiting time with $N = 10, t = 5$

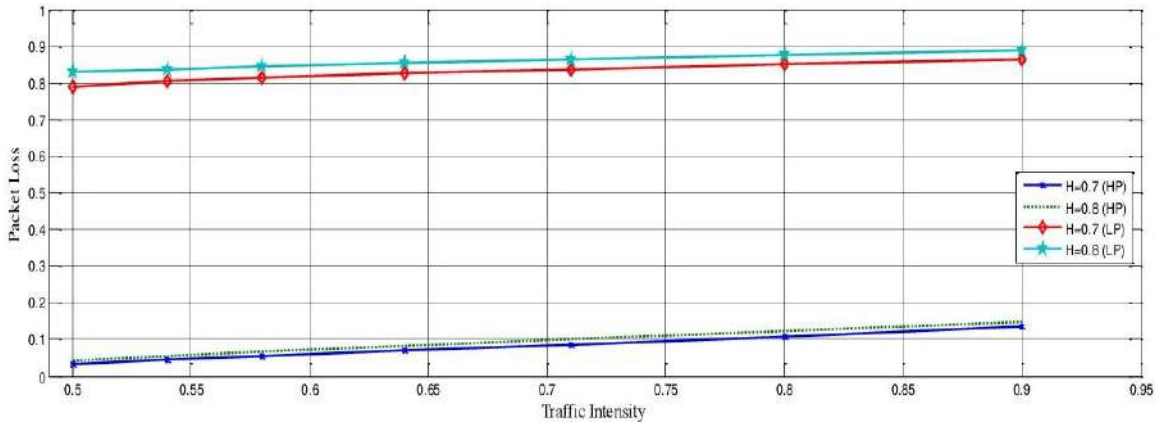


Figure 6: Traffic intensity vs packet loss of Type II packets with $N = 10, t = 5, \Delta = 0.5$

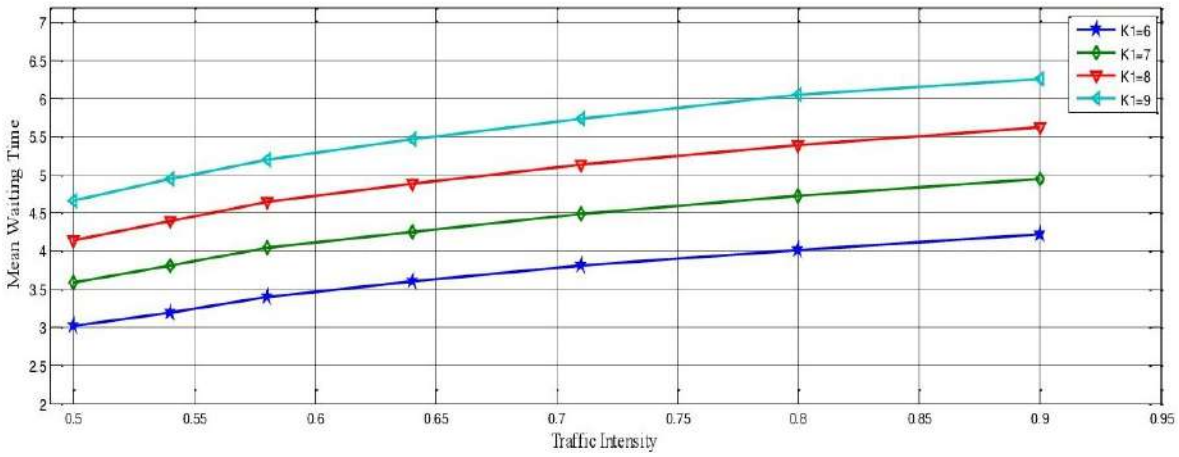


Figure 7: Traffic intensity vs mean waiting time of Type I packets with $t = 3, H = 0.9$

6. Conclusions

In this paper, network nodes with self-similar input priority based traffic are modeled into transient MAP/PH/1 queueing system, and its performance analysis is made by using level dependent quasi-birth and process with preemptive priority mechanism. The system is approximated by a finite buffer and transient state probability vector is obtained by the method of product integrals. Numerical results show that how traffic intensity, Hurst parameter, and threshold effects mean waiting time and packet loss of HP and LP packet arrivals at different time instants.

Acknowledgements

The authors wish to acknowledge Council of Scientific and Industrial Research (CSIR), Government of India, for their funding under the Major Research Project (MRP) scheme (File. No: 25(0301)/19/EMR-II).

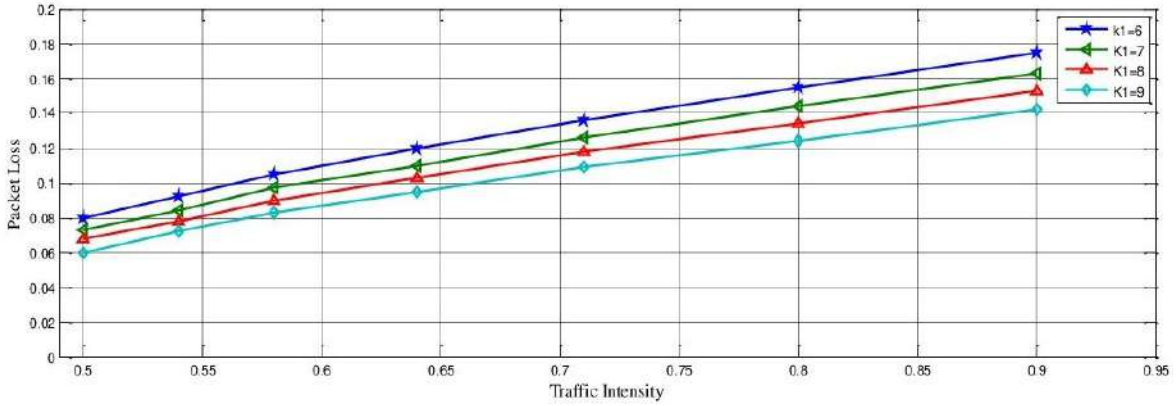


Figure 8: Traffic intensity vs packet loss of Type I packets with $t = 3, H = 0.9, \Delta = 0.5$

References

Abhilash, V. and Malla Reddy, P. (2022). Fitting model for self-similar traffic- time dependent markovian process and second order statistics. *Statistics and Applications*, **20**, 297–309.

Andersen, A. T. and Nielsen, B. F. (1998). A markovian approach for modelling packet traffic with long-range dependence. *IEEE journal of Selected Areas in Communications*, **16**, 719–732.

Boxma, O., Cohen, J., and Deng, Q. (1999). Heavy-traffic analysis of the m/g/1 queue with priority classes. *In: Proceedings of ITC*, **16**, 1157–1167.

Brahimi, M. and Worthington, D. (1991). Queueing models for out-patient appointment systems - a case study. *The Journal of the Operational Research Society*, **42**, 733–746.

Chen, C., Chang, C., Malla Reddy, P., Shao, S., and Wu, J. (2007). Performance analysis of wdm optical packet switches employing wavelength conversion under markovian modeled self-similar traffic input. *Workshop on High Performance Switching and Routing*, **1**, 1–6.

Cohen, M., Kleindorfer, and Lee, H. (1988). Service constrained (s,s) inventory systems with priority demand classes and lost sales. *Management Science*, **34**, 482–499.

Crovella, M. and Bestavros, A. (1997). *Self-similarity in World Wide Web traffic: evidence and possible causes*. IEEE.

Jin, X. and Min, G. (2007). Performance analysis of priority scheduling mechanisms under heterogeneous network traffic. *Journal of Computer and System Sciences*, **73**, 1207–1220.

Kleinrock, L. (1976). *Queueing Systems Volume II: Computer Applications*. John Wiley and Sons, New York.

Leland, W., Taqqu, M., Willinger, W., and Wilson, D. (1994). On the self-similar nature of ethernet traffic (extended version). *IEEE/ACM Transactions on Networking*, **2**, 1–15.

Malla Reddy, P. and Ravi Kumar, G. (2014). Inverstigating priority based performance analysis of optical packet switch under asynchronous self-similar variable length packet traffic input with voids. *Springer Proceedings in Mathematics and Statistics, Switzerland*, **92**, 413–425.

Malla Reddy, P. and Ravi Kumar, G. (2016). Performance analysis of wavelength division multiplexing asynchronous internet router employing space priority mechanism under self-similar traffic input-multi-server queueing system with markovian input and erlang-k services. *Applied Mathematics*, **07**, 1707–1725.

Malla Reddy, P. and Ravi Kumar, G. (2021). Performance analysis of asynchronous priority-based internet router under self-similar traffic input - queueing system with markovian input and hyper-exponential services. *International Journal of Operational Research*, **40**, 239–260.

Marks, B. (1973). State probabilities of M/M/1 priority queues. *Operations Research*, **21**, 974–987.

Miller, R. (1960). Priority queues. *Annals of Mathematical Statistics*, **31**, 86–103.

Paxson, V. and Floyd, S. (1995). Wide area traffic: The failure of poisson modelling. *IEEE/ACM Transactions on Networking*, **3**, 226–244.

Ravi Kumar, G., Raj Kumar, L., and Malla Reddy, P. (2017). Loss behaviour analysis of asynchronous internet switch under self-similar traffic input using MMPP/PH/c/K queueing system employing pbs mechanism. *International Journal of Communication Networks and Distributed Systems*, **19**, 257–269.

Sampath, K., Malla Reddy, P., and Adilakshmi, T. (2013). Performance study of WDM OPS with space priority mechanism under self-similar variable length input traffic. *Proceedings of IEEE ICON-2013 held at Singapore*, **1**, 1–6.

Sandhu, D. and Posner, M. (1989). A priority M/G/1 queue with application to voice/data communication. *European Journal of Operational Research*, **40**, 99–108.

Shao, S. K., Malla Reddy, P., Tsai, M. G., Tsao, H., and Wu, J. (2005). Generalized variance-based markovian fitting for self-similar traffic modelling. *IEEE transactions on communications*, **88**, 1493–1502.

Sharma, V. and Virtamo, J. (2002). A finite buffer queue with priorities. *Performance Evaluation*, **47**, 1–22.

Slavík, A. (2007). *Product Integration, its History and Applications*. Matfyz press.

Stewart, W. J. (1994). *Introduction to the Numerical Solution of Markov Chains*. Princeton University Press, Princeton, New Jersey.

Takada, H. and Miyazawa, M. (2002). A markov modulated fluid queue with batch arrivals and preemptions. *Stochastic Models*, **18**, 529–652.

Takagi, H. (1991). *Queueing Analysis: A Foundation of Performance Evaluation, Volume 1: Vacation and Priority Systems, Part 1*. North-Holland.

Takahashi, Y. and Miyazawa, M. (1994). Relationship between queue-length and waiting time distributions in a priority queue with batch arrivals. *Journal of the Operations Research Society of Japan*, **37**, 48–63.

Takine, T. and Hasegawa, T. (1994). The workload in the MAP/G/1 queue with state dependent services: its application to a queue with pre-emptive resume priority. *Communications in Statistics - Stochastic Models*, **10**, 183–204.

Tarabia, A. (2007). Two-class priority queueing system with restricted number of priority customers. *AEU-International Journal of Electronics and Communications*, **61**, 534–539.

Wang, Y. C., Liu, C. W., and Lu, C. C. (2000). Loss behavior in space priority queue with batch markovian arrival process-discrete-time case. *Performance Evolution*, **41**, 269–293.

White, H. and Christie, L. (1958). Queueing with pre-emptive priorities with breakdown. *Operations Research*, **6**, 79–95.

Yoshihara, T., Shoji, K., and Takahashi, Y. (2001). Practical time-scale fitting of self-similar traffic with markov modulated poisson process. *Telecommunication systems*, **17**, 185–211.

Zhao, N., Zhaotong, L., and Kan, W. (2015). Analysis of a MAP/PH/1 queue with discretionary priority based on service stage. *Asia-Pacific Journal of Operational Research*, **32**, 1–22.

Zhao, Y. and Alfa (1995). Performance analysis of a telephone system with both patient and impatient customers. *Telecommunication Systems*, **4**, 201–215.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 22, No. 2, 2024 (New Series), pp 181–187
http://www.ssca.org.in/journal

Variations of Wholesale Price of Wheat in Different States of India under COVID-19 Pandemic

Rashmi, H. P. Singh and P. K. Singh
Department of Agricultural Economics, Institute of Agricultural Sciences, Banaras Hindu University, Uttar Pradesh - 221005

Received: 18 December 2022; Revised: 27 November 2023; Accepted: 10 December 2023

Abstract

The present study investigates the impact of COVID-19 and restrictions imposed on wheat in different agricultural markets of India. The COVID-19 pandemic has had a significant impact on various sectors worldwide, including the food market. In India, the wheat crop harvest coincided with the lockdown imposed to control the spread of the virus. Monthly wholesale price data of seven states viz. Chhattisgarh, Uttar Pradesh, Madhya Pradesh, West Bengal, and Maharashtra were exercised from agricultural marketing portal of India. We compared monthly prices of April, May and June across 2019, 2020 and 2021. Linear piecewise regression was used to understand the impact COVID-19 on market whole sale price during different phases. The result revealed that wheat prices were at minimum support price in most of the states. Time series analysis showed the immediate impact of lockdown on decreased monthly wholesale price in all the states. Price risk was calculated using Cuddy Della Valle instability index (CDVI). Maharashtra showed the highest average monthly whole sale price and maximum price risk. The findings suggest that the agricultural markets have demonstrated a significant level of resilience in coping with the adverse effects of the COVID-19 pandemic. This is attributed to the provision of adequate policy support that has helped to mitigate the impact of the pandemic on the sector.

Key words: COVID-19; Whole sale Price; Wheat; Agricultural commodity; Price risk.

AMS Subject Classifications: 62K05, 05B05

1. Introduction

Undoubtedly, the COVID-19 pandemic and the subsequent nationwide lockdown enforced in March 2020 have caused a significant economic impact, affecting every sector of the economy, including the agricultural sector and its related markets. The agricultural industry in India, unlike many other countries, is a crucial source of livelihood, accounting for 60% of all rural employment (Varshney *et al.*, 2023). Therefore, it has been severely impacted by the pandemic, just like other sectors of the economy. The greatest current global problem the world has faced since World War II is the COVID-19 pandemic. It has been impacting

Corresponding Author: Rashmi
Email: rashrai.bhu@gmail.com

life and the economy across the world since December 2019. It has brought a threatening challenge to the Indian economy and society. The COVID-19 has affected all walks of life (Cariappa *et al.*, 2021). A nationwide lockdown of 21days was declared by the government of India with an aim to restrict the spread of Coronavirus. The effort of India to contain the spread of Coronavirus has been applauded worldwide (Varshney *et al.*, 2021). It stalled the economy across all enterprises, including agriculture. Consequently, the agricultural value chain during the initial phase of the lockdown faced a huge economic shock. This led to a serious detrimental effect on the health of the rural Indian economy. Agricultural marketing channel was also affected in a way. Unlike different countries across the globe, agricultural enterprises in India account for single largest source of employment generation with more than 60 percent of the population directly depending on agriculture.

Agriculture and allied sectors carry immense importance to the rural economy of India. It contributes nearly one-sixth to the Indian national income and provides employment to nearly 50% of the workforce. It is vital for ensuring food security of the nation and also influences the growth of secondary and tertiary sector of the economy through its forward and backward linkages. The COVID-19 pandemic has occurred at a time when the global and Indian economic growth was already expected to decelerate (NABARD 2020). The economic implications of the novel Coronavirus (COVID-19) pandemic have brought the agricultural sector into sharp focus and heightened its responsibility to feed and employ thousands who might have lost livelihoods. At this time when most sectors of the economy are reported to be under significant stress, the agricultural sector continues to be promising and cushioning the economy. The most important factor of the lockdown was the total breakdown in supply chain both at global and Indian scale. There was a decline in global exports of agricultural goods. During the lockdown there was no proper management of sowing, harvesting and marketing of crops. There were different restrictions: (a) disruptions in procurement of food grain by different agencies; (b) disruptions in assembling of harvests from farms by traders; (c) paucity of farm workers for harvesting of rabi crops; (d) unavailability of truck drivers; (e) barriers in the transport of commodities; (f) inadequate operations of APMC mandis and (g) closures in the retail markets.

In the present context we make effort to access the impact of spread of COVID-19 and lockdown on the wholesale price of wheat across different states of India (Rawal and Verma, 2020). The impact on the price may be conceptualised as combined effect of response from consumers, wholesalers and retailers through stakeholders. Price and quantities traded of different agricultural commodities whose harvesting begin from late march is very crucial to the liquidity of farmer and how their lives are being affected by the pandemic. Several researchers across the globe believe that the empirical evidence of COVID-19 pandemic on food and agricultural market is still evolving with time (Sendhil *et al.*, 2013). For example, Mahajan and Tomar (2020) reported that there was a decline of 10% in the accessibility of various commodity through online mode during the initial phase of lockdown (Ramakumar, 2020). They have also reported a decline of about 20% in market arrival of vegetable and fruits in few cities during lockdown months (March and April). The major reason for the decline in arrival of fruits and vegetables is disruption supply chain in agriculture market (Sharma *et al.*, 2021). A sharp increase in the retail and wholesale prices for various commodities including pulses and edible oils was witnessed immediately after the lockdown (Narayanan and Saha, 2020). They reported that movement restrictions were the prime reason and contributed in increased prices. We have taken 5 different states: Uttar Pradesh,

Madhya Pradesh, Gujarat, Maharashtra. We have focused are study on agricultural commodities. The present paper is organised in four sections. The first section presents a short description of COVID-19 pandemic and its impact on agriculture, few recent studies related to COVID-19 impact on agricultural commodity. The second section describes the data used and methodology applied in present study. The third section presents the result and analysis. The fourth section discusses the result and fifth section ends with conclusion.

2. Data and methodology

2.1. Average monthly wholesale price

Data used for the present research were assessed from Agriculture Marketing Information portal (Source: https://agmarknet.gov.in) of Indian government, which provide commodity wise state wise monthly average wholesale price. It is the price at which the grain markets or mandis sell wheat to the wholesalers, who in turn sell it to retailers or food processing industries. The wholesale price of wheat is influenced by various factors such as demand and supply, production, transportation costs, and government policies. It is an important indicator of the overall performance of the wheat market in India. The study period includes January 2019 to June 2021. In the entire study period first wave and second phase of COVID-19 disrupted the agricultural marketing chain which entirely affected the consumer behaviour and price of different agricultural commodities. For the entire study period wheat wholesale price data for 5 states (Chhattisgarh, Gujarat, Madhya Pradesh, Uttar Pradesh and Rajasthan) were analysed.

3. COVID-19 events in India

Lockdown during COVID-19 has impacted agriculture marketing chain in different ways in different phases of lockdown, which started from the end of march 2020. Table 2 represents the Descriptive statistics of monthly average whole sale price. To analyse the impact of COVID-19 lockdown on the wholesale price of agricultural commodities categorisation of period is very important. The different phases of the lockdown along with the activities exempted during each period is summarised in Table 1.

Table 1: Lockdown and unlock timelines and activities allowed

Lockdown	Duration	Activities allowed
Phase-1	25th March to 14th April 2020	Nearly all activities were suspended
Phase-2	15th April to 3rd May 2020	Allowed agricultural activities starting 20th April 2020
Phase-3	4th May to 17th May 2020	Lockdown in Green, Orange and Red zones
Phase-4	18th May to 31st May 2020	Movement allowed with some conditions across districts and states.
Unlock 1	1st June to 30th June	Reopening phase with an economic focus.

Source: Ministry of Home Affairs, Govt. of India

4. Methodology

The price variation during the lockdown period and normal year has been analysed using piecewise linear regression tool. This analysis is used if the data follows different linear trend over different time segment. The piecewise linear regression can be conceptualised as follows:

$$\begin{aligned} y(x) = n_1\beta_1(x - b_1), \qquad & b_1 < x \leq b_2 \\ n_2 + \beta_2(x - b_2), \qquad & b_2 < x \leq b_3 \\ n_n + \beta_{nb}(x - b_{nb-1}), \qquad & b_{nb-1} < x \leq b_{nb} \end{aligned}$$

Where b_1 is the x location of first break point, b_2 is the x location of second break point, and so forth until the last break point bnb for nb number of break points.

4.1. Price risk

Cuddy Della Valle instability index (CDVI) (Cuddy and Valle, 1978), represented the modified form of coefficient of variation which capture the price risk. CDVI has been used in this study to analyse the risk in monthly wholesale price for before and after lockdown period. It can be computed as follows:

$$CDVI = CV\sqrt{(1 - R^2)}$$

where, CV is coefficient of variation and R^2 is coefficient of determination.

4.2. Percentage change

We used PC to analyse the impact of lockdown on wholesale price of wheat. It is a simple mathematical concept that represents the degree of change over time. The value of percentage change is positive then there is a increase in percentage of that unit over time, while negative value shows a depicts a decrease in percentage over time. It can be calculated using the following formula:

$$\%change = \frac{\text{(price during lockdown-price before lockdown)}}{\text{(price before lockdown)}}\text{x100}$$

5. Results and discussion

5.1. Monthly wholesale price

The monthly time series wholesale price data of wheat for all six states are represented in Figure 1. The wholesale price witnessed decrease after March 2020 for all the states. Maharashtra showed highest wholesale price after Lockdown, while Chhattisgarh showed the least wholesale price trend.

5.2. Descriptive statistics

The descriptive statistics of monthly average whole sale price is summarised in Table 2. The mean monthly wholesale price before lockdown was higher than that of after lockdown

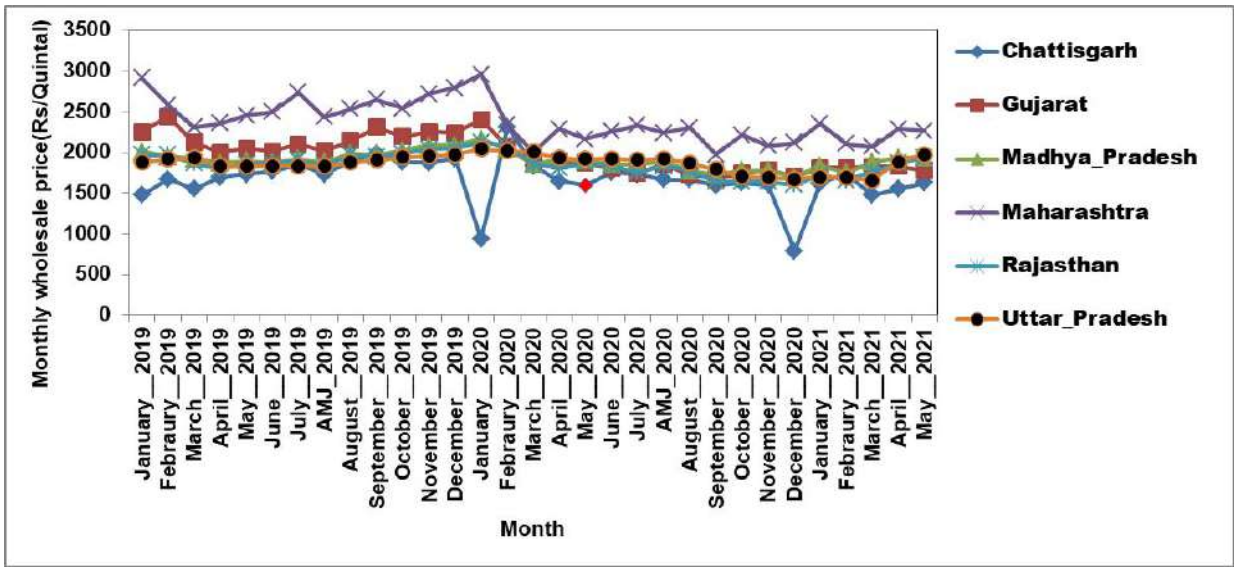


Figure 1: Time series variation of average monthly wholesale price of wheat

for all the states. The highest average monthly whole sale price was found for Maharashtra among different states. Chhattisgarh showed least whole sale price among six states after lockdown. Different states showed different statistics trend in India. In the present study difference in monthly whole sale price for different region is attributed to disruption in supply chain management in different region due to COVID-19 pandemic.

Table 2: Descriptive statistics of monthly average whole sale price

	Lock-down	Chhatt-isgarh	Gujarat	Madhya Pradesh	Mahara-shtra	Rajasthan	Uttar Pradesh
Mean	Before	1769.94	2161.94	1956.85	2576.21	1925.53	1886.45
	After	1596.54	1856.46	1863.10	2248.81	1803.69	1873.65
Maximum	Before	1981.83	2438.54	2088.29	2908.84	2047.86	1975.11
	After	2314.00	2399.09	2163.06	2951.27	2125.06	2039.25
Minimum	Before	1476.88	2000.36	1868.54	2305.86	1793.75	1826.45
	After	788.90	1680.44	1695.45	1969.46	1602.00	1672.04
Sd	Before	147.45	131.60	76.81	176.54	78.36	52.47
	After	378.76	192.65	127.98	242.78	160.32	124.80
Skewness	Before	-0.611	0.572	0.545	0.320	-0.051	0.227
	After	-0.779	2.12	1.101	2.08	0.769	-0.536
Kurtosis	Before	-0.191	-0.142	-0.890	-0.548	-1.021	-1.396
	After	2.030	5.161	1.41	6.18	0.196	-0.922

5.3. Price risk

The price risk was analysed before and after COVID-19 is presented in Figure 2. All states except Uttar Pradesh showed higher price risk after COVID-19. Maharashtra showed highest price risk, while Uttar Pradesh showed least price risk among different states.

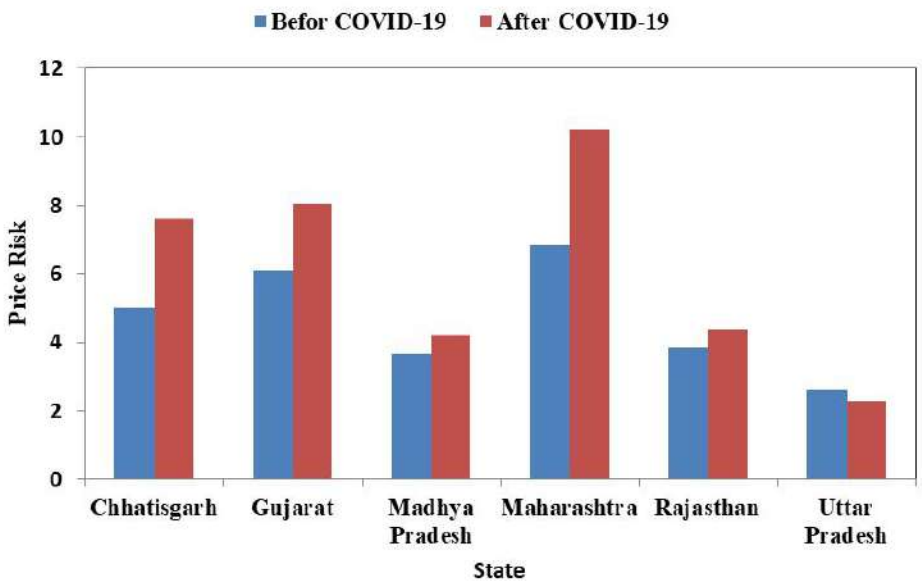


Figure 2: Price risk before and after lockdown

5.4. Percent change (%)

To understand the impact of lockdown on average whole sale price, we have calculated percent change for all states. The percent change analysis is presented in Figure 3. The result revealed that all states showed percent decrease in average monthly whole sale price compared to normal year 2019. Gujarat showed highest percentage decrease in average monthly wholesale price, while Uttar Pradesh showed least percent decrease among different states considered in the present study.

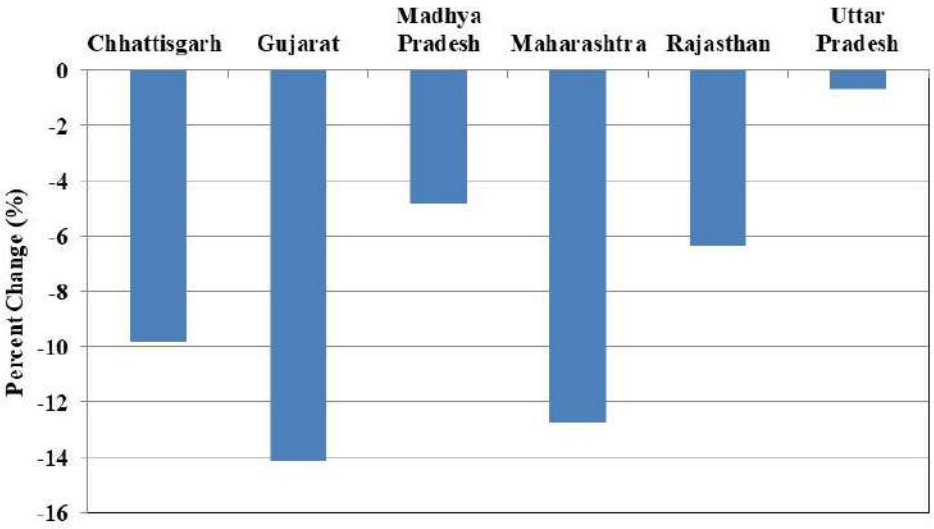


Figure 3: Percent change analysis of different states

6. Conclusion

In the present study an attempt was made to capture the impact of COVID-19 on average monthly whole sale price of wheat across different state in India. Time series analysis showed the immediate impact of lockdown on decreased monthly wholesale price in all the states. Price risk was calculated using Cuddy Della Valle instability index (CDVI) also that there was increase in price risk after lockdown across all the states, except Uttar Pradesh. Measures taken by Government of India after lockdown has been reflected in increase in monthly wholesale price of wheat in post lockdown months. The percentage change analysis showed clear impact of lockdown on average monthly wholesale price of wheat in comparison to normal year (2019).

References

Cariappa, A. A., Acharya, K. K., Adhav, C. A., Sendhil, R., and Ramasundaram, P. (2021). Impact of covid-19 on the indian agricultural system: A 10-point strategy for post-pandemic recovery. *Outlook on Agriculture*, **50**, 26–33.

Cuddy, J. D. and Valle, P. D. (1978). Measuring the instability of time series data. *Oxford Bulletin of Economics and Statistics*, **40**, 79–85.

Mahajan, K. and Tomar, S. (2020). Here today, gone tomorrow: Covid-19 and supply chain disruptions. *Forthcoming American Journal of Agricultural Economics*, **1**, 1–40.

Narayanan, S. and Saha, S. (2020). Urban food markets and the lockdown in india. *Available at SSRN 3599102*, **1**, 1–27.

Ramakumar, R. (2020). Farmers are at their wit's end. The Hindu.

Rawal, V. and Verma, A. (2020). Agricultural supply chains during the covid-19 lockdown. *SSER Monograph*, **20**, 1–26.

Sendhil, R., Kar, A., Mathur, V., and Jha, G. K. (2013). Price discovery, transmission and volatility: Evidence from agricultural commodity futures. *Agricultural Economics Research Review*, **26**, 41–54.

Sharma, J., Tyagi, M., and Bhardwaj, A. (2021). Exploration of covid-19 impact on the dimensions of food safety and security: a perspective of societal issues with relief measures. *Journal of Agribusiness in Developing and Emerging Economies*, **11**, 452–471.

Varshney, D., Kumar, A., Mishra, A. K., Rashid, S., and Joshi, P. K. (2021). India's covid-19 social assistance package and its impact on the agriculture sector. *Agricultural Systems*, **189**, 103049.

Varshney, D., Roy, D., and Meenakshi, J. (2023). Impact of covid-19 on agricultural markets: assessing the roles of commodity characteristics, disease caseload, and market reforms. In *Contextualizing the COVID Pandemic in India: A Development Perspective*, pages 249–271. Springer.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 22, No. 2, 2024 (New Series), pp 189–202
http://www.ssca.org.in/journal

Joint Importance Measures for Repairable Multistate Systems

V. M. Chacko, Ann Sania and Amrutha M.
Department of Statistics
St. Thomas College(Autonomous), Thrissur, Kerala, India-680001.

Received: 21 April 2023; Revised: 25 November 2023; Accepted: 14 December 2023

Abstract

In order to identify the vulnerable components whose joint effect would have changed system performance and ensure the required reliability of various multistate systems, joint importance measures of relevant components are used in the early design of systems. Due to the complexity of multistate systems that have the properties of non-linearity, uncertainty, and randomness, which make it difficult to analyze the reasons of failure mechanisms, model the system, estimate its reliability, and evaluate the joint importance measures of its components. This paper discussed measures of joint importance of three components for repairable multistate systems based on the classical Birnbaum measure. Eight importance measures are studied in detail. These joint importance measures provide a time-dependent analysis of the relevancy of components, thus adding insights on the contributions of the joint effect of three components on the system reliability or performance over time. An illustrative example is given. The results of the study show that joint importance measures can be a valuable decision-support tool for designers and engineers in the design of systems.

Key words: Birnbaum importance; Multistate systems; Repairable components; Joint importance measures.

1. Introduction

The evaluation of joint importance measures for identifying group of relevant components in complex systems is a major concern of reliability engineers and designers. Importance and joint importance measures are widely used to identify the impact and locate the vulnerable spots at the early design stages. The identification of most important component or group of components in a repairable multistate complex system by investigating the improvement resulted in performance measures like reliability or availability or unreliability/risk or unavailability etc with the improvement in corresponding component performance measures, is to be addressed in detail in situation where minor repair or replacement after complete failure of components admits. The concept of the joint importance measures of components or subsystems or modules is crucial, in order to ensure and improve the product

Corresponding Author: Dr. V. M. Chacko
Email: chackovm@gmail.com

quality, reliability, and safety. This is also essential for allocating the limited resources at the design stage, to reduce the cost and providing maintenance to take proper care of crucial components at the operation stage. The information provided by the joint importance measures can be used to give proper repair/replacement activities to the components. Thereby one can ensure the system performance in high level always for the continuous supply of service for the allocated mission. Quantifying the joint importance of components using an efficient method becomes essential in multistate systems, at the early design stage. However, the determination of the relevant components or subsystems at the early stage is challenging, because it is usually difficult to analyze and describe non-linear dependent relationships of components in complex systems, and obtain sufficient reliability information from the joint operational condition of components or system, Borgonovo and Apostolakis (2001).

The development of importance measures and its use can be seen in Birnbaum (1969), Fussell and Vesely (1972), Barlow and Proschan (1975) and Natvig (1985), see also Natvig (1979) and Natvig and Gåsemyr (2009). Since these measures solely depend on the probabilistic characteristics of the system's components and its structure, these traditional measures of importance can be characterized as generic. In power generation system, communication systems, network systems, the multistate approach can be adopted. Fundamental results on multistate system(MSS)s is available in Griffith (1980). The extensions of the Birnbaum measure for binary state systems to the multistate case can be seen in Dui *et al.* (2019). Natvig *et al.* (2011) and Natvig *et al.* (2009) studied on Importance measures for repairable systems. Algorithm for solution of a problem of maximum flow in a network with power estimation is given by Dinic (1970). Dui *et al.* (2015) has given semi-Markov process-based integrated importance measures for multi-state systems. Borgonovo and Apostolakis (2001) discussed a new importance measure for risk-informed decision making. Optimization of linear consecutive-k-out-of-n system with a Birnbaum importance-based genetic algorithm is given by Cai *et al.* (2016). Cai *et al.* (2017) discussed maintenance optimization of continuous state systems. Huseby and Natvig (2009) has introduced advanced discrete simulation methods applied to repairable multi-state systems. Huseby and Natvig (2013) has given discrete event simulation methods applied to advanced importance measures of repairable components in multistate network flow systems. Importance and sensitivity analysis of multistate systems using the universal generating function is carried out by Levitin and Lisnianski (1999). Generalized importance measures for multistate elements based on performance level restrictions can be seen in Levitin *et al.* (2003). Natvig (2011) has given a detailed description of multistate systems reliability theory with applications. Natvig *et al.* (2009) has given application of Natvig measures of component importance in repairable systems. Ramirez-Marquez and Coit (2005) introduced new composite importance measures for multi-state systems with multistate components. Ramirez-Marquez and Coit (2007) explained Multi-state component relevancy analysis for reliability improvement in multi-state systems. Ramirez-Marquez *et al.* (2006) has given new ideas on multi-state component relevancy and importance. Si *et al.* (2012b) proposed Integrated importance measure of component states based on loss of system performance. Si *et al.* (2012a) discussed the integrated importance measure of multistate coherent systems for maintenance processes. Si *et al.* (2013) has introduced component state-based integrated importance measure for multi-state systems. Si *et al.* (2019) proposed system reliability allocation and optimization based on generalized Birnbaum importance measure. Wu and Coolen (2013) has given a cost-based importance measure for system components: an extension of the Birnbaum

importance. Wu *et al.* (2016) used component importance to optimization of preventive maintenance policy.Zhu *et al.* (2017) discussed Birnbaum importance based heuristics for multi-type component assignment problems. Monte-Carlo simulation analysis of the effects on deferent system performance levels on the importance on multistate components is given by Zio and Podofillini (2003). Zio and Podofillini (2006) discussed components interactions in the differential importance measure. Zio *et al.* (2004) described estimation of the importance measures of multistate elements by Monte-Carlo simulation. Zio *et al.* (2007) has given an example in railway industry of importance measures-based prioritization for improving the performance of multi-state systems. Dui *et al.* (2019) proposed system performance-based joint importance analysis guided maintenance for repairable systems. Dui *et al.* (2020) introduced component joint importance measures for maintenances in submarine blowout preventer system. A detailed study on joint importance measures for unrepairable systems can be seen in Chacko and Manoharan (2008, 2011), Chacko (2020, 2023a) and Chacko (2023b).

The investigation of component joint performance with regard to the variation in system performance is crucial for the repair or replacement activities (Chacko (2022)). Existing Joint importance measures for components in multistate systems are used to identify group of components for unrepairable components and systems (Chacko (2022, 2021)). But, sometimes, systems are repairable or its components can be repaired/replaced as a cost effective strategy. The main objective of this paper is to study on joint importance measures for three components of repairable systems which are defined in the Birnbaum sense, a method of observing change in system performance with respect to change in component performance. Moreover, a multistate behavior to the components is assumed.

In the present paper, generic joint importance measures for three components of a repairable systems are studied in detail, which measure the interaction effect of three repairable components. Each component is assumed to follows periodic life cycles, starting out in the top state, say $M_i, i = 1, 2, \ldots, n$ and then moving through the intermediate states $k, M_i > k > 0$, until they reaches down state 0. Then, they are repaired or replaced, and a new life cycle starts. Moreover, repair at intermediate states is also assumed. Component i is allowed to have minor repair at state $k, M_i > k > 0$, to reach to state $k + 1$. If the component reaches the state 0, it will undergo corrective maintenance or replacement to bring the component to as good as new condition.

The present paper includes four sections. In section 2, the new joint importance measures are discussed. Applications are given in section 3. An illustrative example is given in section 4. Conclusions are given in final section.

2. Relevancy and importance in multistate systems

In a binary system setup, the Birnbaum-importance(B-importance) of component i (Birnbaum (1969)) is the probability that i^{th} component is relevant for the system. That is

$$I_B(i; p) = Pr\{\phi(X) = 1|X_i = 1\} - Pr\{\phi(X) = 1|X_i = 0\} \tag{1}$$

This measure is generic since it is defined based on probability and system structure function. Here we consider joint importance measures for three repairable components in MSS setup

based on B –importance. For that, a multistate system of n components is considered. Joint importance measures for three components of the multistate system are discussed.

Let $X(t) = (X_1(t), X_2(t), ..., X_n(t))$ be the state vector of n components and $\phi(X(t))$ represent state of the system, where $X_i(t)$ represent state of the component i at time t, $X_i(t)$ takes the values in $S_i = \{0, 1, \ldots, M_i\}, i \in \{1, 2, \ldots, n\}$. That is,

$$\phi(X(t)) = \phi(X_1(t), ..., X_n(t)) = k, k \in \{0, 1, 2, \ldots, M\}, M = \text{Max}_{\{1 \le i \le n\}}\{M_i\}.$$

where $\phi(x_i, X(t)) = \phi(X_1(t), ..., X_{i-1}(t), x_i, X_{i+1}(t), ..., X_n(t))$,
i.e,$(x_i, X(t)) = (X_1(t), ..., X_{i-1}(t), x_i, X_{i+1}(t), ..., X_n(t))$.

We also define a function $f_i : S_i \to R$ that represents the component's physical state at time t given by $f_i(X_i(t)) = f_i(x_i)$ if $X_i(t) = x_i \in Si, i \in \{1, 2, \ldots, n\}$. For example f_i represents the flow capacity of the component in a network system. It is important to understand that the functions $f_i, i \in \{1, 2, \ldots, n\}$, do not necessarily have to be non-decreasing and hence provide modeling flexibility by avoiding this restriction.

To define the relevancy of the repairable components in system functioning, let us define two functions $X_i^+(t)$ and $X_i^-(t)$, for $i = 1, 2, \ldots, n$.

$X_i^+(t)$ denotes the next state of component i and is defined by

$$X_i^+(t) = X_i(t) - 1, \ if \ X_i(t) > 0 \ \ and \ \ = M_i, \ if \ X_i(t) = 0 \tag{2}$$

Since it's a repairable periodic cycle component, on reaching state 0, the component is repaired. Hence its next state at time t from 0 will be M_i.

Similarly, we define $X_i^-(t)$ as the previous state of component i and is given by

$$X_i^-(t) = X_i(t) + 1, \ if \ X_i(t) < M_i \ \ and \ \ = 0, \ if \ X_i(t) = M_i \tag{3}$$

Here when the component at time t is in the highest possible state M_i, it implies that the previous state will be 0, since the component was repaired. A component is said to be in n-relevant or p-relevant at time t, if there is change in system state when component move either to next state by gradual degradation or to previous state at time t by minor maintenance. That is, component is said to be n-relevant, while component moves to its next state at time t if

$$\phi(X_i(t), \mathbf{X}(t)) \neq \phi(X_i^+(t), \mathbf{X}(t)) \ \ or \ \ \phi(X_i(t), \mathbf{X}(t)) - \phi(X_i^+(t), \mathbf{X}(t)) \neq 0. \tag{4}$$

Similarly, we say that component i is p-relevant while component i moves back to its previous state at time t if

$$\phi(X_i^-(t), \mathbf{X}(t)) \neq \phi(X_i(t), \mathbf{X}(t)) \ \ or \ \ \phi(X_i^-(t), \mathbf{X}(t)) - \phi(X_i(t), \mathbf{X}(t)) \neq 0. \tag{5}$$

We define $\phi_i(t) = \phi(X_i(t), X(t)), \phi_i^+(t) = \phi(X_i^+(t), X(t))$ and $\phi_i^-(t) = \phi(X_i^-(t), X(t))$.

Then component i is n-relevant if, $nREL(i) = \phi_i^+(t) - \phi_i(t)$ is not equal to zero. Hence, component i is n-relevant at time t if it would result in a system state change, while changing the component i to its next state.

Then we denote component i is p-relevant if say, $pREL(i) = \phi_i(t) - \phi_i^-(t)$ is not equal to zero. Hence, component i is p-relevant at time t if it would result in a system state change, while changing the component i back to its previous state.

For an easier representation, define the following functions,

$$\phi_{ij}^{**}(t) = \phi(X_i(t), X_j(t), X(t)), \phi_{ij}^{+*}(t) = \phi(X_i^+(t), X_j(t), X(t)),$$
$$\phi_{ij}^{*+}(t) = \phi(X_i(t), X_j^+(t), X(t)), \phi_{ij}^{++}(t) = \phi(X_i^+(t), X_j^+(t), X(t)),$$
$$\phi_{ij}^{+-}(t) = \phi(X_i^+(t), X_j^-(t), X(t)), \phi_{ij}^{*-}(t) = \phi(X_i(t), X_j^-(t), X(t))$$
$$\phi_{ij}^{-+}(t) = \phi(X_i^-(t), X_j^+(t), X(t)), \phi_{ij}^{-*}(t) = \phi(X_i^-(t), X_j(t), X(t))$$
$$\phi_{ij}^{--}(t) = \phi(X_i^-(t), X_j^-(t), X(t)), \phi_{ijk}^{***}(t) = \phi(X_i(t), X_j(t), X_k(t), X(t))$$
$$\phi_{ijk}^{+**}(t) = \phi(X_i^+(t), X_j(t), X_k(t), X(t)), \phi_{ijk}^{*+*}(t) = \phi(X_i(t), X_j^+(t), X_k(t), X(t))$$
$$\phi_{ijk}^{++*}(t) = \phi(X_i^+(t), X_j^+(t), X_k(t), X(t)), \phi_{ijk}^{**+}(t) = \phi(X_i(t), X_j(t), X_k^+(t), X(t))$$
$$\phi_{ijk}^{+*+}(t) = \phi(X_i^+(t), X_j(t), X_k^+(t), X(t)), \phi_{ijk}^{*++}(t) = \phi(X_i(t), X_j^+(t), X_k^+(t), X(t))$$
$$\phi_{ijk}^{+++}(t) = \phi(X_i^+(t), X_j^+(t), X_k^+(t), X(t)), \phi_{ijk}^{-**}(t) = \phi(X_i^-(t), X_j(t), X_k(t), X(t))$$
$$\phi_{ijk}^{*-*}(t) = \phi(X_i(t), X_j^-(t), X_k(t), X(t)), \phi_{ijk}^{--*}(t) = \phi(X_i^-(t), X_j^-(t), X_k(t), X(t))$$
$$\phi_{ijk}^{**-}(t) = \phi(X_i(t), X_j(t), X_k^-(t), X(t)), \phi_{ijk}^{-*-}(t) = \phi(X_i^-(t), X_j(t), X_k^-(t), X(t))$$
$$\phi_{ijk}^{*--}(t) = \phi(X_i(t), X_j^-(t), X_k^-(t), X(t)), \phi_{ijk}^{---}(t) = \phi(X_i^-(t), X_j^-(t), X_k^-(t), X(t))$$

Suppose, at time t, simultaneously i^{th} component is changing to its next state and j^{th} component is also changing to its next state. Then their states are jointly nn-relevant, if

$$nnREL(i,j) = \phi_{ij}^{**}(t) - \phi_{ij}^{+*}(t) - \phi_{ij}^{*+}(t) + \phi_{ij}^{++}(t) \neq 0 \tag{6}$$

Suppose i^{th} component is changing to its next state and j^{th} component is changing back to its previous state. Then components i and j are jointly np-relevant, if

$$npREL(i,j) = \phi_{ij}^{*-}(t) - \phi_{ij}^{+-}(t) - \phi_{ij}^{**}(t) + \phi_{ij}^{+*}(t) \neq 0 \tag{7}$$

Suppose i^{th} component is changing back to its previous state and j^{th} component is changing to its next state. Then components i and j are jointly pn-relevant, if

$$pnREL(i,j) = \phi_{ij}^{-*}(t) - \phi_{ij}^{**}(t) - \phi_{ij}^{-+}(t) + \phi_{ij}^{*+}(t) \neq 0 \tag{8}$$

Suppose i^{th} component is changing back to its previous state and j^{th} component is also changing back to its previous state. Then components i and j are jointly pp-relevant, if

$$ppREL(i,j) = \phi_{ij}^{--}(t) - \phi_{ij}^{*+}(t) - \phi_{ij}^{-*}(t) + \phi_{ij}^{**}(t) \neq 0 \tag{9}$$

Now, to measure the effect of joint movement of three components, in either direction, let us consider the following statements.

Suppose, at time t, i^{th} component is changing to its next state, j^{th} component is also changing to its next state and k^{th} component is also changing to its next state. Then components i, j and k are jointly *nnn*-relevant, if

$$nnnREL(i,j,k) = \phi_{ijk}^{***}(t) - \phi_{ijk}^{+**}(t) - \phi_{ijk}^{*+*}(t) + \phi_{ijk}^{++*}(t) - \phi_{ijk}^{**+}(t) + \phi_{ijk}^{+*+}(t) + \phi_{ijk}^{*++}(t) - \phi_{ijk}^{+++}(t) \neq 0 \quad (10)$$

Suppose at time t, i^{th} component is changing to its next state, j^{th} component is changing back to its previous state and k^{th} component is changing to its next state. Then components i, j and k are jointly *npn*-relevant, if

$$npnREL(i,j,k) = \phi_{ijk}^{*-*}(t) - \phi_{ijk}^{+-*}(t) - \phi_{ijk}^{***}(t) + \phi_{ijk}^{+**}(t) - \phi_{ijk}^{*-+}(t) + \phi_{ijk}^{+-+}(t) + \phi_{ijk}^{**+}(t) - \phi_{ijk}^{+*+}(t) \neq 0 \quad (11)$$

Suppose i^{th} component is changing back to its previous state, j^{th} component is changing to its next state and k^{th} component is changing to its next state. Then components i, j and k are jointly *pnn*-relevant, if

$$pnnREL(i,j,k) = \phi_{ijk}^{-**}(t) - \phi_{ijk}^{***}(t) - \phi_{ijk}^{-+*}(t) + \phi_{ijk}^{*+*}(t) - \phi_{ijk}^{-*+}(t) + \phi_{ijk}^{**+}(t) + \phi_{ijk}^{-++}(t) - \phi_{ijk}^{*++}(t) \neq 0 \quad (12)$$

Suppose i^{th} component is changing back to its previous state, j^{th} component is also changing back to its previous state and k^{th} component is changing to its next state.Then components i, j and k are jointly *ppn*-relevant, if

$$ppnREL(i,j,k) = \phi_{ijk}^{--*}(t) - \phi_{ijk}^{*-*}(t) - \phi_{ijk}^{-**}(t) + \phi_{ijk}^{***}(t) - \phi_{ijk}^{--+}(t) + \phi_{ijk}^{*-+}(t) + \phi_{ijk}^{-*+}(t) - \phi_{ijk}^{**+}(t) \neq 0 \quad (13)$$

Suppose, at time t, i^{th} component is changing to its next state, j^{th} component is also changing to its next state and k^{th} component is changing back to its previous state. Then components i, j and k are jointly *nnp*-relevant, if

$$nnpREL(i,j,k) = \phi_{ijk}^{**-}(t) - \phi_{ijk}^{+*-}(t) - \phi_{ijk}^{*+-}(t) + \phi_{ijk}^{++-}(t) - \phi_{ijk}^{***}(t) + \phi_{ijk}^{+**}(t) + \phi_{ijk}^{*+*}(t) - \phi_{ijk}^{++*}(t) \neq 0 \quad (14)$$

Suppose, at time t, i^{th} component is changing to its next state, j^{th} component is changing back to its previous state and k^{th} component is changing back to its previous state. Then components i, j and k are jointly *npp*-relevant, if

$$nppREL(i,j,k) = \phi_{ijk}^{*--}(t) - \phi_{ijk}^{+--}(t) - \phi_{ijk}^{**-}(t) + \phi_{ijk}^{+*-}(t) - \phi_{ijk}^{*-*}(t) + \phi_{ijk}^{+-*}(t) + \phi_{ijk}^{***}(t) - \phi_{ijk}^{+**}(t) \neq 0 \quad (15)$$

Suppose i^{th} component is changing back to its previous state, j^{th} component is changing to its next state and k^{th} component is changing back to its previous state. Then components i, j and k are jointly *pnp*-relevant, if

$$pnpREL(i,j,k) = \phi_{ijk}^{-*-}(t) - \phi_{ijk}^{**-}(t) - \phi_{ijk}^{-+-}(t) + \phi_{ijk}^{*+-}(t) - \phi_{ijk}^{-**}(t) + \phi_{ijk}^{***}(t) + \phi_{ijk}^{-+*}(t) - \phi_{ijk}^{*+*}(t) \neq 0 \quad (16)$$

Suppose i^{th} component is changing back to its previous state, j^{th} component is changing back to its previous state and k^{th} component is also changing back to its previous state. Then components i, j and k are jointly *ppp*-relevant, if

$$pppREL(i,j,k) = \phi_{ijk}^{---}(t) - \phi_{ijk}^{*--}(t) - \phi_{ijk}^{-*-}(t) + \phi_{ijk}^{**-}(t) - \phi_{ijk}^{--*}(t) + \phi_{ijk}^{*-*}(t) + \phi_{ijk}^{-**}(t) - \phi_{ijk}^{***}(t) \neq 0 \quad (17)$$

To find the joint importance of three repairable components, in Birnbaum sense, the following measures are proposed, by considering three components, i, j and k.

$$I_{NNNB}^{ijk}(t) = P\{nnnREL(i,j,k) \neq 0\} = \sum_{w=0}^{M_k}\sum_{v=0}^{M_j}\sum_{u=0}^{M_i} P[(\phi(X_i(t)=u, X_j(t)=v, X_k(t)=w, \mathbf{X}(t)) - \phi(X_i(t)=u-1, X_j(t)=v, X_k(t)=w, \mathbf{X}(t))) - (\phi(X_i(t)=u, X_j(t)=v-1, X_k(t)=w, \mathbf{X}(t)) - \phi(X_i(t)=u-1, X_j(t)=v-1, X_k(t)=w,\ \mathbf{X}(t)))] - [(\phi(X_i(t)=u, X_j(t)=v, X_k(t)=w-1, \mathbf{X}(t)) - \phi(X_i(t)=u-1, X_j(t)=v, X_k(t)=w-1,\ \mathbf{X}(t))) - (\phi(X_i(t)=u, X_j(t)=v-1, X_k(t)=w-1,\ \mathbf{X}(t)) - \phi(X_i(t)=u-1, X_j(t)=v-1, X_k(t)=w-1, \mathbf{X}(t))) \neq 0] \quad (18)$$

$$I_{\{NPNB\}}^{ijk}(t) = P\{npnREL(i,j,k) \neq 0\} = \sum_{w=0}^{M_k}\sum_{v=0}^{M_j}\sum_{u=0}^{M_i} P[(\phi(X_i(t)=u, X_j(t)=v+1, X_k(t)=w, \mathbf{X}(t)) - \phi(X_i(t)=u-1, X_j(t)=v+1, X_k(t)=w, \mathbf{X}(t))) - (\phi(X_i(t)=u, X_j(t)=v, X_k(t)=w, \mathbf{X}(t)) - \phi(X_i(t)=u-1, X_j(t)=v, X_k(t)=w, \mathbf{X}(t)))] - [(\phi(X_i(t)=u, X_j(t)=v+1, X_k(t)=w-1, \mathbf{X}(t)) - \phi(X_i(t)=u-1, X_j(t)=v+1, X_k(t)=w-1,\ \mathbf{X}(t))) - (\phi(X_i(t)=u, X_j(t)=v, X_k(t)=w-1, \mathbf{X}(t)) - \phi(X_i(t)=u-1, X_j(t)=v, X_k(t)=w-1, \mathbf{X}(t))) \neq 0] \quad (19)$$

$$I_{\{PNNB\}}^{ijk}(t) = P\{pnnREL(i,j,k) \neq 0\} = \sum_{w=0}^{M_k}\sum_{v=0}^{M_j}\sum_{u=0}^{M_i} P[(\phi(X_i(t)=u+1, X_j(t)=v, X_k(t)=w, \mathbf{X}(t)) - \phi(X_i(t)=u, X_j(t)=v, X_k(t)=w, \mathbf{X}(t))) - (\phi(X_i(t)=u+1, X_j(t)=v-1, X_k(t)=w,\ \mathbf{X}(t)) - \phi(X_i(t)=u, X_j(t)=v-1, X_k(t)=w, \mathbf{X}(t)))] - [(\phi(X_i(t)=u+1, X_j(t)=v, X_k(t)=w-1, \mathbf{X}(t)) - \phi(X_i(t)=u, X_j(t)=v, X_k(t)=w-1, \mathbf{X}(t))) - (\phi(X_i(t)=u+1, X_j(t)=v-1, X_k(t)=w-1, \mathbf{X}(t)) - \phi(X_i(t)=u, X_j(t)=v-1, X_k(t)=w-1, \mathbf{X}(t))) \neq 0] \quad (20)$$

$$I_{\{PPNB\}}^{ijk}(t) = P\{ppnREL(i,j,k) \neq 0\} = \sum_{w=0}^{M_k}\sum_{v=0}^{M_j}\sum_{u=0}^{M_i} P[(\phi(X_i(t)=u+1, X_j(t)=v+1, X_k(t)=w, \mathbf{X}(t)) - \phi(X_i(t)=u, X_j(t)=v+1, X_k(t)=w, \mathbf{X}(t))) - (\phi(X_i(t)=u+1, X_j(t)=v, X_k(t)=w, \mathbf{X}(t)) - \phi(X_i(t)=u, X_j(t)=v, X_k(t)=w, \mathbf{X}(t)))] - [(\phi(X_i(t)=u+1, X_j(t)=v+1, X_k(t)=w-1, \mathbf{X}(t)) - \phi(X_i(t)=u, X_j(t)=v+1, X_k(t)=w-1, \mathbf{X}(t))) - (\phi(X_i(t)=u+1, X_j(t)=v, X_k(t)=w-1, \mathbf{X}(t)) - \phi(X_i(t)=u, X_j(t)=v, X_k(t)=w-1, \mathbf{X}(t))) \neq 0] \quad (21)$$

$$I^{ijk}_{NNPB}(t) = P\{nnpREL(i,j,k) \neq 0\} = \sum_{w=0}^{M_k}\sum_{v=0}^{M_j}\sum_{u=0}^{M_i} P[(\phi(X_i(t)=u, X_j(t)=v, X_k(t)=w+1, \mathbf{X}(t)) - \phi(X_i(t)=u-1, X_j(t)=v, X_k(t)=w+1, \mathbf{X}(t))) - (\phi(X_i(t)=u, X_j(t)=v-1, X_k(t)=w+1, \mathbf{X}(t)) - \phi(X_i(t)=u-1, X_j(t)=v-1, X_k(t)=w+1, \mathbf{X}(t)))] - [(\phi(X_i(t)=u, X_j(t)=v, X_k(t)=w, \mathbf{X}(t)) - \phi(X_i(t)=u-1, X_j(t)=v, X_k(t)=w, \mathbf{X}(t))) - (\phi(X_i(t)=u, X_j(t)=v-1, X_k(t)=w, \mathbf{X}(t)) - \phi(X_i(t)=u-1, X_j(t)=v-1, X_k(t)=w, \mathbf{X}(t))) \neq 0] \tag{22}$$

$$I^{ijk}_{NPPB}(t) = P\{nppREL(i,j,k) \neq 0\} = \sum_{w=0}^{M_k}\sum_{v=0}^{M_j}\sum_{u=0}^{M_i} P[\ (\phi(X_i(t)=u, X_j(t)=v+1, X_k(t)=w+1, \mathbf{X}(t)) - \phi(X_i(t)=u-1, X_j(t)=v+1, X_k(t)=w+1, \mathbf{X}(t))) - (\phi(X_i(t)=u, X_j(t)=v, X_k(t)=w+1, \mathbf{X}(t)) - \phi(X_i(t)=u-1, X_j(t)=v, X_k(t)=w+1, \mathbf{X}(t)))] - [(\phi(X_i(t)=u, X_j(t)=v+1, X_k(t)=w, \mathbf{X}(t)) - \phi(X_i(t)=u-1, X_j(t)=v+1, X_k(t)=w, \mathbf{X}(t))) - (\phi(X_i(t)=u, X_j(t)=v, X_k(t)=w, \mathbf{X}(t)) - \phi(X_i(t)=u-1, X_j(t)=v, X_k(t)=w, \mathbf{X}(t))) \neq 0] \tag{23}$$

$$I^{ijk}_{PNPB}(t) = P\{pnpREL(i,j,k) \neq 0\} = \sum_{w=0}^{M_k}\sum_{v=0}^{M_j}\sum_{u=0}^{M_i} P[(\phi(X_i(t)=u+1, X_j(t)=v, X_k(t)=w+1, \mathbf{X}(t)) - \phi(X_i(t)=u, X_j(t)=v, X_k(t)=w+1, \mathbf{X}(t))) - (\phi(X_i(t)=u+1, X_j(t)=v-1, X_k(t)=w+1, \mathbf{X}(t)) - \phi(X_i(t)=u, X_j(t)=v-1, X_k(t)=w+1, \mathbf{X}(t)))] - [(\phi(X_i(t)=u+1, X_j(t)=v, X_k(t)=w, \mathbf{X}(t)) - \phi(X_i(t)=u, X_j(t)=v, X_k(t)=w, \mathbf{X}(t))) - (\phi(X_i(t)=u+1, X_j(t)=v-1, X_k(t)=w, \mathbf{X}(t)) - \phi(X_i(t)=u, X_j(t)=v-1, X_k(t)=w, \mathbf{X}(t))) \neq 0] \tag{24}$$

$$I^{ijk}_{PPPB}(t) = P\{pppREL(i,j,k) \neq 0\} = \sum_{w=0}^{M_k}\sum_{v=0}^{M_j}\sum_{u=0}^{M_i} P[(\phi(X_i(t)=u+1, X_j(t)=v+1, X_k(t)=w+1, \mathbf{X}(t)) - \phi(X_i(t)=u, X_j(t)=v+1, X_k(t)=w+1, \mathbf{X}(t))) - (\phi(X_i(t)=u+1, X_j(t)=v, X_k(t)=w+1, \mathbf{X}(t)) - \phi(X_i(t)=u, X_j(t)=v, X_k(t)=w+1, \mathbf{X}(t)))] - [(\phi(X_i(t)=u+1, X_j(t)=v+1, X_k(t)=w, \mathbf{X}(t)) - \phi(X_i(t)=u, X_j(t)=v+1, X_k(t)=w, \mathbf{X}(t))) - (\phi(X_i(t)=u+1, X_j(t)=v, X_k(t)=w, \mathbf{X}(t)) - \phi(X_i(t)=u, X_j(t)=v, X_k(t)=w, \mathbf{X}(t))) \neq 0] \tag{25}$$

Clearly, $I^{ijk}_{NNNB}(t)$ is the joint importance measure of three components i, j and k at time t when the three components i, j and k enters its next state, $I^{ijk}_{NPNB}(t)$ is the joint importance measure of three components i, j and k at time t when the components i and k enters its next state and the component j enters its previous state, $I^{ijk}_{PNNB}(t)$ is the joint importance measure of three components i, j and k at time t when the component i enters its previous state and the components j and k enters its next state and $I^{ijk}_{PPNB}(t)$ is the joint importance measure of three components i, j and k at time t when both components i and j enters its previous state and the component k enters the next state, $I^{ijk}_{NNPB}(t)$ is the joint

importance measure of three components i, j and k at time t when the two components i and j enters its next state and the component k enters the previous state, $I^{ijk}_{NPPB}(t)$ is the joint importance measure of three components i, j and k at time t when the component i enters its next state and components j and k enters its previous state, $I^{ijk}_{PNPB}(t)$ is the joint importance measure of three components i, j and k at time t when the components i and k enters its previous state and the component j enters its next state and $I^{ijk}_{PPPB}(t)$ is the joint importance measure of three components i, j and k at time t when the three components i, j and k enters its previous state.

3. Application

In multistate system reliability engineering, the problem of identification of most important component of group of component is required for giving proper repair or maintenance activities to provide the system active for the completion of assigned mission. Most of the existing measures are useful for this purpose if repair or maintenance is not considered. In the proposed measures, the major advantage is that, one can measure importance and joint importance measures when repair or maintenance is applied to the components. Adoption of proper maintenance activity is unavoidable in system engineering. The proposed results are useful to the multistate and binary state systems.

4. Illustration

To illustrate the joint importance of components, we consider a network flow system which is given in Figure 1. In this example, there is a directed network flow system consisting of 6 components represented by edges of the network.

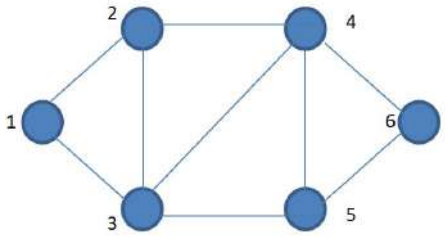


Figure 1: Network flow system

The state functions of the components f_1, f_2, f_3, f_4, f_5 and f_6 represent the flow capacity functions of the components given by

$$f_1(u) = f_6(u) = 2.5\ u,\ u = 0, 1, 2,$$

$$f_2(u) = 1.5\ u,\ u = 0, 1, 2,$$

$$f_3(u) = f_5(u) = 5.0\ u,\ u = 0, 1,$$

$$f_4(u) = 1.0\ I(u = 1) + 2.5\ I(u = 2),\ u = 0, 1, 2,$$

where I is the indicator function. The physical state of the system is the amount of flow that can be sent through the network from the source node 1 to the terminal node 6. In order to express the system state as a function of the component states, we identify the minimal cut sets in the network. These are $K_1 = \{1\}, K_2 = \{2, 3\}, K_3 = \{3, 4\}, K_4 = \{4, 5\}, K_5 = \{6\}$. According to the well-known max-flow-min cut theorem, we then have

$$\phi(X(t)) = \min_{1 \le j \le 5} \sum_{i \in K_j} f_i(X_i(t))$$

The probabilities of each component in its states are given by

$$p_1(u) = p_2(u) = p_4(u) = p_6(u) = \begin{cases} \frac{25}{105} & , u = 0 \\ \frac{35}{105} & , u = 1 \\ \frac{45}{105} & , u = 2 \end{cases}$$

$$p_3(u) = p_5(u) = \begin{cases} \frac{45}{100} & , u = 0 \\ \frac{55}{100} & , u = 1 \end{cases}$$

Here we have computed the physical joint importance $I^{ijk}(t)$ for all the possible combinations of three components. The results are given in Table 1.

It is clear from the example that the component group (3,4,5) is the most important set in any case considered. But the ranking of the rest of the three sets of components keeps changing. The proposed measures give the investigator the ability to look at relevancy from several angles, which is useful in a diagnostic environment as well as when the investigation is done to support decisions for system improvement.

5. Conclusions

In the present paper, a repairable multistate system is considered. The single component Birnbaum importance measure is generalized to three component joint importance measure for multistate systems in eight different ways. The measures gives an insight regarding change in system performance to support decisions regarding improvement of the system, through the movement of components in same/opposite directions. Since the proposed measures are investigating the behavior of components on system performance, they are useful in a diagnostic checking. These joint importance measures are highly appropriate while considering repairable components. In order to locate the weakest group or more consistent group, the proposed measures will be helpful. So more repair activities can be ensured weakest group.

Acknowledgements

The Authors are thankful to Chief Editor, Managing Editor and Reviewers for providing useful comments which helped to improve the paper.

Table 1: Joint importance of three components i, j and k

i,j,k	$I^{(ijk)}_{NNNB}$	$I^{(ijk)}_{NNPB}$	$I^{(ijk)}_{NPNB}$	$I^{(ijk)}_{NPPB}$	$I^{(ijk)}_{PNNB}$	$I^{(ijk)}_{PNPB}$	$I^{(ijk)}_{PPNB}$	$I^{(ijk)}_{PPPB}$
1,2,3	0.1895	0.1895	0.2211	0.2211	0.2211	0.2211	0.2579	0.2579
1,2,4	0.1119	0.1119	0.1306	0.1306	0.1306	0.1306	0.1523	0.1523
1,2,5	0	0	0	0	0	0	0	0
1,2,6	0.0639	0.0746	0.0746	0.0870	0.0746	0.0870	0.0870	0.1015
1,3,4	0.3286	0.3482	0.3286	0.3482	0.3833	0.3999	0.3833	0.3999
1,3,5	0.2487	0.2487	0.2487	0.2487	0.2902	0.2902	0.2902	0.2902
1,3,6	0.2542	0.2965	0.2542	0.2965	0.2965	0.3460	0.2965	0.3460
1,4,5	0.2394	0.3183	0.2394	0.2394	0.2793	0.2793	0.2793	0.2793
1,4,6	0.1927	0.2248	0.1927	0.2248	0.2248	0.2623	0.2248	0.2623
1,5,6	0.1026	0.3912	0.1026	0.1197	0.1197	0.1396	0.0490	0.1396
2,3,4	0.3317	0.3317	0.3317	0.3317	0.3869	0.3869	0.3869	0.3869
2,3,5	0	0	0	0	0	0	0	0
2,3,6	0.1895	0.2211	0.1895	0.2211	0.2211	0.2579	0.2211	0.2579
2,4,5	0	0	0	0	0	0	0	0
2,4,6	0.1119	0.1306	0.1119	0.1306	0.1306	0.1523	0.1306	0.1523
2,5,6	0	0	0	0	0	0	0	0
3,4,5	0.5804	0.5804	0.5804	0.5804	0.5804	0.5804	0.5804	0.5804
3,4,6	0.3286	0.3833	0.3428	0.3999	0.3286	0.3833	0.3428	0.3999
3,5,6	0.2487	0.2902	0.2487	0.2902	0.2487	0.2902	0.2487	0.2902
4,5,6	0.2394	0.2793	0.2394	0.2793	0.2394	0.2793	0.2394	0.2793

References

Amrutkar, K. P. and Kamalja, K. K. (2017). An overview of various importance measures of reliability system. *International Journal of Mathematical, Engineering and Management Sciences*, **2**, 150–171.

Barlow, R. E. and Proschan, F. (1975). Importance of system components and fault tree events. *Stochastic Processes and Their Applications*, **3**, 153–173.

Birnbaum, Z. (1969). On the importance of different components in a multicomponent system. *Multivariate Analysis*, **11**, 581–592.

Borgonovo, E. and Apostolakis, G. E. (2001). A new importance measure for risk-informed decision making. *Reliability Engineering and System Safety*, **72**, 193–212.

Cai, Z., Si, S., Liu, Y., and Zhao, J. (2017). Maintenance optimization of continuous state systems based on performance improvement. *IEEE Transactions on Reliability*, **67**, 651–665.

Cai, Z., Si, S., Sun, S., and Li, C. (2016). Optimization of linear consecutive-k-out-of-n system with a birnbaum importance-based genetic algorithm. *Reliability Engineering and System Safety*, **152**, 248–258.

Chacko, V. M. (2020). New joint importance measures for multistate systems. *International Journal of Statistics and Reliability Engineering*, **7**, 140–148.

Chacko, V. M. (2021). On joint importance measures for multistate reliability systems. *Reliability: Theory and Applications*, **16**, 286–293.

Chacko, V. M. (2022). On joint importance measures for multistate system's reliability. In *Operations Research*, pages 147–166. CRC Press.

Chacko, V. M. (2023a). On birnbaum type joint importance measures for multistate reliability systems. *Communications in Statistics-Theory and Methods*, **52**, 2799–2818.

Chacko, V. M. (2023b). On new joint importance measures for multistate reliability systems. *Computational Intelligence in Sustainable Reliability Engineering*, , 145–158.

Chacko, V. M., Franson, A. S., and Amrutha, M. (2023). Birnbaum joint importance measures for three components of a repairable multistate systems. *Dependability*, **23**, 3–13.

Chacko, V. M. and Manoharan, M. (2008). Joint importance measures for the multistate system. *Advances in Performance and Safety of Complex systems*, , 308–314.

Chacko, V. M. and Manoharan, M. (2011). Joint importance measures for multistate reliability systems. *Opsearch*, **48**, 257–278.

Dinic, E. A. (1970). Algorithm for solution of a problem of maximum flow in networks with power estimation. *Soviet Mathematics – Doklady*, **11**, 1277–1280.

Dui, H., Li, S., Xing, L., and Liu, H. (2019). System performance-based joint importance analysis guided maintenance for repairable systems. *Reliability Engineering and System Safety*, **186**, 162–175.

Dui, H., Si, S., Zuo, M. J., and Sun, S. (2015). Semi-markov process-based integrated importance measure for multi-state systems. *IEEE Transactions on Reliability*, **64**, 754–765.

Dui, H., Zhang, C., and Zheng, X. (2020). Component joint importance measures for maintenances in submarine blowout preventer system. *Journal of loss prevention in the process industries*, **63**, 104003.

Ford, L. R. and Fulkerson, D. R. (1956). Maximal flow through a network. *Canadian journal of Mathematics*, **8**, 399–404.

Fussell, J. and Vesely, W. (1972). New methodology for obtaining cut sets for fault trees. *Transactions of the American Nuclear Society*, **15**, 262–263.

Griffith, W. S. (1980). Multistate reliability models. *Journal of Applied Probability*, **17**, 735–744.

Hosseini, S., Barker, K., and Ramirez-Marquez, J. E. (2016). A review of definitions and measures of system resilience. *Reliability Engineering and System Safety*, **145**, 47–61.

Huseby, A. B., Kalinowska, M., and Abrahamsen, T. (2022). Birnbaum criticality and importance measures for multistate systems with repairable components. *Probability in the Engineering and Informational Sciences*, **36**, 66–86.

Huseby, A. B. and Natvig, B. (2009). Advanced discrete simulation methods applied to repairable multistate systems. In *Reliability, Risk, and Safety, Three Volume Set*, pages 693–700. CRC Press.

Huseby, A. B. and Natvig, B. (2013). Discrete event simulation methods applied to advanced importance measures of repairable components in multistate network flow systems. *Reliability Engineering and System Safety*, **119**, 186–198.

Levitin, G. and Lisnianski, A. (1999). Importance and sensitivity analysis of multi-state systems using the universal generating function method. *Reliability Engineering and System Safety*, **65**, 271–282.

Levitin, G., Podofillini, L., and Zio, E. (2003). Generalised importance measures for multistate elements based on performance level restrictions. *Reliability Engineering and*

System Safety, **82**, 287–298.

Natvig, B. (1979). A suggestion of a new measure of importance of system components. *Stochastic Processes and Their Applications*, **9**, 319–330.

Natvig, B. (1985). New light on measures of importance of system components. *Scandinavian Journal of Statistics*, **12**, 43–54.

Natvig, B. (2010). *Multistate Systems Reliability Theory with Applications*. John Wiley & Sons.

Natvig, B. (2011). Measures of component importance in nonrepairable and repairable multistate strongly coherent systems. *Methodology and Computing in Applied Probability*, **13**, 523–547.

Natvig, B., Eide, K. A., Gåsemyr, J., and et al. (2009). Simulation based analysis and an application to an offshore oil and gas production system of the natvig measures of component importance in repairable systems. *Reliability Engineering and System Safety*, **94**, 1629–1638.

Natvig, B. and Gåsemyr, J. (2009). New results on the barlow–proschan and natvig measures of component importance in nonrepairable and repairable systems. *Methodology and Computing in Applied Probability*, **11**, 603–620.

Natvig, B., Huseby, A. B., and Reistadbakk, M. O. (2011). Measures of component importance in repairable multistate systems-a numerical study. *Reliability Engineering and System Safety*, **96**, 1680–1690.

Ramirez-Marquez, J. E. and Coit, D. W. (2005). Composite importance measures for multi-state systems with multi-state components. *IEEE transactions on Reliability*, **54**, 517–529.

Ramirez-Marquez, J. E. and Coit, D. W. (2007). Multi-state component criticality analysis for reliability improvement in multi-state systems. *Reliability Engineering and System Safety*, **92**, 1608–1619.

Ramirez-Marquez, J. E., Rocco, C. M., Gebre, B. A., Coit, D. W., and Tortorella, M. (2006). New insights on multi-state component criticality and importance. *Reliability Engineering and System Safety*, **91**, 894–904.

Ross, S. M. (2014). *Introduction to Probability Models*. Academic press.

Si, S., Dui, H., Cai, Z., and Sun, S. (2012a). The integrated importance measure of multi-state coherent systems for maintenance processes. *IEEE Transactions on Reliability*, **61**, 266–273.

Si, S., Dui, H., Zhao, X., Zhang, S., and Sun, S. (2012b). Integrated importance measure of component states based on loss of system performance. *IEEE Transactions on Reliability*, **61**, 192–202.

Si, S., Levitin, G., Dui, H., and Sun, S. (2013). Component state-based integrated importance measure for multi-state systems. *Reliability Engineering and System Safety*, **116**, 75–83.

Si, S., Liu, M., Jiang, Z., Jin, T., and Cai, Z. (2019). System reliability allocation and optimization based on generalized birnbaum importance measure. *IEEE Transactions on Reliability*, **68**, 831–843.

Skutlaberg, K. and Natvig, B. (2016). Minimization of the expected total net loss in a stationary multistate flow network system. *Applied Mathematics*, **7**, 793–817.

Todinov, M. (2013). *Flow Networks*. Oxford, UK: Elsevier Insights.

Wu, S., Chen, Y., Wu, Q., and Wang, Z. (2016). Linking component importance to optimisation of preventive maintenance policy. *Reliability Engineering and System Safety*, **146**, 26–32.

Wu, S. and Coolen, F. P. (2013). A cost-based importance measure for system components: An extension of the birnbaum importance. *European Journal of Operational Research*, **225**, 189–195.

Zhu, X., Fu, Y., Yuan, T., and Wu, X. (2017). Birnbaum importance based heuristics for multi-type component assignment problems. *Reliability Engineering and System Safety*, **165**, 209–221.

Zio, E., Marella, M., and Podofillini, L. (2007). Importance measures-based prioritization for improving the performance of multi-state systems: application to the railway industry. *Reliability Engineering and System Safety*, **92**, 1303–1314.

Zio, E. and Podofillini, L. (2003). Monte carlo simulation analysis of the effects of different system performance levels on the importance of multi-state components. *Reliability Engineering and System Safety*, **82**, 63–73.

Zio, E. and Podofillini, L. (2006). Accounting for components interactions in the differential importance measure. *Reliability Engineering and System Safety*, **91**, 1163–1174.

Zio, E., Podofillini, L., and Levitin, G. (2004). Estimation of the importance measures of multi-state elements by monte carlo simulation. *Reliability Engineering and System Safety*, **86**, 191–204.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 22, No. 2, 2024 (New Series), pp 203–216
http://www.ssca.org.in/journal

A Comprehensive Study of the Power Modified Lindley-Geometric Distribution in the T-X Family: COVID-19 Applications

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: 08 October 2023; Revised: 30 December 2023; Accepted: 05 January 2024

Abstract

In this paper we develop a superior and ideal statistical model to provide optimal modelling for the number of deaths resulting from COVID-19 infections. This paper introduces the power modified Lindley-geometric distribution, a novel versatile three-parameter discrete model built on the T-X methodology. In addition, to providing a generalized geometric distribution we offer a thorough list of its mathematical characteristics. The parameter of the new model is estimated using four different estimation techniques: maximum likelihood, Cramer-von Mises, least-square, and weighted least-square. The simulation experiment uses four distinct estimating approaches to test the accuracy of the model parameters. Additionally, we applied two datasets to the COVID-19 mortality data for the United Kingdom and Egypt. These two instances of actual data were used to highlight the significance of our distribution for modelling and fitting this particular kind of discrete data.

Key words: T-X family, Maximum likelihood; Cramer-von Mises; Least-square; Weighted least-square; Data analysis.

AMS Subject Classifications: 62K05, 05B05

1. Introduction

In our modern period, the abundance of data coming in from all fields has submerged the interest in defining new flexible distributions. "Thoroughly changing" a baseline distribution is an easy and quick way to define these mathematical objects. The study of tail properties and improving the goodness-of-fit of the associated models have both been demonstrated to benefit from the addition of parameter(s). The most well-liked distribution among those that have been suggested is the T-X family of distributions by Alzaatreh *et al.* (2013). The following peculiar transformation is one of the most practical transformers for T-X family of distributions $W(F(x)) = -log(1 - F(x))$, where the cumulative density function *(CDF)* of random variable X is represented by the notation $F(x)$. To put it another way, $W(F(x))$ is used to modify the distribution described by $F(x)$ and define a new family

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

of distributions based on a changed *CDF*. With the help of the T-X family, one may quickly build discrete distributions in addition to continuous distributions. The T-geometric family, which are the discrete analogues of the distribution of the random variable T, was defined and explored by Alzaatreh *et al.* (2012) as a result. The *CDF* of T-geometric family is given by

$$G(x;\phi,b) = \int_0^{-log(p^{(x+1)})} r(t;\phi)dt = R[-log(p^{(x+1)});\phi,p)] = R[b(x+1);\phi,b], \quad x = 0,1,2,..., \tag{1}$$

where $b = -log\ p > 0$ and ϕ the parameters of the *CDF* $R(x;\phi)$. Some of the families available in the modern literature are the Pareto-geometric, Weibull-geometric, Burr-geometric and exponentiated exponential -
geometric distribution by Alzaatreh *et al.* (2012), Kumaraswamy-geometric distribution by Akinsete *et al.* (2014) and exponentiated Weibull-geometric distribution by Famoye (2019). The comprehensive review of T-X family of distributions may be found in Tomy *et al.* (2019).

Chesneau *et al.* (2021b) introduce a novel two-parameter lifetime distribution that is the power version of the modified Lindley distribution and call it as power modified Lindley (PML) distribution. It offers a compelling substitute for the Weibull and power Lindley distributions as its primary goal. Let T be a random variable with the PML distribution. The probability density function (*PDF*) and the *CDF* are each defined as

$$r(t;\alpha,\theta) = \frac{\theta\alpha}{1+\theta}t^{\alpha-1}e^{-2\theta t^{\alpha}}\left[(1+\theta)e^{\theta t^{\alpha}} + 2\theta t^{\alpha} - 1\right], \quad t > 0, \tag{2}$$

$$R(t;\alpha,\theta) = 1 - \left[1 + \frac{\theta t^{\alpha}}{1+\theta}e^{-\theta t^{\alpha}}\right]e^{-\theta t^{\alpha}}, \quad t > 0, \tag{3}$$

where α >0 and θ >0. This distribution is derived by using the power parameter α in modified Lindely distribution, has been proposed by Chesneau *et al.* (2021a).

This paper introduces a flexible three parameter discrete distribution called power modified Lindley-geometric, which is based on T-geometric family of distribution and power modified Lindley distribution. The main driving force behind the development of this new discrete distribution was the fact that, in contrast to the amount of literature on continuous cases, there was a dearth of research on the discrete families of distributions. Another fact is that there are lots of researchers work to understand the patterns of the COVID-19 epidemic and offer models that better suit the data and can be used to estimate the anticipated number of cases and deaths to assist the government in making decisions on preventative measures. And the new distribution is suitable for fitting COVID-19 data sets, which is the main goal of this study. Another motivator is the characteristics of the suggested distribution itself. In other words, the newly proposed discrete distribution features a probability mass function (*PMF*) that is right-skewed, symmetric and left-skewed. Additionally, the new distribution features hazard rate functions (*HRF*) that are increasing, decreasing and upside-down bathtub-shaped. Additionally, we provided a comparison of the various estimation techniques.

Following is an outline of the remaining content: In Section 2, we provide a brand-new discrete family of distributions. Section 3, a special case of the obtained new discrete discrete family of distribution and its probabilistic characteristics are studied in detail. Section 4 discussed least-squares, weighted least squares, and the Cramer von Mises technique in addition to maximum likelihood estimation. A thorough simulation analysis is employed in Section 5 to evaluate the performance of these estimators. Applications to the two COVID-19 data sets used to demonstrate how the new distribution performs are detailed in Section 6. A few closing thoughts are provided in Section 7.

2. Discrete power modified Lindley-X family of distribution

In this section, we introduce the discrete power modified Lindley-X (PML-X) family of distributions, a new discrete family of distributions. Utilising the Alzaatreh *et al.* (2013) T-X generalization technique, we enable the transformed random variable T to have the PML distribution and the transformer random variable X is a discrete random variable, with $W(F(x)) = -log(1 - F(x))$. Then the *CDF* of new family is given by

$$\begin{aligned} G(x;\alpha,\theta,\Im) &= \int_0^{-log(1-F(x;\Im))} r(t;\alpha,\theta)dt = R(-log(1-F(x;\Im))) \\ &= 1 - \left[1 + \frac{\theta[-log(1-F(x;\Im))]^{\alpha}}{1+\theta} e^{-\theta[-log(1-F(x;\Im))]^{\alpha}}\right] e^{-\theta[-log(1-F(x;\Im))]^{\alpha}} \end{aligned} \tag{4}$$

The corresponding *PMF* of the PML-X family of discrete distributions becomes.

$$\begin{aligned} g(x;\alpha,\theta,\Im) &= G(x) - G(x-1) \\ &= \left[1 + \frac{\theta[-log(1-F(x-1;\Im))]^{\alpha}}{1+\theta} e^{-\theta[-log(1-F(x-1;\Im))]^{\alpha}}\right] e^{-\theta[-log(1-F(x-1;\Im))]^{\alpha}} \\ &\quad - \left[1 + \frac{\theta[-log(1-F(x;\Im))]^{\alpha}}{1+\theta} e^{-\theta[-log(1-F(x;\Im))]^{\alpha}}\right] e^{-\theta[-log(1-F(x;\Im))]^{\alpha}} \end{aligned} \tag{5}$$

where $\alpha > 0$, $\theta > 0$ and $\Im$ the parameters of the *CDF* $F(x;\Im)$, and the range of variation of PML-X family of distribution depends on the random variable X with *CDF* $F(x;\Im)$.

In the following section, we examine one member of this family, the power modified Lindley-geometric distribution, and provide its detailed features. The geometric distribution was chosen because it has a simplified *CDF* form.

3. Power modified Lindley-geometric distribution

Let's make the assumption that the transformed distribution is geometric with parameter p, $0 < p < 1$, and that the survival function $S(x) = 1 - F(x) = p^{(x+1)}$.Then, the

PMF of the new model using Equation (5) is given by

$$g(x;\alpha,\theta,b) = \left[1+\frac{\theta(bx)^{\alpha}}{1+\theta}e^{-\theta(bx)^{\alpha}}\right]e^{-\theta(bx)^{\alpha}} - \left[1+\frac{\theta(b(x+1))^{\alpha}}{1+\theta}e^{-\theta(b(x+1))^{\alpha}}\right]e^{-\theta(b(x+1))^{\alpha}};\quad x=0,1,2,.... \tag{6}$$

where $b = -log\ p > 0$, $\alpha > 0$ and $\theta > 0$. We call this new distribution the power modified Lindley-geometric (PMLG) distribution with parameters b, α and θ. Note that,

$$\lim_{x\to+\infty} g(x;\alpha,\theta,b) = 0,\ \lim_{x\to 0} g(x;\alpha,\theta,b) = 0\ when\ b\to 0\ and$$

$$\lim_{x\to 0} g(x;\alpha,\theta,b) = 1\ when\ b\to\infty.$$

The corresponding *CDF* is given by

$$G(x;\alpha,\theta,b) = 1 - \left[1+\frac{\theta(b(x+1))^{\alpha}}{1+\theta}e^{-\theta(b(x+1))^{\alpha}}\right]e^{-\theta(b(x+1))^{\alpha}};\quad x=0,1,2,... \tag{7}$$

and the hazard rate function (*HRF*) corresponding to the *CDF* is provided by

$$h(x;\alpha,\theta,b)) = \frac{\left[1+\frac{\theta(bx)^{\alpha}}{1+\theta}e^{-\theta(bx)^{\alpha}}\right]e^{-\theta(bx)^{\alpha}} - \left[1+\frac{\theta(b(x+1))^{\alpha}}{1+\theta}e^{-\theta(b(x+1))^{\alpha}}\right]e^{-\theta(b(x+1))^{\alpha}}}{\left[1+\frac{\theta(b(x+1))^{\alpha}}{1+\theta}e^{-\theta(b(x+1))^{\alpha}}\right]e^{-\theta(b(x+1))^{\alpha}}}$$

A graphic illustration of the *PMF* of the PMLG distribution in various forms is shown in Figure 1. These graphs demonstrate the possibility of right-skewed, symmetric, left-skewed, increasing decreasing curves for the *PMF* of the PMLG distribution. The *HRF* of the PMLG distribution in Figure 2 is depicted in some of its potential shapes for various parameter values. Figures show that the *HRF* can have a variety of shapes, including increasing, decreasing and upside-down bathtub shapes. As a result, the PMLG distribution is excellent at modelling a variety of data sets.

3.1. Probability generating function, r^{th} moment function, mean and variance

The probability generating function (*PGF*) of PMLG distribution is given by

$$p(s) = 1 + (s-1)\sum_{x=1}^{\infty} s^{x-1}\left[1+\frac{\theta(bx)^{\alpha}}{1+\theta}e^{-\theta(bx)^{\alpha}}\right]e^{-\theta(bx)^{\alpha}}. \tag{8}$$

Using Equation (6), the non-central r^{th} moment of the PMLG distribution can be calculated as follows:

$$\mu'_r = \sum_{x=0}^{\infty} x^r g(x;\alpha,\theta,b)$$
$$= \sum_{x=0}^{\infty} x^r\left[1+\frac{\theta(bx)^{\alpha}}{1+\theta}e^{-\theta(bx)^{\alpha}}\right]e^{-\theta(bx)^{\alpha}} - \sum_{x=0}^{\infty} x^r\left[1+\frac{\theta(b(x+1))^{\alpha}}{1+\theta}e^{-\theta(b(x+1))^{\alpha}}\right]e^{-\theta(b(x+1))^{\alpha}}.$$

In particular, the first two moments of the PMLG distribution are given by

$$\mu_1' = E(X) = \sum_{x=1}^{\infty} \left[1 + \frac{\theta(bx)^\alpha}{1+\theta} e^{-\theta(bx)^\alpha}\right] e^{-\theta(bx)^\alpha}. \tag{9}$$

$$\mu_2' = \sum_{x=1}^{\infty} (2x-1) \left[1 + \frac{\theta(bx)^\alpha}{1+\theta} e^{-\theta(bx)^\alpha}\right] e^{-\theta(bx)^\alpha}. \tag{10}$$

The variance of PMLG distribution is given as

$$V(X) = \sum_{x=1}^{\infty} (2x-1) \left[1 + \frac{\theta(bx)^\alpha}{1+\theta} e^{-\theta(bx)^\alpha}\right] e^{-\theta(bx)^\alpha} - \left[\sum_{x=1}^{\infty} \left[1 + \frac{\theta(bx)^\alpha}{1+\theta} e^{-\theta(bx)^\alpha}\right] e^{-\theta(bx)^\alpha}\right]^2. \tag{11}$$

Table 1 shows the mean and variance of the PMLG distribution for different values of b, α and θ using statistical software. From this, we are able to understand that the variance decreases with α and θ for different values of b. Furthermore, based on the values of b, α and θ, the mean can be equal, lower or larger than its variance. As a result, many data sets can be modelled using the characteristics of the PMLG distribution.

Table 1: The mean (variance) of PMLG for various choices of parameters

	$\alpha \longrightarrow$	0.5	1	2
	$\theta \downarrow$			
$b = 0.25$	0.5	15.1411(337.0099)	8.8333(61.9832)	5.1041(6.0108)
	1.5	3.3854(49.6625)	2.4523(7.0107)	2.5990(2.2115)
	2.5	1.0478(7.7934)	1.2522(2.4986)	1.8552(1.3915)
$b = 1$	0.5	7.4605(143.5667)	1.8484(3.8987)	0.9007(0.4569)
	1.5	0.6757(3.7225)	0.3203(0.3896)	0.2555(0.1952)
	2.5	0.1470(0.3978)	0.0943(0.1015)	0.0869(0.0795)
$b = 1.75$	0.5	4.6069(72.9978)	0.8633(1.2649)	0.2662(0.1998)
	1.5	0.3096(1.1287)	0.0837(0.0890)	0.0103(0.0102)
	2.5	0.0532(0.1039)	0.0129(0.0131)	0.0005(0.0005)

3.2. Infinite divisibility

The Central Limit Theorem and waiting time distributions are closely related to infinite divisibility. In accordance with Steutel and Van Harn (2003), If $p(x), x \in \mathbb{N}_0$ is infinitely divisible, then $p(x) < e^{-1}$ for all $x \in \mathbb{N}$. We can observe that for the PMLG distribution with parameters $b = 0.4$, α =2 and $\theta = 5$, $r(1) = 0.4346001 > e^{-1} = 0.367$. It follows that the PMLG distribution is not infinitely divisible. Additionally, since the discrete concepts of self-decomposable and stable distributions are subclasses of infinitely divisible distributions, we are able to conclude that the PMLG distribution cannot be either of these properties.

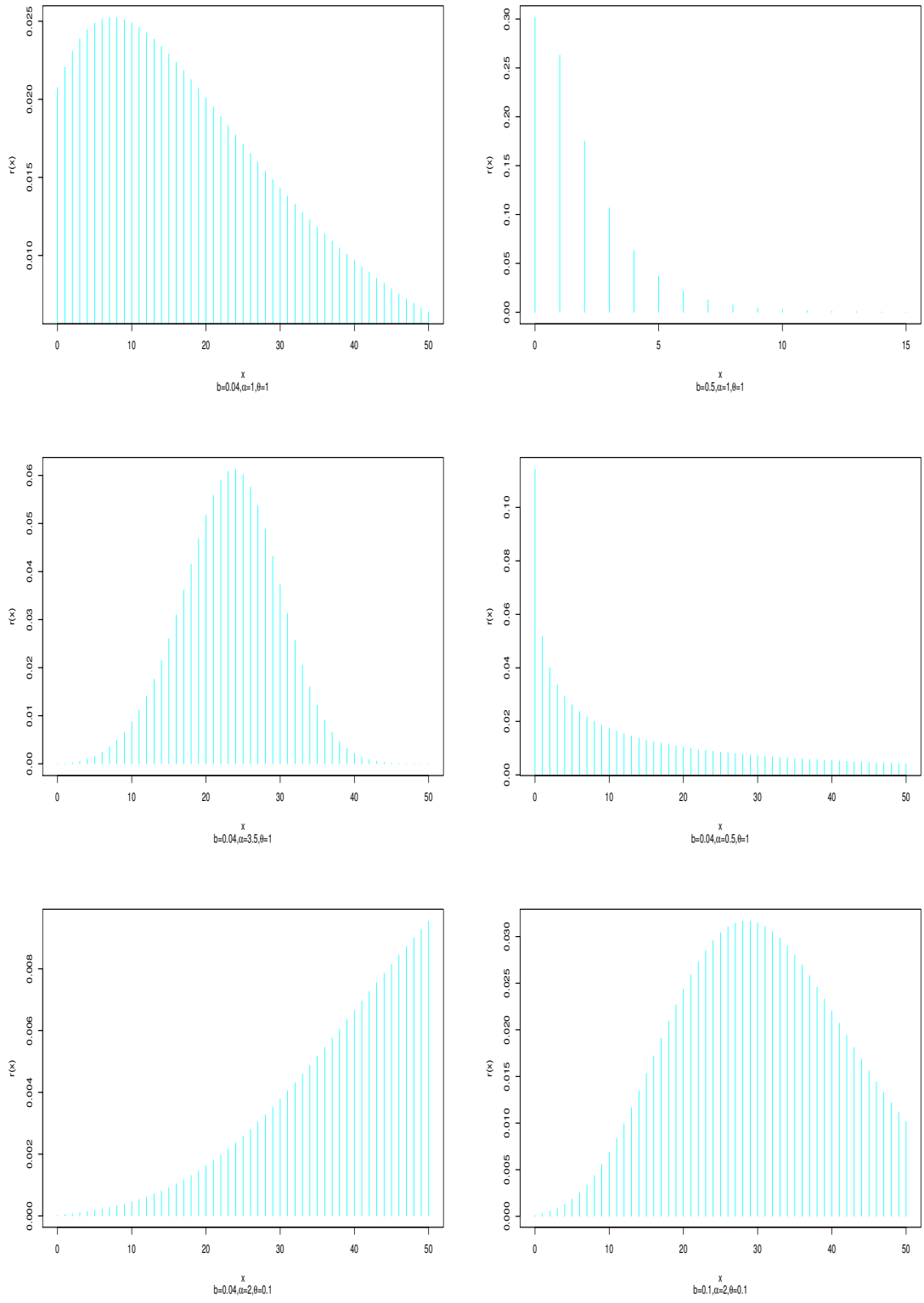


Figure 1: *PMF*s of some parameter values for the PMLG distribution

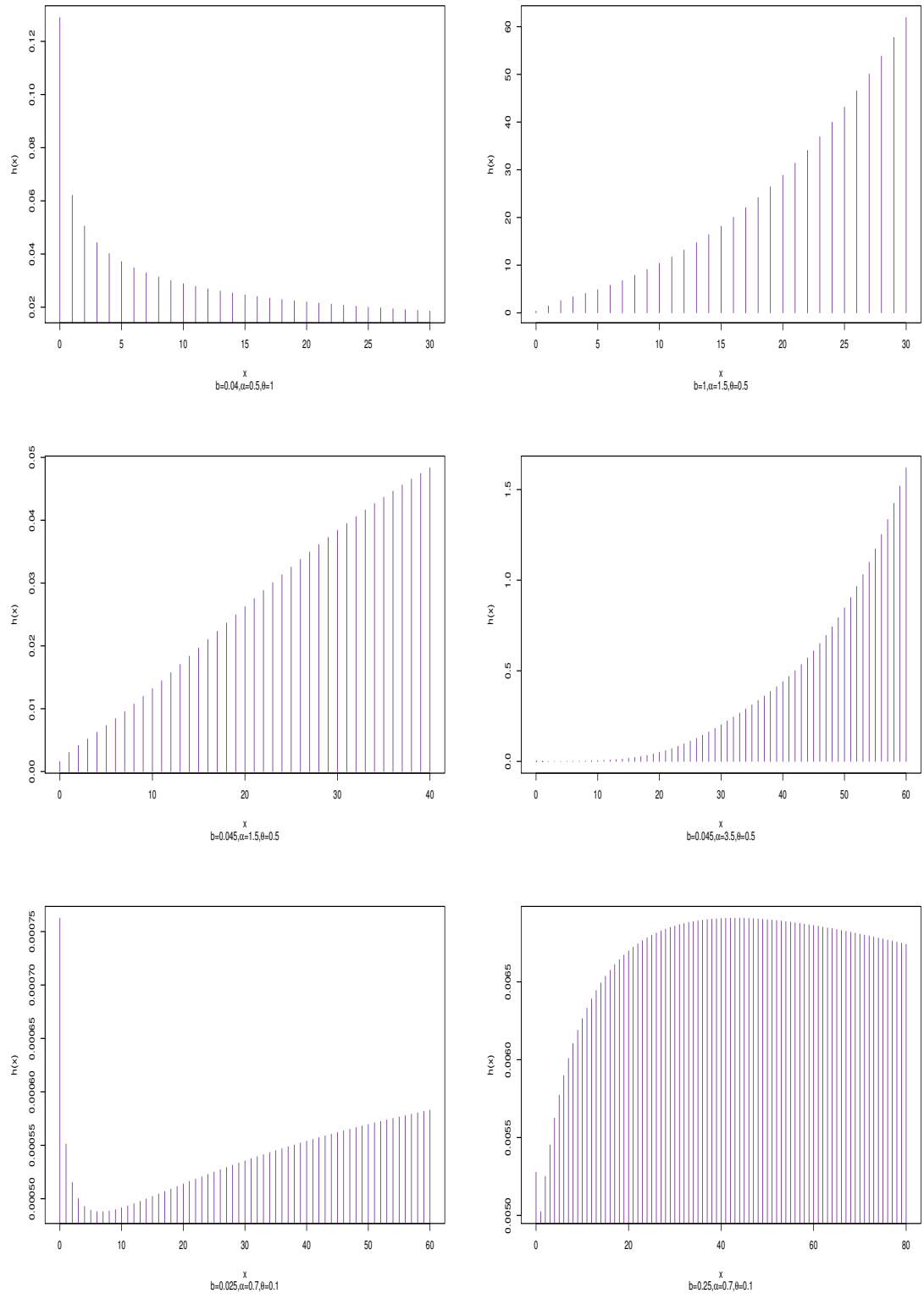


Figure 2: *HRF*s of some parameter values for the PMLG distribution

4. Parameter estimation

In this section, we focus on the many classical estimating techniques. There are numerous and different classical approaches, some of which rely on the theory of maximisation and others on the theory of minimization. This section includes, maximum likelihood, Cramer-von-Mises, least squares and weighted least squares approaches of estimation as part of four classical estimation methods.

4.1. Maximum likelihood approach of estimation

If we choose $x_1, x_2, \ldots, x_n$ to be a random sample from the PMLG distribution with unknown parameters b, α and θ and, the likelihood function is given by

$$\begin{aligned} L(\alpha,\theta,b) &= \prod_{i=1}^{n} g(x_i;\alpha,\theta,b) \\ &= \prod_{i=1}^{n}\left[1+\frac{\theta(bx_i)^{\alpha}}{1+\theta}e^{-\theta(bx_i)^{\alpha}}\right]e^{-\theta(bx_i)^{\alpha}} - \left[1+\frac{\theta(b(x_i+1))^{\alpha}}{1+\theta}e^{-\theta(b(x_i+1))^{\alpha}}\right]e^{-\theta(b(x_i+1))^{\alpha}}. \end{aligned}$$

The log-likelihood function follows immediately as

$$\begin{aligned} \ell(b,\alpha,\theta) &= \log\left[L(b,\alpha,\theta)\right] \\ &= \sum_{i=1}^{n} log\Bigg\{\left[1+\frac{\theta(bx_i)^{\alpha}}{1+\theta}e^{-\theta(bx_i)^{\alpha}}\right]e^{-\theta(bx_i)^{\alpha}} \\ &\quad - \left[1+\frac{\theta(b(x_i+1))^{\alpha}}{1+\theta}e^{-\theta(b(x_i+1))^{\alpha}}\right]e^{-\theta(b(x_i+1))^{\alpha}}\Bigg\}. \end{aligned}$$

The first derivatives of $\ell(b,\alpha,\theta)$ with respect to b, α and θ are

$$\begin{aligned} \frac{\partial\ell(b,\alpha,\theta)}{\partial b} &= \sum_{i=1}^{n}\frac{\frac{\alpha\theta\Delta_1}{b}\left\{\frac{e^{-\theta(bx_i)^{\alpha}}}{1+\theta}\left[1-\theta(bx_i)^{\alpha}\right]-\left[1+\frac{\theta}{1+\theta}\Delta_1\right]\right\}}{\left[1+\frac{\theta}{1+\theta}\Delta_1\right]e^{-\theta(bx_i)^{\alpha}}-\left[1+\frac{\theta}{1+\theta}\Delta_2\right]e^{-\theta(b(x_i+1))^{\alpha}}} \\ &\quad - \sum_{i=1}^{n}\frac{\frac{\alpha\theta\Delta_2}{b}\left\{\frac{e^{-\theta(b(x_i+1))^{\alpha}}}{1+\theta}\left[1-\theta(b(x_i+1))^{\alpha}\right]-\left[1+\frac{\theta}{1+\theta}\Delta_2\right]\right\}}{\left[1+\frac{\theta}{1+\theta}\Delta_1\right]e^{-\theta(bx_i)^{\alpha}}-\left[1+\frac{\theta}{1+\theta}\Delta_2\right]e^{-\theta(b(x_i+1))^{\alpha}}} \end{aligned}$$

$$\begin{aligned} \frac{\partial\ell(b,\alpha,\theta)}{\partial\alpha} &= \sum_{i=1}^{n}\frac{\theta\Delta_1 log(bx_i)\left\{\frac{e^{-\theta(bx_i)^{\alpha}}}{1+\theta}\left[1-\theta(bx_i)^{\alpha}\right]-\left[1+\frac{\theta}{1+\theta}\Delta_1\right]\right\}}{\left[1+\frac{\theta}{1+\theta}\Delta_1\right]e^{-\theta(bx_i)^{\alpha}}-\left[1+\frac{\theta}{1+\theta}\Delta_2\right]e^{-\theta(b(x_i+1))^{\alpha}}} \\ &\quad - \sum_{i=1}^{n}\frac{\theta\Delta_2 log(b(x_i+1))\left\{\frac{e^{-\theta(b(x_i+1))^{\alpha}}}{1+\theta}\left[1-\theta(b(x_i+1))^{\alpha}\right]-\left[1+\frac{\theta}{1+\theta}\Delta_2\right]\right\}}{\left[1+\frac{\theta}{1+\theta}\Delta_1\right]e^{-\theta(bx_i)^{\alpha}}-\left[1+\frac{\theta}{1+\theta}\Delta_2\right]e^{-\theta(b(x_i+1))^{\alpha}}} \end{aligned}$$

$$\frac{\partial \ell(b,\alpha,\theta)}{\partial \theta} = \sum_{i=1}^{n} \frac{\Delta_1 \left\{ \frac{e^{-\theta(bx_i)^\alpha}}{(1+\theta)^2} \left[(1+\theta)(1-\theta(bx_i)^\alpha) - \theta\right] - \left[1 + \frac{\theta}{1+\theta}\Delta_1\right] \right\}}{\left[1 + \frac{\theta}{1+\theta}\Delta_1\right] e^{-\theta(bx_i)^\alpha} - \left[1 + \frac{\theta}{1+\theta}\Delta_2\right] e^{-\theta(b(x_i+1))^\alpha}}$$
$$- \sum_{i=1}^{n} \frac{\Delta_2 \left\{ \frac{e^{-\theta(b(x_i+1))^\alpha}}{(1+\theta)^2} \left[(1+\theta)(1-\theta(b(x_i+1))^\alpha) - \theta\right] - \left[1 + \frac{\theta}{1+\theta}\Delta_2\right] \right\}}{\left[1 + \frac{\theta}{1+\theta}\Delta_1\right] e^{-\theta(bx_i)^\alpha} - \left[1 + \frac{\theta}{1+\theta}\Delta_2\right] e^{-\theta(b(x_i+1))^\alpha}}$$

Where, $\Delta_1 = e^{-\theta(bx_i)^\alpha}(bx_i)^\alpha$and $\Delta_2 = e^{-\theta(b(x_i+1))^\alpha}(b(x_i+1))^\alpha$.

Setting $\frac{\partial \ell(b,\alpha,\theta)}{\partial b} = 0$, $\frac{\partial \ell(b,\alpha,\theta)}{\partial \alpha} = 0$ and $\frac{\partial \ell(b,\alpha,\theta)}{\partial \theta} = 0$, and then solving the equations iteratively will yield the maximum likelihood (ML) estimators of b, α and θ. These equations are complicated to solve analytically. One can use mathematical software to get numerical solutions.

4.2. Cramer-von-Mises approach of estimation

The Cramer-von-Mises (CVM) estimation approach is a significant estimation method that was discussed in Macdonald (1971). The CVM estimation technique's parameters can be calculated by minimising the function CVM in respect to the unknown parameters.

$$CVM = \frac{1}{12} + \sum_{i=1}^{n} \left\{ G(x_i;\alpha,\theta,b) - \frac{2i-1}{2n} \right\}$$
$$= \frac{1}{12} + \sum_{i=1}^{n} \left\{ 1 - \left[1 + \frac{\theta(b(x_i+1))^\alpha}{1+\theta} e^{-\theta(b(x_i+1))^\alpha} \right] e^{-\theta(b(x_i+1))^\alpha} - \frac{2i-1}{2n} \right\}$$

4.3. Least square approach of estimation

Assume that $x_1, x_2, \ldots, x_n$ is a randomly selected sample of size n from the PMLG distribution and that $x_{1:n}, x_{2:n}, \ldots, x_{n:n}$ signifies a corresponding ordered sample. Consequently, the following quantity can be minimized to produce least squares (LS) estimators for PMLG parameters

$$LS = \sum_{i=1}^{n} \left\{ G(x_{i:n};\alpha,\theta,b) - \frac{i}{n+1} \right\}^2$$
$$= \sum_{i=1}^{n} \left\{ 1 - \left[1 + \frac{\theta(b(x_{i:n}+1))^\alpha}{1+\theta} e^{-\theta(b(x_{i:n}+1))^\alpha} \right] e^{-\theta(b(x_{i:n}+1))^\alpha} - \frac{i}{n+1} \right\}^2$$

with respect to b, α and θ respectively.

4.4. Weighted least square approach of estimation

The weighted least square (WLS) estimators of the unknown parameters for the PMLG distribution are derived in this subsection. Let $x_1, x_2, \ldots, x_n$ be a random sample and $x_{1:n}, x_{2:n}, \ldots, x_{n:n}$ be the corresponding ordered sample of size n from the PMLG distribution. The following sum of squares errors can be minimised to generate the PMLG estimators

$$WLS = \sum_{i=1}^{n} \frac{(n+1)^2(n+2)}{i(n-i+1)} \left\{ G(x_{i:n}; \alpha, \theta, b) - \frac{i}{n+1} \right\}^2$$
$$= \sum_{i=1}^{n} \frac{(n+1)^2(n+2)}{i(n-i+1)} \left\{ 1 - \left[1 + \frac{\theta(b(x_{i:n}+1))^\alpha}{1+\theta} e^{-\theta(b(x_{i:n}+1))^\alpha} \right] e^{-\theta(b(x_{i:n}+1))^\alpha} - \frac{i}{n+1} \right\}^2$$

with respect to b, α and θ respectively.

5. Simulation

Here, a simulation study is used to examine how well various estimates of the PMLG distribution work. Using the PMLG distribution, we produce random data with varying sample sizes and parameter values. The simulation research is run N= 1000 times with n = 50, 100, 150, and 200 as the sample size and the chosen parameter values. We compute the *ML*, *CVM*, *LS* and *WLS* estimates of b, α and θ. Based on the calculated results estimates, average biases (*Bias*) and mean squared errors (*MSE*s) measurements are calculated. The results of this simulation are shown in Tables 2 and 3. We can draw the following interpretations from the tables:

- With larger sample sizes, all estimates experience a decreasing trend in *MSE*s and *Bias* decays towards zero.
- The *LS* estimates *MSE*s are lower than those for the *ML*, *WLS*, and *CVM* estimates.

Table 2: The *Bias* and *MSE* of the *ML*, CVM, LS and WLS estimates for b=0.5, α=0.4 and θ=0.045

n		$Bias(\hat{b})$	$MSE(\hat{b})$	$Bias(\hat{\alpha})$	$MSE(\hat{\alpha})$	$Bias(\hat{\theta})$	$MSE\ (\hat{\theta})$
50	*ML*	0.0322	1.3889	-0.4728	0.2365	-0.3687	0.2769
	CVM	-0.0348	0.0829	-0.4892	0.2447	-0.3869	0.1582
	LS	-0.0005	0.2704e-04	-0.0001	0.4885e-06	-0.0001	0.1155e-05
	WLS	-0.0303	0.2046	-0.4945	0.2471	-0.3941	0.1595
100	*ML*	-0.0168	0.5318	-0.4434	0.2217	-0.3553	0.1421
	CVM	-0.0050	0.0335	-0.1889	0.0945	-0.1284	0.0627
	LS	-0.0002	0.2899e-05	-0.3202e-04	0.6908e-07	-0.4317e-04	0.1450e-06
	WLS	-0.0178	0.1858	-0.4163	0.2080	-0.3249	0.1388
150	*ML*	0.0267	0.4623	0.0003	0.4505e-04	0.0093	0.0537
	CVM	0.0009	0.0001	-0.0045	0.0023	0.0042	0.0020
	LS	-0.6792e-05	0.4697e-06	-0.2010e-05	0.1184e-07	-0.2032e-06	0.2626e-07
	WLS	-0.0084	0.0997	-0.2404	0.1202	-0.1848	0.0779
200	*ML*	0.0003	0.6853e-04	-0.1957e-04	0.3828e-06	-0.0002	0.5283e-04
	CVM	0.0003	0.5130e-04	-0.0010	0.0005	0.0009	0.0005
	LS	-0.3016e-06	0.9838e-07	-0.6776e-06	0.2865e-08	-0.7719e-06	0.590e-08
	WLS	0.37691e-04	0.1421e-05	-0.0005	0.0002	0.0005	0.0002

Table 3: The *Bias* and *MSE* of the *ML*, *CVM*, *LS* and *WLS* estimates for b=0.6, α=0.5 and θ=0.05

n		$Bias(\hat{b})$	$MSE(\hat{b})$	$Bias(\hat{\alpha})$	$MSE(\hat{\alpha})$	$Bias(\hat{\theta})$	$MSE\ (\hat{\theta})$
50	*ML*	-0.0143	0.3723	-0.5719	0.3436	-0.4532	0.5205
	CVM	-0.0172	0.1487	-0.4736	0.2841	-0.3554	0.19245
	LS	-0.0007	0.1823e-04	-0.0003	0.383e-04	0.0001	0.4367e-04
	WLS	-0.0459	0.0101	-0.5860	0.3519	-0.4846	0.2451
100	*ML*	-0.0043	-0.0043	-0.3523	0.2123	-0.2616	0.4132
	CVM	-0.0107	0.0010	-0.1783	0.1071	-0.1120	0.0706
	LS	-0.0002	0.2613e-05	-0.1761e-04	0.2295e-07	-0.2939e-04	0.7558 e-07
	WLS	-0.0220	0.0301	-0.3597	0.2156	-0.282	0.1592
150	*ML*	0.0028	0.0070	0.1515e-04	0.21181e-06	-0.0007	0.0003
	CVM	-0.0030	0.0002	-0.0389	0.0233	-0.0291	0.0157
	LS	-0.3115e-04	0.2720e-06	-0.3128e-05	0.2699e-08	-0.5645e-05	0.8846e-08
	WLS	0.0030	0.0095	-0.0027	0.0015	0.0026	0.0017
200	*ML*	0.0007	0.0002	-0.6898e-05	0.1166e-06	-0.0007	0.0002
	CVM	0.9024e-04	0.4348e-05	-0.0011	0.0006	0.0006	0.0002
	LS	-0.1707e-04	0.1457e-06	-0.1788e-05	0.1598e-08	-0.3128e-06	0.4894e-08
	WLS	0.0009	0.0007	-0.0012	0.0007	0.0007	0.0003

6. Application

This section uses two actual count data sets to demonstrate the significance of the PMLG distribution over the existing models, namely exponentiated exponential-geometric (EEG) distribution and Kumaraswamy-geometric (KG) distribution, in modelling count data from the field of medicine. We used the maximum likelihood method to estimate the values of the unknown parameters in order to compare these distributions. Additionally, the estimated log-likelihood function ($\hat{\ell}$), Akaike Information Criterion (AIC), correct Akaike information criterion ($AICc$), Anderson-Darling statistic (A), Cramér von Mises statistic (W) and Kolmogorov-Smirnov (K-S) statistic with p-value (p-V) are used to compare the fitted distributions. The following displays the considered data sets.

Data set I:The first data set shows the number of COVID-19-related deaths that occurred on a daily basis in the United Kingdom from August 1 through August 28, 2021. This information is obtained from the website
$https://www.worldometers.info/coronavirus/country/uk/$, which lists the number of deaths caused by COVID-19 in the United Kingdom on a daily basis. The data set is provided below.
{65, 24, 138, 119, 86, 92, 103, 39, 37, 140, 104, 94, 100, 91, 61, 26, 170, 111, 113, 114, 104, 49, 40, 174, 149, 140, 100, 133}

Data set II:The second data set, which has 42 observations and is available on the Worldometer website through
$https://www.worldometers.info/coronavirus/country/Egypt/$, shows the number of daily COVID-19 infection-related deaths that occurred in Egypt from 13 March to 30 April 2020. The data are as follows.
{1, 2, 4, 5, 1, 1, 3, 6, 6, 4, 1, 5, 6, 6, 8, 5, 7, 7, 9, 9, 15, 17, 11, 13, 5, 14, 5, 13, 9, 19, 15, 11,

14, 12, 11, 7, 13, 10, 20, 22, 21, 12}

Tables 4, 5, 6 and 7, contain the MLEs, $(-\hat{\ell})$, AIC and goodness-of-fit tests for COVID-19 data sets. The analysis yields the PMLG distribution with the lowest $-\hat{\ell}$, AIC, $AICc$, $HQIC$, A, W, K-S statistic, and highest p-Vs. The PMLG distribution is the appropriate one based on these results. We can say from the two applications that the PMLG distribution is the best model for capturing the daily deaths by COVID-19.

Figure 3 gives the total time test (TTT)-plots of PMLG distribution for the COVID-19 data sets. The TTT-plots shows increasing HRF, allowing us to fit PMLG distribution. Figure 4 display the probability-probability (PP) plots for the two data sets, respectively. The PMLG distribution offers a better fit for the COVID-19 data sets, which support the findings in Tables 4, 5, 6 and 7.

Table 4: Estimated values, $-\hat{\ell}$, AIC, and $AICc$ for the data set I

Distribution	Estimates	$-\hat{\ell}$	AIC	AICc
PMLG	$\hat{\boldsymbol{\alpha}} = \mathbf{2.5303}$	**143.5036**	**293.0072**	**294.0072**
	$\hat{\boldsymbol{\theta}} = \mathbf{8.6018}$			
	$\hat{\boldsymbol{b}} = \mathbf{0.0039}$			
EEG	$\hat{\alpha} = 4.8971$	146.008	296.0161	296.4961
	$\hat{\theta} = 0.9774$			
KG	$p = 0.9941$	144.3038	294.6075	295.6075
	$\hat{\alpha} = 3.1638$			
	$\hat{\theta} = 10.4923$			

Table 5: A, W and K-S with p-Vs for the data set I

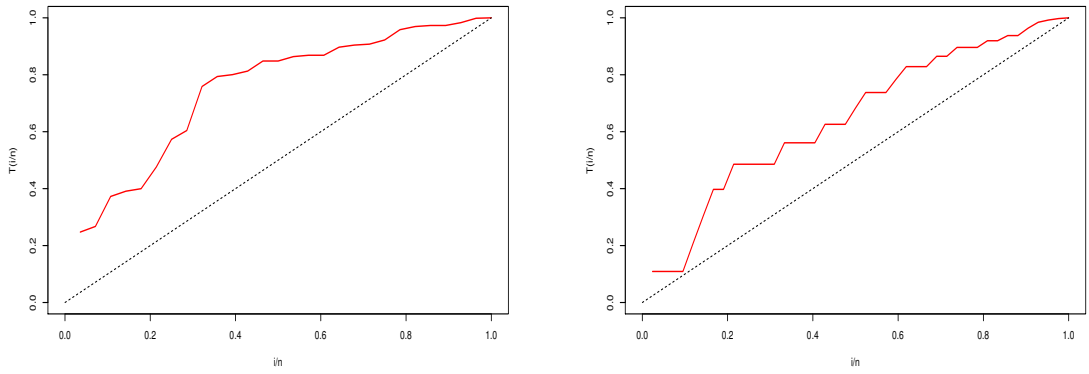
Distribution	A	W	K-S	p-Vs
PMLG	**0.5679**	**0.0941**	**0.14277**	**0.6179**
EEG	1.0461	0.2023	0.2072	0.1805
KG	0.7393	0.1351	0.1720	0.3785

Table 6: Estimated values, $-\hat{\ell}$ and AIC and $AICc$ for the data set II

Distribution	Estimates	$-\hat{\ell}$	AIC	AICc
PMLG	$\hat{\boldsymbol{\alpha}} = \mathbf{1.7357}$	**129.0244**	**264.0489**	**264.6805**
	$\hat{\boldsymbol{\theta}} = \mathbf{11.6800}$			
	$\hat{\boldsymbol{b}} = \mathbf{0.0227}$			
EEG	$\hat{\alpha} = 2.6088$	130.1956	264.3913	264.699
	$\hat{\theta} = 0.8385$			
KG	$p = 0.9734$	129.243	264.486	265.1176
	$\hat{\alpha} = 1.9437$			
	$\hat{\theta} = 14.2437$			

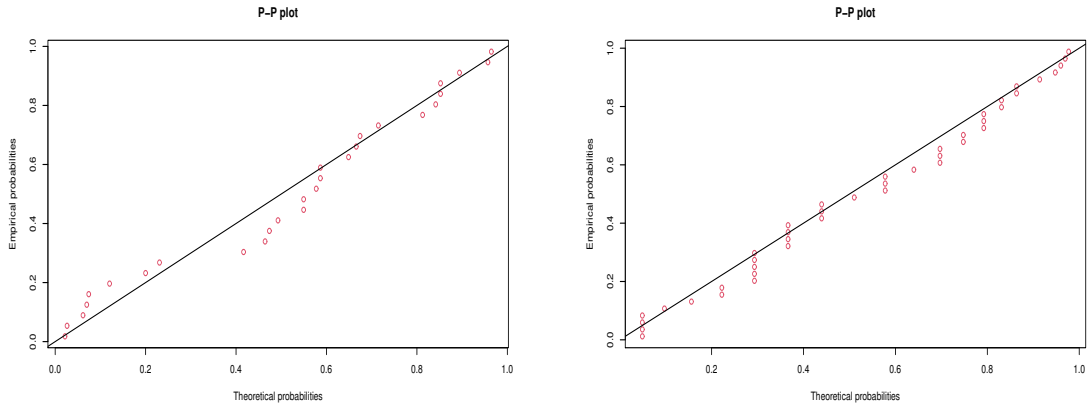
Table 7: *A*, *W* and *K-S* with *p-V*s for the data set II

Distribution	A	W	K-S	p-Vs
PMLG	**0.4625**	**0.0725**	**0.10275**	**0.767**
EEG	0.8018	0.1447	0.13774,	0.403
KG	0.5264	0.0871	0.10979	0.692



(a) (b)

Figure 3: *TTT*-plots for the COVID-19 (a) data set I and (b) data set II



(a) (b)

Figure 4: *PP*-plots for the COVID-19 (a) data set I and (b) data set II

7. Conclusion

In this study, we suggested an entirely novel family of discrete PML-X distributions. PMLG distribution is a specific instance of this family that is thoroughly researched. *ML*, *CVM*, *OLS* and *WLS* techniques have been used to estimate the model parameters. A simulation study is conducted to evaluate the effectiveness of the various estimating techniques. In order to demonstrate the significance and adaptability of defined distribution, two real data sets are analysed at the end. We anticipate that the suggested model will replace

various types of discrete distributions found in the statistical literature.

References

Akinsete, A., Famoye, F., and Lee, C. (2014). The Kumaraswamy-geometric distribution. *Journal of Statistical Distributions and Applications*, **1**, 1–21.

Alzaatreh, A., Lee, C., and Famoye, F. (2012). On the discrete analogues of continuous distributions. *Statistical Methodology*, **9**, 589–603.

Alzaatreh, A., Lee, C., and Famoye, F. (2013). A new method for generating families of continuous distributions. *Metron*, **71**, 63–79.

Chesneau, C., Tomy, L., and Gillariose, J. (2021a). A new modified Lindley distribution with properties and applications. *Journal of Statistics and Management Systems*, **24**, 1383–1403.

Chesneau, C., Tomy, L., and Jose, M. (2021b). Power modified Lindley distribution: Theory and applications. *Journal of Mathematical Extension*, **16**, 1–32.

Famoye, F. (2019). Exponentiated Weibull-geometric distribution and its application to count data. *Journal of Data Science*, **17**, 712–725.

Macdonald, P. (1971). Comments and queries comment on "an estimation procedure for mixtures of distributions" by choi and bulgren. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, **33**, 326–329.

Steutel, F. W. and Van Harn, K. (2003). *Infinite Divisibilty of Probability Distributions on the Real Line*. CRC Press.

Tomy, L., Jose, M., and Jose, M. (2019). The T-X family of distributions: A retrospect. *Think India Journal*, **22**, 9407–9420.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 22, No. 2, 2024 (New Series), pp 217–229
http://www.ssca.org.in/journal

Wavelet-ARIMA-TDNN Model for Agricultural Commodity Price Forecasting

Sathees Kumar K., Banjul Bhattacharyya, Gowthaman T. and Elakkiya N.
Department of Agricultural Statistics, Bidhan Chandra Krishi Viswavidyalaya, Mohanpur, Nadia, West Bengal, India.

Received: 10 November 2022; Revised: 01 November 2023; Accepted: 08 January 2024

Abstract

In every agricultural market, accurate agricultural commodity price forecasting is essential for farmers, traders, policymakers, and government sectors. Decomposition of the price series has sufficiently increased the forecast accuracy. In the past years, wavelet analysis has been widely used for the decomposition of price series, where it converted time series into high and low frequencies. Often, without accounting for the linearity of the frequencies in wavelet-hybrid models, those frequencies are modeled directly. A major problem arises when wavelet-hybrid models contain both linear and non-linear frequencies. Hence, a type of wavelet-hybrid model was developed to solve this problem. Tomato's monthly wholesale price in the Mumbai market was used in this study. First, linear, and non-linear frequencies are separated by the McLeod and Li test after the wavelet decomposition of the tomato price series. Autoregressive Integrated Moving Average (ARIMA) and Time Delay Neural Network (TDNN) were applied to linear and non-linear frequencies, respectively. Forecasts of ARIMA and TDNN were reconstructed to obtain forecasts of the tomato price series. Finally, our proposed wavelet-ARIMA-TDNN model was compared to ARIMA, TDNN, and Wavelet-ARIMA, Wavelet-TDNN. The result revealed that our proposed method outperformed other models.

Key words: McLeod and Li test; Non-linearity; Decomposition; Wavelet analysis; Wavelet-ARIMA-TDNN.

AMS Subject Classifications: 62K05, 05B05

1. Introduction

In a populous country like India, satisfying people's daily food demands is cumbersome. Thus, farmers have the responsibility to increase food production, especially for vegetables. It is due to the perishable nature of vegetables, which causes their prices to fluctuate and affects farmers' revenue. Therefore, predicting this price fluctuation is essential for farmers, traders, policymakers, government sectors, *etc.* Forecasting this highly volatile price is a very challenging task for forecasters. Understanding the nature of the price series

Corresponding Author: Sathees Kumar K.
Email: satheesamma55@gmail.com

is important for forecasting it. Generally, time series follow either a linear or non-linear pattern.

The ARIMA model is one of the most important and widely used models for linear time series. Due to its inherent statistical properties and use of the Box George *et al.* (1976) approach, the ARIMA model is popular. On the other hand, ANNs provide good self-learning and non-linear approximation skills when dealing with non-linear complex data sets. ANN has some success with predicting applications with a lag, particularly for non-linear time series. However, there is no specific model to handle all the circumstances. But we can get good results through the appropriate application of suitable models.

The ARIMA model performs well as a predictor for linear time series. Singla *et al.* (2021) found that ARIMA model outperformed wavelet-hybrid models for the onion price series. But, the ARIMA model's precision is insufficient to address complex non-linear situations. Jha and Sinha (2014) showed that ANN models provide better prediction accuracy for non-linear patterns than ARIMA models. Although ANNs are effective against non-linear time series, they might produce inconsistent results against linear models. Additionally, it shows that the sampling size and noise level affect the performance of linear regression model using ANN Markham and Rakes (1998).

Decomposition of time series is an essential process in modelling. Wavelet analysis Antoniadis (1997) is extensively used for decomposition and converts the time series into high and low frequencies. These frequencies are fitted using time series models, which are known as wavelet-hybrid models. Generally, without considering the linearity or non-linearity of the data's frequencies, time series models are used for modelling the frequencies in wavelet-hybrid models. Also, Paul *et al.* (2020) found that wavelet-ANN outperformed wavelet-ARIMA for modelling sub-divisional rainfall data. Nury *et al.* (2017) reported that wavelet-ARIMA was performed better than wavelet-ANN for temperature time series data. The above studies indicate that the linearity or non-linearity of the frequencies is a significant factor in wavelet-hybrid models. For instance, Anjoy *et al.* (2017) fitted the ANN model to all frequencies due to their non-linearity. Similarly,Ray *et al.* (2020) were fitted WNN to high frequencies and ANN to low frequency due to their non-linear pattern. Also, it is possible to get both linear and non-linear frequencies after the wavelet decomposition of a time series. In such a situation, it is not optimal to fit all frequencies using the same time series model.

In this research, the problem of containing both linear and non-linear frequencies was addressed. According to this problem, the wavelet-ARIMA-TDNN model was developed to obtain reliable and accurate forecasting.

2. Materials and Methods

Monthly wholesale prices of Tomato for the Mumbai market (Jan-2011 to Dec-2021) was collected from AGMARKNET (https://agmarknet.gov.in/) website.

2.1. Wavelet Analysis

Wavelets are underlying building block functions like trigonometric sine and cosine functions. A wavelet function (Equation 1) oscillates about zero.

$$\psi_{\tau,s} = \frac{1}{\sqrt{|s|}}\,\psi\left(\frac{t-\tau}{s}\right) \qquad \tau, s \in R,\ s \neq 0 \tag{1}$$

Here, τ - Translation parameter s - Scaling parameter.

Wavelets are well-described in Daubechies *et al.* (1992), Ogden (1997), and Percival and Walden (2000). Based on scaling and translation parameters, two types of wavelet transforms (continuous and discrete) exist. The continuous wavelet transform (CWT) provides coefficients for the entire real axis, which are more than necessary for extracting frequencies. On the other hand, due to scaling proportional steps of translation parameters, the dyadic discrete wavelet transform (DWT) requires a sample size of multiples of two. If the sample size is $2^J, J$ is known as the maximum level of decomposition. Equation (2) is the dyadic DWT.

$$\psi_{m,n}(t) = \frac{1}{\sqrt{2}}\psi\left(\frac{t-n2^m}{2^m}\right) \tag{2}$$

where, $\frac{1}{\sqrt{2}}$ - variance preserving factor; m- scaling parameter; n- translation parameter (ranges from 1 to 2^{J-m}).

These reasons lead to the requirement of a modified wavelet transform, which is known as a maximal overlap discrete wavelet transform (MODWT).

2.2. Maximal Overlap Discrete Wavelet Transform

A MODWT (Equation 3) can be obtained from a slight modification of dyadic DWT. In MODWT, the translation parameter is not proportional to the scaling parameter where wavelets are convoluted in each time interval for all the dyadic scales, so there is no restriction on sample size. For N sample size,

$$\psi_{m,n}(t) = \frac{1}{\sqrt{2}}\psi\left(\frac{t-n}{2^m}\right) \tag{3}$$

where n ranges from 1 to N.

It produces an overlapping tile in the time-frequency plane, so the transform is not orthogonal Percival and Walden (2000). Because of the non-orthogonality, it demands an orthogonal filter for perfect reconstruction. Based on linear filter operation, MODWT gives high frequencies and low frequencies using synthesis filters. MODWT provides J high frequencies and one low frequency at the J^{th} decomposition level. The maximum level of decomposition for the N sample size is $J = \log_2(N)$. It ranges from 1 to J. This transform partitions variance across the scale. Frequencies are reconstructed by inverse maximal overlap discrete wavelet transform (IMODWT). The variance of reconstructed series at any J^{th} level and variance of actual time series are always equal, which explains that MODWT is the variance-preserving transform.

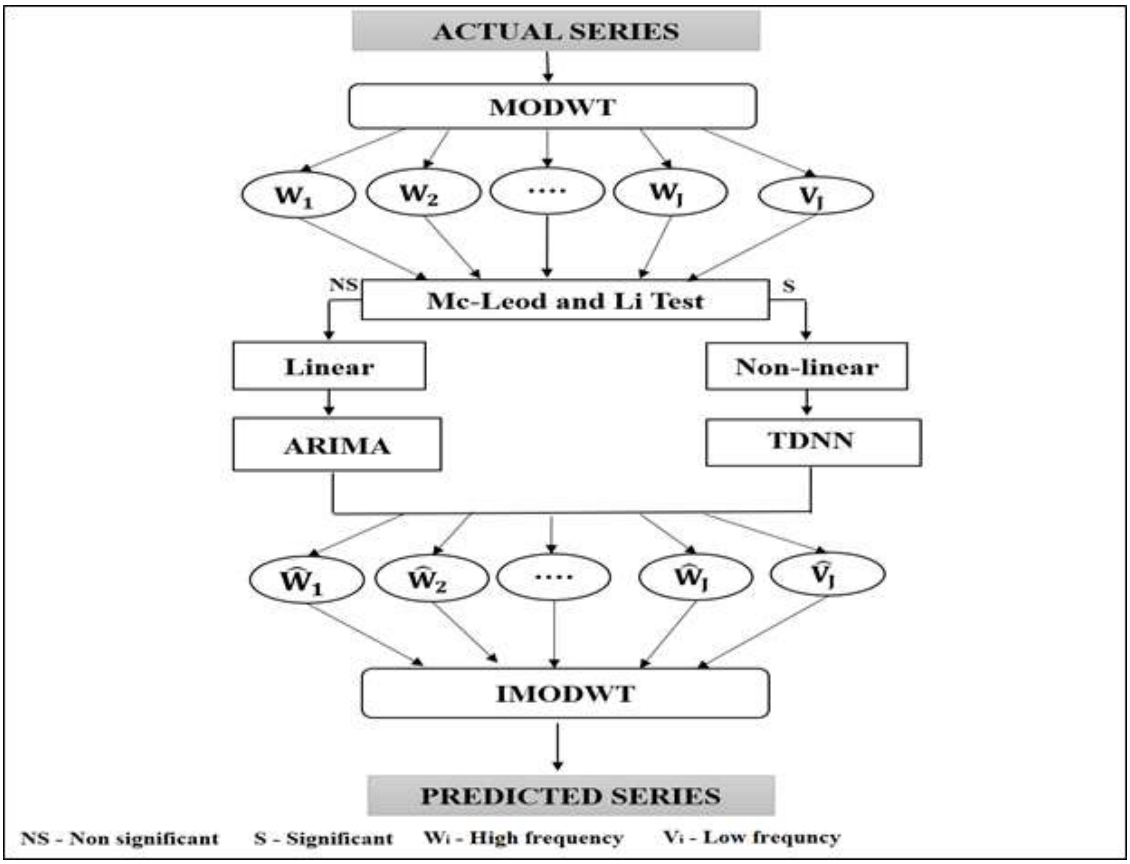


Figure 1: Schematic representation of the proposed wavelet-ARIMA-TDNN model

2.3. Wavelet-ARIMA-TDNN

Wavelet-hybrid models are the combination of wavelet analysis and time series analysis, in which time series are converted into high and low frequencies using wavelet analysis and then fitted by any time series model to increase the forecast accuracy. In this research, we developed a hybrid time series forecasting method that combines features of wavelet transformation, ARIMA, and TDNN based on the non-linearity test.

Since some of the time series data contain both linear and non-linear frequencies, the following method is developed:

Step 1: MODWT divides time series into high and low-frequency components.

Step 2: Test the non-linearity for each frequency using the McLeod and Li test.

Step 3: Identify the linear and non-linear frequencies which are fitted by ARIMA and TDNN, respectively.

Step 4: Reconstruct the forecast value of frequencies obtained from fitted models by IMODWT.

Figure 1 shows the schematic representation of the wavelet-ARIMA-TDNN model. The Haar

filter was used in this study. Among all types of filters, only the Haar filter has the property of discontinuity. So, it can capture sudden changes in the signal.

2.4. Non-linearity test

McLeod and Li (1983) is the Ljung-Box test for squared time series data.

$$Q\left(m\right) = n(n+2)\sum_{j=1}^{m}\frac{r_j^2}{n-j} \tag{4}$$

where, r_j- autocorrelation at j^{th} lag; m-number of lags. Under the null hypothesis of linearity, the statistic (Q) is asymptotically distributed as a Chi-square distribution with m degrees of freedom.

2.5. ARIMA

The combination of Autoregressive and Moving Average processes and the integration is more efficient for achieving higher adaptability of actual time series data. It is denoted as ARIMA (p, d, q). It is one of the linear nonstationary time series models, defined in equation 5. For seasonal time series, ARIMA expanded into SARIMA $(p, d, q)(P, D, Q)$, which stands for Seasonal Autoregressive Integrated Moving Average. It is stated in equation 6.

$$\left(1-\sum_{i=1}^{p}\emptyset_i L^i\right)(1-L)^d y_t = c+\left(1+\sum_{k=1}^{q}\theta_k L^k\right)\varepsilon_t \tag{5}$$

$$\left(1-\sum_{i=1}^{p}\emptyset_i L^i\right)\left(1-\sum_{j=1}^{P}\Phi_j L^j\right)(1-L)^d\ (1-L^D)y_t = c+\left(1+\sum_{k=1}^{q}\theta_k L^k\right)\left(1+\sum_{r=1}^{Q}\Theta_r L^r\right)\varepsilon_t \tag{6}$$

where, L- lag operator; y_t- time series; p- Autoregressive order; P- Seasonal autoregressive order; d- No. of. Differences; D- No. of. Seasonal differences; q- Moving average order; Q- Seasonal moving average order; ε_t- white noise.

2.6. Time-Delay Neural Network

An Artificial Neural Network is for modelling non-linear data sets Ogden (1997), especially unknown relations between input and output datasets, through a data-driven and self-adaptive approach. Over the last few decades, neural modelling systems have been used to deal with a variety of prediction difficulties. The primary theoretical guideline for resolving problematic situations with ANNs is based on the learning principle Valiant (1984). ANN is inspired by human neurological science.

A network of basic processing nodes or neurons that are connected in a certain order to carry out basic arithmetic manipulations is known as a neural network and can be used to forecast future values of potentially noisy time series based on historical data Adamowski and Chan (2011). A Time-delay neural network is an illustration of such a design (TDNN). The number of layers and the total number of nodes in each layer must be selected to create the neural network structure that is appropriate for a given application in time-series prediction. A feed-forward neural network with a single hidden layer and an output node

has been employed in the present investigation. In the hidden layer, the sigmoid function has been used as an activation function with form for the y time series,

$$f(y) = \frac{1}{1+e^{-y}} \tag{7}$$

For g input lag, h hidden nodes in the hidden layer, and one output node, the total number of parameters in a three-layer feed-forward neural network is $h(g+2)+1$.

3. Evaluation criteria

It is necessary to verify the model's accuracy to choose the most suitable model for forecasting. Root Mean Square Error (RMSE) is the standard deviation of the residuals of the model; Mean Absolute Error (MAE) is the average difference of residuals of the model; and Mean Absolute Percentage Error (MAPE) is the percentage of average absolute error which give a way to compare the performance of the different models.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(Y_t - \widehat{Y}_t\right)^2} \tag{8}$$

$$\text{MAE} = \frac{1}{n}\sum_{t=1}^{n}\left|Y_t - \widehat{Y}_t\right| \tag{9}$$

$$\text{MAPE} = \frac{1}{n}\left(\sum_{t=1}^{n}\left|\frac{Y_t - \widehat{Y}_t}{Y_t}\right|\right) * 100 \tag{10}$$

Finally, the Tomato price series forecast accuracy of the developed model (Wavelet-ARIMA-TDNN) was compared to that of Wavelet-ARIMA and Wavelet-TDNN, and the single ARIMA, TDNN in this investigation.

4. Results and Discussion

The tomato price series of the Mumbai market was used to apply the developed hybrid methodology. Descriptive statistics of tomato price series of Mumbai market is given in Table 1. The data set was separated into training data (Jan-2011 to Dec-2020) and validation data groups (Jan-2021 to Dec-2021). The validation data set is used to determine the predictive accuracy after model fitting. The predicting outcomes of various methods, including ARIMA, TDNN, Wavelet-ARIMA, and Wavelet-TDNN, were examined to compare the performance of the suggested methodology with other related techniques in the field. The Ljung-Box (LB) test Ljung and Box (1978) was used to test residual series.

For ARIMA fitting, ACF and PACF plots of stationary series were used to get the possible orders for model fitting. Among all possible models, ARIMA (1,0,1) $(2,1,1)_{[12]}$ gave low AIC and BIC values. Parameter estimates are given in Table 2. The performance of the ARIMA model and its residual test is shown in Table 3.

Table 1: Descriptive statistics of tomato price series in Mumbai market

Mean	Standard deviation	Minimum	Maximum	Range	Skewness	Kurtosis	CV (%)
1439.09	835.67	459.17	4159.38	3700.21	1.18	0.57	58.07

Table 2: Fitted ARIMA model parameter estimates

Parameters	C	AR (1)	MA (1)	SAR (1)	SAR (2)	SMA (1)
Coefficient	7.54**	0.32*	0.41**	0.46**	0.22	-0.76
(S.E)	(2.31)	(0.14)	(0.15)	(0.16)	(0.14)	(0.22)
AIC	1698.66					
BIC	1717.43					

Table 3: Results of fitted ARIMA and TDNN models

Models	Training set			Validation set			Ljung-Box test	
	RMSE	MAE	MAPE	RMSE	MAE	MAPE	Statistic	P value
ARIMA	502.44	339.56	24.92	973.02	808.02	47.04	7.12	0.85
TDNN	594.04	419.45	26.51	855.03	783.06	46.63	20.96	0.06

This study applied an TDNN with a hidden layer, along with sigmoid and identity as activation functions at the hidden and output layers, respectively as per prior studies Jha and Sinha (2014). The backpropagation algorithm can be used to train feed-forward networks in several different ways. In this study, the second-tier training speed was obtained using the Levenberg-Marquardt algorithm Hagen and Menhaj (1994). Rapid convergence into the modestly sized feed forward neural network is provided by this algorithm. Thus, functional approximation issues were addressed by this technique Demuth and Beale (2002). For model fitting, several combinations of input lags and hidden node sizes were tested. Input delays ranged from 1 to 8, whereas hidden neurons ranged from 1 to 10. In the validation data set for the tomato price series, three tapped delay and two hidden nodes (3:2s:1l) provided the lowest RMSE, MAE and MAPE values. Table 3 shows the performance of the TDNN model.

Table 4: Results of Wavelet-ARIMA model for all the decomposition levels

Decomposition Level	Training set			Validation set			Ljung-Box test	
	RMSE	MAE	MAPE	RMSE	MAE	MAPE	Statistic	P value
1	649.52	449.16	42.84	824.55	674.95	34.27	13.64	0.32
2	625.69	431.87	39.04	822.93	650.61	31.80	9.53	0.66
3	628.95	433.14	40.19	844.59	810.80	45.87	8.77	0.72
4	630.34	433.98	44.07	884.13	863.94	56.37	8.04	0.78
5	629.78	434.24	43.97	883.40	862.96	56.25	8.29	0.76
6	629.02	433.50	43.38	884.64	865.01	56.50	8.27	0.76
7	628.46	433.37	42.65	880.66	851.98	54.49	8.30	0.76

Actual tomato price series were decomposed through MODWT from one to seven ($\log_2[120] = 6.9$) decomposition levels. In each J^{th} level of decomposition, tomato price series were separated into J high frequencies (W_1, W_2, ... ,W_J) and one low frequency (V_J) by the Haar mother wavelet, which is a frequently used wavelet, especially for the price series. In the Wavelet-ARIMA model, all the high and low frequencies were fitted by ARIMA for all the decomposition levels without conducting the non-linearity test, and the results of Wavelet-ARIMA are reported in Table 4. Similarly, TDNN was used to fit all high and low frequencies at every decomposition level for the Wavelet-TDNN model without taking linearity into account. The results of Wavelet-TDNN are given in Table 5. Results of the non-linearity test for all the high and low frequencies are shown in Table 6. In our developed model, ARIMA was used to predict linear frequencies (W_1 and W_2) and TDNN was used for modelling non-linear frequencies (W_3, W_4, W_5, W_6, W_7, V_1, V_2, V_3, V_4, V_5, V_6, V_7). Table 7 gives that the models were used to forecast the frequencies at each level of decomposition in the developed hybrid model. Finally, the predicted values of different frequencies from fitted ARIMA and TDNN were used to reconstruct the data series at each level of decomposition.

Table 5: Results of Wavelet-TDNN model for all the decomposition levels

Decomposition Level	**Training set**			**Validation set**			**Ljung-Box test**	
	RMSE	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**	**Statistic**	**P value**
1	342.76	246.59	19.22	595.83	462.56	30.49	11.84	0.46
2	422.48	306.10	23.37	487.13	393.36	26.16	35.19	<0.01
3	440.05	319.89	24.57	651.53	554.93	37.66	36.65	<0.01
4	454.30	325.86	24.96	788.35	735.23	48.05	34.93	<0.01
5	458.38	330.33	25.32	877.61	820.73	52.22	47.10	<0.01
6	462.07	333.82	25.61	770.39	698.75	44.37	36.38	<0.01
7	459.32	331.50	25.42	769.76	694.30	44.04	46.13	<0.01

McLeod and Li's test for the actual series shows (Table 6) that the tomato price series is non-linear. Both ARIMA and TDNN models were fitted for the tomato price series. But TDNN gave better results than the ARIMA model. Therefore, TDNN performed well for non-linear time series.

Table 6: Results of McLeod and Li test

Actual Series		**Statistic**		**P value**	
Y		80.46		<0.01	
Decomposed series					
High frequency	**Statistic**	**P value**	**Low frequency**	**Statistic**	**P value**
W_1	11.95	0.98	V_1	125.15	<0.01
W_2	28.67	0.23	V_2	216.08	<0.01
W_3	97.03	<0.01	V_3	324.69	<0.01
W_4	127.27	<0.01	V_4	806.21	<0.01
W_5	297.19	<0.01	V_5	667.16	<0.01
W_6	589.54	<0.01	V_6	924.31	<0.01
W_7	697.13	<0.01	V_7	226.14	<0.01

Next, without considering the non-linearity, the Wavelet-ARIMA model was used for model fitting, which was fitted at all the levels of decomposition. The 2^{nd} decomposition level gave better results than other decomposition levels. Although Wavelet-ARIMA was fitted, Table 9 shows that it gave less RMSE, MAE and MAPE than the ARIMA model, which confirm that wavelet analysis improves the performance of the ARIMA. Similarly, Wavelet-TDNN was also tried for model fitting, which was fitted only at the first level of decomposition. But Table 6 shows that the tomato price series consists of both significant linear and non-linear frequencies. Due to modelling the linear high frequencies (W_1 and W_2) using TDNN in Wavelet-TDNN, the Ljung-Box test shows that Wavelet-TDNN was not fitted for other levels of decomposition. But Wavelet-TDNN enhanced the performance of TDNN at single level decomposition. To overcome these contrasted applications of linear and non-linear models, the developed hybrid model was applied to all the levels of decomposition. Models used for Wavelet-ARIMA-TDNN at each decomposition level are given in Table 7. The developed hybrid model (Wavelet-ARIMA-TDNN) was given better forecasts than the Wavelet-ARIMA and Wavelet-TDNN models at every decomposition level. Finally, Wavelet-ARIMA-TDNN gave a better forecast at the 2^{nd} level of decomposition than at any other level of decomposition (Table 8). In, W_1, W_2, and V_2 are the outcome frequencies where high frequencies are linear and a low frequency is non-linear. Two level decomposition of tomato price series is given in Figure 2.

Table 7: Models used for Wavelet-ARIMA-TDNN at each decomposition level

Frequencies	**Decomposition level**						
	1	**2**	**3**	**4**	**5**	**6**	**7**
W1	ARIMA	ARIMA	ARIMA	ARIMA	ARIMA	ARIMA	ARIMA
W2	-	ARIMA	ARIMA	ARIMA	ARIMA	ARIMA	ARIMA
W3	-	-	TDNN	TDNN	TDNN	TDNN	TDNN
W4	-	-	-	TDNN	TDNN	TDNN	TDNN
W5	-	-	-	-	TDNN	TDNN	TDNN
W6	-	-	-	-	-	TDNN	TDNN
W7	-	-	-	-	-	-	TDNN
V1	TDNN	-	-	-	-	-	-
V2	-	TDNN	-	-	-	-	-
V3	-	-	TDNN	-	-	-	-
V4	-	-	-	TDNN	-	-	-
V5	-	-	-	-	TDNN	-	-
V6	-	-	-	-	-	TDNN	-
V7	-	-	-	-	-	-	TDNN

When some time series have both linear and non-linear frequencies, it is very difficult to detect the relationship between such a series using either ARIMA or TDNN. The developed hybrid method has captured this complicated relationship significantly. It is important to note that the authors attempted to model this complicated relationship using Wavelet-ARIMA and Wavelet-TDNN, which proved to be less effective than the developed model. Hence, the model fitting and forecast accuracy became worse because ARIMA was unable to model non-linear frequencies and TDNN was unable to capture the linear relationship of the linear frequencies.

Model performance for the validation set (Table 9) shows that a combination of ARIMA and TDNN models based on the non-linearity test along with the MODWT can improve the overall accuracy. Finally, our developed hybrid method can give more appropriate results than the other methods such as ARIMA, TDNN, Wavelet-ARIMA, and Wavelet-TDNN models, especially for the series that contain linear and non-linear frequencies. Forecasts of the tomato price series from the developed model are given in Figure 3.

Table 8: Results of Wavelet-ARIMA-TDNN model for all the decomposition levels

Decomposition Level	**Training set**			**Validation set**			**Ljung-Box test**	
	RMSE	**MAE**	**MAPE**	**RMSE**	**MAE**	**MAPE**	**Statistic**	**P value**
1	463.48	331.44	25.05	524.19	415.35	23.74	6.00	0.92
2	453.52	314.16	23.45	458.34	353.15	18.93	11.54	0.48
3	481.20	341.22	26.83	587.90	474.78	24.38	12.52	0.41
4	505.52	351.16	27.38	751.53	624.70	32.02	9.84	0.63
5	510.13	356.81	28.49	861.84	726.78	37.73	12.68	0.39
6	511.64	359.27	28.57	754.63	611.60	30.34	10.21	0.60
7	526.08	367.91	27.51	814.63	637.90	31.83	15.28	0.23

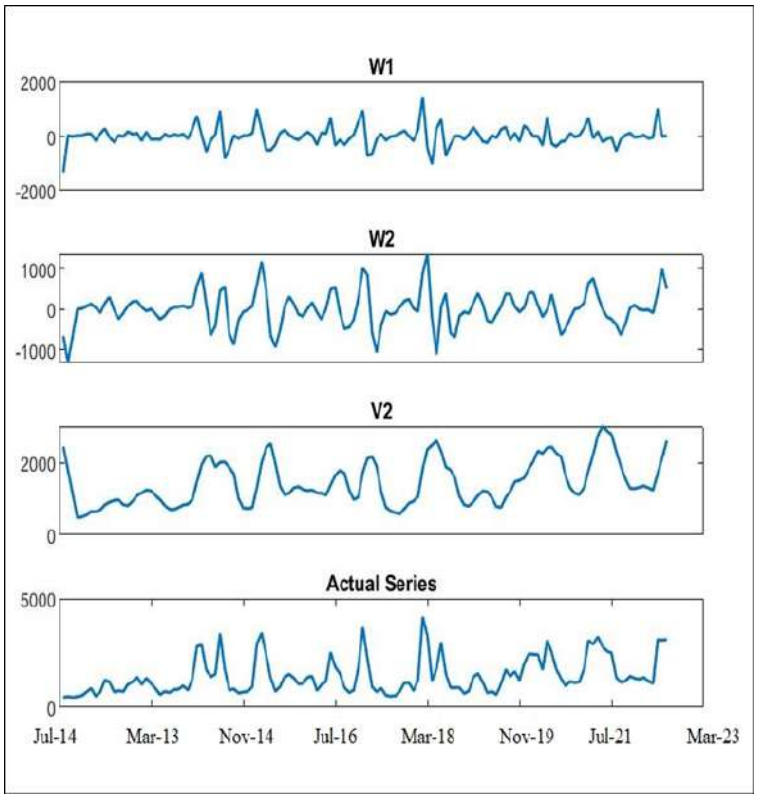


Figure 2: MODWT of tomato price series at level 2

Table 9: Forecasting ability of five different models in the validation set

Models	RMSE	MAE	MAPE
ARIMA	973.02	808.02	47.04
Wavelet-ARIMA	822.93	650.61	31.80
TDNN	855.03	783.06	46.63
Wavelet-TDNN	595.83	462.56	30.49
Wavelet-ARIMA-TDNN	458.34	353.15	18.93

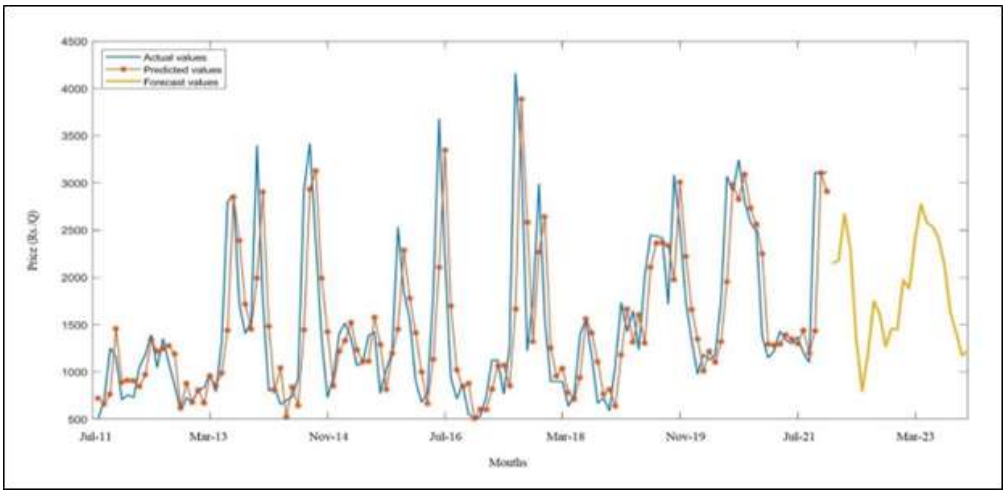


Figure 3: Actual and predicted tomato price series with its forecasts

5. Conclusion

This paper has developed a wavelet-ARIMA-TDNN model for forecasting the tomato price series in the Mumbai market. In the developed model, the McLeod and Li test was used to separate the frequencies into linear and non-linear frequencies, whereas ARIMA and TDNN were applied to model the linear and non-linear frequencies, respectively. Additionally, the developed model is confirmed to be superior to Wavelet-ARIMA, Wavelet-TDNN, ARIMA, and TDNN for modelling the series consisting of linear and non-linear frequencies. The choice of the best model was determined by forecast accuracy in the validation data set.

Finally, this study supports the following statements: (1) the linear time series model (non-linear time series model) is not appropriate for modelling the non-linear time series (linear time series); (2) wavelet decomposition (MODWT) improves the performance of the both time series models; and (3) whenever some time series contain both linear and non-linear frequencies, logical application of linear and non-linear models to the respective frequencies helps to enhance the wavelet-hybrid model fitting and forecasting.

Future research on other important non-linear models (LSTM, SVR, WNN, and so on) for modelling non-linear frequencies, as well as the use of different mother wavelets, is expected to improve our hybrid model.

Acknowledgements

The authors are grateful to AGMARKNET for providing data for analysis.

References

Adamowski, J. and Chan, H. F. (2011). A wavelet neural network conjunction model for groundwater level forecasting. *Journal of Hydrology*, **407**, 28–40.

Anjoy, P., Paul, R. K., Sinha, K., Paul, A., and Ray, M. (2017). A hybrid wavelet based neural networks model for predicting monthly WPI of pulses in india. *Indian Journal of Agricultural Science*, **87**, 834–839.

Antoniadis, A. (1997). Wavelets in statistics: a review. *Journal of the Italian Statistical Society*, **6**, 97–130.

Box George, E., Jenkins Gwilym, M., Reinsel Gregory, C., and Ljung Greta, M. (1976). *Time Series Analysis: Forecasting and Control*. San Francisco: Holden Bay.

Daubechies, I. et al. (1992). Ten lectures on wavelets (siam, philadelphia, 1992). *MR 93e*, **42045**.

Demuth, H. and Beale, M. (2002). Neural network toolbox for use with Matlab: User's guide, Natick, USA: Mathworks.

Hagen, M. and Menhaj, M. (1994). Training multilayer networks with the Marquardt algorithm. *IEEE Transactions on Neural Networks*, **5**, 989–993.

Jha, G. K. and Sinha, K. (2014). Time-delay neural networks for time series prediction: an application to the monthly wholesale price of oilseeds in india. *Neural Computing and Applications*, **24**, 563–571.

Ljung, G. M. and Box, G. E. (1978). On a measure of lack of fit in time series models. *Biometrika*, **65**, 297–303.

Markham, I. S. and Rakes, T. R. (1998). The effect of sample size and variability of data on the comparative performance of artificial neural networks and regression. *Computers & Operations Research*, **25**, 251–263.

McLeod, A. I. and Li, W. K. (1983). Diagnostic checking arma time series models using squared-residual autocorrelations. *Journal of Time Series Analysis*, **4**, 269–273.

Nury, A. H., Hasan, K., and Alam, M. J. B. (2017). Comparative study of wavelet-arima and wavelet-ann models for temperature time series data in northeastern bangladesh. *Journal of King Saud University-Science*, **29**, 47–61.

Ogden, R. T. (1997). *Essential Wavelets for Statistical Applications and Data Analysis*. Springer.

Paul, R. K., Paul, A., and Bhar, L. (2020). Wavelet-based combination approach for modeling sub-divisional rainfall in india. *Theoretical and Applied Climatology*, **139**, 949–963.

Percival, D. B. and Walden, A. T. (2000). *Wavelet Methods for Time Series Analysis*, volume 4. Cambridge University Press.

Ray, M., Singh, K. N., Ramasubramanian, V., Paul, R. K., Mukherjee, A., and Rathod, S. (2020). Integration of wavelet transform with ann and wnn for time series forecasting: an application to indian monsoon rainfall. *National Academy Science Letters*, **43**, 509–513.

Singla, S., Paul, R. K., and Shekhar, S. (2021). Modelling price volatility in onion using wavelet based hybrid models. *Indian Journal of Economics and Development*, **17**, 256–265.

Valiant, L. G. (1984). A theory of the learnable. *Communications of the ACM*, **27**, 1134–1142.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 22, No. 2, 2024 (New Series), pp 231–258
http://www.ssca.org.in/journal

Bayesian Analysis of Exponentiated Exponential Power Distribution under Hamiltonian Monte Carlo Method

Laxmi Prasad Sapkota and Vijay Kumar
Department of Mathematics and Statistics
Deen Dayal Upadhyaya Gorakhpur University, Gorakhpur (UP)-273009, India

Received: 10 November 2022; Revised: 01 November 2023; Accepted: 09 January 2024

Abstract

In this article, we present a new univariate probability distribution containing three parameters named exponentiated exponential power distribution. The density function and failure rate function of this new distribution accommodate broad varieties of shapes. Some mathematical and statistical properties of the proposed model are provided. Also, we have performed a full Bayesian analysis of the proposed model. Using Stan software whose Markov chain Monte Carlo (MCMC) techniques are based on a No-U-Turn sampler (NUTS) which is an adaptive variant of Hamiltonian Monte Carlo (HMC); a more robust and efficient sampler. We have presented the numerical as well as graphical analysis of the EEP model and found that all chains are well mixed and conversed. Further, we have estimated the parameters of the model and performed posterior predictive checks, and found that the underlying model can be used to generate reliable samples. The developed techniques are applied to a real data set, thus we can apply for full Bayesian analysis for the proposed model using these Bayesian techniques. Hence it is expected that the EEP model will be a choice in the fields of the theory of probability, applied statistics, bayesian inferences, and survival analysis.

Key words: Exponential power distribution; Posterior distribution; Bayesian analysis; HMC.

AMS Subject Classifications: 62F10, 62F15, 62E10

1. Introduction

Lifetime distributions are typically adapted to study the length of the lifetime of parts of a system, or a device, and usually, we conduct the survival and reliability analysis. Generally, lifetime models are extensively employed in fields like bioscience, medicine, demography, engineering, biology, insurance, etc. Several continuous probability distributions like exponential, Weibull, Cauchy, gamma, etc. are generally found in the literature of probability and applied statistics to study real-life data. Since the last decade, most scientists have been paying attention to the family of exponential models for their capability to model real-life data, and it has been observed that this model has performed well in several applications because of the bearing of closed-type solutions to several survival analyses. It will simply be

Corresponding Author: Laxmi Prasad Sapkota
Email: laxmisapkota75@gmail.com

even underneath the assumption of a constant failure rate however in practice, the failure rates don't seem to be continuously constant. Hence, the chaotic use of the exponential model appears to be insufficient and inappropriate. In this paper, we have presented a new model by extending the exponential power (EP) distribution defined by (Smith and Bain, 1975). The shape of the hazard function of this distribution depends on the value of the shape parameter α. For $\alpha \geq 1$, the hazard function is increasing and for $\alpha < 1$, it has a U-shape and exponentially increasing (towards the right) hazard function (Chen, 1999; Barriga *et al.*, 2011). The distribution and density function of EP distribution having parameters α and λ are as follows

$$F_{EP}(x) = 1 - \exp\left\{1 - e^{(\lambda\, x)^{\alpha}}\right\};\ (\alpha,\ \lambda) > 0, \quad x \geq 0. \tag{1}$$

$$f_{EP}(x) = \alpha\, \lambda^{\alpha}\, x^{\alpha-1}\, e^{(\lambda\, x)^{\alpha}} \exp\left\{1 - e^{(\lambda\, x)^{\alpha}}\right\} \quad ;\ (\alpha,\ \lambda) > 0, \quad x \geq 0. \tag{2}$$

Using the EP distribution (Barriga *et al.*, 2011) has defined a flexible lifetime model named complementary exponential power (CEP) distribution. The CDF of CEP is

$$F\left(t; \alpha, \beta, \theta\right) = \left[1 - \exp\left(1 - \exp\left\{\left(\frac{t}{\alpha}\right)^{\beta}\right\}\right)\right]^{\theta};\ t > 0.$$

To define the proposed new lifetime distribution we have used the technique presented by (Gupta and Kundu, 1999). They defined the generalized exponential (GE) distribution by inserting a shape parameter to the exponential distribution and it is superior to an exponential distribution, having decreasing and increasing failure rate hazard function. The cumulative density function (CDF) and its probability density function (PDF) of GE distribution are

$$F_{GE}\left(x;\ \alpha,\ \lambda\right) = \left\{1 - e^{-\lambda\, x}\right\}^{\alpha}; (\alpha,\ \lambda) > 0,\ x > 0.$$

$$f_{GE}\left(x;\ \alpha,\ \lambda\right) = \alpha\ \lambda\, e^{-\lambda\, x}\left\{1 - e^{-\lambda\, x}\right\}^{\alpha-1} (\alpha,\ \lambda) > 0,\ x > 0.$$

Using the same technique (Mudholkar and Srivastava, 1993) developed a three-parameter exponentiated Weibull (EW) distribution by inserting one additional shape parameter to the Weibull distribution. The CDF of EW is

$$F\left(x\right) = \left\{1 - \exp\left(-\alpha x^{\beta}\right)\right\}^{\lambda}; x > 0.$$

Another extension of exponential distribution has been developed by (Nadarajah and Haghighi, 2011) which can be taken alternative to exponentiated exponential, Weibull, and gamma distributions. Similarly exponentiated Chen (EC) distribution was defined by (Chaubey and Zhang, 2015) with either unimodel or decreasing density shape and decreasing or bathtub hazard shape. Dey *et al.* (2017) have redefined the exponentiated Chen distribution and extensively investigated the properties, and estimated the parameters using different methods. The CDF of EC is

$$F_{EC}\left(x; \alpha, \delta, \lambda\right) = \left\{1 - \exp\left[\lambda\left(1 - e^{x^{\delta}}\right)\right]\right\}^{\alpha}; (\alpha, \delta, \lambda) > 0,\ x > 0.$$

The exponentiated exponential Poisson (EEP) was introduced by (Ristić and Nadarajah, 2014) with a flexible hazard function. Using the same technique (Ashour and Eltehiwy, 2015) has defined the exponentiated power Lindley having CDF as

$$F\left(t; \alpha, \beta, \theta\right) = \left[1 - \left(1 + \frac{\theta t^{\beta}}{\theta + 1}\right) e^{-\theta t^{\beta}}\right]^{\alpha};\ t > 0.$$

Another extension of exponential distribution has been defined by (Almarashi *et al.*, 2019) whose hazard function can have a variety of shapes. Similarly, EP distribution is also used by (Joshi *et al.*, 2020) and generated a flexible model named logistic-exponential power distribution that can have decreasing or increasing or bathtub-shaped hazard function. Sapkota (2020) has defined exponentiated exponential logistic distribution and introduced a flexible hazard function. Hence we are motivated to generalize the EP distribution to get a versatile model by inserting only one shape parameter.

Further in this study, we have analyzed the suggested new model under the Bayesian approach. It is a fundamental framework for reasoning about uncertainty in statistical modeling and decision-making. It is a flexible and coherent approach that can handle various statistical problems, ranging from simple parameter estimation to complex hierarchical modeling and machine learning tasks. It provides a principled way to incorporate prior knowledge, update beliefs based on data, and quantify uncertainty in the results (Lambert, 2018; McElreath, 2020). Under this approach, we have used the HMC algorithm which is a powerful MCMC algorithm used to sample from complex probability distributions, especially in Bayesian statistics and machine learning. Unlike traditional MCMC methods, which often suffer from slow exploration of high-dimensional spaces, HMC leverages the concept of Hamiltonian dynamics from physics to efficiently explore the target distribution. HMC treats the probability distribution as a potential energy surface, and the Markov chain as a particle moving through this surface (Neal, 2011; Sapkota, 2022). HMC is more efficient than traditional MCMC methods because it generates less correlated samples and requires fewer evaluations of the target distribution's gradient (Carpenter *et al.*, 2017).

The remaining sections of this article are structured as follows: In the second section, we introduce the new distribution and examine its statistical characteristics. Moving on to the third section, we present some statistical properties of the EEP model. Section 4 is dedicated to discussing the application of the suggested model under the classical approach. Under the Bayesian approach, we formulate the proposed model, and its posterior analysis is presented in sections 5, 6, and 7, respectively. In section 8, we showcase the compatibility of the model, while section 9 delves into concluding remarks.

2. Exponentiated exponential power (EEP) distribution

Let $X \sim EEP(\alpha, \lambda, \theta)$ then the CDF of EEP distribution can be obtained by using Equation (1) and written as

$$F(x) = \left[1 - \exp\left\{1 - \exp\left(\lambda x^{\alpha}\right)\right\}\right]^{\theta}; x > 0, (\alpha, \lambda, \theta) > 0. \tag{3}$$

The PDF of EEP is obtained using Equation (2) as

$$f(x) = \alpha\lambda\theta x^{\alpha-1} \exp\left\{1 + \lambda x^{\alpha} - e^{\lambda x^{\alpha}}\right\} \left[1 - \exp\left(1 - e^{\lambda x^{\alpha}}\right)\right]^{\theta-1}; x > 0. \tag{4}$$

Different shapes of PDF curves of $EEP(\alpha, \lambda, \theta)$ distribution are presented in Figure 1.

2.1. Some special cases

- When $\theta = 1$, obviously EEP distribution reduces to EP distribution (Smith and Bain, 1975).

- When $\lambda = 1$, the EEP distribution reduces to EC distribution (Chaubey and Zhang, 2015).
- When $\lambda = 1$ and $\theta = 1$, the EEP distribution reduces to Chen distribution (Chen, 2000).
- If $\lambda = \frac{1}{\alpha^{\beta}}$, then the EEP distribution reduces to CEP distribution (Barriga *et al.*, 2011).

2.2. Survival function of EEP distribution

The survival function for the time t is

$$S(t) = \overline{F}(t) = 1 - \left[1 - \exp\left\{1 - \exp\left(\lambda t^{\alpha}\right)\right\}\right]^{\theta}; t > 0. \tag{5}$$

2.3. The hazard function of EEP distribution

Suppose t be the time of an item or component or an event that will survive and we would like to compute the probability of failing at time $t + \Delta t$ then the hazard function can be defined as

$$h(t) = \alpha\lambda\theta t^{\alpha-1} \exp\left\{1 + \lambda t^{\alpha} - e^{\lambda t^{\alpha}}\right\} \frac{\left[1 - \exp\left(1 - e^{\lambda t^{\alpha}}\right)\right]^{\theta-1}}{1 - \left[1 - \exp\left\{1 - \exp\left(\lambda t^{\alpha}\right)\right\}\right]^{\theta}}; t > 0. \tag{6}$$

2.4. Reverse hazard function of EEP distribution

The reverse hazard function of EEP distribution is

$$\begin{aligned} P_{rev}(x) &= \frac{f(x; \alpha, \lambda, \theta)}{F(x; \alpha, \lambda, \theta)} \\ &= \alpha\lambda\theta x^{\alpha-1} \exp\left\{1 + \lambda x^{\alpha} - e^{\lambda x^{\alpha}}\right\} \left[1 - \exp\left(1 - e^{\lambda x^{\alpha}}\right)\right]^{\theta-1}, x > 0. \end{aligned}$$

2.5. Quantile function

Suppose X be a non-negative continuous random variable with a CDF $F_X(x)$ and $U \in (0, 1)$, then the u^{th} quantile of X is,

$$Q(u) = \left[\frac{1}{\lambda} \ln\left\{1 - \ln\left(1 - u^{1/\theta}\right)\right\}\right]^{1/\alpha}; 0 < u < 1. \tag{7}$$

We can also calculate the median through Equation (7) as

$$Median = \left[\frac{1}{\lambda} \ln\left\{1 - \ln\left(1 - 2^{-1/\theta}\right)\right\}\right]^{1/\alpha}$$

To generate the random numbers for EEP distribution we can use

$$x = \left[\frac{1}{\lambda} \ln\left\{1 - \ln\left(1 - v^{1/\theta}\right)\right\}\right]^{1/\alpha}; 0 < v < 1. \tag{8}$$

2.6. Skewness of EEP distribution

Using quartiles, Bowley's coefficient of skewness can be computed as,

$$S_k(B) = \frac{Q(1/4) - 2Q(1/2) + Q(3/4)}{Q(0.75) - Q(0.25)}.$$

2.7. Kurtosis of EEP distribution

The coefficient of kurtosis using octiles (Moors, 1988) is

$$K_u(M) = \frac{Q(0.875) + Q(0.375) - Q(0.625) - Q(0.125)}{Q(0.75) - Q(0.25)}.$$

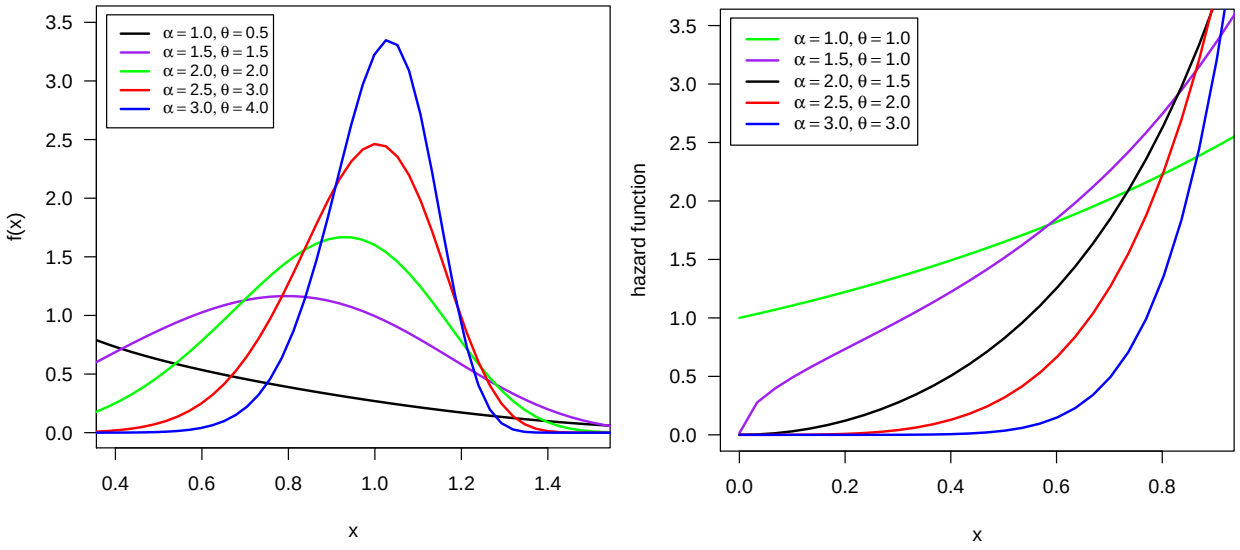


Figure 1: Graphs of PDF (left panel) and HRF (right panel) for different values of α and θ and fixed $\lambda = 1$

3. Some statistical properties of EEP distribution

3.1. Moments

The K^{th} moment about origin using the quantile function see for details (Balakrishnan and Cohen, 2014) and (Dey *et al.*, 2017) can be computed as

$$\begin{aligned}\mu_k^{raw} = E(X^r) &= \int_0^\infty x^k f(x)\,dx = \int_0^1 [Q(v)]^k dv \\ &= \int_0^1 \lambda^{-k/\alpha}\left[\log\left\{1 - \log\left(1 - v^{1/\theta}\right)\right\}\right]^{k/\alpha} dv \qquad (9) \\ &= \lambda^{-k/\alpha}\sum_{p=0}^{\infty}\sum_{q=0}^{\infty} W_p(k) W_q\left(\frac{k}{\alpha} + p\right)(-1)^{\frac{2k}{\alpha}+p} \int_0^1 v^{\frac{1}{\theta}\left(\frac{k}{\alpha}+p\right)+q} dv.\end{aligned}$$

$$\therefore \quad \mu_k^{raw} = \lambda^{-k/\alpha} \sum_{p=0}^{\infty} \sum_{q=0}^{\infty} W_p(k) W_q \left(\frac{k}{\alpha} + p\right) (-1)^{\frac{2k}{\alpha}+p} (C_{pq})^{-1}. \tag{10}$$

here $W_p(k)$ is the coefficient of $\log\left(1 - v^{1/\theta}\right)$ after the expansion of $\left[\log\left\{1 - \log\left(1 - v^{1/\theta}\right)\right\}\right]^{k/\alpha}$, $W_q\left(\frac{k}{\alpha} + p\right)$ is the coefficient of v^q after the expansion of $\left[\log\left(1 - v^{1/\theta}\right)\right]^{k/\alpha}$ and $C_{pq} = \frac{1}{\theta}\left(\frac{k}{\alpha} + p\right) + q + 1$.

Remark: $Mean(\mu_1)$ and $Variance(\mu_2)$ of EEP distribution can be computed as

$$\mu_1 = \lambda^{-1/\alpha} \sum_{p=0}^{\infty} \sum_{q=0}^{\infty} W_p(1) W_q \left(\frac{1}{\alpha} + p\right) (-1)^{\frac{2}{\alpha}+p} \left(\frac{1}{\theta}\left(\frac{1}{\alpha} + p\right) + q + 1\right)^{-1}.$$

$$\mu_2 = \lambda^{-2/\alpha} \sum_{p=0}^{\infty} \sum_{q=0}^{\infty} W_p(2) W_q \left(\frac{2}{\alpha} + p\right) (-1)^{\frac{4}{\alpha}+p} \left(\frac{1}{\theta}\left(\frac{2}{\alpha} + p\right) + q + 1\right)^{-1} - (\mu_1)^2.$$

3.2. Moment generating function (MGF)

The expression for MGF of EEP distribution can be obtained by using the Equation (10) as

$$M_X(t) = \sum_{s=0}^{\infty} \frac{t^s}{s!} \mu_k^{raw} = \lambda^{-k/\alpha} \sum_{p=0}^{\infty} \sum_{q=0}^{\infty} \sum_{s=0}^{\infty} W_p(k) W_q \left(\frac{k}{\alpha} + p\right) (-1)^{\frac{2k}{\alpha}+p} \frac{t^s}{s!} (C_{pq})^{-1}. \tag{11}$$

3.3. Conditional moments (CM)

Let Y be a random variable from the EEP distribution, and then the CM for Y can be expressed as

$$\begin{aligned} E\left(Y^k / Y > y\right) &= \frac{1}{S(y)} \int_x^{\infty} y^k f(y)\, dy \\ &= \frac{1}{S(y)} \int_v^1 Q^k(v) dv \\ &= \frac{1}{S(y)} \lambda^{-k/\alpha} \sum_{p=0}^{\infty} \sum_{q=0}^{\infty} \sum_{s=0}^{\infty} W_p(k) W_q \left(\frac{k}{\alpha} + p\right) (-1)^{\frac{2k}{\alpha}+p} \frac{1 - F(y)^{C_{pq}}}{C_{pq}}, \end{aligned} \tag{12}$$

here $S(y)$ and $F(y)$ are survival functions and CDF of EEC distribution.

3.4. Average residual life (ARL) function

ARL function of a component using quantile function can be calculated by

$$\mu_{ARL}(x) = \frac{1}{S(x)} \lambda^{-1/\alpha} \sum_{p=0}^{\infty} \sum_{q=0}^{\infty} W_p(1) W_q \left(\frac{1}{\alpha} + p\right) (-1)^{\frac{2}{\alpha}+p} \frac{1 - F^{\left(\frac{1}{\theta}\left(\frac{1}{\alpha}+p\right)+q+1\right)}(x)}{\left(\frac{1}{\theta}\left(\frac{1}{\alpha} + p\right) + q + 1\right)} - x.$$

3.5. Mean deviation (MD)

Let μ and $F(.)$ denote the mean and CDF of EEP distribution then MD can be expressed as

$$\mu_{MD\_Mean}(x) = 2\mu F(\mu) - 2\mu + 2\int_{\mu}^{1} Q^{k}(v)dv. \tag{13}$$

$$= 2\mu F(\mu) - 2\mu + 2\lambda^{-1/\alpha}\sum_{p=0}^{\infty}\sum_{q=0}^{\infty} W_p(1)W_q\left(\frac{1}{\alpha}+p\right) \times (-1)^{\frac{2}{\alpha}+p}\frac{1-F^{\left(\frac{1}{\theta}\left(\frac{1}{\alpha}+p\right)+q+1\right)}(\mu)}{\left(\frac{1}{\theta}\left(\frac{1}{\alpha}+p\right)+q+1\right)}.$$

4. Classical analysis of the proposed model

4.1. Parameter estimation

In this subsection, we have used the maximum likelihood estimation (MLE) method which is the most frequently used method for the point and interval estimation of the parameters of the model. Let $\underline{x} = (x_1, \ldots, x_n)$ be a non-negative observed sample of size 'n' following the $EEP(\alpha, \lambda, \theta)$ then we can define the likelihood function for the parameter vector $\chi = (\alpha, \lambda, \theta\ )^T$ as

$$L(\chi) = (\alpha\lambda\theta)^n \prod_{i=1}^{n} x_i^{\alpha-1} \exp\left\{1 + \lambda x_i^{\alpha} - e^{\lambda x_i^{\alpha}}\right\}\left[1 - \exp\left(1 - \mathrm{e}^{\lambda x_i^{\alpha}}\right)\right]^{\theta-1}. \tag{14}$$

Taking the logarithm to (14) we get the log-likelihood function as

$$\ell(\chi) = n\ln(\alpha\lambda\theta) + (\alpha-1)\sum_{i=1}^{n}\ln x_i + n + \lambda\sum_{i=1}^{n} x_i^{\alpha} - \sum_{i=1}^{n} e^{\lambda x_i^{\alpha}} + (\theta-1)\sum_{i=1}^{n}\ln[1-K(x_i)]. \tag{15}$$

Differentiating (15) with respect to parameters $\alpha,\ \lambda\ and\ \theta$, we get

$$\frac{\partial\ell}{\partial\alpha} = \frac{n}{\alpha} + \sum_{i=1}^{n}\ln x_i + \lambda\sum_{i=1}^{n}\left[x_i^{\alpha}\ln x_i\left\{1 - \ln x_i\left\{1 - (\theta-1)K(x_i)\left\{1-K(x_i)\right\}^{-1}\right\}\right\}\right],$$

$$\frac{\partial\ell}{\partial\lambda} = \frac{n}{\lambda} + \sum_{i=1}^{n} x_i^{\alpha} - \sum_{i=1}^{n} x_i^{\alpha} e^{\lambda x_i^{\alpha}}\left\{1 - (\theta-1)\left\{1-K(x_i)\right\}^{-1}K(x_i)\right\},$$

$$\frac{\partial\ell}{\partial\theta} = \frac{n}{\theta} + \sum_{i=1}^{n}\ln[1-K(x_i)],$$

where $K(x_i) = \exp\left(1 - e^{\lambda x_i^{\alpha}}\right)$. Manually it is quite difficult to solve these equations for the parameters $\alpha,\ \lambda\ and\ \theta$. Using the appropriate software like R, Python, Matlab, etc. we can solve them manually. Let $\chi = (\alpha, \lambda, \theta\ )^T$ be the parameter vector and MLEs of χ is $\hat{\chi} = (\hat{\alpha}, \hat{\lambda}, \hat{\theta})$, then $(\hat{\chi} - \chi) \to N_3\left[0, (M(\chi))^{-1}\right]$ distributed as the normal distribution, here

$M(\chi)$ is known as Fisher's information matrix computed as,

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

where,

$$M_{11} = \frac{\partial^2 l}{\partial \alpha^2},\ \mathrm{M}_{12} = \frac{\partial^2 l}{\partial \alpha \partial \lambda},\ \mathrm{M}_{13} = \frac{\partial^2 l}{\partial \alpha \partial \theta},$$
$$M_{21} = \frac{\partial^2 l}{\partial \lambda \partial \alpha},\ \mathrm{M}_{22} = \frac{\partial^2 l}{\partial \lambda^2},\ \mathrm{M}_{23} = \frac{\partial^2 l}{\partial \theta \partial \lambda},$$
$$M_{31} = \frac{\partial^2 l}{\partial \lambda \partial \alpha},\ \mathrm{M}_{32} = \frac{\partial^2 l}{\partial \theta \partial \lambda},\ \mathrm{M}_{33} = \frac{\partial^2 l}{\partial \theta^2},$$

which can be calculated as,

$$\frac{\partial^2 \ell}{\partial \alpha^2} = -\frac{n}{\alpha^2} + \lambda \sum_{i=1}^{n} \left\{ x_i^{\alpha} (\ln x_i)^2 \right\} - (\theta - 1) \sum_{i=1}^{n} \left[1 + \lambda \ln x_i + \alpha x_i \right] \left(\lambda x_i^{\alpha} e^{\lambda x_i^{\alpha}} \ln x_i \right)^2$$
$$K(x_i) \{1 - K(x_i)\}^{-1} \left[1 + K(x_i) \{1 - K(x_i)\}^{-1} \right].$$

$$\frac{\partial^2 \ell}{\partial \lambda^2} = -\frac{n}{\lambda^2} - \sum_{i=1}^{n} x_i^{2\alpha} e^{\lambda x_i^{\alpha}} (\theta - 1) K(x_i) \left\{ \{1 - K(x_i)\}^{-1} K(x_i) - \{1 - K(x_i)\}^{-2} K(x_i) e^{\lambda x_i^{\alpha}} \right\}.$$

$$\frac{\partial^2 \ell}{\partial \theta^2} = -\frac{n}{\theta^2}.$$

$$\frac{\partial^2 \ell}{\partial \theta \partial \alpha} = \lambda \sum_{i=1}^{n} \left[x_i^{\alpha} e^{\lambda x_i^{\alpha}} \ln(x_i) K(x_i) \{1 - K(x_i)\}^{-1} \right].$$

$$\frac{\partial^2 \ell}{\partial \theta \partial \lambda} = \sum_{i=1}^{n} \left[x_i^{\alpha} e^{\lambda x_i^{\alpha}} \{1 - K(x_i)\}^{-1} K(x_i) \right].$$

$$\frac{\partial^2 \ell}{\partial \alpha \partial \lambda} = -\sum_{i=1}^{n} x_i^{\alpha} \ln x_i e^{\lambda x_i^{\alpha}} \left\{ (1 + \lambda x_i^{\alpha}) \left\{ 1 - (\theta - 1) \{1 - K(x_i)\}^{-1} K(x_i) \right\} + Z_i \right\},$$

where

$$Z_i = (\theta - 1) \lambda x_i^{\alpha} e^{\lambda x_i^{\alpha}} K(x_i) \{1 - K(x_i)\}^{-1} \{1 + K(x_i)\}.$$

Now the observed information matrix can be calculated through algorithms like Newton-Raphson and can be computed as,

$$[M(\chi)]^{-1} = \begin{pmatrix} V(\hat{\alpha}) & \mathrm{cov}(\hat{\alpha}, \hat{\lambda}) & \mathrm{cov}(\hat{\alpha}, \hat{\theta}) \\ \mathrm{cov}(\hat{\lambda}, \hat{\alpha}) & V(\hat{\lambda}) & \mathrm{cov}(\hat{\lambda}, \hat{\theta}) \\ \mathrm{cov}(\hat{\theta}, \hat{\alpha}) & \mathrm{cov}(\hat{\theta}, \hat{\lambda}) & V(\hat{\theta}) \end{pmatrix}.$$

Hence, estimated $100(1-\delta)\%$ CI for $\alpha, \lambda \, and \, \theta$ can be created as, $\hat{\alpha} \pm z_{\delta/2} SE(\hat{\alpha})$, $\hat{\lambda} \pm z_{\delta/2} SE(\hat{\lambda})$, and $\hat{\theta} \pm z_{\delta/2} SE(\hat{\theta})$.

4.2. Illustration with real dataset

In this section, we examine a real dataset previously utilized by various researchers to showcase the capabilities and applicability of the EEP distribution. Additionally, we present the EEP distribution alongside several competing distributions, listed below, to offer a comprehensive comparison.

- Exponential power (EP) distribution by (Smith and Bain, 1975).
- Power Lindley distribution (PL) by (Ghitany *et al.*, 2015).
- Generalized Rayleigh (GR) distribution by (Kundu and Raqab, 2005).
- Marshall-Olkin Extended Exponential (MOEE) distribution by (Marshall and Olkin, 1997).

To compare the proposed distribution with the distributions as mentioned above we have computed the Bayesian information criterion (BIC), Akaike information criterion (AIC), negative log-likelihood (-LL), Hannan-Quinn information criterion (HQIC), and Corrected Akaike Information criterion (CAIC) statistic. These statistics are obtained by using the following expressions

$$\begin{aligned} AIC &= -2\ell(\hat{\chi}) + 2d. \\ BIC &= -2\ell(\hat{\chi}) + d\log(n). \\ CAIC &= \frac{2d(d+1)}{n-d-1} + AIC. \\ HQIC &= -2\ell(\hat{\chi}) + 2d\log[\log(n)]. \end{aligned}$$

Here $\hat{\chi}$ denotes estimated parameter space, n is the size of the sample and d is the number of parameters of the model under study. In addition, to judge the goodness-of-fit of EEP distribution Cramer-Von Mises (A^2), Kolmogorov-Smirnov (KS), and Anderson-Darling (W) statistics are presented and calculated as

$$\begin{aligned} KS &= \max_{1\le j\le n}\left(d_j - \frac{j-1}{n}, \frac{j}{n} - d_j\right). \\ W &= -n - \frac{1}{n}\sum_{j=1}^{n}(2j-1)\left[\ln d_j + \ln\left(1 - d_{n+1-j}\right)\right]. \\ A^2 &= \frac{1}{12n} + \sum_{j=1}^{n}\left[\frac{2j-1}{2n} - d_j\right]^2. \end{aligned}$$

where the d_j's are the ordered observations, and $d_j = CDF\left(x_j\right)$.

4.2.1. Dataset

The dataset represents the waiting time (in minutes) of 100 clients (Ghitany *et al.*, 2008) before the client received service in a bank. The data set is,
"0.8, 0.8, 1.3, 1.5, 1.8, 1.9, 1.9, 2.1, 2.6, 2.7, 2.9, 3.1, 3.2, 3.3, 3.5, 3.6, 4.0, 4.1, 4.2, 4.2, 4.3,

4.3, 4.4, 4.4, 4.6, 4.7, 4.7, 4.8, 4.9, 4.9, 5.0, 5.3, 5.5, 5.7, 5.7, 6.1, 6.2, 6.2, 6.2, 6.3, 6.7, 6.9, 7.1, 7.1, 7.1, 7.1, 7.4, 7.6, 7.7, 8.0, 8.2, 8.6, 8.6, 8.6, 8.8, 8.8, 8.9, 8.9, 9.5, 9.6, 9.7, 9.8, 10.7, 10.9, 11.0, 11.0, 11.1, 11.2, 11.2, 11.5, 11.9, 12.4, 12.5, 12.9, 13.0, 13.1, 13.3, 13.6, 13.7, 13.9, 14.1, 15.4, 15.4, 17.3, 17.3, 18.1, 18.2, 18.4, 18.9, 19.0, 19.9, 20.6, 21.3, 21.4, 21.9, 23.0, 27.0, 31.6, 33.1, 38.5"

4.2.2. Exploratory study of the dataset

The main aim of the exploratory data analysis is to explore more information about the data. The latest statistical tools for data analysis incorporate exploratory data analysis. The descriptive statistics of the dataset are presented in Table 1. The basic exploratory

Table 1: Summary statistics of the dataset

Minimum	Q1	Median	Mean	Q3	Maximum	Skewness	Kurtosis
0.800	4.675	8.100	9.877	13.025	38.500	1.451	2.430

data analysis technique is applied to study the data and results are displayed respectively in Table 1. Efficient modeling requires an excellent understanding of the properties of different types of models. The parameters of the proposed model are estimated using the maximum likelihood (ML) estimation method. To evaluate the validity of the model, we calculate the Kolmogorov-Smirnov (KS) distance between the fitted distribution function and empirical distribution function where the parameters are estimated by the ML estimation method. The probability–probability (PP) plot and quantile-quantile (QQ) plot are used to check the suitability of the proposed model.

4.2.3. Computation of MLE

The MLEs of EEP distribution are calculated with the help of R programming software using `maxLik()` an R package developed by (Henningsen and Toomet, 2011) and they are uniquely determined (see Figure 2). In Table 2, the MLEs with 95% confidence interval (CI) and standard errors (SE) are presented.

Table 2: MLE and SE for α, λ, and θ of EEP distribution

Parameter	MLE	SE	95% CI
α	0.3407	0.0590	(0.2252, 0.4562)
λ	0.6068	0.1259	(0.3600, 0.8535)
θ	7.6150	2.6998	(2.3234, 12.9065)

4.2.4. Model validation

To check the validity of the proposed model we performed the Kolmogorov-Smirnov (KS) test and we found that $KS = 0.0358$ and p-value $= 0.9995$ which indicates that the proposed model can fit the data well. Further, we have presented the K-S plot (right panel) and quantile-quantile (Q-Q) plot (left panel) to evaluate the validity of the model in Figure 3 and it also verifies the validity of the model.

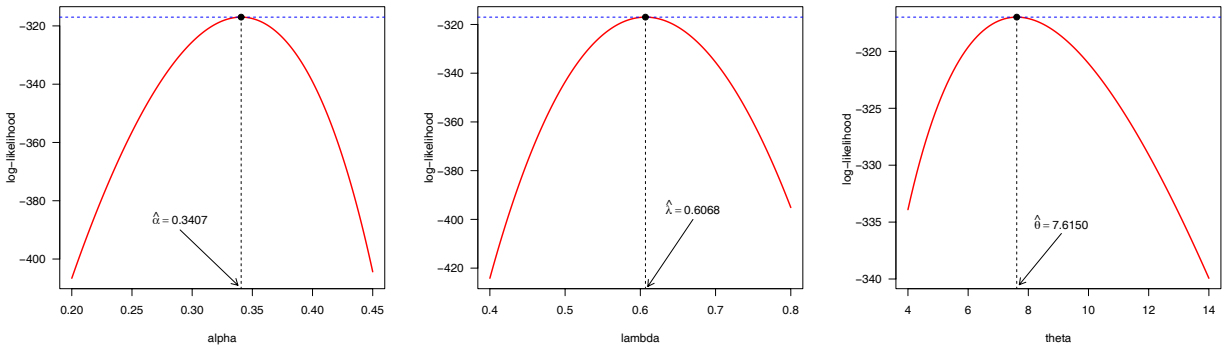


Figure 2: The graph of profile log-likelihood for α, λ, and θ

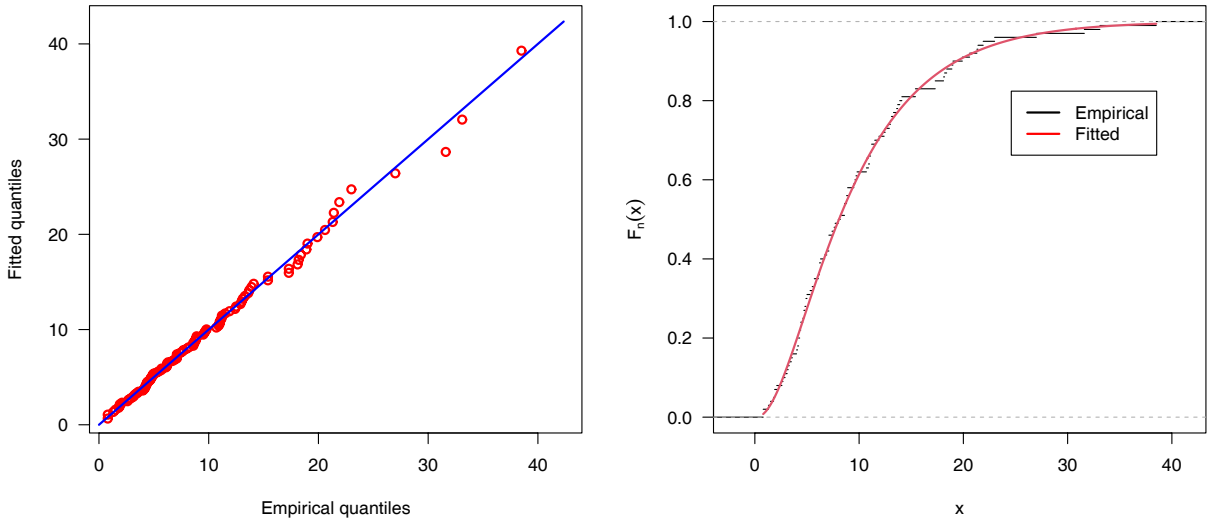


Figure 3: The graph of Q-Q (left side) and K-S (right side) of the EEP distribution

4.2.5. Model selection

To select a good model we have computed the AIC, CAIC, BIC, and HQIC for the proposed model as well as the other four models taken for comparison, and it is observed that the EEP distribution has the highest value of LL and lowest values of AIC, CAIC, BIC, and HQIC, hence we confirmed that the proposed distribution is better than the competing distribution (Table 3) for more detail see (Lambert, 2018). To evaluate the fit attained by EEP distribution among challenging distributions, the Anderson-Darling (W), Kolmogorov-Smirnov (KS), and the Cramer-Von Mises (A^2) tests are conducted and results are reported in Table 4. We have seen that the EEP distribution gets the smallest test statistic with a higher p-value which indicates the EEP distribution gets consistently better fit than those models taken under consideration. We have also presented the graphs to evaluate the goodness-of-fit of EEP distribution with distributions that are taken for comparison in Figure 4 (left panel) and the Kaplan-Meier (KM) estimate which is used to estimate the

reliability function of EEP distribution in Figure 4 (right panel) and exhibits good fit.

Table 3: AIC, CAIC, BIC, and HQIC, Log-likelihood (LL) statistics

Model	AIC	CAIC	BIC	HQIC	LL
EEP	639.9793	640.2293	647.7948	643.1420	-316.9897
PL	640.6372	640.7609	645.8475	642.7460	-318.3186
MOEE	645.4241	645.5453	650.6344	647.5330	-320.7120
GR	647.0364	647.1601	652.2467	649.1450	-321.5182
EP	654.0395	654.1607	659.2499	656.1480	-325.0198

Table 4: Value of W, KS and A^2 statistics with p-value

Model	*W(p-value)*	*KS(p-value)*	A^2*(p-value)*
EEP	0.0173(0.9990)	0.0358(0.9995)	0.1274(0.9997)
PL	0.0458(0.9025)	0.0520(0.9498)	0.3028(0.9359)
MOEE	0.0760(0.7164)	0.0596(0.8690)	0.6351(0.6150)
GR	0.2043(0.2595)	0.0945(0.3337)	1.0911(0.3126)
EP	0.2549(0.1822)	0.0930(0.3532)	1.6490(0.1447)

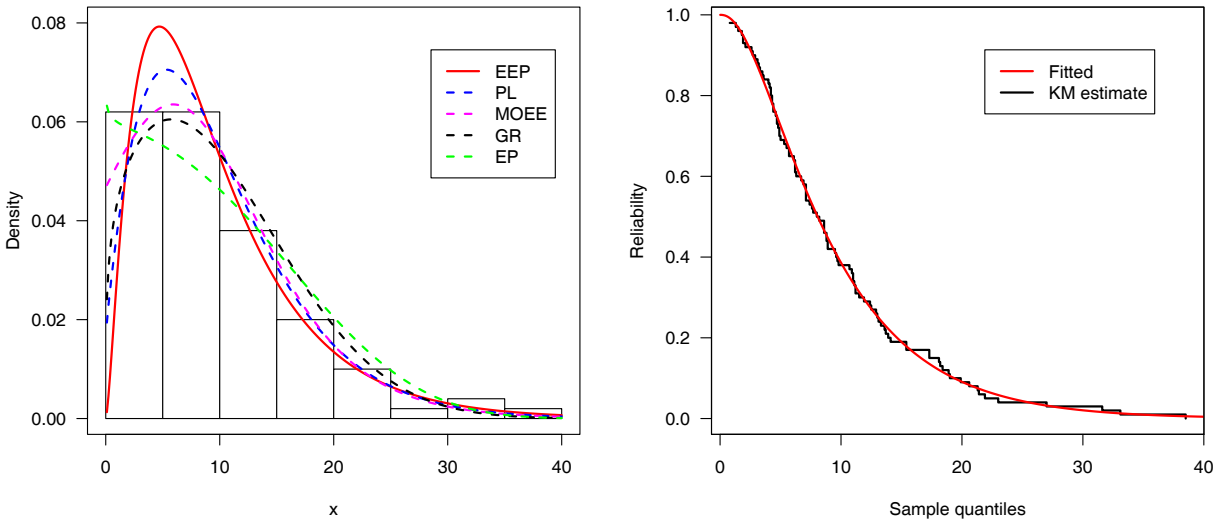


Figure 4: PDF plot with a histogram of fitted distributions (left side) and KM estimate with fitted quantiles (right side) of EEP distribution

5. Model formulation under the Bayesian approach

We usually assume the parameters $\Theta = (\alpha, \lambda, \theta)$ (for our study) as a constant in the classical approach and the goal is to investigate the distribution of the observed data set given Θ using the likelihood of the data sample. But the parameter Θ is considered as a random variable whereas the observed data set is taken as constant in the Bayesian approach (Lambert, 2018). In this type of modeling, prior information is used to support

our assumption about the parameters of the distribution (Gelman *et al.*, 2013). In Bayesian modeling, the posterior distribution function is obtained by multiplying the prior distribution function and the likelihood function of the model under consideration for more detail see (McElreath, 2020). For Bayesian inference, we need the following elements

- The probability distribution function:$f\left(x/\Theta\right)$
- Prior distribution: $p\left(\Theta\right)$
- Likelihood: $p\left(Data/\Theta\right)$
- Data:$(x_1, ..., x_n)$

5.1. Prior distribution $p\left(\Theta\right)$

In Bayesian inference, a prior distribution (simply called prior) is the unconditional probability distribution that is used to express our beliefs about the true value of the parameters before the data is taken into account. The term $p\left(\Theta\right)$ denotes the probability distribution which represents our pre-data beliefs depending upon the different values of the parameters $\Theta = (\alpha, \lambda, \theta)$ of our model. In this study, we have taken the weakly informative Gamma prior for the parameters $\Theta = (\alpha, \lambda, \theta)$ as $\alpha \sim G\left(a_1, b_1\right)$, $\lambda \sim G\left(a_2, b_2\right)$ and $\theta \sim G\left(a_3, b_3\right)$. Particularly we have chosen $(a_1 = 0.001, b_1 = 0.001)$, $(a_2 = 0.001, b_2 = 0.001)$, and $(a_3 = 0, b_3 = 0.001)$ respectively for Gamma prior and most commonly used as weak prior on variance which is nearly flat as in Figure (5). The prior distributions can be written as

$$p\left(\alpha\right) = \frac{b_1^{a_1}}{\Gamma(a_1)}\alpha^{a_1-1}\exp\left(-b_1\alpha\right);\ \alpha > 0,\ \ (a_1, b_1) > 0.$$

$$p\left(\lambda\right) = \frac{b_2^{a_2}}{\Gamma\left(a_2\right)}\lambda^{a_2-1}\exp\left(-b_2\lambda\right);\ \lambda > 0,\ \ (a_2, b_2) > 0.$$

$$p\left(\theta\right) = \frac{b_3^{a_3}}{\Gamma(a_3)}\theta^{a_3-1}\exp\left(-b_3\theta\right);\ \theta > 0,\ \ (a_3, b_3) > 0.$$

5.2. Likelihood $p\left(Data/\Theta\right)$

Given a set of data $(x_1, ..., x_n)$, the likelihood function of EEP distribution can be computed as

$$L\left(x\right) = (\alpha\lambda\theta)^n \prod_{i=1}^{n} x_i^{\alpha-1}\exp\left(1 + \lambda x_i^{\alpha} - e^{\lambda x_i^{\alpha}}\right)\left[1 - \exp\left(1 - e^{\lambda x_i^{\alpha}}\right)\right]^{\theta-1}. \tag{16}$$

5.3. Posterior distribution $p\left(\Theta/Data\right)$

Let $p\left(\alpha, \lambda, \theta/\underline{x}\right)$ denote the posterior distribution and it can be obtained by using Bayes' rule as

$$p\left(\alpha, \lambda, \theta/\underline{x}\right) \propto L\left(\alpha, \lambda, \theta/\underline{x}\right) \times p\left(\alpha, \lambda, \theta\right).$$

In the Bayesian inference technique, we use Bayes' rule to estimate probability distribution called posterior distribution which can be obtained as

$$p\left(\Theta/data\right) \propto p\left(data/\Theta\right) p\left(\Theta\right).$$

$$p\left(\alpha,\lambda,\theta/\underline{x}\right) \propto \alpha^{n+a_1-1}\theta^{n+a_3-1}\lambda^{n+a_2-1}\prod_{i=1}^{n} e^{-b_1\alpha-b_2\lambda-b_3\theta}x_i^{\alpha-1}\times \exp\left(1+\lambda x_i^{\alpha}-e^{\lambda x_i^{\alpha}}\right)\left[1-\exp\left(1-e^{\lambda x_i^{\alpha}}\right)\right]^{\theta-1}. \tag{17}$$

All the information needed for Bayesian analyses is contained in the posterior distribution

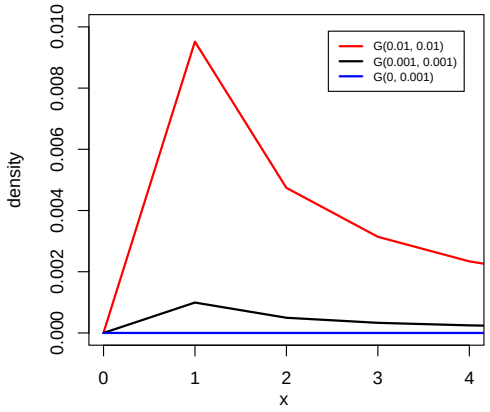


Figure 5: Graph of Gamma prior for various values of the parameters

and the aim is to compute the numeric as well as graphic summaries of it through integration. But the posterior distribution is quite complicated and could not draw any inferences. Hence we propose an alternative technique known as the simulation technique. This technique is based on the Markov Chain Monte Carlo (MCMC) method. MCMC draws samples by running a cleverly constructed Markov Chain that eventually converges to the target distribution i.e. posterior distribution $p\left(\alpha,\lambda,\theta/\underline{x}\right)$ (Brooks, 1998).

There are many different techniques to construct such chains some of them are, Gibbs sampler (Geman and Geman, 1984; Gelfand and Smith, 1990) are special cases of the general framework of (Metropolis *et al.*, 1953) and (Hastings, 1970). In this article, we implement MCMC algorithms through Stan (a probabilistic programming language) (Stan Development Team, 2022), the HMC algorithm, and its adaptive variant the NUTS for more detail see (Hoffman *et al.*, 2014; Carpenter *et al.*, 2017). Also Chaudhary and Kumar (2020) presented the Bayesian estimate of Gompertz extension distribution having three parameters. Also, Alizadeh *et al.* (2020) discussed the technique for estimating the model parameters of the odd log-logistic Lindley-G family of distribution.

6. MCMC method

6.1. HMC method

HMC is computationally a bit costly as compared to Metropolis and Gibbs sampling but its proposals are much more efficient (Gelman *et al.*, 2015). As a result, HMC doesn't require as many samples to explore the posterior distribution. For more detail about the HMC algorithm see (Beskos *et al.*, 2013).

6.1.1. No-U-Turn Sampler (NUTS)

NUTS engine routinely selects a suitable value for leapfrog step L in every iteration to maximize the distance at every L and control the random walk behavior. Let ω_1 and ω_0 be the current position and initial position of a particle and D be half of the distance between the positions ω_1 and ω_0 at each leapfrog step. The aim is to run leapfrog steps until ω_1 starts to move backward towards ω_0, which is achieved using the following algorithm, where leapfrog steps are run until the derivative of D with respect to time becomes less than 0.

$$\frac{\partial D}{\partial t} = \frac{\partial}{\partial t}\left[\frac{1}{2}\left(\omega_1 - \omega_0\right)^T\left(\omega_1 - \omega_0\right)\right] = \left(\omega_1 - \omega_0\right)^T p < 0.$$

However, this algorithm doesn't assure convergence or reversibility to the target distribution. The NUTS solves this type of problem by performing a doubling method for slice sampling (Neal, 2003). To generate the samples using NUTS, see (Hoffman *et al.*, 2014). For more details about NUTS, readers can go through (Nishio and Arakawa, 2019) and Devlin *et al.* (2021).

6.1.2. Defining the model in STAN

For the Bayesian analysis of the EEP model, we have used the latest Bayesian analysis software called Stan a high-level programming language that uses NUTS which is a variant of HMC simulation (Hoffman *et al.*, 2014). We have used the Rstan package (Stan Development Team, 2020) to run STAN in R software (R Core Team, 2022). The Stan scripts in R for the EEP model for the Bayesian analysis are presented in the appendix. We run the Stan using the algorithm HMC and engine NUTS having 4 chains for 2000 iterations. By default, Stan generates 1000 warm-up samples and 1000 real samples for a chain which are used for inferences.

6.2. Convergence and efficiency diagnostics for NUTS/ HMC and Markov chains

In the convergence diagnostic, we monitor the performance of NUTS/ HMC and MCMC sampling as

NUTS/ HMC: Here we study the information about divergence, energy, `tree-depth`, `step-size`, and `acceptance statistic`. Figure 6 (left panel) is the plot of the overlaid histograms of the marginal energy distribution π_E and the energy transition distribution $\pi_{\Delta E}$ for all 4 chains. The plot shows the histograms that look well-matched and indicate that the Hamiltonian Monte Carlo has performed robustly and Figure 6 (right panel) indicates that there are no divergent transitions. In Figure 7 we have displayed the performance of the NUTS sampling algorithm and Figure 8 are plots of the histogram of Rhat statistic, the ratio of effective sample size and sample size, and the ratio of Monte Carlo Standard Error (MCSE) and posterior SD. These plots show the good efficiency of the sampling algorithm NUTS for detail see Betancourt (2017).

MCMC: The MCMC draws can be monitored by plotting the following graphs autocorrelation plots, rank plots, trace plots, ergodic mean plots, and pairs plots.

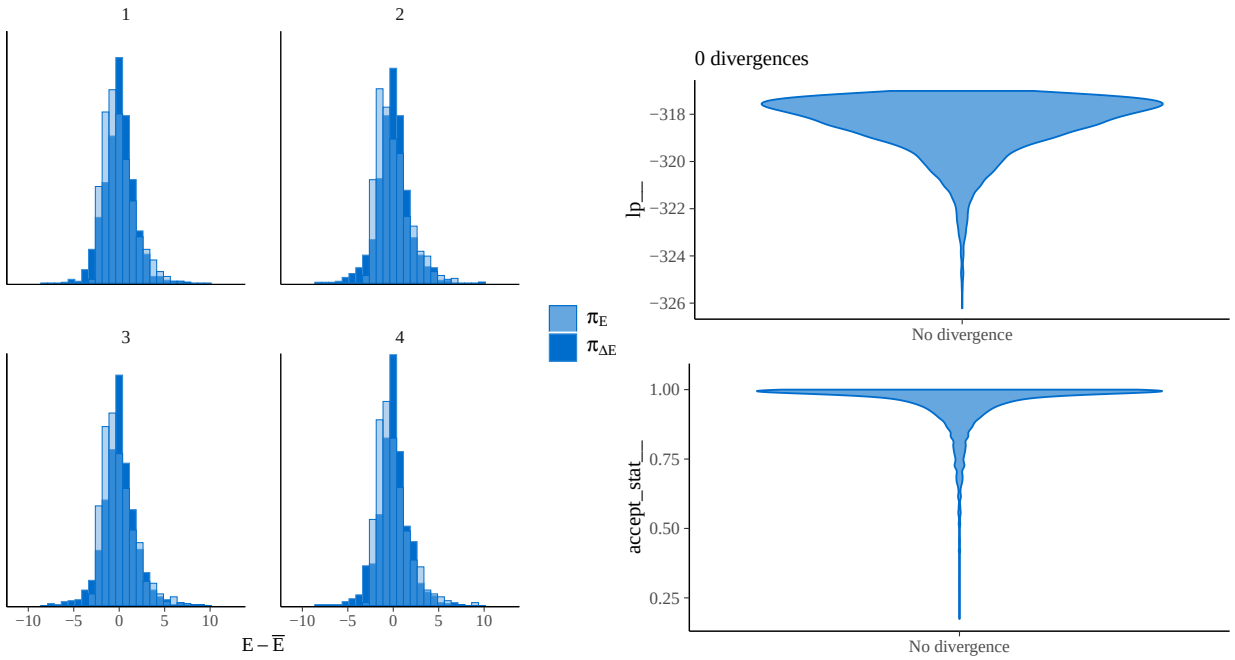


Figure 6: Histograms of π_E and $\pi_{\Delta E}$ for all 4 chains (left panel) and the divergent transition status (x-axis) against the log-posterior and the acceptance statistic (right panel) of the sampling algorithm for all chains

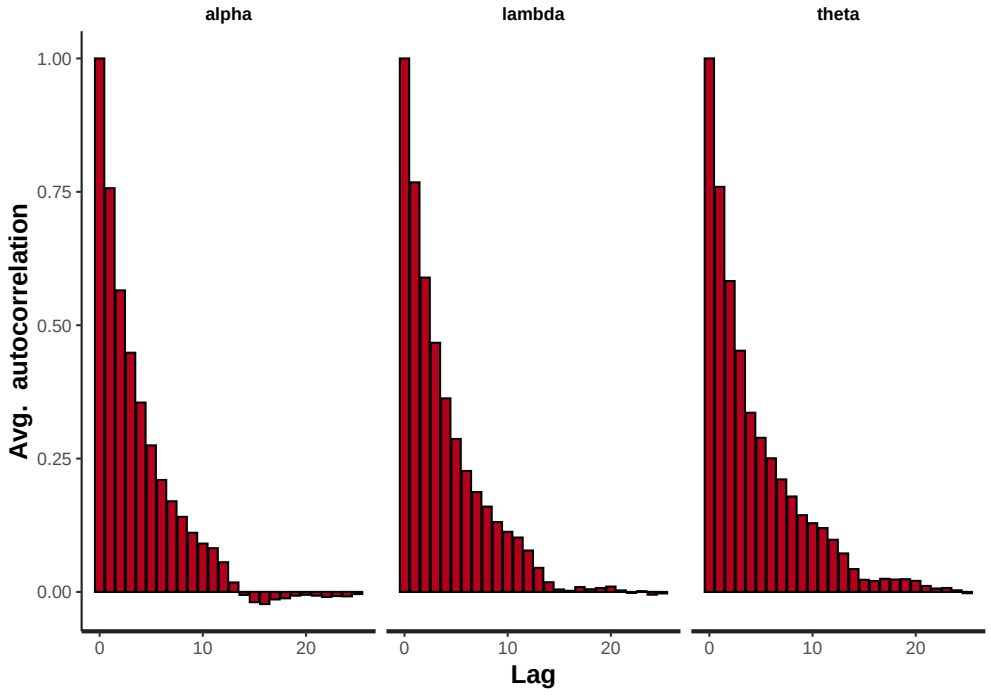


Figure 9: Autocorrelation plots of the parameters α, λ, and θ for all chains

These are autocorrelation plots for all chains and indicate that the samples of a Monte Carlo simulation are independent (Figure, 9). In Figure 10 we have displayed the histogram

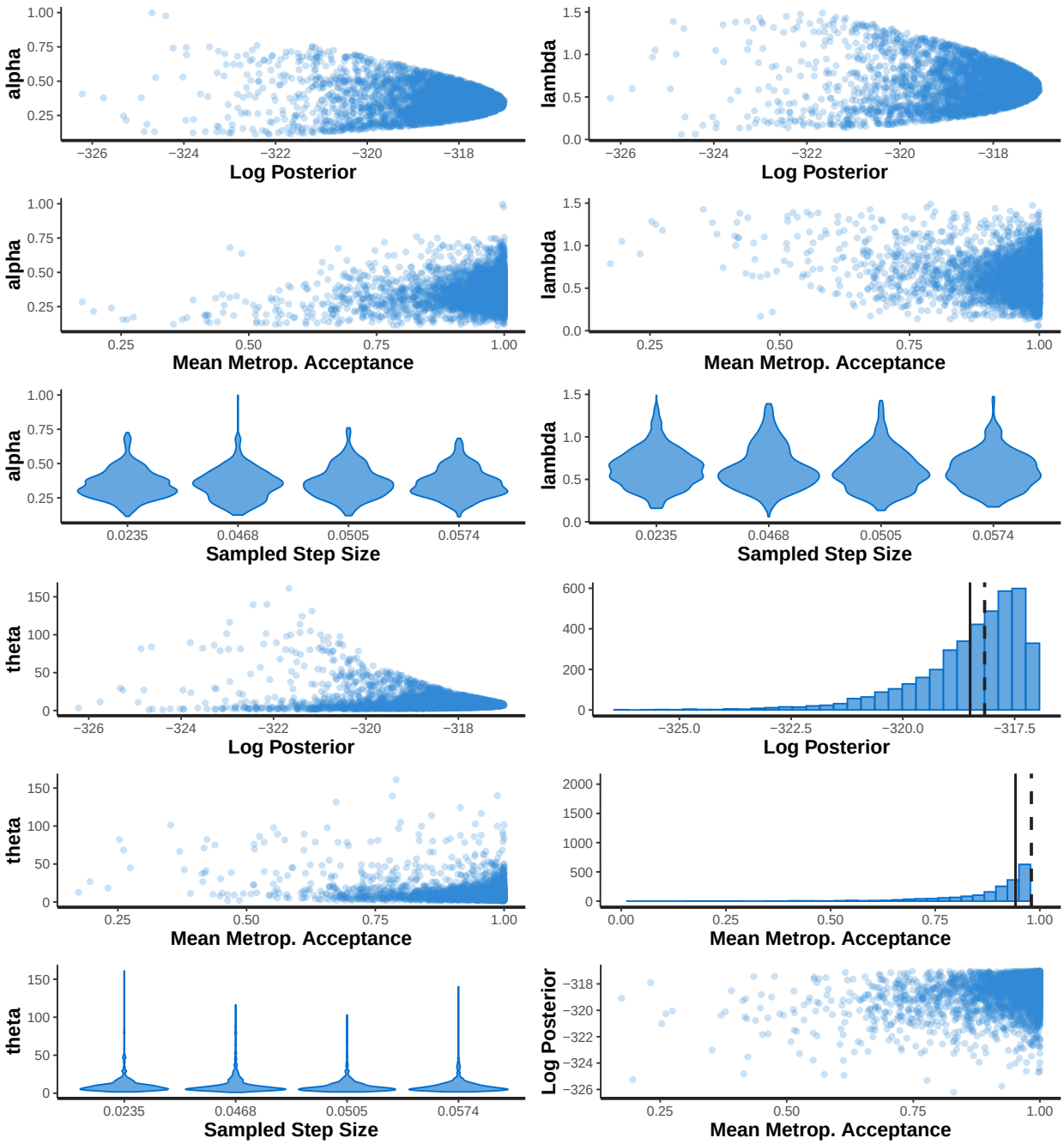


Figure 7: Average metropolis acceptance and step size for the parameters α, λ, θ and log posterior

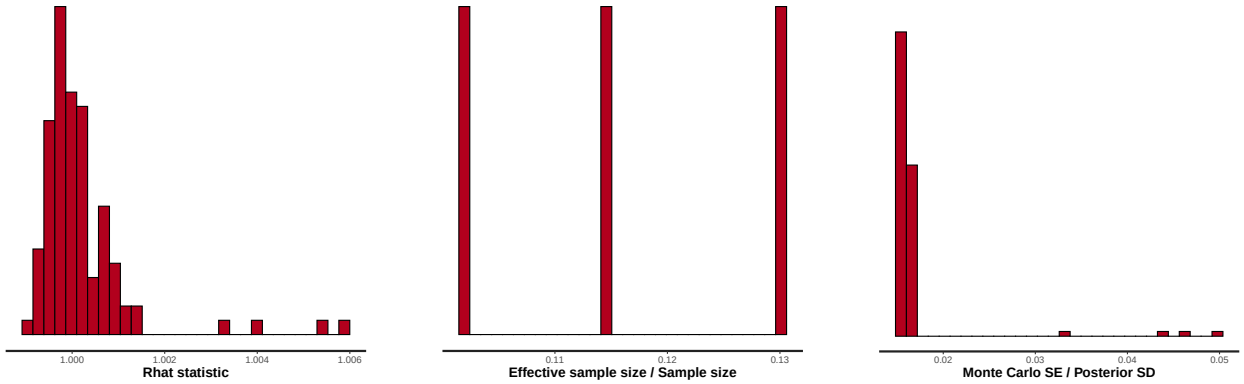


Figure 8: Histogram of Rhat statistic, the ratio of effective sample size and sample size, and the ratio of MCSE and posterior SD

of rank plots of α, λ, and θ for all four chains. Rank histograms visualize how the values from the chains mix in terms of ranking. An ideal plot would show the rankings mixing or overlapping in a uniform distribution. See (Vehtari *et al.*, 2021) for details. In general, we look for three possessions in the trace plots good mixing, stationarity, and convergence.

Good mixing implies that the chain quickly explores the full posterior region. It doesn't slowly wander, but rather rapidly zig-zags around, as a good Hamiltonian chain should. Stationarity indicates the path of each chain staying within the same high probability portion of the posterior distribution. Another way to imagine this is that the average value of the chain is relatively stable from start to end. Convergence represents that independent, multiple chains attach around the same area of high probability. Figure 11 shows the trace plots for alpha, theta, and lambda are well mixing and convergent for all chains.

The Ergodic mean is computed as the average of all values of the samples for all chains corresponding iterations. Figure 12 indicates that all chains converged smoothly around the mean value. Figure 13 is a pairs plot of MCMC draws of α, λ, and θ. Univariate marginal posteriors are shown along the diagonal as histograms. Bivariate plots are displayed above and below the diagonal as scatter plots. The red colored draws represent, if present, the divergent transitions. Divergent transitions can indicate problems with the validity of the results. A good plot would show no divergent transitions. A bad plot would show divergent transitions in a systematic pattern. We have also presented a detailed numerical summary of the HMC and NUTS algorithm in Table 5 and statistics related to the posterior summary are presented in Table 6.

Table 5: Informational statistic of NUTS/HMC for convergence of chains

	accept_stat	stepsize	treedepth	n_leapfrog	divergent	energy
All chains	0.9419	0.0446	3.9813	36.2755	0	319.991
chain1	0.9443	0.0574	3.8890	33.0920	0	319.913
chain2	0.9294	0.0505	3.9140	32.6420	0	320.035
chain3	0.9536	0.0468	4.0110	36.7580	0	320.111
chain4	0.9403	0.0235	4.1110	42.6100	0	319.907

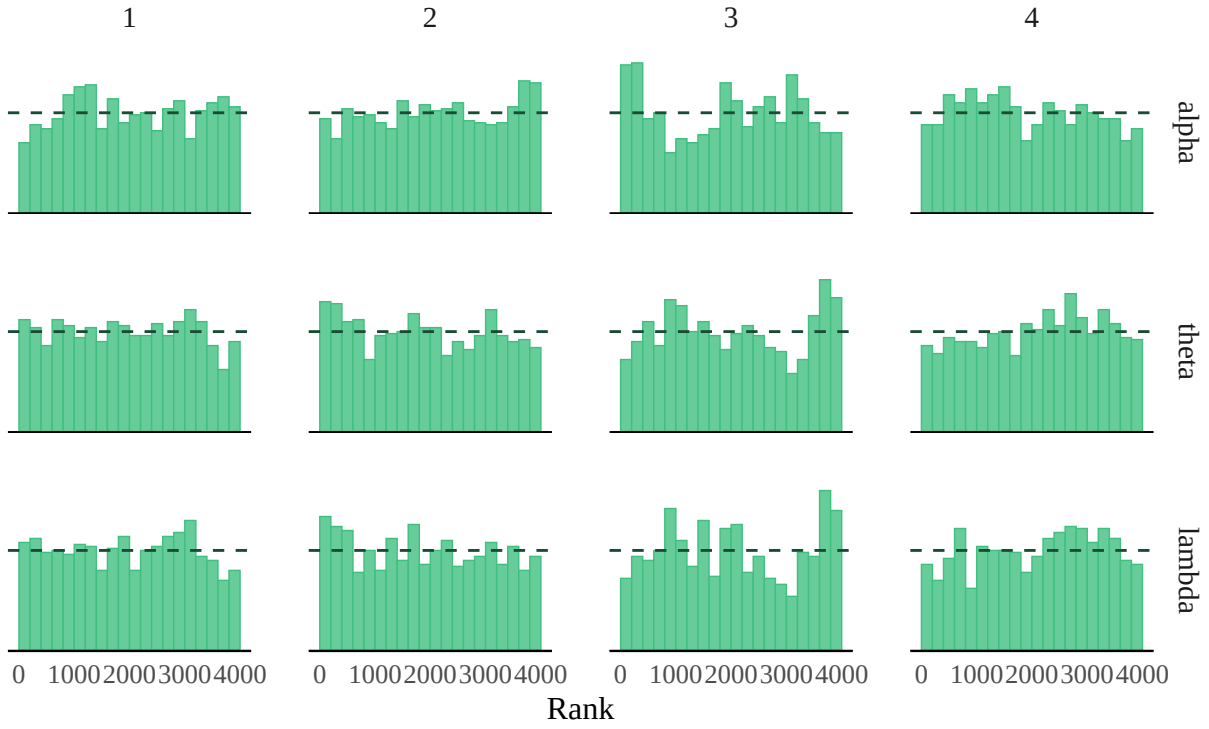


Figure 10: Rank histograms of α, λ, and θ

7. Posterior analysis

7.1. Numerical summary

Using the `stan()` function in R-Software we have estimated the posterior density of the fitted EEP model. The numerical summaries of the posterior distribution after fitting the EEP model for the data taken under study for all merged chains are reported in Table 6. The MCMC estimate for α is 0.36 ± 0.109 which is statistically significant. Similarly, the estimate for λ is 0.62 ± 0.233, which is statistically significant. The estimate for θ is 10.926 ± 12.646 which is also statistically significant. No parameters have an effective sample size (`n_eff`) for estimating the posterior mean less than 10 % of the total sample size indicating that the samples are efficient and `Rhat`($\hat{R}$) (estimated potential scale reduction statistic) provides the analysis of sampling and its efficiency. Here `Rhat` is less than 1.01 indicating convergence of all chains. Also, we have depicted the highest posterior density (HPD) credible interval and credible interval in Table 7.

Table 6: Output summary of posterior samples for the EEP model

Parameters	mean	`se_mean`	sd	2.50%	50%	97.50%	`n_eff`	`Rhat`
alpha	0.3553	0.0048	0.1094	0.1716	0.3454	0.6085	520	1.0039
lambda	0.6212	0.0109	0.2334	0.2339	0.5966	1.1632	458	1.0055
theta	10.9263	0.6267	12.6463	2.4253	7.2978	45.8048	407	1.0059
Log-posterior	-318.4981	0.0410	1.2386	-321.6890	-318.1700	-317.1100	911	1.0033

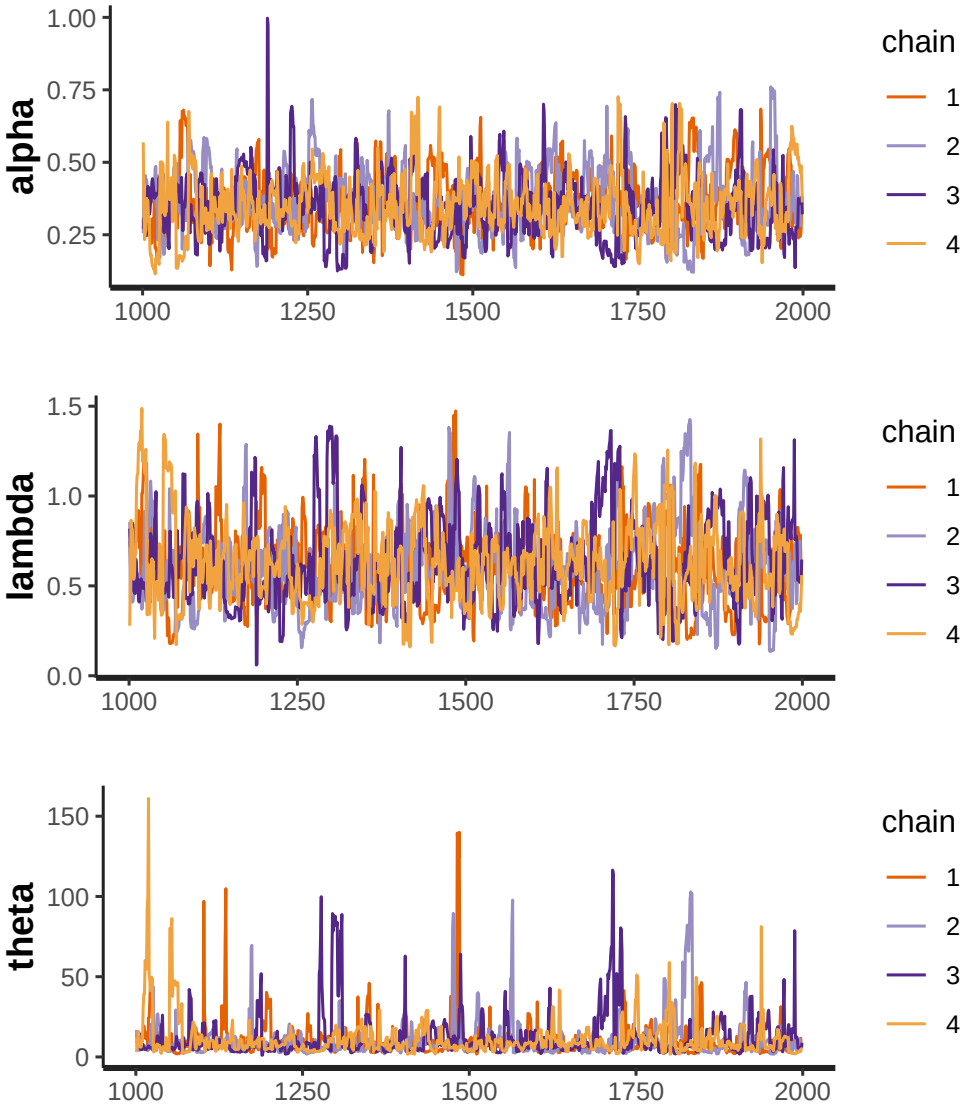


Figure 11: Trace plot of the parameters α, λ, and θ for all chains

7.2. Visual summary

Various graphical representations can be employed to visually summarize the posterior distribution, such as histograms, boxplots, caterpillar plots, and density plots. In this study, we utilized Gamma priors to plot histograms and kernel density estimates for α, λ, and θ (Figure, 14), based on a total of 4000 posterior samples. These graphical presentations offer comprehensive insights into the parameters' posterior distribution. Histograms are particularly useful for understanding the distribution's tail behavior, skewness, kurtosis, the presence of outliers, and whether multi-modal behavior exists. Our analysis reveals that α and λ exhibit almost symmetrical distributions, whereas θ demonstrates positive skewness under Gamma prior. Furthermore, in Figure (15), we present histograms of posterior parameters using a Uniform prior. It is evident that the choice of prior significantly impacts the resulting posterior distribution.

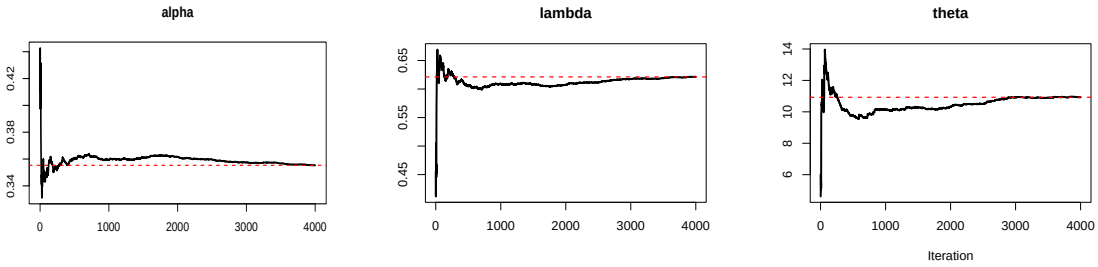


Figure 12: The Ergodic mean plots for α, λ, and θ

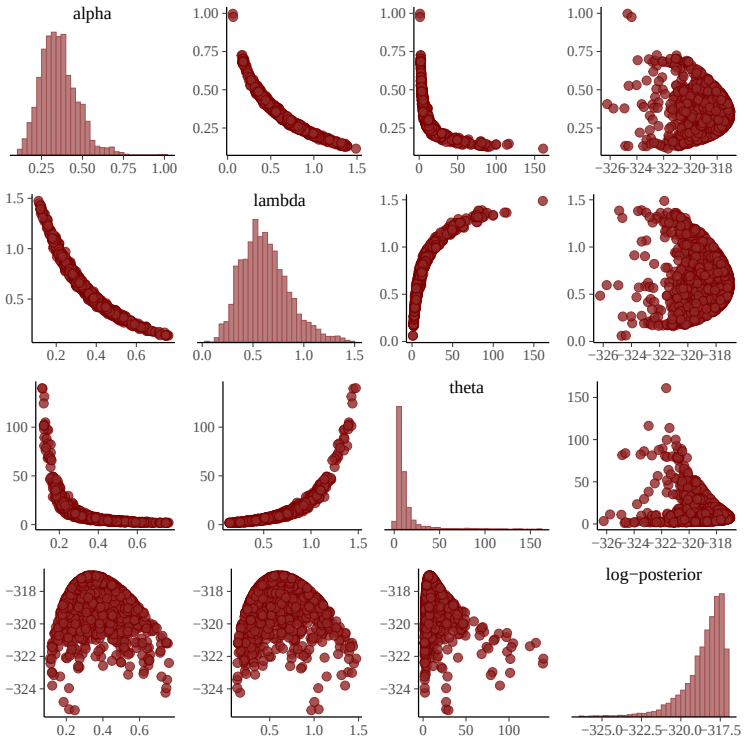


Figure 13: Pairs plot of α, λ, θ and log-posterior

8. Model compatibility

8.1. Posterior predictive checks (PPCs)

A usual way to access the fit of a Bayesian model is to observe how well the predictions can be made from the model that agrees with the observed data (Gelman, 2003; Gelman *et al.*, 2004). If our model is capable of fitting the data then it should generate data that are quite similar to the observed data. The data that are used for posterior predictive checks (PPCs) we can generate them by simulating the posterior predictive distribution. The R package bayesplot presents different plotting functions for visual posterior predictive checking; using observed data and simulated data from the posterior predictive distribution, we can generate these graphical displays (Gabry *et al.*, 2017).

Table 7: HPD interval and credible interval for model parameters α, λ, and θ

Parameters	HPD interval	Credible Interval
alpha	(0.141, 0.554)	(0.1716, 0.6085)
lambda	(0.168, 1.070)	(0.2339, 1.1632)
theta	(1.650, 31.20)	(2.4253, 45.8048)

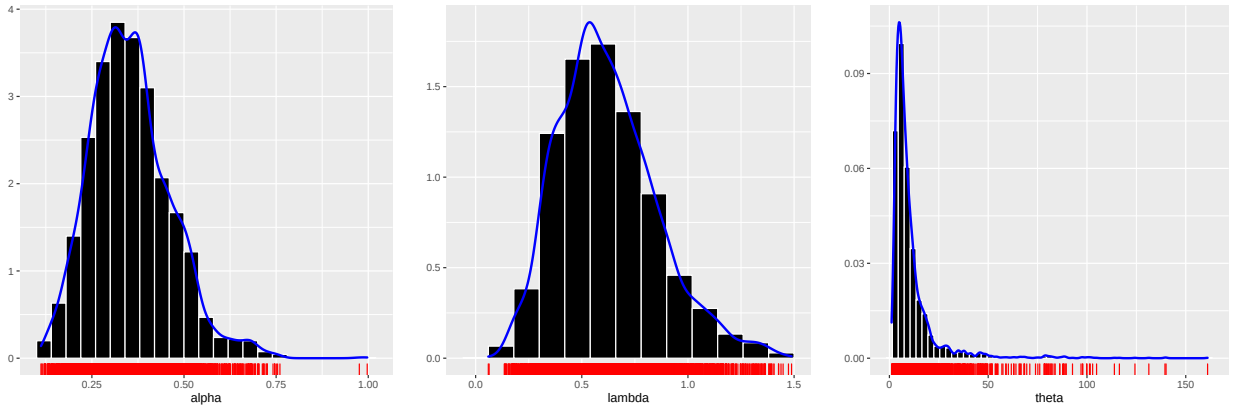


Figure 14: Histogram with kernel density estimates of posterior samples for the parameters α, λ, and θ respectively under a gamma prior

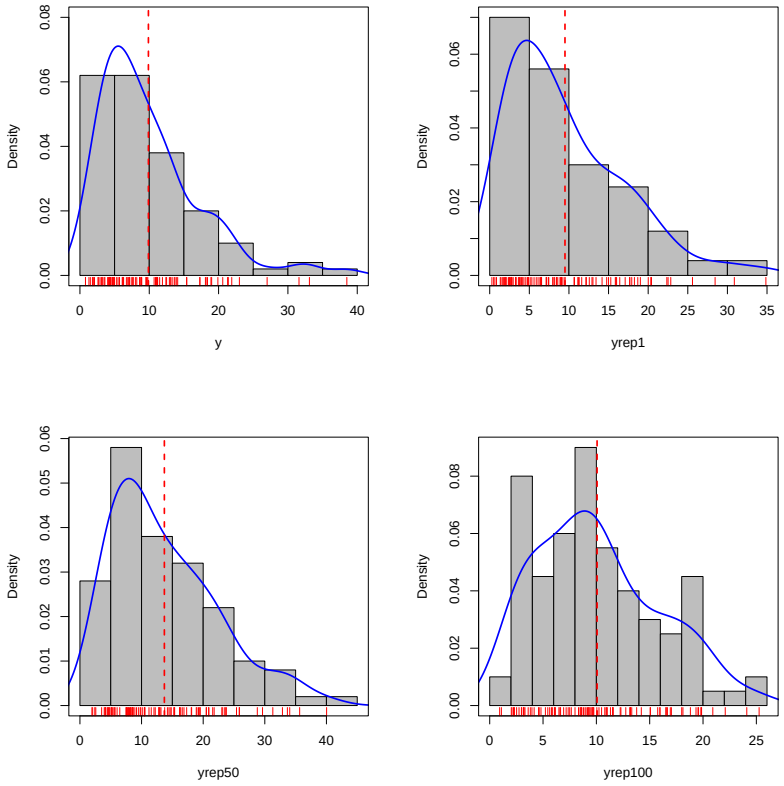


Figure 17: Histogram of observed data (y) and replicated data $y_{rep}[1]$, $y_{rep}[50]$ *and* $y_{rep}[100]$ with point estimate mean (red vertical line)

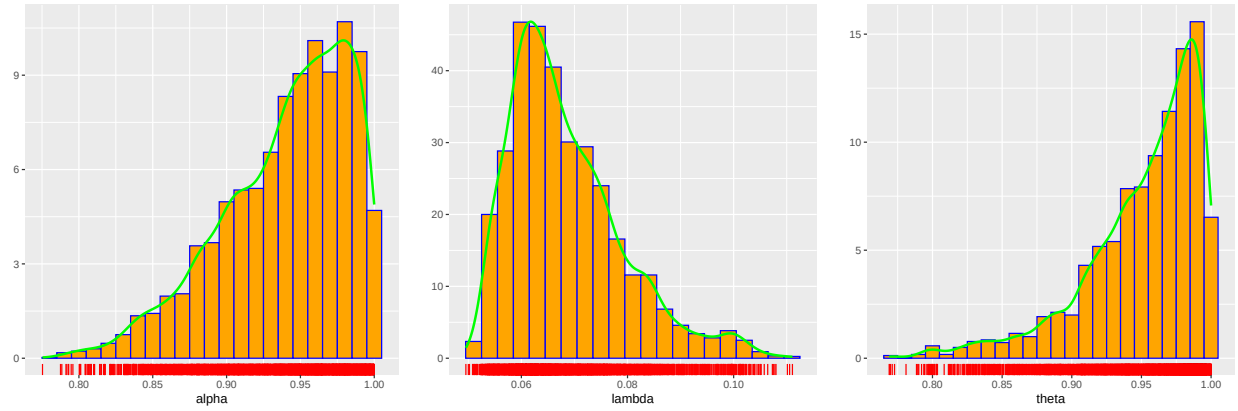


Figure 15: Histogram with kernel density estimates of posterior samples for the parameters α, λ, and θ respectively under a uniform prior

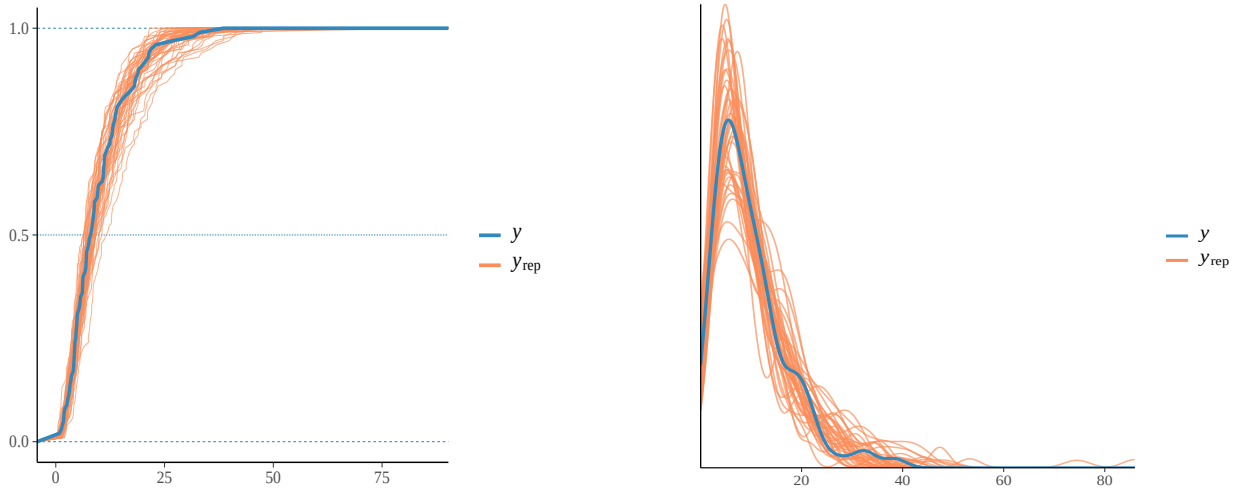


Figure 16: CDF plot of the observed dataset y (blue), with 40 simulated datasets y_{rep} (left panel) and kernel density estimate of the observed dataset y (blue), with density estimates for 40 simulated datasets y_{rep} drawn from the posterior predictive distribution (left panel)

The posterior predictive distribution is the distribution of the outcome variable implied by a model after using the observed data y (a vector of length $N = 100$) to update our beliefs about unknown parameters $\Theta = (\alpha, \lambda, \theta)$ of the model. The posterior predictive distribution for observation y_{rep} can be written as,

$$p\left(\tilde{y}/y\right) = \int p\left(\tilde{y}/\Theta\right) p\left(\Theta/y\right) d\Theta.$$

For every simulation (draw) $s = 1, ..., S$ of the parameters from the posterior distribution $\Theta(s) \sim p(\Theta|y)$, we generate a vector of N outcomes $\tilde{y}^{(s)}$ using the posterior predictive distribution by simulating from the data model conditional on parameters $\Theta_{(s)}$.

The result is an $S \times N(4000 \times 100)$ matrix of draws $\tilde{y}$. We have denoted the resulting simulation matrix by y_{rep}, this matrix is the replication of the observed data y rather than predictions for future observations. To attain further clarity on our decision for the study

of the posterior predictive checks we have taken the smallest, middle, and largest, i.e. ($y_{rep}[1],\ y_{rep}[50]$ *and* $y_{rep}[100]$) replicated observations. We have presented a wide variety of graphical model checks based on comparing observed data to draws from the posterior (or prior) predictive distribution. To Compare the empirical distribution of the data y to the distributions of simulated/replicated data y_{rep} from the posterior predictive distribution an empirical CDF estimate of each dataset (row) in y_{rep} are overlaid with the distribution of y (blue curve) is displayed in (Figure, 16, left panel) and kernel density estimate of the observed dataset y (blue), with density estimates for 40 simulated datasets y_{rep} drawn from the posterior predictive distribution (Figure, 16, right panel). To analyze the predicting capacity of posterior samples we have presented the visual summaries such as a histogram with kernel density plot for observed data y and simulated data $y_{rep}[1],\ y_{rep}[50]$ *and* $y_{rep}[100]$ (Figure 17).

8.2. Model selection

The WAIC (Widely Applicable Information Criterion) is used to compare different statistical models based on their out-of-sample predictive accuracy. The lower the WAIC value, the better the model's predictive performance. Hence EEP model is better than the EP model see (Table, 8). Where `elpd_waic` is the estimated log pointwise predictive density using the WAIC. It represents the model's fit to the data and is measured in terms of log-likelihood. `p_waic` is the effective number of parameters computed from the WAIC. It takes into account both the actual number of parameters in the model and the model's complexity and `waic` is the value of the WAIC itself, which is a combination of the model fit (`elpd_waic`) and the effective number of parameters `p_waic`. A lower WAIC indicates better predictive performance.

Table 8: Model selection statistics

Estimate	EEP distribution	EP distribution
`elpd_waic`	-319.6	-326.9
`p_waic`	1.5	1.2
`waic`	639.1	653.9

9. Conclusion

In this research work, we put forward a new distribution using the exponential power model as a baseline distribution and named it exponentiated exponential power (EEP) distribution. We have explored some properties including the hazard rate function, cumulative distribution function, survival function, probability density function, cumulative hazard function, order statistics, quantiles, the measures of skewness based on quartiles, and median, and kurtosis based on octiles.

Also we have performed a full Bayesian analysis for the proposed model. Using Stan software whose MCMC techniques are based on the NUTS which is an adaptive variant of HMC; a more robust and efficient sampler. We have presented the numerical as well as graphical analysis of the EEP model and found that all chains are well mixed and conversed. Further, we have estimated the parameters of the model and performed posterior predictive checks, and found that the underlying model can be used to generate reliable samples. The

developed techniques are applied to a real data set, thus we can apply for full Bayesian analysis for the proposed model using these Bayesian techniques. Hence it is expected that the EEP model will be a choice in the fields of the theory of probability, applied statistics, bayesian inferences, and survival analysis.

Acknowledgements

We are very grateful to the Chair Editor and the reviewer for valuable comments and suggestions which have improved considerably the first version of the manuscript.

References

Alizadeh, M., Afify, A. Z., Eliwa, M. S., and Ali, S. (2020). The odd log-logistic Lindley-G family of distributions: properties, Bayesian and non-Bayesian estimation with applications. *Computational Statistics*, **35**, 281–308.

Almarashi, A. M., Elgarhy, M., Elsehetry, M. M., Kibria, B., and Algarni, A. (2019). A new extension of exponential distribution with statistical properties and applications. *Journal of Nonlinear Sciences and Applications (JNSA)*, **12**, 135–145.

Ashour, S. K. and Eltehiwy, M. A. (2015). Exponentiated power Lindley distribution. *Journal of Advanced Research*, **6**, 895–905.

Balakrishnan, N. and Cohen, A. C. (2014). *Order Statistics and Inference: Estimation Methods*. Elsevier.

Barriga, G. D., Louzada-Neto, F., and Cancho, V. G. (2011). The complementary exponential power lifetime model. *Computational Statistics & Data Analysis*, **55**, 1250–1259.

Beskos, A., Pillai, N., Roberts, G., Sanz-Serna, J.-M., and Stuart, A. (2013). Optimal tuning of the hybrid Monte Carlo algorithm. *Bernoulli*, **19**, 1501–1534.

Betancourt, M. (2017). A conceptual introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434*, **22**.

Brooks, S. (1998). Markov chain Monte Carlo method and its application. *Journal of the Royal Statistical Society: series D (the Statistician)*, **47**, 69–100.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, **76**.

Chaubey, Y. P. and Zhang, R. (2015). An extension of chen's family of survival distributions with bathtub shape or increasing hazard rate function. *Communications in Statistics-Theory and Methods*, **44**, 4049–4064.

Chaudhary, A. K. and Kumar, V. (2020). A Bayesian estimation and prediction of Gompertz extension distribution using the MCMC method. *Nepal Journal of Science and Technology*, **19**, 142–160.

Chen, Z. (1999). Statistical inference about the shape parameter of the exponential power distribution. *Statistical Papers*, **40**, 459–468.

Chen, Z. (2000). A new two-parameter lifetime distribution with bathtub shape or increasing failure rate function. *Statistics and Probability Letters*, **49**, 155–161.

Devlin, L., Horridge, P., Green, P. L., and Maskell, S. (2021). The No-U-Turn sampler as a proposal distribution in a sequential Monte Carlo sampler with a near-optimal L-kernel. *arXiv preprint arXiv:2108.02498*, **Preprint**, 1–5.

Dey, S., Kumar, D., Ramos, P. L., and Louzada, F. (2017). Exponentiated Chen distribution: Properties and estimation. *Communications in Statistics-Simulation and Computation*, **46**, 8118–8139.

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., and Gelman, A. (2017). Visualization in Bayesian workflow. *arXiv preprint arXiv:1709.01449*, **182**.

Gelfand, A. E. and Smith, A. F. (1990). Sampling-based approaches to calculating marginal densities. *Journal of the American Statistical Association*, **85**, 398–409.

Gelman, A. (2003). A Bayesian formulation of exploratory data analysis and goodness-of-fit testing. *International Statistical Review*, **71**, 369–382.

Gelman, A., Carlin, J. B., Stern, H. S., and Rubin, D. B. (2004). Bayesian data analysis chapman & hall. *CRC Texts in Statistical Science*, **136**.

Gelman, A., J. B., C., Hal S., S., David B., D., Vehtari, A., and Donald B., R. (2013). *Bayesian data analysis*. Chapman and Hall/CRC.

Gelman, A., Lee, D., and Guo, J. (2015). Stan: A probabilistic programming language for bayesian inference and optimization. *Journal of Educational and Behavioral Statistics*, **40**, 530–543.

Geman, S. and Geman, D. (1984). Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 721–741.

Ghitany, M., Al-Mutairi, D. K., and Aboukhamseen, S. (2015). Estimation of the reliability of a stress-strength system from power Lindley distributions. *Communications in Statistics-Simulation and Computation*, **44**, 118–136.

Ghitany, M. E., Atieh, B., and Nadarajah, S. (2008). Lindley distribution and its application. *Mathematics and Computers in Simulation*, **78**, 493–506.

Gupta, R. D. and Kundu, D. (1999). Theory and methods: Generalized exponential distributions. *Australian and New Zealand Journal of Statistics*, **41**, 173–188.

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

Henningsen, A. and Toomet, O. (2011). maxlik: A package for maximum likelihood estimation in R. *Computational Statistics*, **26**, 443–458.

Hoffman, M. D., Gelman, A., et al. (2014). The no-u-turn sampler: adaptively setting path lengths in hamiltonian monte carlo. *Journal of Machine Learning Research*, **15**, 1593–1623.

Joshi, R. K., Sapkota, L. P., and Kumar, V. (2020). The logistic-exponential power distribution with statistical properties and applications. *International Journal of Emerging Technologies and Innovative Research*, **7**, 629–641.

Kundu, D. and Raqab, M. Z. (2005). Generalized Rayleigh distribution: different methods of estimations. *Computational Statistics and Data Analysis*, **49**, 187–200.

Lambert, B. (2018). *A Student's Guide to Bayesian Statistics*. SAGE Publications Ltd.

Marshall, A. W. and Olkin, I. (1997). A new method for adding a parameter to a family of distributions with application. *Biometrika*, **84**, 641–652.

McElreath, R. (2020). *Statistical Rethinking: A Bayesian Course with Examples in R and Stan*. Chapman and Hall/CRC.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The journal of Chemical Physics*, **21**, 1087–1092.

Moors, J. (1988). A quantile alternative for kurtosis. *Journal of the Royal Statistical Society: Series D (The Statistician)*, **37**, 25–32.

Mudholkar, G. S. and Srivastava, D. K. (1993). Exponentiated weibull family for analyzing bathtub failure-rate data. *IEEE Transactions on Reliability*, **42**, 299–302.

Nadarajah, S. and Haghighi, F. (2011). An extension of the exponential distribution. *Statistics*, **45**, 543–558.

Neal, R. (2011). *MCMC Using Hamiltonian Dynamics*. Chapman & Hall/CRC.

Nishio, M. and Arakawa, A. (2019). Performance of hamiltonian monte carlo and no-u-turn sampler for estimating genetic parameters and breeding values. *Genetics Selection Evolution*, **51**, 1–12.

R Core Team (2022). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ristić, M. M. and Nadarajah, S. (2014). A new lifetime distribution. *Journal of Statistical Computation and Simulation*, **84**, 135–150.

Sapkota, L. P. (2020). Exponentiated–exponential logistic distribution: Some properties and application. *Janapriya Journal of Interdisciplinary Studies*, **9**, 100–108.

Sapkota, L. P. (2022). A Bayesian analysis and estimation of weibull inverse rayleigh distribution using hmc method. *Nepal Journal of Mathematical Sciences*, **3**, 39–58.

Smith, R. M. and Bain, L. J. (1975). An exponential power life-testing distribution. *Communications in Statistics-Theory and Methods*, **4**, 469–481.

Stan Development Team (2020). RStan: the R interface to Stan. R package version 2.21.1.

Stan Development Team (2022). The Stan Core Library. Version 2.31.0.

Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., and Bürkner, P.-C. (2021). Rank-normalization, folding, and localization: An improved $\hat{R}$ for assessing convergence of MCMC (with discussion). *Bayesian Analysis*, **16**, 667–718.

ANNEXURE

```
functions{real expexp_power_lpdf(real y, real alpha,
        real lambda, real theta){
        return log(alpha*lambda*theta) + (alpha-1)*
        log(y)+1+lambda*y^alpha - exp(lambda*y^alpha)+
                (theta-1)*log(1-exp(1-exp(lambda*y^alpha)));
        }
        real expexp_power(real alpha, real lambda, real theta){
        return ((1/lambda)*log(1-log(1-
        (uniform_rng(0,1))^(1/theta))))^(1/alpha);
        }
}
```

```
        data{
                int N;
                real y[N];
        }

        parameters{
                real <lower=0> alpha;
                real <lower=0> lambda;
                real <lower=0> theta;
        }

        model{
                for(i in 1 : N) {
                        y[i] ~ expexp_power(alpha, lambda, theta);
                }
                alpha ~ gamma(0.001, 0.001);
                lambda ~ gamma(0.001, 0.001);
                theta ~ gamma(0, 0.001);
        }

        generated quantities{
        vector [N] yrep;
        for(i in 1 : N)
        {
                yrep[i]= expexp_power_rng(alpha, lambda, theta);
        }
}
```

Data Creation in R software

```
y = c(0.8, 0.8, 1.3, 1.5, 1.8, 1.9, 1.9, 2.1, 2.6, 2.7, 2.9,
3.1, 3.2, 3.3, 3.5, 3.6, 4.0, 4.1, 4.2, 4.2, 4.3, 4.3,
4.4, 4.4, 4.6, 4.7, 4.7, 4.8, 4.9, 4.9, 5.0, 5.3, 5.5,
5.7, 5.7, 6.1, 6.2, 6.2, 6.2, 6.3, 6.7, 6.9, 7.1, 7.1,
7.1, 7.1, 7.4, 7.6, 7.7, 8.0, 8.2, 8.6, 8.6, 8.6, 8.8,
8.8, 8.9, 8.9, 9.5, 9.6, 9.7, 9.8, 10.7, 10.9, 11.0,
11.0, 11.1, 11.2, 11.2, 11.5, 11.9, 12.4, 12.5, 12.9,
13.0, 13.1, 13.3, 13.6, 13.7, 13.9, 14.1, 15.4, 15.4,
17.3, 17.3, 18.1, 18.2, 18.4, 18.9, 19.0, 19.9, 20.6,
21.3, 21.4, 21.9, 23.0, 27.0, 31.6, 33.1, 38.5)
N <- length(y)
Data = list(y=y, N=N).
```

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 22, No. 2, 2024 (New Series), pp 259–265
http://www.ssca.org.in/journal

Linear Trend-Free Group Divisible Design

Longjam Roshini Chanu and K. K. Singh Meitei
Department of Statistics
Manipur University, Canchipur, Imphal, Manipur, 795003, India

Received: 05 August 2023; Revised: 19 January 2024; Accepted: 24 January 2024

Abstract

In this paper, we extend the construction method of Srivastava for a linear trend-free balanced incomplete block design of size k=2 into a linear trend-free group divisible design. Another construction method for linear trend-free group divisible design has also been developed.

Key words: Linear trend free; Association scheme; Block design; Group divisible.

AMS Subject Classifications: 62K10

1. Introduction

In specific experiments where several treatments are compared in blocks and within blocks, the treatments are applied to the experimental units sequentially over time or space; there is a possibility that a systematic effect or trend effect influences the observations in addition to the block and the treatment effects. In such a situation, a common polynomial trend in one or more dimensions is assumed to exist over the plots in each block of a classical experimental design. One may think of a suitable design that is orthogonal to trend effects, in the sense that the analysis of the design could be done in the usual manner as if no trend effects were present. Bradley and Yeh (1980) have called such designs as Trend Free Block (TFB) designs. The idea is that starting from a block design, a good design is chosen by permuting the treatments to plot positions within blocks. For example, Latin square and Youden square designs with blocks formed by their column are trend-free designs. TFB design has been extensively studied in the literature by Yeh and Bradley (1983), Chai and Majumdarn(1993), Lal *et al.* (2005), Gupta *et al.* (2020), Srivastava R. (accessed on 21.11.2023) gave on the construction of TFB designs.

2. Notation and preliminary results

We assume that within blocks there is a common polynomial trend of order p on the k periods that can be expressed by the orthogonal polynomials $\phi_\alpha(l), 1 \leq \alpha \leq p$, on $l = 1, 2, \ldots, k$, where $\phi_\alpha(l)$ is a polynomial of degree α. The polynomials $\phi_1(l), \ldots, \phi_p(l)$

Corresponding Author: Longjam Roshini Chanu
Email: roshini.longjam.96@gmail.com

satisfy

$$\sum_{l=1}^{k} \phi_\alpha(l) = 0, \sum_{l=1}^{k} \phi_\alpha(l)\phi_{\alpha'}(l) = \delta_{\alpha\alpha'}$$

where $\delta_{\alpha\alpha'}$ denotes the Kronecker delta, $\alpha, \alpha' = 1, 2, \ldots, p$.

If the trend is linear then, $p = 1$.

Let a design d will be represented by a $k \times b$ array of symbols $1, \ldots, \nu$, with columns denoting blocks and row periods. Thus, if the entry in cell (l, j) of d is i, it means that under d, treatment i, has to be applied in period l of block j. Let $D(\nu, b, k)$ be all connected designs in b blocks, k periods based on ν treatments.

Let $d \in D(\nu, b, k)$ and S_{dil} denote the number of times treatment i appears in row (period) l. It has been shown by Chai and Majumdar(1993) that a design is linear trend-free block (LTFB) design iff

$$\sum_{l=1}^{k} S_{dil}\phi_1(l) = 0, \ \ i = 1, \ldots, \nu \tag{1}$$

where $\phi_1(l)$ is the orthogonal polynomials of degree 1, $l = 1, 2, \ldots, k$ and S_{dil} denotes the number of times treatment i appears in row (period) l.

Condition (1) holds for binary as well as non-binary designs, and also irrespective of whether k is large, equal or smaller than ν, see Lin and Dean (1991). The polynomials $\phi_1(l)$ satisfy the condition

$$\phi_1(l) = -\phi_1(k - l + 1) \tag{2}$$

In addition,

$$\phi_1\left(\frac{k+1}{2}\right) = 0, \text{when } k \text{ is odd.}$$

3. Construction of linear trend-free group divisible (LTFGD) designs

3.1. Extension of Srivastava construction

Srivastava proposed a construction method of linear trend-free (LTF) balanced incomplete block design (BIBD) with parameters $v^* = 2q + 1, b^* = v^*(v^* - 1)/2, r^* = v^* - 1, k^* = 2, \lambda^* = 1$ for q positive integer. Then, such designs can be converted into LTF group divisible (GD) designs by augmenting some more treatments and blocks.

Theorem 1: The existence of an LTFBIBD with parameter $v^* = 2q + 1, b^* = v^*(v^* - 1)/2, r^* = v^* - 1, k^* = 2, \lambda^* = 1$ implies that an LTFGD design with parameters $v = v^*m, b = b^*m, r = r^*, k = 2, \lambda_1 = 1, \lambda_2 = 0, m, n = v^*$.

Proof: Let D be an LTFBIB design. Consider a group divisible association scheme (GDAS) on m different groups each of $n = v^*$ different treatments.

By using all treatments of every group of the GDAS as the treatment of the design,

then b^*m blocks are constructed, resulting in a new incomplete block design, d with $v = v^*m, b = b^*m$.

Obviously, $r = r^*, k = k^*, m, n = v^*$.

As all the treatments in a group of the GDAS are treated as treatments of the given BIBD, D, with $\lambda^* = 1$ any two treatments in the same group of GDAS occur once in a block of d, *i.e.*, $\lambda_1 = 1$.

By the construction method of LTFBIBD, no two treatments from different groups can occur together in any block of d. It follows that $\lambda_2 = 0$.

From (2), $\phi_1(1) = -\phi_1(2)$

By the construction method of LTFBIBD, $\sum_{l=1}^{2} S_{dil}\phi_1(l) = 0,\ i = 1, \ldots, v^*$.

Now,

$$\sum_{l=1}^{2} S_{dil}\phi_1(l) = \sum_{i=1}^{v^*m}[S_{dil}\phi_1(1) + S_{dil}\phi_1(2)]$$

As each period of the LTFBIB design are replicated m times by the construction method of LTFGD design.

$$\sum_{l=1}^{2} S_{dil}\phi_1(l) = m\sum_{i=1}^{v^*m}[S_{dil}\phi_1(1) + S_{dil}\phi_1(2)] = 0$$

Hence, proof of the theorem is complete. □

Starting from an LTFBIB design with the parameters $v^* = 5, b^* = 10, r^* = 4, k^* = 2, \lambda^* = 1$ when every treatment occupies all the period (*viz.* 1st and 2nd) the same number of times, *i.e.*, twice, a LTFGD is constructed as an example of the theorem 1.

Example 1: Given a group divisible association scheme (m=2, n=5) as follows

1st group: 0, 1, 2, 3, 4;

2nd group: 5, 6, 7, 8, 9,

using the LTFBIB design with the parameters $v^* = 5, b^* = 10, r^* = 4, k^* = 2, \lambda^* = 1$,

θ	-1	1
B_1	a	b
B_2	a	c
B_3	d	a
B_4	e	a
B_5	b	c
B_6	d	e
B_7	b	d
B_8	e	b
B_9	c	d
B_{10}	c	e

considering $0, 1, 2, 3, 4$ and again $5, 6, 7, 8, 9$, the elements in the 1st group and in the 2nd group respectively as treatments of the LTFGD design, 20 blocks of LTFGD design with the parameters $v = 10, b = 20, r = 4, k = 2, \lambda_1 = 1, \lambda_2 = 0, m = 2, n = 5$, and the first row represent the orthogonal trend component of degree one without normalization,

θ	-1	1
B_1	0	1
B_2	0	2
B_3	3	0
B_4	4	0
B_5	1	2
B_6	3	4
B_7	1	3
B_8	4	1
B_9	2	3
B_{10}	2	4
B_{11}	5	6
B_{12}	5	7
B_{13}	8	5
B_{14}	9	5
B_{15}	6	7
B_{16}	8	9
B_{17}	6	8
B_{18}	9	6
B_{19}	7	8
B_{20}	7	9

3.2. LTFGD designs for $k \geq 2$

Consider a GDAS with m groups each of n elements where the i^{th} group is given by

$$G_i = \{(i-1)n + 1, (i-1)n + 2, \cdots, in\}$$

Consider m latin square arrays of the same order n (whether they are the same or not, but the order should be the same).

Treating all the n elements of the i^{th} group as the elements of the i^{th} latin square and considering each column of the resulting i^{th} latin square array with elements from G_i, as block for each group, n blocks are constructed as given by

$$B_j^{(i)} = l_j^{(i)} \tag{3}$$

where $l_j^{(i)}$ is the j^{th} column of the i^{th} resulting latin square array L_i, say, with elements $(i-1)n + 1, (i-1)n + 2, \ldots, in$ from the i^{th} group G_i. Continuing the same process for i, we have mn blocks.

Taking p (positive integer) copies of these mn blocks $B_j^{(i)}$ where $i = 1, 2, \ldots, m; j = 1, 2, \ldots, n$, the configuration yields an LTFGD as shown in the following theorem.

Theorem 2: A series of LTFGD design with parameters $v = mn, b = mnp, r = np, k = n; m, n, \lambda_1 = r, \lambda_2 = 0$ for p positive integer can always be constructed.

Proof: As the GDAS under consideration is on m different groups, each of n different elements, so $v = mn$.

By the construction method of blocks given in the relation (3), from each resulting latin square array L_i, n blocks $l_j^{(i)}$, are constructed. Counting the p copies of n blocks from the resulting latin square array L_i for all $i; i = 1, 2, \ldots, m$, the configuration has mnp blocks. Further, any treatment of the i^{th} group G_i gets replicated once in each of the columns of the resulting latin square array L_i and gets replicated n times in those n blocks $B_j^{(i)}$ given by the relation (3); $j = 1, 2, \ldots, n$. By the process of taking p copies of each block, $r = np$.

Since each column of these m latin square designs has n distinct treatments, then $k = n$.

The construction method of blocks given in the relation (3), it can be seen that any two treatments from the i^{th} group of the GDAS occurs together exactly once in each column of the latin square array, under consideration, *i.e.*, the i^{th} latin square array, as any element in a latin square array occurs exactly once in each column of the latin square array. So, from those n blocks constructed based on the i^{th} latin square array, any two treatments from the i^{th} group of the GDAS occurs together in n blocks which have been constructed based on that i^{th} latin square array. Treating of p copies of each of the constructed blocks by the construction method given in the relation (3) gives as $\lambda_1 = np = r$.

From the construction method of blocks given in the relation (3), it is known that no two treatments from different groups occur together in any block. Thus, $\lambda_2 = 0$.

Since every treatment of the i^{th} group appears n times in each position l.

Then,

$$\begin{aligned} S_{dil} &= \text{number of times treatment } i \text{ appears in position } l \\ &= n \end{aligned}$$

By, $\phi_1(l) = -\phi_1(k - l + 1)$; where $\phi_1(l)$ is the orthogonal polynomial of degree 1 and $\phi_1[(k + 1)/2] = 0$; when k is odd,

We get, $\phi_1(l) = -\phi_1(k); \phi_1(2) = -\phi_1(k - 1)$ and so on.

Now, for $k =$ even

$$\begin{aligned} \sum_{i=1}^{k} S_{dil}\phi_1(l) &= n \sum_{i=1}^{k} \phi_1(l) \\ &= n \left[\phi_1(1) + \phi_1(2) + \cdots + \phi_1\left(\frac{k}{2} - 1\right) + \phi_1\left(\frac{k}{2}\right) + \phi_1\left(\frac{k}{2} + 1\right) \cdots + \phi_1(k-1) \right. \\ &\left. + \phi_1(k) \right] \\ &= n \left[\phi_1(1) + \phi_1(2) + \cdots + \phi_1\left(\frac{k}{2}\right) - \phi_1\left(\frac{k}{2}\right) - \cdots - \phi_1(2) - \phi_1(1) \right] \\ &= n \times 0 \\ &= 0 \end{aligned}$$

Again, for $k =$ odd

$$\begin{aligned}\sum_{i=1}^{k} S_{dil}\phi_1(l) &= n\sum_{i=1}^{k}\phi_1(l)\\ &= n\Bigg[\phi_1(1)+\phi_1(2)+\cdots+\phi_1\left(\frac{k+1}{2}-1\right)+\phi_1\left(\frac{k+1}{2}\right)\\ &\qquad +\phi_1\left(\frac{k+1}{2}+1\right)+\cdots+\phi_1(k-1)+\phi_1(k)\Bigg]\\ &= n\Bigg[\phi_1(1)+\phi_1(2)+\cdots+\phi_1\left(\frac{k+1}{2}-1\right)+\phi_1\left(\frac{k+1}{2}\right)\\ &\qquad -\phi_1\left(\frac{k+1}{2}-1\right)-\cdots-\phi_1(2)-\phi_1(1)\Bigg]\\ &= n\times 0\\ &= 0\end{aligned}$$

Hence, proof of the theorem is complete. □

An example of Theorem 2 is shown as an illustration below,

Consider GDAS($m = 2, n = 3$) such that $G_1 = (1,2,3); G_2 = (4,5,6)$ and also consider 2 latin square arrays of order 3.

$$L_1 = \begin{pmatrix} a & b & c\\ b & c & a\\ c & a & b\end{pmatrix}, L_2 = \begin{pmatrix} \beta & \gamma & \alpha\\ \alpha & \beta & \gamma\\ \gamma & \alpha & \beta\end{pmatrix}$$

From these Latin squares L_1 and L_2, by the construction method given in the relation (3), using the elements $(1,2,3)$ and $(4,5,6)$, respectively given below

$$L_1^* = (l_1^1, l_2^1, l_3^1) = \begin{pmatrix} 1 & 2 & 3\\ 2 & 3 & 1\\ 3 & 1 & 2\end{pmatrix}; L_2^* = (l_1^2, l_2^2, l_3^2) = \begin{pmatrix} 5 & 6 & 4\\ 4 & 5 & 6\\ 6 & 4 & 5\end{pmatrix}$$

Considering each column of L_1^* and L_2^* as blocks for each group

$$B_1^{(1)} = l_1^1 = \begin{pmatrix}1\\2\\3\end{pmatrix}; B_2^{(1)} = l_2^1 = \begin{pmatrix}2\\3\\1\end{pmatrix}; B_3^{(1)} = l_3^1 = \begin{pmatrix}3\\1\\2\end{pmatrix};$$

$$B_1^{(2)} = l_1^2 = \begin{pmatrix}4\\5\\6\end{pmatrix}; B_2^{(2)} = l_2^2 = \begin{pmatrix}5\\6\\4\end{pmatrix}; B_3^{(2)} = l_3^2 = \begin{pmatrix}6\\4\\5\end{pmatrix}.$$

Taking 2 copies of these 6 blocks, the configuration yields an LTFGD design, as shown in the example given below.

Example 2: Following is a plan of LTFGD design with the parameters $v = 6, b = 12, r = 6, k = 3, m = 2, n = 3, \lambda_1 = 6, \lambda_2 = 0$ and 1^{st} row represents orthogonal trend component of degree one without normalization.

θ	-1	0	1
B_1	1	2	3
B_2	2	3	1
B_3	3	1	2
B_4	4	5	6
B_5	5	6	4
B_6	6	4	5
B_7	1	2	3
B_8	2	3	1
B_9	3	1	2
B_{10}	4	5	6
B_{11}	5	6	4
B_{12}	6	4	5

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.

References

Bradley, R. A. and Yeh, C. M. (1980). Trend-free block designs: Theory. *The Annals of Statistics*, **8**, 883–893.

Chai, F. S. and Majumdar, D. (1993). On the yeh-bradley conjecture on linear trend-free block designs. *The Annals of Statistics*, **21**, 2087–2097.

Gupta, R. K., Bhowmik, A., Jaggi, S., Varghese, C., Harun, M., and Datta, A. (2020). Trend free block designs in three plots per block. *RASHI*, **4**, 1–6.

Lal, K., Parsad, R., and Gupta, V. K. (2005). *A Study on Trend Free Designs.* IASRI, New Delhi. I.A.S.R.I./P.R.-02/2005.

Lin, M. and Dean, A. M. (1991). Trend-free block designs for varietal and factorial design. *Annals of Statistics*, **19**, 1582-1598.

Srivastava, R. (Accessed on 21.11.2023). Trend free block designs. In Parsad, R., Srivasatava, R., and Gupta, V. K. Eds: Design and analysis of agricultural experiments - a teaching manual hosted at *Design Resources Server, Indian Agricultural Statistics Research Institute, New Delhi-110012*, 437-444. https://drs.icar.gov.in/ebook/EDBDAT/index.htm.

Yeh, C. M. and Bradley, R. A. (1983). Trend-free block designs: existence and construction results. *Communications in Statistics-Theory and Methods*, **12**, 1–24.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 22, No. 2, 2024 (New Series), pp 267–283
http://www.ssca.org.in/journal

A Systematic Literature Review of Sustainable Probabilistic Inventory Models

Khimya Tinani and Anuja Sarangale
P. G. Department of Statistics, Sardar Patel University, Vallabh Vidyanagar, Gujarat, 388120, India.

Received: 11 July 2023; Revised: 25 December 2023; Accepted: 04 February 2024

Abstract

Incorporation of changing environmental needs in the daily businesses of the life are becoming the essential element for the long-term sustenance of the humanity considering the degrading natural environment. Brundtland report by the United Nations signifies the inclusion of sustainable practices to fulfil the needs of the present generation at the same time preserving resources for the future generations. Realizing the presence of number of polluting factors in the inventory management practices, this research attempts to give extensive systematic review of the available literature which is also incorporating the uncertainty of the components by considering their stochastic or probabilistic behaviour. This review assessed 32 research articles to write a comprehensive review where all the articles have incorporated at least one probability distribution. This study identifies that transportation, storage and production are the main contributors of carbon emissions and normal distribution is the most preferred probability distribution and hence future research can be extended by incorporation of other probability distributions in the model building of sustainable inventory management.

Key words: Carbon emission; Sustainable inventory management; Probability distributions; Demand; Normal distribution; Lead time.

1. Introduction

Sustainable inventory management has become an increasingly important topic in recent years, as businesses look to minimize their environmental footprint and promote social responsibility. One of the key benefits of sustainable inventory management is the potential for cost savings. By reducing waste and increasing efficiency, businesses can lower their costs and improve their bottom line. In addition to cost savings, sustainable inventory management also has the potential to improve a company's reputation and increase customer loyalty. As consumers become more environmentally conscious, they are increasingly looking for companies that are committed to sustainability. However, implementing sustainable inventory management can also present challenges. One of the main challenges is the lack of clear and consistent definitions and metrics for sustainability. This can make it difficult for businesses to know exactly what they need to do in order to be considered sustainable.

Corresponding Author: Khimya Tinani
Email: khimya27@yahoo.coml

Additionally, sustainable inventory management can require a significant investment in new technologies and processes, which can be a barrier for small and medium-sized businesses. To overcome these challenges, researchers have proposed a number of different methods and technologies that can be used to achieve sustainability goals in inventory management. One approach is the use of green supply chain management, which involves integrating environmental considerations into all aspects of the supply chain, including inventory management.

A systematic literature review of sustainable inventory management reveals a growing body of research that explores the various aspects of this topic including the benefits and challenges of implementing sustainable practices as well as the various methods and technologies that can be used to achieve sustainability goals. However, it also points to the challenges that businesses may face in implementing sustainable inventory management, and the importance of clear definitions and metrics, as well as the use of advanced technologies.

Variability in demand, lead time or any other component of inventory problems navigates the entire procedure of decision making of the practitioner and hence inclusion of it will help to understand more diverse and realistic scenarios of the inventory problem. There is enough literature where probabilistic nature of the demand or lead time has benefitted to develop inventory models for real life scenarios. Among which research work on normal distribution and gamma distribution is available for past many decades. Such as Burgin and Wild (1967) developed a procedure to obtain the reorder level and reorder quantity when lead time demand have probabilistic nature and obtained numerical expressions to particular case for gamma distribution. In another study by Burgin (1972) where demand is normal and lead times is gamma distributed exact expressions for reorder level and lost sales was obtained. Another interesting use of probability distributions can be found in the study by Lee *et al.* (2007) where mixture of two normal distributions is considered to obtain the order quantity.

Examining such diverse applications of probability distributions in inventory management this paper aims: (i) to lay out extensive analysis of the sustainable inventory management problems which include probabilistic nature of components. (ii) to discover different probability distributions used in the inventory management. (iii) to find the future research directions with incorporation of various probability distributions in inventory management.

The research paper further subdivided into four sections where Section 2 provides the review methodology followed for the inclusion of the articles for systematic literature review which itself divided into different subsections based on analysis techniques used. Section 3 concludes on the overall study and Section 4 ends the research paper by giving limitations and possible future research direction to the study.

2. Review methodology

For conducting systematic literature review, a review methods proposed by Becerra *et al.* (2021, 2022), Pattnaik *et al.* (2021), Tinani and Kandpal (2017) based on which this study can be broadly divided into two phases as articles selection phase and analysis phase. Articles selection phase involves identification of keywords, searching through database, abstract screening and full article screening based on the objectives. Whereas analysis phase involves thorough examination of articles to provide valuable insights.

2.1. Article selection phase

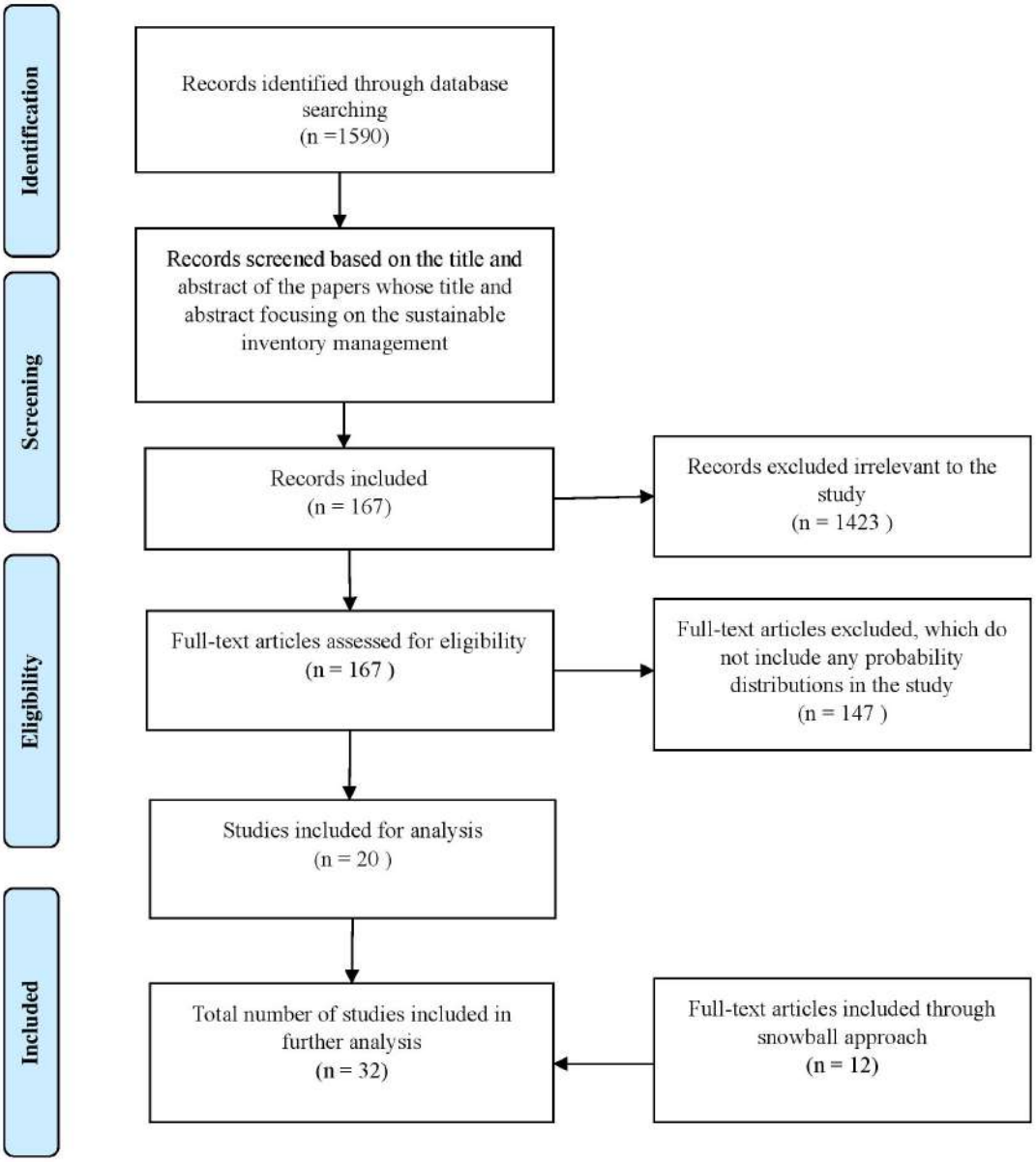


Figure 1: Flowchart of literature review

The search for the articles was conducted using Web of Science database for which initially keywords were identified such as “sustainability”, “environment”, “carbon emission”, “carbon tax”, “carbon footprint” and “carbon cap and trade” which were then combined with the inventory management to prepare a search string as TS=(sustainab* OR green OR environment* OR carbon OR ”carbon tax” OR ”carbon emission” OR “carbon footprint” OR “carbon cap and trade”) AND TS= (”inventory management” OR ”inventory model*”

OR "inventory control"). According to Andriolo *et al.* (2014) significant number of papers in sustainable inventory management are published after 2011, hence for this study articles published before 2011 were filtered out. Further articles published in peer-reviewed journals and available in English language are only considered. The above provided string displays 1590 publications carried out in March 2023 whose titles and abstracts were screened further to identify the relevant articles based on the objectives, which were trimmed down to 20 articles. To make the review more comprehensive and inclusion of more papers into the review as through the database only 20 research articles were shortlisted, therefore snowball approach is adopted where references of the 20 articles were extensively explored which helped to include 12 more relevant publications as per previously described criteria. Figure 1 depicts the flowchart of the literature review.

2.2. Analysis phase

2.2.1. Descriptive analysis of selected papers

(a) Distribution of articles based on Journal

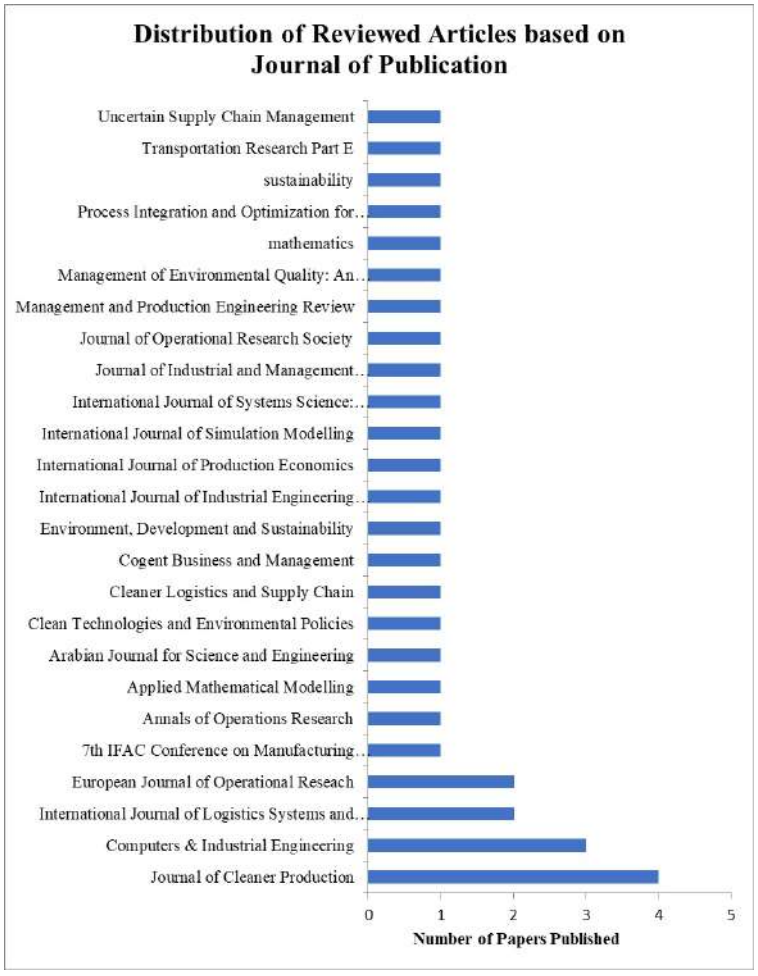


Figure 2: Distribution of reviewed articles based on Journal

From the Figure 2 it can be observed that around 34% of the articles are published in

the journals like 'Journal of Cleaner Production', 'Computers & Industrial Engineering', 'International Journal of Logistics Systems and Management' and 'European Journal of Operational Research'. Other articles are distributed in other journals and one conference proceeding each having one article.

(b) Distribution of articles based on publication year

Figure 3: Distribution of articles based on publication year

From the Figure 3 though there are articles publishing year on year from the inception of the topic, approximately 70% are published during 2018-2023 which depicts the importance of the topic in recent years.

2.2.2. Bibliographic coupling

When two research publications have a common third publication in their reference lists, they are said to be bibliographically coupled. Third common publication is considered as a coupling unit between the two (Kessler, 1963). The bibliographic method helps to create the cluster of articles with common thread between them and further helped to identify the major areas of study.

Figure 4 depicts the bibliographic network for the finally included 30 publications excluding other unrelated publications. The Figure 4 provided below was created using the VOSviewer software. In the following Figure 4 each circle presents the research publication and the line between two publications indicates the number of common references. Publications having higher number of citations are shown by large circles and publications with smaller circles have relatively less number of citations. Bibliographic coupling helps to assigning different publications to different clusters based on important attributes. From the Figure 4 generated using bibliographic technique where 30 publications can be grouped into 4 clusters. These 4 clusters can be summarized as provided below. The 4 different clusters can be identified by observing their colours in the bibliometric network.

Cluster 1: Transportation and perishability inventory models

This is the largest cluster in the network with 12 documents depicted with red colour

in the figure provided below. Primary focus of the articles from this cluster is to develop inventory models considering the environmental considerations but the prominent components included are perishability of products, deteriorating items, costs and emissions related to transportation, carbon emission policies, *etc.*

Cluster 2: Imperfect quality items inventory models

This cluster consists of 7 articles including the earliest article published which is Wahab *et al.*(2011) which tried to provide optimum inventory policies for international supply chain for both vendor and buyer while taking into account imperfect quality of items and impact on the environment. Overall articles from this clusters attempted to incorporate the imperfect products and production processes for inventory management. Articles of this are represented with the blue colour in the bibliometric network.

Cluster 3: Hybrid production process inventory models

The cluster shown with the green colour in the figure, consisting of 7 articles this cluster is dominated by the author Wakhid Jauhari having 4 articles. Articles from this cluster focused on the inclusion of hybrid production processes while considering the range of other scenarios which include imperfect production, multiple retailers, single manufacturer, energy usage, *etc.*

Cluster 4: Demand and lead time uncertainty inventory models

This is the smallest cluster in the network consisting only 4 articles represented by the yellow colour. Digiesi *et al.* (2013a, 2013b) proposed the order quantity models for uncertain demand and lead time where as other two articles Kaur *et al.* (2020) and Kaur and Singh (2018) attempted to assimilate linear programming approach to get the better managerial insights.

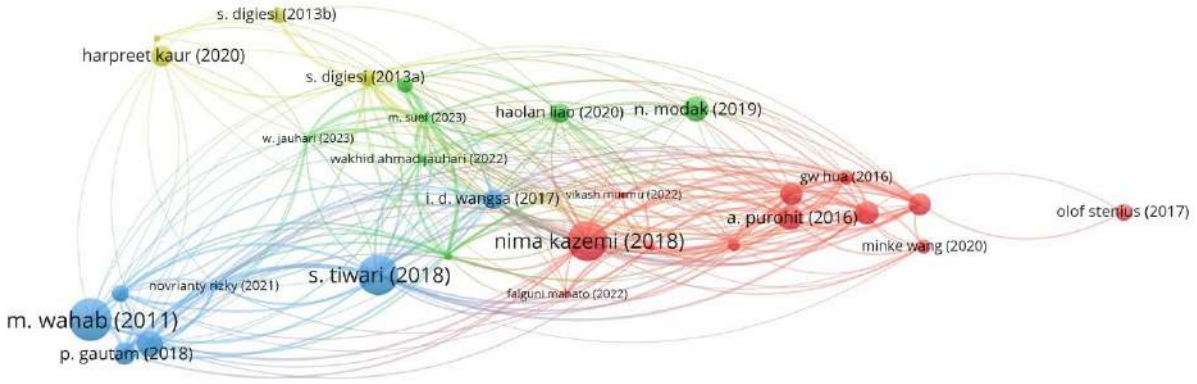


Figure 4: Bibliographic network of selected articles

2.2.3. Classification of articles based on probability distributions of different components

In a review conducted by Tinani and Kandpal (2017) , articles were classified based on two types of uncertainty problems namely yield uncertainty and random yield diversifica-

tion. Using the similar approach reviewed articles can be classified based on the probability distributions for demand being the crucial component and other components of inventory models as provided below.

(a) Probability distributions for demand component

The very first study in the reviewed articles where demand is considered uncertain with some probability distribution *i.e.* Digiesi *et al.* (2013a), where demand for the product is considered a stochastic independent variable which follows normal probability distribution with some mean and standard deviation. Proceeding further it can be observed that Purohit *et al.* (2016) also considered stochastic demand to formulate lot-sizing model using the inverse cumulative distribution function assuming the normal distribution for the demand. Further review of the articles also suggests that the choice of the normal distribution for the demand is more frequent than any other probability distributions. Also other studies such as Ahmad Jauhari (2022), Darma Wangsa (2017), Jauhari (2018), Jauhari *et al.* (2021), Jauhari *et al.* (2023), Jauhari *et al.* (2023), Kaur *et al.* (2020), Kaur and Singh (2018), Liao and Li (2021), Manupati *et al.* (2019), Modak and Kelle (2021), Rizky *et al.* (2021), Suef *et al.* (2023), Tang *et al.* (2018) have all preferred demand to be normally distributed. Interestingly, some studies have also incorporated other scenarios or probability distributions like Ghosh *et al.*(2017) assumed normal distribution jointly for lead time and demand termed as a lead time demand distribution. Further Stenius *et al.* (2018) developed a model where demand follows Poisson distribution whereas Wang *et al.* (2020) assumed that each demand zone has Poisson-normal compound demand during particular time period. Therefore, it can be concluded that the normal distribution is the most popular choice among the authors for formulation of inventory problems.

(b) Probability distributions for other components

Though demand is very decisive component of the inventory modelling, authors are also interested to incorporate uncertainty in the inventory problems by considering them as a random variable with some probability distributions. Among which contemplation of fractions or percentage of defectives or imperfect items is very often. For instance Gautam and Khanna (2018) and Kazemi *et al.* (2018) incorporated rate of defective items in the model without specifying any probability distribution for the same. However, Gautam *et al.* (2019), Mishra and Mishra (2022), Rizky *et al.* (2021) and Tiwari *et al.* (2018) preferred Uniform distribution for the same in the prescribed inventory problem. De-la-Cruz-Márquez *et al.* (2022) initially did not specify any probability distribution for imperfect items percentage at model development stage but for numerical example, author assumed it has Uniform Distribution. It is also worthwhile to notice that many of the studies does not specify any probability distributions while developing the theory. For example Lee *et al.* (2017) embodied unspecified distribution for lead time, Tang *et al.* (2018) for inventory level, Gautam *et al.* (2019) for number of items like scrap, repairable and non-repairable proportion of items, Mahato *et al.* (2023) for time between process to go out of control which shows that this component can also be expanded with experimenting with different probability distributions. Further for inclusion of deterioration concept Hua *et al.* (2016) considered Exponential distribution of time to deterioration where as Murmu *et al.* (2023) considered two parameter Weibull distributed deterioration rate. There are many such other components which have assumed different probability distributions which are summarized in the table.

2.2.4. Detailed analysis of the selected papers

This section aims to provide in depth analysis of selected papers by providing the overview, sustainability factors incorporated, other factors considered for model building with probability distribution and proposed future directions. Limited number of articles in the final inclusion stage of this study allows starting the review and analysis procedure from the earliest article published and going till the recent one. Considering the significance of EOQ models at the international level Wahab *et al.* (2011) developed a model for vendor and buyer when they are located in different countries while incorporating imperfect quality items in terms of percentage of defective items in three different scenarios also rate of exchange between two countries follows stochastic behaviour and to cater the needs of sustainability, carbon emission costs are considered. This model can be further extended as given by author by the inclusion of uncertainty in demand, lead time, credit, *etc.* and other environmental factors such as green packaging, remanufacturing, cleaner production, recycling, *etc.* Digiesi *et al.* (2013a) proposed the sustainable order quantity model when the product demand is uncertain jointly considering logistic costs like safety stock, shortage cost and environmental cost of transportation. Demand at the particular period is considered as independent stochastic variable with equal expected demand and standard deviation. The model building started with stating cost function adding associated costs. Further the loss factor component was considered which was only related to the technological advancement related to the energy sector in transportation. Lastly the cost function was optimised to obtain the order quantity levels and optimal safety stock. Proposed model then applied to automotive case study to obtain the insights on optimal solution.

On the similar lines Digiesi *et al.*(2013b) developed another sustainable order quantity models when lead time is uncertain and demand is deterministic in nature following the same approach as the previous article including logistic and environmental transportation cost and developed model applied to real industrial case. Jauhari *et al.* (2014) arrived at a model for vendor and buyer by integrating defective items produced in the production process and unequal size of the shipment with the environmental carbon emission cost for both vendor and buyer. The papers considered the probability distribution of defect rate and further model building was formulated. They further analyzed that if the probability of defects increases then it also increases the carbon emission cost. Authors also suggested that the proposed model can be further extended by incorporating of defective raw material, inspection and rework process for raw material, inspection error and application of other distribution models of defective rate on vendor-buyer problem.

Purohit *et al.* (2016) studied the lot sizing inventory problem using mixed integer linear programming approach with constraints on emission and service levels when demand from the buyer is uncertain and dynamic which is normally distributed. This model considered wide ranging emissions generated during ordering, storage and purchasing and their corresponding costs. The objective function for the proposed problem is to minimise the total cost for the prescribed time period consisting of the four components. Additionally various constraints such as cycle service level, identification of the optimum replenishment schedule and emission constraints are incorporated. This study analysed the impact of various emission factors and features related to product and system under the carbon cap-and-trade policy assuming constant carbon price. This study can be extended by considering variable carbon price, applying to real life cases and different supply structures. A detailed overview

of the remaining articles is summarized in the Table 1 provided below mentioning the authors, name of the publishing journal, probabilistic components used, sustainability factors and proposed future research.

3. Conclusion and future research direction

This paper attempts to provide the comprehensive review of the inventory models which included sustainability criteria where probability distributions are taken into consideration for at least one of the components of inventory management. This study tries to provide the detailed analysis of selected papers with sustainability criteria included, incorporated probability distributions and possible future prospects proposed by the reviewed articles. It can be observed that carbon emissions due to transportation and costs associated with it are the frequent components enclosed in the articles reviewed. Other emission factors such as emissions due to production and storage are also the prominent one. Therefore incorporation of investments made to reduce emissions from the transportation, production and storage could be exhaustive topic to move towards sustainability. Also it would be great extension to consider the other carbon regulations more than carbon tax like cross border adjustment mechanism, carbon penalty, *etc.* Another important concept identified that is incorporation of uncertainty in demand where Normal distribution is the very much preferred distribution. As normal distribution has some limitations such as it is symmetric in nature and assumes negative values as well. Hence consideration of other possible probability distributions for demand as well as other components of the inventory model can provide enormous and varied opportunities for the extension of the sustainable inventory models.

4. Limitations of study

This review can be extended by the considering the papers from the other databases like Scopus and Google Scholar. Major focus of this research paper is on the order quantity models and relevant studies with few papers with production inventory models. Hence the present study can be explored with inclusion of more production problems and other supply chain scenario models. This study does not focus much on the various quantitative and qualitative methods employed in the prescribed paper. Though the database was thoroughly examined there might be some possibility that some articles may have slipped and further excluded from the process. Inclusion of such articles will help to broaden all the horizons of study.

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.

Table 1: Summary of the articles

Authors	Journal Name	Sustainability Factors	Inventory components with their Probability Distributions	Proposed Future research work
Hua *et al.* (2016)	International Journal of Simulation Modelling	Carbon cap and trade mechanism, emission cost	Time to deterioration – Exponential Distribution	Incorporation of partial backlogging, lost sales, freshness having deterioration rate
Darma Wangsa (2017)	International Journal of Industrial Engineering Computations	Direct and indirect emissions from transport and industry sector, carbon emission tax	Demand – Normal Distribution	Consideration of multi-manufacturer and multi-buyer system, other indirect emissions like waste disposal, cleaner production. Green manufacturing, remanufacturing, recycling, emission reduction investment costs
Ghosh *et al.* (2017)	Applied Mathematical Modelling	Carbon emissions from production, inventory, transportation	Lead time demand – Normal Distribution	Model can be extended to multi-echelon or reverse supply chain with defective and waste item, inclusion of perishable products,
Lee *et al.* (2017)	Sustainability	Carbon emission cost of warehouse, collection and disposal of inventory waste	Lead time – Unspecified probability distribution	Extension can be done considering processing time at custom, waiting time at border and terminal handling activities
Kazemi *et al.* (2018)	International Journal of Systems Science: Operations & Logistics	Carbon emission cost and tax of holding inventory, warehousing, obsolete items	Fraction of imperfect items - Unspecified	Consideration of multiple manufacturer and supplier scenario, emissions from transportations or energy usage, imperfect supply process
Stenius *et al.* (2018)	European Journal of Operational Research	Emission cost of transportation	Demand – Poisson Distribution Shipment quantity to each retailer group – Binomial Distribution	Generalization of model by considering other demand distributions like compound Poisson
Tang *et al.* (2018)	European Journal of Operational Research	Emissions from storage and transportation	Demand – Normal Distribution Inventory level – Unspecified Distribution	Carbon reduction consideration, other sources of operational emissions

Kaur and Singh (2018)	Management of Environmental Quality: An International Journal	Carbon emission and its price and quota	Demand, Supplier capacity, Carrier capacity – Normal Distribution	Study can be extended by considering uncertain lead time, purchasing and transportation cost, other qualitative factors
Jauhari (2018)	International Journal of Logistics Systems and Management	Carbon emission cost due to transportation and production	Demand – Normal Distribution	Inclusion of inspection process, consideration of different supply chain structures, periodic review policy
Gautam and Khanna (2018)	Uncertain Supply Chain Management	Fixed and variable cost of carbon emissions at vendor	Defective rate – Unspecified Distribution	Further extended for the case of multiple buyer and items, also considering partial backordering, pricing discounts, inspection errors
Tiwari *et al.* (2018)	Journal of Cleaner Production	Costs of carbon emissions from transportation, warehousing and holding deteriorating items, emission tax	Percentage of Defective items – Uniform Distribution	Impact of rework and recycle activities on carbon emission, Multi-product and delay in payments can be very good extension
Manupati *et al.* (2019)	Computers & Industrial Engineering	Carbon cap-and-trade, carbon tax	Demand – Normal Distribution	Extension can be done for reverse closed loop supply chain, some model restrictions can be relaxed
Gautam *et al.* (2019)	Journal of Cleaner Production	Costs of carbon emission	Repairable proportion, non-repairable proportion and scrap proportion – Unspecified Distribution Defect Percentage – Uniform Distribution	Inspection error can be incorporated, also some permissible delays can also considered. Model can be extended in fuzzy environment
Kaur *et al.* (2020)	Computers & Industrial Engineering	Carbon emissions during ordering, transportation and holding	Demand, Machine capacity, carrier capacity, supplier capacity - Normally distributed	Flexibility of supplier and carrier selection, fuzzy environments can be applied

Wang *et al.* (2020)	Transportation Research Part - E	Carbon emissions due to transportation, carbon cap	Demand zone – Poisson-Normal compound demand Random order inter arrival time – Exponential Distribution Random order sizes – Normal Distribution Carbon price – Uniform Distribution	Consideration of vehicle routing decisions, other carbon regulation schemes, more efficient heuristics
Modak and Kelle (2021)	Journal of Operational Research Society	Carbon footprint, carbon tax, Emission reduction due to recycling	Demand – Normal Distribution	Can be extended to multiple products, manufacturers and retailers case, also internet marketing, recycling activity influence can be incorporated
Liao and Li (2021)	Computers & Industrial Engineering	Carbon emissions in logistics and storage	Demand – Normal Distribution, Exponential Distribution	Study in hybrid manufacturing system, forward and reverse production system simultaneous consideration
Rizky *et al.* (2021)	Clean Technologies and Environmental Policies	Carbon emission costs, energy consumption	Demand – Normal Distribution Defective Percentage – Uniform Distribution	Model can be extended by considering imperfect raw material and impact of returned product
Jauhari *et al.* (2021)	Journal of Cleaner Production	Carbon emission due to storage, production, transportation, carbon tax	Demand – Normal Distribution	Extension can be done by considering inspection errors, variable production rate, more parties like supplier and distributors
Ahmad Jauhari (2022)	Cleaner Logistics and Supply Chain	Carbon emissions due to production, transportation and storage, investment in green technology	Demand – Normal Distribution	Inclusion of imperfect reworking process and investment in quality, other carbon regulations like carbon penalty, three party logistic supply chain

Mishra and Mishra (2022)	Arabian Journal for Science and Engineering	Carbon emissions and its cost from electricity generation, deterioration, fuel consumption by vehicle, energy consumption from Warehouse	Defective Percentage – Uniform Distribution	Consideration of imperfect screening with errors, demand variations like stochastic and fuzzy, can also be extended for manufacturer
De-la-Cruz-Márquez *et al.* (2022)	Mathematics	Carbon emissions and costs, carbon tax	Percentage of imperfect items – Unspecified in the model (Uniform Distribution in Numerical Example)	Model can be extended by incorporating investment in preservation technology to reduce deterioration
Mahato *et al.* (2023)	Environment, Development and Sustainability	Pollution control costs and scenarios	Time elapsed after which production becomes out of control – Unspecified Distribution in the model (Exponential Distribution in Numerical example)	Model can be extended by considering stochastic demand and default risk rate, partial backlogging, investment in low carbon technologies, transportation costs, screening errors, recycling, inflation
Murmu *et al.* (2023)	Journal of Industrial and Management Optimization	Carbon emission, emission cap, carbon tax, investment in green technology	Deterioration rate – Two parameter Weibull Distribution	Study can be expanded by considering three parameter weibull distributed deterioration rate, stochastic demand, non linear programming approach for fuzzy environment, trade credit, manufacturing process reliability
Jauhari *et al.* (2023)	Annals of Operations Research	Carbon emissions from transportation, storage, investment in green technology, carbon tax	Demand – Normal Distribution	Incorporation of human errors in inspection, other carbon reduction policies like carbon penalty, carbon cap and trade

Suef *et al.* (2023)	Process Integration and Optimization for Sustainability	Carbon emission from storage, production and transportation, investment in green technology, carbon tax	Demand – Normal Distribution	Investigation of influence of routes of transportation on emissions and costs, consideration of imperfect production and green transporters
Jauhari *et al.* (2023)	Cogent Business and Management	Carbon emissions, green investment and incentives, energy consumptions	Demand – Normal Distribution	Model can be extended by considering imperfect production process, other carbon policies like cap and trade, carbon offset, carbon cap

References

Ahmad Jauhari, W. (2022). Sustainable inventory management for a closed-loop supply chain with energy usage, imperfect production, and green investment. *Cleaner Logistics and Supply Chain*, **4**, 100055.

Andriolo, A., Battini, D., Grubbström, R. W., Persona, A., and Sgarbossa, F. (2014). A century of evolution from Harris's basic lot size model: Survey and research agenda. *International Journal of Production Economics*, **155**, 16–38.

Becerra, P., Mula, J., and Sanchis, R. (2021). Green supply chain quantitative models for sustainable inventory management: A review. *Journal of Cleaner Production*, **328**, 129544.

Becerra, P., Mula, J., and Sanchis, R. (2022). Sustainable Inventory Management in Supply Chains: Trends and Further Research. *Sustainability*, **14(5)**, 2613.

Burgin, T. A. (1972). Inventory control with normal demand and gamma lead times. *Operational Research Quarterly (1970-1977)*, **23(1)**, 73.

Burgin, T. A. and Wild, A. R. (1967). Stock control-experience and usable theory. *OR*, **18(1)**, 35.

Darma Wangsa, I. (2017). Greenhouse gas penalty and incentive policies for a joint economic lot size model with industrial and transport emissions. *International Journal of Industrial Engineering Computations*, **8(4)**, 453–480.

De-la-Cruz-Márquez, C. G., Cárdenas-Barrón, L. E., Mandal, B., Smith, N. R., Bourguet-Díaz, R. E., Loera-Hernández, I. D. J., Céspedes-Mota, A., and Treviño-Garza, G. (2022). An inventory model in a three-echelon supply chain for growing items with imperfect quality, mortality, and shortages under carbon emissions when the demand is price sensitive. *Mathematics*, **10(24)**, 4684.

Digiesi, S., Mossa, G., and Mummolo, G. (2013a). A sustainable order quantity model under uncertain product demand. *IFAC Proceedings Volumes*, **46(9)**, 664–669.

Digiesi, S., Mossa, G., and Mummolo, G. (2013b). Supply lead time uncertainty in a sustainable order quantity inventory model. *Management and Production Engineering Review*, **4(4)**, 15–27.

Gautam, P. and Khanna, A. (2018). An imperfect production inventory model with setup cost reduction and carbon emission for an integrated supply chain. *Uncertain Supply Chain Management*, **6(3)**, 271–286.

Gautam, P., Kishore, A., Khanna, A., and Jaggi, C. K. (2019). Strategic defect management for a sustainable green supply chain. *Journal of Cleaner Production*, **233**, 226–241.

Ghosh, A., Jha, J. K., and Sarmah, S. P. (2017). Optimal lot-sizing under strict carbon cap policy considering stochastic demand. *Applied Mathematical Modelling*, **44**, 688–704.

Hua, G. C., T. C. E., Zhang, Yi, Zhang, Juliang, and Wang, S. Y. (2016). Carbon-constrained perishable inventory management with freshness-dependent demand. *International Journal of Simulation Modelling*, **15(3)**, 542–552.

Jauhari, W. A. (2018). A collaborative inventory model for vendor-buyer system with stochastic demand, defective items and carbon emission cost. *International Journal of Logistics Systems and Management*, **29(2)**, 241.

Jauhari, W. A., Pamuji, A. S., and Rosyidi, C. N. (2014). Cooperative inventory model for vendor-buyer system with unequal-sized shipment, defective items and carbon emission cost. *International Journal of Logistics Systems and Management*, **19(2)**, 163.

Jauhari, W. A., Pujawan, I. N., and Suef, M. (2021). A closed-loop supply chain inventory model with stochastic demand, hybrid production, carbon emissions, and take-back incentives. *Journal of Cleaner Production*, **320**, 128835.

Jauhari, W. A., Pujawan, I. N., and Suef, M. (2023). Sustainable inventory management with hybrid production system and investment to reduce defects. *Annals of Operations Research*, **324(1–2)**, 543–572.

Jauhari, W. A., Wangsa, I. D., Hishamuddin, H., and Rizky, N. (2023). A sustainable vendor-buyer inventory model with incentives, green investment and energy usage under stochastic demand. *Cogent Business & Management*, **10(1)**, 2158609.

Kaur, H. and Singh, S. P. (2018). Environmentally sustainable stochastic procurement model. *Management of Environmental Quality: An International Journal*, **29(3)**, 472–498.

Kaur, H., Singh, S. P., Garza-Reyes, J. A., and Mishra, N. (2020). Sustainable stochastic production and procurement problem for resilient supply chain. *Computers & Industrial Engineering*, **139**, 105560.

Kazemi, N., Abdul-Rashid, S. H., Ghazilla, R. A. R., Shekarian, E., and Zanoni, S. (2018). Economic order quantity models for items with imperfect quality and emission considerations. *International Journal of Systems Science: Operations & Logistics*, **5(2)**, 99–115.

Kessler, M. M. (1963). Bibliographic coupling between scientific papers. *American Documentation*, **14(1)**, 10–25.

Lee, S.-K., Yoo, S., and Cheong, T. (2017). Sustainable EOQ under lead-time uncertainty and multi-modal transport. *Sustainability*, **9(3)**, 476.

Lee, W.-C., Wu, J.-W., and Lei, C.-L. (2007). Optimal inventory policy involving back-order discounts and variable lead time demand. *The International Journal of Advanced Manufacturing Technology*, **34(9–10)**, 958–967.

Liao, H. and Li, L. (2021). Environmental sustainability EOQ model for closed-loop supply chain under market uncertainty: A case study of printer remanufacturing. *Computers & Industrial Engineering*, **151**, 106525.

Mahato, F. M., Mahata, C, and Mahata, G. C. (2023). Sustainable optimal production policies for an imperfect production system with trade credit under different carbon emission regulations. *Environment, Development and Sustainability*, **25**, 10073–10099 (2023).

Manupati, V. K., Jedidah, S. J., Gupta, S., Bhandari, A., and Ramkumar, M. (2019). Optimization of a multi-echelon sustainable production-distribution supply chain system with lead time consideration under carbon emission policies. *Computers & Industrial Engineering*, **135**, 1312–1323.

Mishra, R. K. and Mishra, V. K. (2022). An optimum sustainable inventory model for non-instantaneous deterioration and quality assessment under carbon emissions and complete backordering shortage. *Arabian Journal for Science and Engineering*, **47(3)**, 3929–3944.

Modak, N. M. and Kelle, P. (2021). Using social work donation as a tool of corporate social responsibility in a closed-loop supply chain considering carbon emissions tax and demand uncertainty. *Journal of the Operational Research Society*, **72(1)**, 61–77.

Murmu, V., Kumar, D., Sarkar, B., Mor, R. S., and Jha, A. K. (2023). Sustainable inventory management based on environmental policies for the perishable products under first or last in and first out policy. *Journal of Industrial and Management Optimization*, **19(7)**, 4764–4803.

Pattnaik, S., Nayak, M. M., Abbate, S., and Centobelli, P. (2021). Recent trends in sustainable inventory models: A literature review. *Sustainability*, **13(21)**, 11756.

Purohit, A. Kr., Shankar, R., Dey, P. K., and Choudhary, A. (2016). Non-stationary stochastic inventory lot-sizing with emission and service level constraints in a carbon cap-and-trade system. *Journal of Cleaner Production*, **113**, 654–661.

Rizky, N., Wangsa, I. D., Jauhari, W. A., and Wee, H. M. (2021). Managing a sustainable integrated inventory model for imperfect production process with type one and type two errors. *Clean Technologies and Environmental Policy*, **23(9)**, 2697–2712.

Stenius, O., Marklund, J., and Axsäter, S. (2018). Sustainable multi-echelon inventory control with shipment consolidation and volume dependent freight costs. *European Journal of Operational Research*, **267(3)**, 904–916.

Suef, M., Jauhari, W. A., Pujawan, I. N., and Dwicahyani, A. R. (2023). Investigating carbon emissions in a single-manufacturer multi-retailer system with stochastic demand and hybrid production facilities. *Process Integration and Optimization for Sustainability*, **7**, 743–764 (2023).

Tang, S., Wang, W., Cho, S., and Yan, H. (2018). Reducing emissions in transportation and inventory management: (R, Q) Policy with considerations of carbon reduction. *European Journal of Operational Research*, **269(1)**, 327–340.

Tinani, K. S. and Kandpal, D. H. (2017). Literature review on supply uncertainty problems: yield uncertainty and supply disruption. *Journal of the Indian Society for Probability and Statistics*, **18(2)**, 89–109.

Tiwari, S., Daryanto, Y., and Wee, H. M. (2018). Sustainable inventory management with deteriorating and imperfect quality items considering carbon emission. *Journal of Cleaner Production*, **192**, 281–292.

Wahab, M. I. M., Mamun, S. M. H., and Ongkunaruk, P. (2011). EOQ models for a coordinated two-level international supply chain considering imperfect items and environmental impact. *International Journal of Production Economics*, **134(1)**, 151–158.

Wang, M., Wu, J., Kafa, N., and Klibi, W. (2020). Carbon emission-compliance green location-inventory problem with demand and carbon price uncertainties. *Transportation Research Part E: Logistics and Transportation Review*, **142**, 102038.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 22, No. 2, 2024 (New Series), pp 285–305
http://www.ssca.org.in/journal

Extropy Properties of Ranked Set Sample for Sarmanov Family of Distributions

Manoj Chacko[1] **and Varghese George**[1,2]
[1]*Department of Statistics, University of Kerala, Trivandrum, Kerala 695581, India*
[2]*St. Stephen's College, Maloor College P. O.,Pathanapuram, Kerala 689695, India*

Received: 09 May 2023; Revised: 11 November 2023; Accepted: 25 February 2024

Abstract

In this paper, the extropy of ranked set sample from Sarmanov family of distributions is considered. By deriving the expression for extropy of concomitants of order statistics, the expression for extropy of ranked set sample of the study variable Y in which an auxiliary variable X is used to rank the units in each set, under the assumption that (X, Y) follows Sarmanov family of distributions is obtained.

Key words: Ranked set sampling; Sarmanov family of distributions; Concomitants of order statistics; Extropy.

AMS Subject Classifications: 62B10, 94A20, 62D05

1. Introduction

Let (X, Y) be a random vector with joint probability density function (PDF) $f(x, y)$ and cumulative distribution function (CDF) $F(x, y)$. Let $f_X(x)$ and $f_Y(y)$ be the marginal PDFs and $F_X(x)$ and $F_Y(y)$ be the marginal CDFs of X and Y respectively. Let (X_i, Y_i), $i = 1, 2, ...n$ be a random sample of size n from the population with cdf $F(x, y)$. If these observations are arrange in increasing order of magnitude based on X_i's, then the rth largest observation $X_{r:n}$ is the rth order statistic of X_i's. Then the Y variable associated with $X_{r:n}$ is called concomitant of rth order statistic and it is denoted by $Y_{[r:n]}$. David (1973) introduced the concept of concomitants of statistics which is applicable in various areas like ranked set sampling, double sampling, correlation analysis and in certain selection procedures. More details on this idea was given in David and Nagaraja (1998).

McIntyre (1952) introduced an efficient sampling scheme named ranked set sampling, as an alternative to simple random sampling (see, Chen *et al.* (2004)). The procedure of ranked set sampling is as follows. Select n^2 units randomly from the population. These units are randomly alloted into n sets, each of size n. Then the units in each set are ranked visually, judgement method or using some inexpensive methods. From the first set of n units, choose the unit which has the lowest rank for actual measurement. From the second set of n units the unit ranked second lowest is chosen. The process is continued until choose

Corresponding Author: Varghese George
Email: varghesessc@gmail.com

the unit which has the highest rank in the nth set. Then make measurement on variable of interest of the selected units, which constitute the ranked set sample(RSS).

Ranked set sampling as described in McIntyre (1952) is applicable whenever sample size is small and ranking of a set of sampling units can be done easily by a judgment method. Suppose the variable of interest, say Y, is expensive to measure and difficult to rank the units. In this case as an alternative method, Stokes (1977) modified the method by using an auxiliary variable for ranking the sampling units in each set. Stokes (1977) explained the ranked set sampling procedure as follows. Choose n^2 units randomly from a bivariate population. Arrange these units into n sets, each of size n and measure the auxiliary variable X. In the first set, that unit for which smallest measurment on the auxiliary variable X is chosen and take the measurement of the study variable Y, denoted by $Y_{[1]}$. In the second set, that unit for which second smallest measurement on the auxiliary variable X is chosen and take the measurement of the study variable Y, denoted by $Y_{[2]}$. Finally, in the nth set, that unit for which largest measurment on the auxiliary variable X is chosen and take the measurement of the study variable Y, denoted by $Y_{[n]}$. Clearly $Y_{[r]}$, $r = 1, 2, ...n$ are concomitants of order statistics of the given random sample and are independent.

Bain (2017) give an example for the application of RSS as proposed by Stokes (1977). Here the study variable Y represents the oil pollution of sea water and auxiliary variable X represents the tar deposit in the nearby sea shore. Clearly collecting sea water sample and measuring the oil pollution in it is difficult and costly. However the prevalence of pollution in sea water is much reflected by the tar deposit in the surrounding terminal sea shore. In this example ranking the pollution level of sea water based on the tar deposit in the sea shore is more natural and scientific than ranking it visually or by judgement method. Applying the concepts of concomitant of order statistics in ranked set sampling, Chacko and Thomas (2007, 2008, 2009), Chacko (2017) and Mehta (2022) estimated the parameters of different distributions belonging to Morgenstern family of distributions.

As an alternative to entropy defined by Shannon (1948), Lad *et al.* (2015) introduced a new measure of uncertainty called extropy . Let X be a random variable with PDF $f_X(x)$ and CDF $F_X(x)$. Then the extropy of X is defined as

$$J(X) = \frac{-1}{2}\int_{-\infty}^{\infty}(f_X(x))^2 dx \tag{1}$$

$$= \frac{-1}{2}\int_{0}^{1} f_X(F^{-1}(u))du, \tag{2}$$

where $F^{-1}(u) = inf\{x; F_X(x) \geq u\}, u \in [0, 1]$ is the quantile function of $F_X(x)$.

Lad *et al.* (2015) gave some properties and applications of extropy measure. Qiu (2017) discussed the characterization results, monotone properties, and lower bounds of extropy of order statistics and record values. Zamanzade and Mahdizadeh (2019) discussed the nonparametric estimation of extropy based on ranked set sampling. Eftekharian and Qiu (2022) considered the information content of statified ranked set sampling in terms of extropy. Qiu and Raqab (2022) discussed the properties of weighted extropy using Ranked Set Samples.

Morgenstern (1956)introduced a bivariate family of distributions which can be con-

structed with specific marginal distributions and the PDF is given by

$$f(x,y) = f_X(x)f_Y(y)[1+\delta(2F_X(x)-1)(2F_Y(y)-1)], -1 \le \delta \le 1,$$

where δ is the association parameter, $f_X(x)$ and $f_Y(y)$ are the marginal PDFs and $F_X(x)$ and $F_Y(y)$ are the marginal CDFs of X and Y respectively. One of the important limitations of the Morgenstern family of distributions (MFD) is that the correlation coefficient lies between -1/3 and 1/3. Several authors have modified the MFD to enhance the range of correlation and extended the domain of applications. One of the important modifications of MFD was given by Sarmanov (1966) in the sense that it provides the best improvement in correlation level with only one parameter as in the MFD. The PDF of family of distributions of Sarmanov (1966) is given by

$$\begin{aligned} f(x,y) = f_X(x)f_Y(y)\Big[1 &+ 3\alpha(2F_X(x)-1)(2F_Y(y)-1) \\ &+ \frac{5}{4}\alpha^2\Big(3(2F_X(x)-1)^2-1\Big)\Big(3(2F_Y(y)-1)^2-1\Big)\Big], |\alpha| \le \frac{\sqrt{7}}{5} \end{aligned} \tag{3}$$

where α is the association parameter. When the marginal distributions follow uniform, the distribution attain its maximum correlation coefficient, α.

Alemany *et al.* (2020) give an example for application for Sarmanov family of distributions given in (3). Here the study variable Y follows the average claim cost per insured and X represents the number of claims of individual. This model can be used to obtain the distribution of the total cost of claims based on the collective model, for a policyholder with specific characteristics. If the profiles have larger dependency, the Sarmanov distribution can be used to fit a non- linear dependence between frequency and severity (cost random variable). The different applications of Sarmanov family of distributions are given in Abdallah *et al.* (2016) and Bolancé *et al.* (2020). Barakat *et al.* (2022) discussed the properties of concomitants of order statistics of Sarmanov family of distributions.

It is well known that ranked set sample provides more information than simple random sample(SRS) of the same size about the unknown parameters of the underlying distribution in parametric inferences (see, Chen *et al.* (2004)). Jozani and Ahmadi (2014) explained the concept of information content of RSS data and compared them with their counterparts in SRS data. Raqab and Qiu (2019) described the monotone properties and stochastic orders of ranked set sample and compared the results with their counterpart under SRS design. Husseiny *et al.* (2022) discussed information measures in records and their concomitants arising from Sarmanov family of distributions. Chacko and George (2024, 2023) discussed the extropy properties of RSS for MFD and Cambanis type bivariate distributions. George and Chacko (2023) considered the cumulative residual extropy properties of ranked set samples for Cambanis type bivariate distributions.

In this paper, we derive the extropy of concomitant of order statistic $Y_{[r:n]}$ of a random sample of size n from Sarmanov family of distributions. Since observations of a ranked set sample, in which an auxiliary variable X is used to rank the units in each set, are nothing but concomitant of order statistics, we derive the extropy of RSS when (X,Y) follows Sarmanov family of distributions. The properties and bounds for extropy of RSS are also derived. We also consider the joint extropy of $(X_{RSS}, Y_{[RSS]})$, where $X_{RSS} = \Big(X_{(1)}, X_{(2)}, ..., X_{(n)}\Big)$ is the

RSS of X observations in which ranking in each unit is perfect and $Y_{[RSS]} = \Big(Y_{[1]}, Y_{[2]}, ..., Y_{[n]}\Big)$ is the RSS of Y observations in which ranking in each unit is based on X observations.

The paper is organized as follows. In section 2, the expression for extropy of $Y_{[r:n]}$ and also obtain upper and lower bounds of it. In section 3, we obtain the extropy of the RSS arising from Sarmanov family of distributions and study its properties. Section 4 devotes to obtain extropy of $(X_{r:n}, Y_{[r:n]})$ and thereby obtain the extropy of $(X_{RSS}, Y_{[RSS]})$, where $X_{RSS} = \Big(X_{(1)}, X_{(2)}, ..., X_{(n)}\Big)$ is the ranked set sampling based on X observations in which ranking in each unit is perfect and $Y_{[RSS]} = \Big(Y_{[1]}, Y_{[2]}, ..., Y_{[n]}\Big)$. Finally, in section 5 we give the conclusion.

2. Extropy of concomitant of rth order statistic

Let $Y_{[r:n]}$ $r = 1, 2, ...n$ be the concomitant of rth order statistic of a bivariate random sample arising from Sarmanov family of distributions. If $f_{r:n}(x)$ is the pdf of rth order statistic and $f_{Y|X}(y/x)$ is the conditional pdf of Y given X, then the pdf of concomitant of rth order statistic, $Y_{[r:n]}$ is

$$\begin{aligned} f_{Y[r:n]}(y) &= \int_{-\infty}^{\infty} f_{Y|X}(y/x) f_{r:n}(x) dx \\ &= \int_{-\infty}^{\infty} f_Y(y) \Big[1 + 3\alpha(2F_X(x) - 1)(2F_Y(y) - 1) \\ &\quad + \frac{5}{4}\alpha^2 \Big(3(2F_X(x) - 1)^2 - 1\Big)\Big(3(2F_Y(y) - 1)^2 - 1\Big)\Big] \\ &\quad \times \frac{n!}{(r-1)!(n-r)!} (F_X(x))^{r-1} (1 - F_X(x))^{n-r} dx \\ &= f_Y(y) \left[1 + d_1(2F_Y(y) - 1) + d_2\Big(3(2F_Y(y) - 1)^2 - 1\Big)\right], \end{aligned} \tag{4}$$

where

$$d_1 = 3\alpha \frac{2r - n - 1}{n + 1} \tag{5}$$

and

$$d_2 = \frac{5}{2}\alpha^2 \Big(1 - \frac{6r(n - r + 1)}{(n + 1)(n + 2)}\Big). \tag{6}$$

Then by using (1) the extropy of $Y_{[r:n]}$ is given by

$$\begin{aligned} J(Y_{[r:n]}) &= \frac{-1}{2}\int_y (f_{Y[r:n]}(y))^2 dy \\ &= \frac{-1}{2}\int_y (f_Y(y))^2 \left[1 + d_1(2F_Y(y)-1) + d_2\Big(3(2F_Y(y)-1)^2 - 1\Big)\right]^2 dy \\ &= \frac{-1}{2}\int_{u=0}^{1} f_Y(F^{-1}(u)) \left[1 + d_1(2u-1) + d_2\Big(3(2u-1)^2 - 1\Big)\right]^2 du \\ &= \frac{-1}{2}\int_{u=0}^{1} f_Y(F^{-1}(u)) \Big(\rho_{(r,n,\alpha)}(u)\Big)^2 du, \end{aligned} \tag{7}$$

where

$$\rho_{(r,n,\alpha)}(u) = 1 + d_1(2u-1) + d_2\Big(3(2u-1)^2 - 1\Big). \tag{8}$$

Theorem 1: Let $Y_{[r:n]}$ be the concomitant of rth order statistic of a random sample of size n arising from Sarmanov family of distributions, then the extropy of $Y_{[r:n]}$ can be written as

$$J(Y_{[r:n]}) = \frac{-1}{2}\sum_{k=0}^{4} \frac{a_k}{k+1} E\Big(F^{-1}(U_k)\Big), \tag{9}$$

where $a_0 = (1 - d_1 + 2d_2)^2$, $a_1 = 2(1 - d_1 + 2d_2)(2d_1 - 12d_2)$, $a_2 = (2d_1 - 12d_2)^2 + 24d_2(1 - d_1 + 2d_2)$, $a_3 = 24d_2(2d_1 - 12d_2)$, $a_4 = 144d_2^2$ and

$$E\Big(F^{-1}(U_k)\Big) = \int_0^1 (k+1)u^k f_Y(F^{-1}(u))du$$

with U_k follows Beta $(k+1, 1)$.

Proof: Since $Y_{[r:n]}$ is the concomitant of rth order statistic of a random sample of size n arising from Sarmanov family of distributions, we have

$$\begin{aligned} (f_{Y[r:n]}(y))^2 &= (f_Y(y))^2 \left[1 + d_1(2F_Y(y)-1) + d_2\Big(3(2F_Y(y)-1)^2 - 1\Big)\right]^2 \\ &= (f_Y(y))^2 \sum_{k=0}^{4} a_k \Big(F_Y(y)\Big)^k, \end{aligned}$$

where $a_0 = (1 - d_1 + 2d_2)^2$, $a_1 = 2(1 - d_1 + 2d_2)(2d_1 - 12d_2)$, $a_2 = (2d_1 - 12d_2)^2 + 24d_2(1 - d_1 + 2d_2)$, $a_3 = 24d_2(2d_1 - 12d_2)$ and $a_4 = 144d_2^2$.

Therefore, the extropy of $Y_{[r:n]}$ is given by

$$\begin{aligned} J(Y_{[r:n]}) &= \frac{-1}{2}\int (f_{Y[r:n]}(y))^2 dy \\ &= \frac{-1}{2}\int (f_Y(y))^2 \sum_{k=0}^{4} a_k \Big(F_Y(y)\Big)^k dy \\ &= \frac{-1}{2}\sum_{k=0}^{4} a_k \int_0^1 u^k f_Y(F^{-1}(u)) du \\ &= \frac{-1}{2}\sum_{k=0}^{4} \frac{a_k}{k+1} E\Big(F^{-1}(U_k)\Big), \end{aligned}$$

where U_k follows Beta $(k+1, 1)$. Hence the theorem. □

Remark 1: If $r = 1$ and $r = n$ in (4), we get the concomitant of first order statistic and largest order statistic of a random sample of size n. Then the extropy of concomitant of first order statistic $Y_{[1:n]}$ and concomitant of largest order statistic $Y_{[n:n]}$ are given by

$$J(Y_{[1:n]}) = \frac{-1}{2}\sum_{k=0}^{4} \frac{a_k^{(1)}}{k+1} E\Big(F^{-1}(U_k)\Big),$$

where $a_0^{(1)} = (1+q_1+2q_2)^2$, $a_1^{(1)} = -2(1+q_1+2q_2)(2q_1+12q_2)$, $a_2^{(1)} = (2q_1+12q_2)^2 + 24q_2(1+q_1+2q_2)$, $a_3^{(1)} = -24q_2(2q_1+12q_2)$ and $a_4^{(1)} = 144q_2^2$
and

$$J(Y_{[n:n]}) = \frac{-1}{2}\sum_{k=0}^{4} \frac{a_k^{(n)}}{k+1} E\Big(F^{-1}(U_k)\Big),$$

where $a_0^{(n)} = (1-q_1+2q_2)^2$, $a_1^{(n)} = 2(1-q_1+2q_2)(2q_1-12q_2)$, $a_2^{(n)} = (2q_1-12q_2)^2 + 24q_2(1-q_1+2q_2)$, $a_3^{(n)} = 24q_2(2q_1-12q_2)$ and $a_4^{(n)} = 144q_2^2$
with $q_1 = 3\alpha\frac{n-1}{n+1}$ and $q_2 = \frac{5}{2}\alpha\Big(1-\frac{6n}{(n+1)(n+2)}\Big)$.

Remark 2: If $\alpha = 0$, that is X and Y are independent, then $d_1 = 0$ and $d_2 = 0$ and hence $J(Y_{[r:n]}) = \frac{-1}{2}E\Big(F^{-1}(U_0)\Big) = J(Y)$.

Corollary 1: Let (X_i, Y_i), $i = 1, 2, ...n$ be a bivariate sample of size n arising from Sarmanov family of distributions. Then the extropy of concomitant of rth order statistic for $\alpha > 0$ is same as the extropy of concomitant of $(n-r+1)$th order statistic for $\alpha < 0$.

Proof: Let $J^{(\alpha)}(Y_{[r:n]})$ be the extropy of concomitant of rth order statistic for any α. We have by (5) and (6), $d_{1(n,\alpha)} = d_{1(n-r+1,-\alpha)}$ and $d_{2(n,\alpha)} = d_{2(n-r+1,-\alpha)}$. Therefore by (9),

$$J^{(\alpha)}(Y_{[r:n]}) = J^{(-\alpha)}(Y_{[n-r+1:n]}).$$

□

Example 1: If (X, Y) follows Sarmanov family of distributions given in (3) with $f_X(x) = 1, 0 \le x \le 1$ and $f_Y(y) = 1, 0 \le y \le 1$, then

$$J(Y_{[r:n]}) = \frac{-1}{2}\sum_{k=0}^{4} \frac{a_k}{k+1}.$$

Example 2: If (X, Y) follows Sarmanov family of distributions given in (3) with $f_X(x) = \theta_1 e^{-\theta_1 x}, x \geq 0$ and $f_Y(y) = \theta_2 e^{-\theta_2 y}, y \geq 0$, then

$$J(Y_{[r:n]}) = \frac{-\theta_2}{2} \sum_{k=0}^{4} \frac{a_k}{(k+1)(k+2)}.$$

Theorem 2: Let $Y_{[r:n]}$ be the concomitant of rth order statistic of a random sample of size n arising from Sarmanov family of distributions, the upper bound of $\mathrm{J}(Y_{[r:n]})$ can be written as

$$J(Y_{[r:n]}) \leq \frac{-1}{2} \sum_{k=1}^{3} \frac{a_k}{k+1} E\Big(F^{-1}(U_k)\Big), \tag{10}$$

where U_k follows Beta $(k+1, 1)$.

Proof: Since $a_0 \geq 0$ and $a_4 \geq 0$, by using Theorem 1 we can obtain the inequality (10) directly . Hence the proof. □

Example 3: If (X, Y) follows Sarmanov family of distributions given in (3) with $f_X(x) = 1, 0 \leq x \leq 1$ and $f_Y(y) = 1, 0 \leq y \leq 1$, then

$$J(Y_{[r:n]}) \leq \frac{-1}{2} \sum_{k=1}^{3} \frac{a_k}{k+1}.$$

Example 4: If (X, Y) follows Sarmanov family of distributions given in (3) with $f_X(x) = \theta_1 e^{-\theta_1 x}, x \geq 0$ and $f_Y(y) = \theta_2 e^{-\theta_2 y}, y \geq 0$, then

$$J(Y_{[r:n]}) \leq \frac{-\theta_2}{2} \sum_{k=1}^{3} \frac{a_k}{(k+1)(k+2)}.$$

Theorem 3: Let $Y_{[r:n]}$ be the concomitant of rth order statistic of a random sample of size n arising from Sarmanov family of distributions, then the lower bound of $J(Y_{[r:n]})$ is given by

$$J(Y_{[r:n]}) \geq \frac{-1}{2} \left(E[(f_Y(y))^2] \right)^{\frac{1}{2}} \left(\int_0^1 \Big(\sum_{k=0}^{4} a_k u^k \Big)^2 du \right)^{\frac{1}{2}}. \tag{11}$$

Proof: From (7), we have

$$J(Y_{[r:n]}) = \frac{-1}{2} \int_{u=0}^{1} f_Y(F^{-1}(u)) \Big(\rho_{(r,n,\alpha)}(u) \Big)^2 du$$

By applying Cauchy - Schwarz inequality, we have

$$J(Y_{[r:n]}) \geq \frac{-1}{2} \left(\int_{u=0}^{1} (f_Y(F^{-1}(u)))^2 du \right)^{\frac{1}{2}} \left(\int_{u=0}^{1} \Big(\rho_{(r,n,\alpha)}(u) \Big)^4 du \right)^{\frac{1}{2}}. \tag{12}$$

Therefore

$$\begin{aligned}\int_{u=0}^{1}(f_Y(F^{-1}(u)))^2du &= \int_y(f_Y(y))^3dy \\ &= E[(f_Y(y))^2]. \end{aligned} \tag{13}$$

Also

$$\Big(\rho_{(r,n,\alpha)}(u)\Big)^4 = \Big(\sum_{k=0}^{4}a_ku^k\Big)^2. \tag{14}$$

On substituting (13) and (14) in (12) we get (11). Hence the proof. □

Example 5: If (X,Y) follows Sarmanov family of distributions given in (3) with $f_X(x) = 1, 0 \le x \le 1$ and $f_Y(y) = 1, 0 \le y \le 1$, then

$$J(Y_{[r:n]}) \ge \frac{-1}{2}\left(\int_0^1\Big(\sum_{k=0}^{4}a_ku^k\Big)^2du\right)^{\frac{1}{2}}.$$

Example 6: If (X,Y) follows Sarmanov family of distributions given in (3) with $f_X(x) = \theta_1e^{-\theta_1x}, x \ge 0$ and $f_Y(y) = \theta_2e^{-\theta_2y}, y \ge 0$, then

$$J(Y_{[r:n]}) \ge \frac{-1}{2}\Big(\frac{\theta_2^2}{3}\Big)^{\frac{1}{2}}\left(\int_0^1\Big(\sum_{k=0}^{4}a_ku^k\Big)^2du\right)^{\frac{1}{2}}.$$

3. Extropy of ranked set sample

Let $Y_{[1]}$,$Y_{[2]}$,...$Y_{[n]}$ be the RSS of size n arising from Sarmanov family of distributions in which X observations are used to rank the units in each set. Clearly $Y_{[r]}$, $r = 1, 2, ..., n$ are independent and $Y_{[r]} \stackrel{d}{=} Y_{[r:n]}$. If $Y_{[RSS]} = \{Y_{[r]}, r = 1, 2, ...n\}$, then the extropy of $Y_{[RSS]}$ can be written as

$$\begin{aligned}J(Y_{RSS}) &= \frac{-1}{2}\prod_{r=1}^{n}\int_y(f_{Y[r:n]}(y))^2dy \\ &= \frac{-1}{2}\prod_{r=1}^{n}[-2J(Y_{[r:n]})].\end{aligned}$$

Therefore,

$$J(Y_{RSS}) = \frac{-1}{2}\prod_{r=1}^{n}\sum_{k=0}^{4}\frac{a_k}{k+1}E\Big(F^{-1}(U_k)\Big).$$

Example 7: If (X,Y) follows Sarmanov family of distributions given in (3) with $f_X(x) = 1, 0 \le x \le 1$ and $f_Y(y) = 1, 0 \le y \le 1$,then

$$J(Y_{RSS}) = \frac{-1}{2}\prod_{r=1}^{n}\sum_{k=0}^{4}\frac{a_k}{k+1}.$$

Example 8: If (X, Y) follows Sarmanov family of distributions given in (3) with $f_X(x) = \theta_1 e^{-\theta_1 x}, x \geq 0$ and $f_Y(y) = \theta_2 e^{-\theta_2 y}, y \geq 0$, then

$$J(Y_{RSS}) = \frac{-1}{2}\theta_2^n \prod_{r=1}^{n} \sum_{k=0}^{4} \frac{a_k}{(k+1)(k+2)}.$$

Definition 1: (Shaked and Shanthikumar (2007)) Let X_1 and X_2 be two random variables with cdfs F_1 and F_2 and pdfs f_1 and f_2 respectively. The left continuous inverses of F_1 and F_2 are given by $F_1^{-1}(u) = inf\{t : F_1(t) \geq u\}$ and $F_2^{-1}(u) = inf\{t : F_2(t) \geq u\}$, $0 \leq u \leq 1$. Then X_1 is said to be smaller than X_2 in dispersive order denoted by $X_1 \leq_{disp} X_2$ if $F_2^{-1}(F_1(x)) - x$ is increasing in $x \geq 0$. Clearly if $X_1 \leq_{disp} X_2$, then $f_1(F_1^{-1}(u)) \leq f_2(F_2^{-1}(u))$, for $0 \leq u \leq 1$.

Theorem 4: Let (X, Y) follows Sarmanov family of distributions given in (3) with marginal cdfs $F_X(x)$ and $F_Y(y)$ and pdfs $f_X(x)$ and $f_Y(y)$ respectively. Let $Y_{RSS} = \{Y_{[r]}, r = 1, 2, ...n\}$ be the ranked set sample of size n arising from Sarmanov family of distributions in which X observations are used to rank the units. Let (V, W) be another pair of random variables follows Sarmanov family of distributions given in (3) with marginal cdfs $G_V(v)$and $G_W(w)$ and pdfs $g_V(v)$ and $g_W(w)$ respectively. Let $W_{RSS} = \{W_{[r]}, r = 1, 2, ...n\}$ be the ranked set sample of size n arising from (V, W) in which V observations are used to rank the units. If $Y \leq_{disp} W$, then $J(Y_{RSS}) \leq J(W_{RSS})$.

Proof: We have

$$J(Y_{RSS}) = \frac{-1}{2} \prod_{r=1}^{n} \int_{u=0}^{1} f_Y(F^{-1}(u)) \Big(\rho_{(r,n,\alpha)}(u)\Big)^2 du.$$

Since $Y \leq_{disp} W$, we have $f_Y(F^{-1}(u)) \geq g_W(G^{-1}(u))$ for all u in $(0, 1)$.
Therefore

$$\begin{aligned} J(Y_{RSS}) &\leq \frac{-1}{2} \prod_{r=1}^{n} \int_{u=0}^{1} g_W(G^{-1}(u)) \Big(\rho_{(r,n,\alpha)}(u)\Big)^2 du \\ &= J(W_{RSS}). \end{aligned}$$

Hence the proof. □

3.1. Bounds of $J(Y_{RSS})$

In this subsection, we obtain some lower bounds and upper bounds for $J(Y_{RSS})$. Before that we give some properties of $\rho_{(r,n,\alpha)}(u)$ given in (8). We have tabulated the value of $\rho_{(r,n,\alpha)}(u)$ for $r = 1, 2, ...10$ and $\alpha = -0.5, -0.25, 0.25$, and 0.5 and are given in Table 1 and Table 2. We have also drawn the graphs of $\rho_{(r,n,\alpha)}(u)$ for $n = 10$ and for $\alpha > 0$ and $\alpha < 0$ and are given in Figure 1 to Figure 4.

Remark 3: From Table 1 and Table 2, we have for a fixed α , $\rho_{(r,n,\alpha)}(u) = \rho_{(n-r+1,n,\alpha)}(1-u)$. The above inference also be seen from Figures 1, 2, 3 and 4.

Remark 4: From Figures 1 and 2 we have for $\alpha > 0$, $\rho_{(r,n,\alpha)}(u)$ is decreasing in r if $0 \leq u < 0.5$ and is increasing in r if $0.5 < u \leq 1$. Again for $\alpha < 0$, $\rho_{(r,n,\alpha)}(u)$ is increasing in r if $0 \leq u < 0.5$ and is decreasing in r if $0.5 < u \leq 1$.

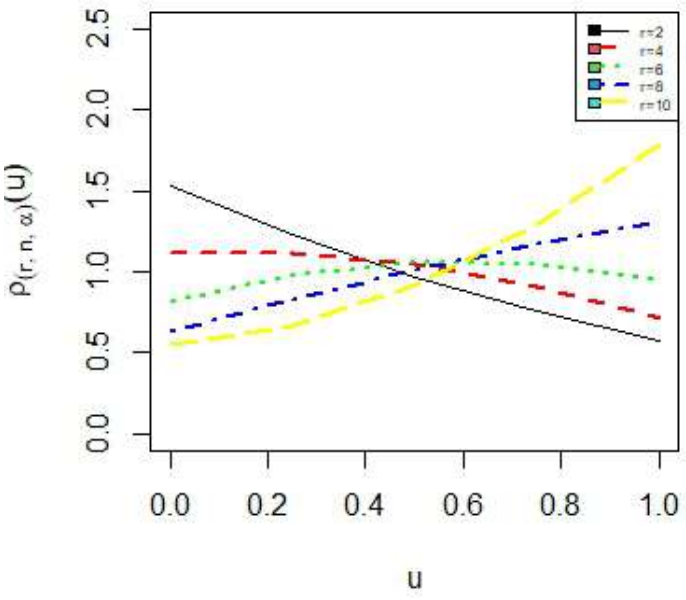


Figure 1: Graph of $\rho_{(r,n,\alpha)}(u)$ against u when $\alpha > 0$

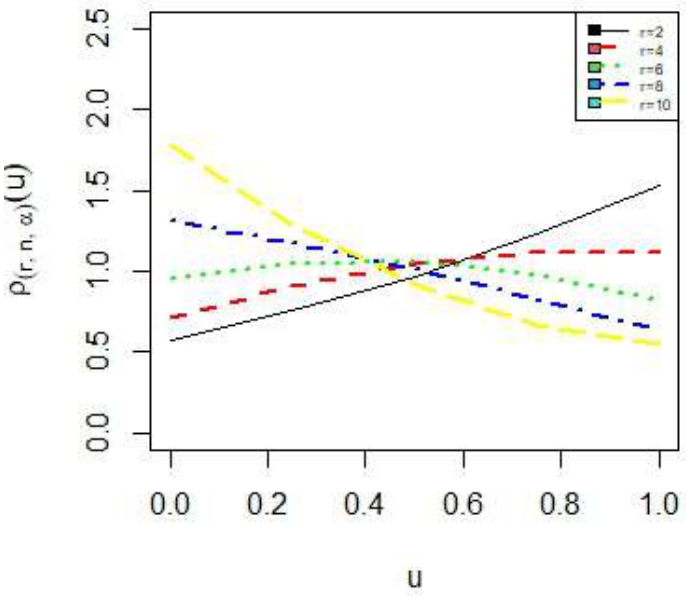


Figure 2: Graph of $\rho_{(r,n,\alpha)}(u)$ against u when $\alpha < 0$

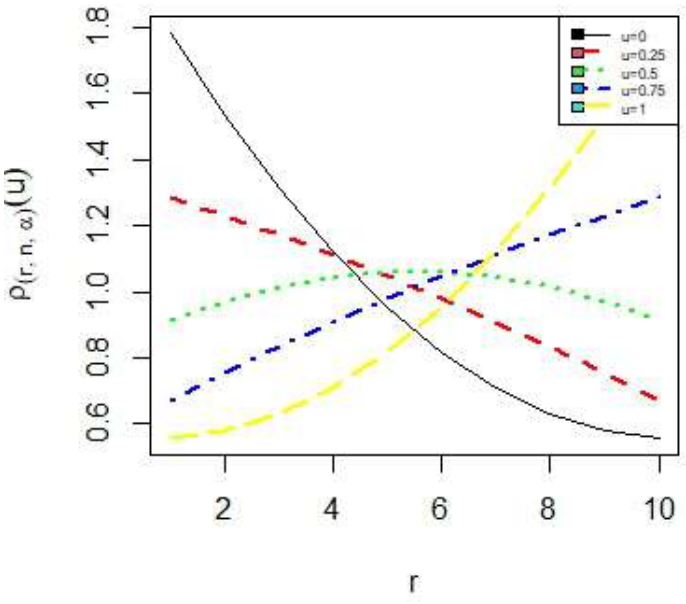


Figure 3: Graph of $\rho_{(r,n,\alpha)}(u)$ against r when $\alpha > 0$

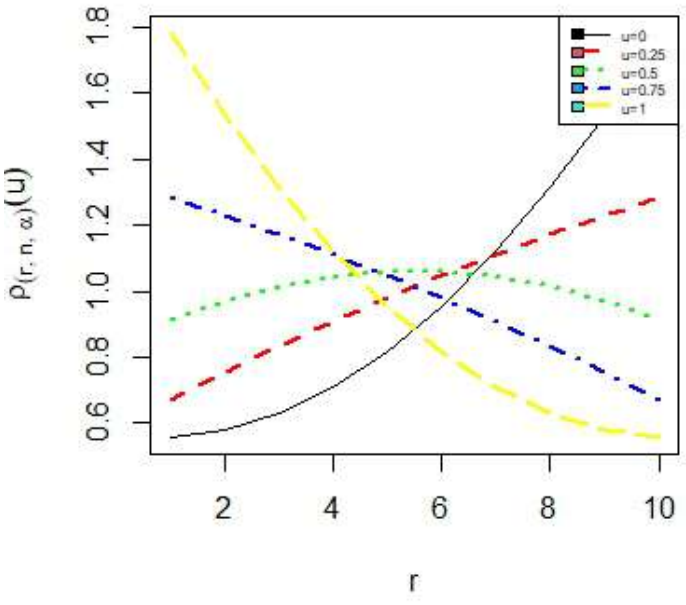


Figure 4: Graph of $\rho_{(r,n,\alpha)}(u)$ against r when $\alpha < 0$

Table 1: $\rho_{(r,n,\alpha)}(u)$ **when** α **is positive for** $n = 10$

	α=0.25				
r	u=0	u=0.25	u=0.5	u=0.75	u=1
1	1.7841	1.2855	0.9148	0.6719	0.5568
2	1.5341	1.2315	0.9716	0.7543	0.5795
3	1.3125	1.1740	1.0142	0.8331	0.6307
4	1.1193	1.1129	1.0426	0.9084	0.7102
5	0.9545	1.0483	1.0568	0.9801	0.8182
6	0.8182	0.9801	1.0568	1.0483	0.9545
7	0.7102	0.9084	1.0426	1.1129	1.1193
8	0.6307	0.8331	1.0142	1.1740	1.3125
9	0.5795	0.7543	0.9716	1.2315	1.5341
10	0.5568	0.6719	0.9148	1.2855	1.7841
	α=0.5				
r	u=0	u=0.25	u=0.5	u=0.75	u=1
1	2.9091	1.5284	0.6591	0.3011	0.4545
2	2.1818	1.4489	0.8864	0.4943	0.2727
3	1.5682	1.3551	1.0568	0.6733	0.2045
4	1.0682	1.2472	1.1705	0.8381	0.2500
5	0.6818	1.1250	1.2273	0.9886	0.4091
6	0.4091	0.9886	1.2273	1.1250	0.6818
7	0.2500	0.8381	1.1705	1.2472	1.0682
8	0.2045	0.6733	1.0568	1.3551	1.5682
9	0.2727	0.4943	0.8864	1.4489	2.1818
10	0.4545	0.3011	0.6591	1.5284	2.9091

Theorem 5: Let $Y_1, Y_2, ...Y_n$ be a simple random sample from a distribution with cdf $F_Y(y)$ and pdf $f_Y(y)$. Let $\{Y_{[r]}, r = 1, 2, ...n\}$ be the RSS of size n arising from Sarmanov family of distributions in which X observations are used to rank the units. If $Y_{SRS} = \{Y_1, Y_2, ..Y_n\}$ and $Y_{[RSS]} = \{Y_{[1]}, Y_{[2]}, ...Y_{[n]}\}$, then for $n \geq 1$,

$$\frac{J(Y_{RSS})}{J(Y_{SRS})} \leq \prod_{r=1}^{n} \Big(\rho_{(r,n,\alpha)}(u_0)\Big)^2,$$

where u_0 is the value of u which maximise $\rho_{(r,n,\alpha)}(u)$.

Proof: We have

$$\begin{aligned} J(Y_{SRS}) &= \frac{-1}{2} \prod_{r=1}^{n} \int_y \Big(f_Y(y)\Big)^2 dy \\ &= \frac{-1}{2} \prod_{r=1}^{n} \int_0^1 f_Y(F^{-1}(u)) du. \end{aligned}$$

Then,

$$J(Y_{RSS}) = \frac{-1}{2} \prod_{r=1}^{n} \int_0^1 f_Y(F^{-1}(u)) \Big(\rho_{(r,n,\alpha)}(u)\Big)^2 du.$$

Table 2: $\rho_{(r,n,\alpha)}(u)$ **when** α **is negative for** $n = 10$

	α=-0.5				
r	u=0	u=0.25	u=0.5	u=0.75	u=1
1	0.4545	0.3011	0.6591	1.5284	2.9091
2	0.2727	0.4943	0.8864	1.4489	2.1818
3	0.2045	0.6733	1.0568	1.3551	1.5682
4	0.2500	0.8381	1.1705	1.2472	1.0682
5	0.4091	0.9886	1.2273	1.1250	0.6818
6	0.6818	1.1250	1.2273	0.9886	0.4091
7	1.0682	1.2472	1.1705	0.8381	0.2500
8	1.5682	1.3551	1.0568	0.6733	0.2045
9	2.1818	1.4489	0.8864	0.4943	0.2727
10	2.9091	1.5284	0.6591	0.3011	0.4545
	α=-0.25				
r	u=0	u=0.25	u=0.5	u=0.75	u=1
1	0.5568	0.6719	0.9148	1.2855	1.7841
2	0.5795	0.7543	0.9716	1.2315	1.5341
3	0.6307	0.8331	1.0142	1.1740	1.3125
4	0.7102	0.9084	1.0426	1.1129	1.1193
5	0.8182	0.9801	1.0568	1.0483	0.9545
6	0.9545	1.0483	1.0568	0.9801	0.8182
7	1.1193	1.1129	1.0426	0.9084	0.7102
8	1.3125	1.1740	1.0142	0.8331	0.6307
9	1.5341	1.2315	0.9716	0.7543	0.5795
10	1.7841	1.2855	0.9148	0.6719	0.5568

Let u_0 be the value of u which maximise $\rho_{(r,n,\alpha)}(u)$. Then,

$$\begin{aligned} J(Y_{RSS}) &\geq \frac{-1}{2}\prod_{r=1}^{n}\int_0^1 \left(f_Y(F^{-1}(u))(\rho_{(r,n,\alpha)}(u_0))^2\right)du \\ &= \frac{-1}{2}\prod_{r=1}^{n}\left(\int_0^1 f_Y(F^{-1}(u))du\right)\prod_{r=1}^{n}\left(\rho_{(r,n,\alpha)}(u_0)\right)^2 \\ &= J(Y_{SRS})\prod_{r=1}^{n}\left(\rho_{(r,n,\alpha)}(u_0)\right)^2. \end{aligned}$$

Since $J(Y_{SRS}) < 0$,

$$\frac{J(Y_{RSS})}{J(Y_{SRS})} \leq \prod_{r=1}^{n}\left(\rho_{(r,n,\alpha)}(u_0)\right)^2.$$

Hence the proof. □

Theorem 6: Let $Y_{RSS} = \{Y_{[r]}, r = 1, 2, ...n\}$ be the RSS of size n arising from Sarmanov family of distributions in which X observations are used to rank the units then for all $n \geq 1$,

then the lower bound of extropty of Y_{RSS} is given by

$$J(Y_{RSS}) \geq \frac{-1}{2}\left(Ef_Y(y)^2\right)^{\frac{n}{2}} \prod_{r=1}^{n}\left(\int_0^1 \Big(\sum_{k=0}^{4} a_k u^k\Big)^2 du\right)^{\frac{1}{2}}.$$

Proof: We have

$$J(Y_{RSS}) = \frac{-1}{2}\prod_{r=1}^{n}\int_{u=0}^{1} f_Y(F^{-1}(u))\Big(\rho_{(r,n,\alpha)}(u)\Big)^2 du.$$

Using Cauchy-Schwarz inequality

$$J(Y_{RSS}) \geq \frac{-1}{2}\prod_{r=1}^{n}\Big(\int_{u=0}^{1} f_Y(F^{-1}(u))^2 du\Big)^{\frac{1}{2}}\Big(\int_{u=0}^{1}\Big(\rho_{(r,n,\alpha)}(u)\Big)^4 du\Big)^{\frac{1}{2}}.$$

We have $\Big(\rho_{(r,n,\alpha)}(u)\Big)^2 = \sum_{k=0}^{4} a_k u^k$.
Therefore

$$J(Y_{RSS}) \geq \frac{-1}{2}\Big(Ef_Y(y)^2\Big)^{\frac{n}{2}} \prod_{r=1}^{n}\left(\int_0^1 \Big(\sum_{k=0}^{4} a_k u^k\Big)^2 du\right)^{\frac{1}{2}}.$$

Hence the proof. □

4. Extropy of $(X_{RSS}, Y_{[RSS]})$

If $X_{(r)}$ is the observation measured on the auxiliary variable X of the unit chosen from the rth set then $X_{(r)}$ is the rth order statistic of a random sample of size n. Since $Y_{[r]}$ is the concomitant of $X_{(r)}$, the joint pdf of $(X_{(r)}, Y_{[r]})$ is given by

$$h(X_{(r)}, Y_{[r]}) = \frac{n!}{(r-1)!(n-r)!} f(x,y)(F_X(x))^{r-1}(1-F_X(x))^{(n-r)}. \tag{15}$$

Then the extropy of $(X_{(r)}, Y_{[r]})$ can be defined as

$$\begin{aligned}
J(X_{(r)}, Y_{[r]}) &= \frac{-1}{2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\Big(h(X_{(r)}, Y_{[r]})\Big)^2 dydx \\
&= \frac{-1}{2}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\Big(\frac{n!}{(r-1)!(n-r)!}\Big)^2\Big(f_X(x)\Big)^2\Big(f_Y(y)\Big)^2 \\
&\quad\times\Bigg[1+3\alpha(2F_X(x)-1)(2F_Y(y)-1) \\
&\quad+\frac{5}{4}\alpha^2\Big(3(2F_X(x)-1)^2-1\Big)\Big(3(2F_Y(y)-1)^2-1\Big)\Bigg]^2 \\
&\quad\times\Big(F_X(x)\Big)^{2(r-1)}\Big(1-F_X(x)\Big)^{2(n-r)}dxdy \\
&= -\frac{1}{2}\Big(\frac{n!}{(r-1)!(n-r)!}\Big)^2\Bigg[M_{00}N_{00}+9\alpha^2M_{20}N_{20}+\frac{25}{16}\alpha^2M_{02}N_{02} \\
&\quad+6\alpha M_{10}N_{10}+\frac{5}{2}\alpha^2M_{01}N_{01}+\frac{15}{2}\alpha^3M_{11}N_{11}\Bigg], \qquad (16)
\end{aligned}$$

where M_{ij} and N_{ij} for $i = 0, 1$ and 2 are given below.

$$\begin{aligned}
M_{ij} &= \int (f_X(x))^2(F_X(x))^{2(r-1)}(1-F_X(x))^{2(n-r)}(2F_X(x)-1)^i(3(2F_X(x)-1)^2-1)^j dx \\
&= \frac{(2r-2)!(2n-2r)!}{(2n-1)!}E\Big[f_X(F^{-1}(U))(2U-1)^i(3(2U-1)^2-1)^j\Big], \qquad (17)
\end{aligned}$$

where U follows beta distribution with parameters $(2r-1, 2n-2r+1)$ and

$$\begin{aligned}
N_{ij} &= \int (f_Y(y))^2(2F_Y(y)-1)^i(3(2F_Y(y)-1)^2-1)^j dy \\
&= E\Big[f_Y(F^{-1}(V))(2V-1)^i(3(2V-1)^2-1)^j\Big], \qquad (18)
\end{aligned}$$

where V follows uniform distribution over $(0, 1)$.
If $X_{RSS} = \{X_{(1)}, X_{(2)}, .., X_{(n)}\}$, then X_{RSS} is the RSS of X observations in which ranking of units in each set is perfect. Let $(X_{RSS}, Y_{[RSS]}) = \{(X_{(r)}, Y_{[r]}), r = 1, 2, 3...n\}$ then extropy of

$(X_{RSS}, Y_{[RSS]})$ is given by

$$\begin{aligned} J(X_{RSS}, Y_{[RSS]}) &= \frac{-1}{2}\prod_{r=1}^{n}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\Big(h(X_{(r)}, Y_{[r]})\Big)^2 dydx \\ &= \frac{-1}{2}\prod_{r=1}^{n} -2J(X_{(r)}, Y_{[r]}) \\ &= \frac{-1}{2}\prod_{r=1}^{n}\Big(\frac{n!}{(r-1)!(n-r)!}\Big)^2 \\ &\quad \times \Bigg[M_{00}N_{00} + 9\alpha^2 M_{20}N_{20} + \frac{25}{16}\alpha^2 M_{02}N_{02} \\ &\quad + 6\alpha M_{10}N_{10} + \frac{5}{2}\alpha^2 M_{01}N_{01} + \frac{15}{2}\alpha^3 M_{11}N_{11}\Bigg]. \end{aligned} \tag{19}$$

Example 9: If (X, Y) follows Sarmanov family of distributions given in (3) with marginal pdfs of X and Y are $f_X(x) = 1, 0 \le x \le 1$ and $f_Y(y) = 1, 0 \le y \le 1$ respectively, then

$$\begin{aligned} M_{ij} &= \frac{(2r-2)!(2n-2r)!}{(2n-1)!}E\Big[(2U-1)^i(3(2U-1)^2-1)^j\Big] \\ &= \int_0^1 (2u-1)^i(3(23-1)^2-1)^j u^{2r-2}(1-u)^{2n-2r}du \end{aligned}$$

and

$$\begin{aligned} N_{ij} &= E\Big[(2V-1)^i(3(2V-1)^2-1)^j\Big] \\ &= \int_0^1 (2v-1)^i(3(2v-1)^2-1)^j dv. \end{aligned}$$

Therefore,

$$M_{00} = \frac{(2r-2)!(2n-2r)!}{(2n-1)!},$$

$$M_{10} = \frac{(2r-2)!(2n-2r)!}{(2n-1)!}\left[\frac{(2r-1)}{n} - 1\right],$$

$$M_{20} = \frac{(2r-2)!(2n-2r)!}{(2n-1)!}\left[\frac{4(2r)(2r-1)}{(2n)(2n+1)} - \frac{4(2r-1)}{2n} + 1\right],$$

$$M_{01} = \frac{(2r-2)!(2n-2r)!}{(2n-1)!}\left[\frac{12(2r)(2r-1)}{(2n)(2n+1)} - \frac{12(2r-1)}{2n} + 2\right],$$

$$\begin{aligned} M_{02} = 4\frac{(2r-2)!(2n-2r)!}{(2n-1)!}\Bigg[&\frac{36(2r+2)(2r+1)(2r)(2r-1)}{(2n+3)(2n+2)(2n+1)(2n)} \\ &- \frac{72(2r+1)(2r)(2r-1)}{(2n+2)(2n+1)(2n)} + \frac{48(2r)(2r-1)}{(2n+1)(2n)} - \frac{12(2r-1)}{2n} + 1\Bigg], \end{aligned}$$

and

$$M_{11} = 2\frac{(2r-2)!(2n-2r)!}{(2n-1)!}\left[\frac{12(2r+1)(2r)(2r-1)}{(2n+2)(2n+1)(2n)} - \frac{18(2r)(2r-1)}{(2n+1)(2n)} + \frac{8(2r-1)}{2n} - 1\right]$$

Also $N_{00} = 1$, $N_{10} = 0$, $N_{20} = \frac{1}{3}$, $N_{01} = 0$, $N_{02} = \frac{4}{5}$ and $N_{11} = 0$. Then from (16),

$$\begin{aligned} J(X_{(r:n)}, Y_{[r:n]}) = {} & \frac{-1}{2}\Big(\frac{n!}{(r-1)!(n-r)!}\Big)^2 \frac{(2r-2)!(2n-2r)!}{(2n-1)!} \\ & \times \left[\frac{180\alpha^2(2r+2)(2r+1)(2r)(2r-1)}{(2n+3)(2n+2)(2n+1)(2n)}\right. \\ & - \frac{360\alpha^2(2r+1)(2r)(2r-1)}{(2n+2)(2n+1)(2n)} + \frac{252\alpha^2(2r)(2r-1)}{(2n+1)(2n)} \\ & \left. - \frac{72\alpha^2(2r-1)}{2n} + 8\alpha^2 + 1\right]. \end{aligned}$$

Therefore,

$$\begin{aligned} J(X_{RSS}, Y_{[RSS]}) = {} & \frac{-1}{2}\prod_{r=1}^{n}\Big(\frac{n!}{(r-1)!(n-r)!}\Big)^2 \frac{(2r-2)!(2n-2r)!}{(2n-1)!} \\ & \times \left[\frac{180\alpha^2(2r+2)(2r+1)(2r)(2r-1)}{(2n+3)(2n+2)(2n+1)(2n)}\right. \\ & - \frac{360\alpha^2(2r+1)(2r)(2r-1)}{(2n+2)(2n+1)(2n)} + \frac{252\alpha^2(2r)(2r-1)}{(2n+1)(2n)} \\ & \left. - \frac{72\alpha^2(2r-1)}{2n} + 8\alpha^2 + 1\right]. \end{aligned}$$

Example 10: If (X, Y) follows Sarmanov family of distributions given in (3) with marginal pdfs of X and Y are $f_X(x) = \theta_1 e^{-\theta_1 x}, x \geq 0$ and $f_Y(y) = \theta_2 e^{-\theta_2 y}, y \geq 0$ respectively,then

$$\begin{aligned} M_{ij} &= \frac{(2r-2)!(2n-2r)!}{(2n-1)!}\theta_1 E\Big[(1-U)(2U-1)^i(3(2U-1)^2-1)^j\Big] \\ &= \theta_1 \int_0^1 (1-u)(2u-1)^i(3(23-1)^2-1)^j u^{2r-2}(1-u)^{2n-2r}du \end{aligned}$$

and

$$\begin{aligned} N_{ij} &= \theta_2 E\Big[(1-V)(2V-1)^i(3(2V-1)^2-1)^j\Big] \\ &= \theta_2 \int_0^1 (1-v)(2v-1)^i(3(2v-1)^2-1)^j dv. \end{aligned}$$

Therefore,

$$M_{00} = \frac{(2r-2)!(2n-2r+1)!}{(2n)!}\theta_1,$$

$$M_{10} = \frac{(2r-2)!(2n-2r+1)!}{(2n)!}\theta_1\left[\frac{2(2r-1)}{(2n+1)} - 1\right],$$

$$M_{20} = \frac{(2r-2)!(2n-2r+1)!}{(2n)!}\theta_1\left[\frac{4(2r)(2r-1)}{(2n+1)(2n+2)} - \frac{4(2r-1)}{(2n+1)} + 1\right],$$

$$M_{01} = \frac{(2r-2)!(2n-2r+1)!}{(2n)!}\theta_1\left[\frac{12(2r)(2r-1)}{(2n+1)(2n+2)} - \frac{12(2r-1)}{(2n+1)} + 2\right],$$

$$M_{02} = \frac{4(2r-2)!(2n-2r+1)!}{(2n)!}\theta_1\left[\frac{36(2r+2)(2r+1)(2r)(2r-1)}{(2n+4)(2n+3)(2n+2)(2n+1)}\right.$$
$$\left. - \frac{72(2r+1)(2r)(2r-1)}{(2n+3)(2n+2)(2n+1)} + \frac{48(2r)(2r-1)}{(2n+2)(2n+1)} - \frac{12(2r-1)}{2n+1} + 1\right]$$

and

$$M_{11} = \frac{2(2r-2)!(2n-2r+1)!}{(2n)!}\theta_1\left[\frac{122(2r+1)(2r)(2r-1)}{(2n+3)(2n+2)(2n+1)}\right.$$
$$\left. - \frac{18(2r)(2r-1)}{(2n+2)(2n+1)} - \frac{8(2r-1)}{2n+1} - 1\right].$$

Also, $N_{00} = \frac{\theta_2}{2}$, $N_{10} = \frac{-\theta_2}{6}$, $N_{20} = \frac{\theta_2}{6}$, $N_{01} = 0$, $N_{02} = \frac{2\theta_2}{5}$ and $N_{11} = \frac{-2\theta_2}{15}$. Then from (16),

$$J(X_{(r:n)}, Y_{[r:n]}) = \frac{-1}{2}\left(\frac{n!}{(r-1)!(n-r)!}\right)^2\frac{(2r-2)!(2n-2r+1)!}{(2n)!}\theta_1\theta_2$$
$$\times\left[\frac{45\alpha^2(2r+2)(2r+1)(2r)(2r-1)}{4(2n+4)(2n+3)(2n+2)(2n+1)}\right.$$
$$- \frac{(2r+1)(2r)(2r-1)(45\alpha^2-24\alpha^3)}{(2n+3)(2n+2)(2n+1)} + \frac{36(2r)(2r-1)(\alpha^2+\alpha^3)}{(2n+2)(2n+1)}$$
$$\left. - \frac{(32\alpha^3+27\alpha^2+4\alpha)}{2}\frac{\alpha^2(2r-1)}{2n+1} + \alpha^3 + \frac{51}{24}\alpha^2 + \alpha + \frac{1}{2}\right].$$

Therefore,

$$J(X_{RSS}, Y_{[RSS]}) = \frac{-\theta_1^n \theta_2^n}{2} \prod_{r=1}^{n} \left(\frac{n!}{(r-1)!(n-r)!} \right)^2 \frac{(2r-2)!(2n-2r+1)!}{(2n)!}$$
$$\times \left[\frac{45\alpha^2(2r+2)(2r+1)(2r)(2r-1)}{4(2n+4)(2n+3)(2n+2)(2n+1)} \right.$$
$$- \frac{(2r+1)(2r)(2r-1)(45\alpha^2 - 24\alpha^3)}{(2n+3)(2n+2)(2n+1)} + \frac{36(2r)(2r-1)(\alpha^2+\alpha^3)}{(2n+2)(2n+1)}$$
$$\left. - \frac{(32\alpha^3 + 27\alpha^2 + 4\alpha)}{2} \frac{\alpha^2(2r-1)}{2n+1} + \alpha^3 + \frac{51}{24}\alpha^2 + \alpha + \frac{1}{2} \right].$$

5. Conclusion

In this work, we considered the extropy of concomitants of order statistic arising from Sarmanov family of distributions when ranking is subject to error. If we considered a ranked set sampling in which an auxiliary variable is used to rank the units in each set, then the observation of RSS are nothing but concomitants of order statistics. Hence by using the results for extropy of concomitants of order statistics $Y_{[r:n]}$, we derived the extropy of RSS in which units are ranked based on measurements made on an easily and exactly measurable auxiliary variable X which is correlated with the study variable Y, under the assumption that (X, Y) follows Sarmanov family of distributions. The lower and upper bounds of extropy of $Y_{[r:n]}$ were obtained. Moreover, we obtained the lower and upper bound of extropy of RSS. The upper bound for the ratio of extropy of ranked set sample to that of simple random sample were obtained. The extropy of $(X_{RSS}, Y_{[RSS]})$ were also obtained for Sarmanov family of distributions, where X_{RSS} is the RSS of the X observations and $Y_{[RSS]}$ is the RSS of the Y observations in which X observations are used to rank.

Acknowledgements

The authors are thankful to the referee for valuable comments and constructive criticism which lead to an improved version of the manuscript.

References

Abdallah, A., Boucher, J.-P., and Cossette, H. (2016). Sarmanov family of multivariate distributions for bivariate dynamic claim counts model. *Insurance: Mathematics and Economics*, **68**, 120–133.

Alemany, R., Bolancé, C., Rodrigo, R., and Vernic, R. (2020). Bivariate mixed poisson and normal generalised linear models with sarmanov dependence-an application to model claim frequency and optimal transformed average severity. *Mathematics*, **9**, 73.

Bain, L. (2017). *Statistical Analysis of Reliability and Life-testing Models: Theory and Methods*. Routledge.

Barakat, H., Alawady, M., Husseiny, I., and Mansour, G. (2022). Sarmanov family of bivariate distributions: statistical properties-concomitants of order statistics-information measures. *Bulletin of the Malaysian Mathematical Sciences Society*, **45**, 49–83.

Bolancé, C., Guillen, M., and Pitarque, A. (2020). A sarmanov distribution with beta marginals: An application to motor insurance pricing. *Mathematics*, **8**, 2020.

Chacko, M. (2017). Bayesian estimation based on ranked set sample from morgenstern type bivariate exponential distribution when ranking is imperfect. *Metrika*, **80**, 333–349.

Chacko, M. and George, V. (2023). Extropy properties of ranked set sample for cambanis type bivariate distributions. *Journal of the Indian Society for Probability and Statistics*, **24**, 111–133.

Chacko, M. and George, V. (2024). Extropy properties of ranked set sample when ranking is not perfect. *Communications in Statistics-Theory and Methods*, **53**, 3187–3210.

Chacko, M. and Thomas, P. Y. (2007). Estimation of a parameter of bivariate pareto distribution by ranked set sampling. *Journal of Applied Statistics*, **34**, 703–714.

Chacko, M. and Thomas, P. Y. (2008). Estimation of a parameter of morgenstern type bivariate exponential distribution by ranked set sampling. *Annals of the Institute of Statistical Mathematics*, **60**, 301–318.

Chacko, M. and Thomas, P. Y. (2009). Estimation of parameters of morgenstern type bivariate logistic distribution by ranked set sampling. *Journal of the Indian Society of Agricultural Statistics*, **63**, 77–83.

Chen, Z., Bai, Z., and Sinha, B. K. (2004). *Ranked Set Sampling: Theory and Applications*, volume 176. Springer.

David, H. A. (1973). Concomitants of order statistics. *Bulletin of the International Statistical Institute*, **45**, 295–300.

David, H. A. and Nagaraja, H. N. (1998). 18 concomitants of order statistics. *Handbook of Statistics*, **16**, 487–513.

Eftekharian, A. and Qiu, G. (2022). On extropy properties and discrimination information of different stratified sampling schemes. *Probability in the Engineering and Informational Sciences*, **36**, 644–659.

George, V. and Chacko, M. (2023). Cumulative residual extropy properties of ranked set sample for cambanis type bivariate distributions: Cumulative residual extropy properties of ranked set sample. *Journal of the Kerala Statistical Association*, **33**, 50–70.

Husseiny, I., Barakat, H., Mansour, G., and Alawady, M. (2022). Information measures in records and their concomitants arising from sarmanov family of bivariate distributions. *Journal of Computational and Applied Mathematics*, **408**, 114120.

Jozani, M. J. and Ahmadi, J. (2014). On uncertainty and information properties of ranked set samples. *Information Sciences*, **264**, 291–301.

Lad, F., Sanfilippo, G., and Agro, G. (2015). Extropy: Complementary dual of entropy. *Statistical Science*, **30**.

McIntyre, G. (1952). A method for unbiased selective sampling, using ranked sets. *Australian Journal of Agricultural Research*, **3**, 385–390.

Mehta, V. (2022). An improved estimation of parameter of morgenstern-type bivariate exponential distribution using ranked set sampling. In *Ranked Set Sampling Models and Methods*, pages 1–25. IGI Global.

Morgenstern, D. (1956). Einfache beispiele zweidimensionaler verteilungen. *Mitteilingsblatt fur Mathematische Statistik*, **8**, 234–235.

Qiu, G. (2017). The extropy of order statistics and record values. *Statistics & Probability Letters*, **120**, 52–60.

Qiu, G. and Raqab, M. Z. (2022). On weighted extropy of ranked set sampling and its comparison with simple random sampling counterpart. *Communications in Statistics-Theory and Methods*, **53**, 1–18.

Raqab, M. Z. and Qiu, G. (2019). On extropy properties of ranked set sampling. *Statistics*, **53**, 210–226.

Sarmanov, O. V. (1966). Generalized normal correlation and two-dimensional fréchet classes. In *Doklady Akademii Nauk*, volume 168, pages 32–35. Russian Academy of Sciences.

Shaked, M. and Shanthikumar, J. G. (2007). *Stochastic Orders*. Springer.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, **27**, 379–423.

Stokes, S. (1977). Ranked set sampling with concomitant variables. *Communications in Statistics-Theory and Methods*, **6**, 1207–1211.

Zamanzade, E. and Mahdizadeh, M. (2019). Extropy estimation in ranked set sampling with its application in testing uniformity. In *Ranked set sampling*, pages 259–267. Elsevier.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 22, No. 2, 2024 (New Series), pp 307–322
http://www.ssca.org.in/journal

Construction of Nearly Orthogonal Arrays Mappable to Tight Orthogonal Arrays of Strength Two Using Projective Geometry

Poonam Singh[1], Mukta D. Mazumder[2] and Santosh Babu[3]
[1]*Department of Statistics, University of Delhi, Delhi 110007.*
[2]*Department of Statistics, Ram Lal Anand College, University of Delhi, New Delhi 110021.*
[3]*Department of Satistics, University of Delhi, Delhi 110007.*

Received: 18 October 2023; Revised: 05 February 2024; Accepted: 29 February 2024

Abstract

Bose and Bush (1952) used projective geometry to construct orthogonal arrays of strength two and three. Mukerjee *et al.* (2014) constructed some series of mappable nearly orthogonal arrays (MNOAs) of strength two by using resolvable orthogonal arrays. This paper proposes a method to construct nearly orthogonal arrays that are mappable to tight orthogonal arrays of strength two by using projective geometry. The method is illustrated through examples and the constructed MNOAs are tabulated for 1-flat, 2-flat, and 3-flat of the projective geometry. Many new arrays are constructed with a better degree of orthogonality.

Key words: Orthogonal arrays; Tight orthogonal array; Mappable nearly orthogonal arrays; Projective geometry.

AMS Subject Classifications: 05B15

1. Introduction

Orthogonal arrays have been widely used in scientific, agricultural, and industrial investigations and in computer experiments. The concept of orthogonal arrays was introduced by Rao (1946). Rao (1947) obtained the upper bound on the maximum number of factors for a symmetric orthogonal array. Bose and Bush (1952) constructed orthogonal arrays of strength two and three by using Galois field, difference schemes and projective geometry. For more details on the construction and applications of orthogonal arrays see Hedayat *et al.* (1999).

Wang and Wu (1992) systematically studied Nearly Orthogonal Arrays (NOAs) and proposed some general combinatorial methods for their construction. Nguyen (1996) proposed an algorithm for constructing NOAs by adding two level columns to the existing OAs. These arrays are economic in run size but sacrifice the column orthogonality.

Corresponding Author: Santosh Babu
Email: santosh.yadav0506@gmail.com

Mukerjee *et al.* (2014) introduced the concept of Mappable Nearly Orthogonal Array (MNOA) and developed a method for construction of these arrays by using resolvable orthogonal arrays. In these arrays, each column is orthogonal to a large proportion of the other columns and easily convertible to fully orthogonal array via a mapping of symbols in each column to a possibly smaller set of symbols. The importance and applications of this type of array have been of considerable interest because of their inherent better space filling properties.

Mukerjee *et al.* (2014) have also illustrated through an example how an MNOA with 81 runs, 40 factors each with 9 symbols can achieve stratification on a 9 x 9 grid in 720 out of 780 two-dimensions and on a 3 x 3 grid in the remaining 60 two-dimensions. Thus, having a better space filling properties than an OA with 81 runs, 40 factors each at 3 levels and accommodating more factors than an OA with 81 runs, 10 factors with 9 symbols. An important property of MNOAs is that an another MNOA can be obtained with the same number of runs but less columns after deleting one or more columns from a MNOA Mukerjee *et al.* (2014). However, the main intent is to increase the number of groups for attaining a better degree of orthogonality instead of obtaining other orthogonal array after deleting columns. Li *et al.* (2023) constructed mappable nearly orthogonal arrays with column-orthogonality and enhance the projection uniformity on any one dimension by using the constructed nearly column orthogonal MNOAs and rotation matrices. Singh *et al.* (2023 a) constructed many new mappable nearly orthogonal arrays using difference matrix. Singh *et al.* (2023 b) constructed mappable nearly orthogonal array using projective geometry.

In this paper, we propose a method to construct mappable nearly orthogonal arrays of strength two using projective geometry. The constructed nearly orthogonal arrays are mappable to tight symmetric orthogonal arrays of strength two.

Section 2 gives notations and definitions of orthogonal array, MNOA and projective geometry. In Section 3, the steps of proposed method of construction of MNOA using projective geometry are given. Some newly constructed mappable nearly orthogonal arrays of strength two using proposed method are also given in this section. The constructed MNOAs are listed in Table 3 to Table 5.

2. Preliminaries

The following results and definitions are important for the present study.

2.1. Orthogonal array

An $N \times k$ matrix A, with entries from a set G of $s(\geq 2)$ elements, is called a symmetric orthogonal array of strength t, size N, k constraints and s levels if every $(N \times t)$ submatrix of A contains all possible $(1 \times t)$ row vectors with the same frequency λ. The array is denoted by $OA[N, k, s, t]$ and the number λ is called index of the array. For a symmetric orthogonal array $N = \lambda s^t$.

Theorem 1: (Rao, 1947) In an $OA[N, k, s, t]$ the following inequalities must hold:

$$N-1 \geq \binom{k}{1}(s-1)+\cdots+\binom{k}{u}(s-1)^u \qquad \text{if } t=2u$$

and

$$N-1 \geq \binom{k}{1}(s-1)+\cdots+\binom{k}{u}(s-1)^u+\binom{k-1}{u}(s-1)^{u+1} \qquad \text{if } t=2u+1$$

An orthogonal array is said to be tight orthogonal array if the equality holds in Theorem 1. Theorem 1 gives the lower bound on the minimum number of runs required for the existence of $OA[N,k,s,t]$.

2.2. Mappable nearly orthogonal array

A mappable nearly orthogonal array $MNOA\Big[N,\prod_{i=1}^{m}s_i^{c_i},\prod_{i=1}^{m}\prod_{j=1}^{c_i}r_{ij}\Big]$ is an $N\times\tilde{c}$ arrays whose $\tilde{c}=c_1+c_2+\cdots+c_m$ columns can be partitioned into m disjoint groups of $c_1,c_2,\ldots,c_m$ columns with the following properties:

I. for $i=1,\ldots,m$ every column of the ith group is populated by s_i symbols;

II. any two columns from different groups are orthogonal;

III. for $i=1,\ldots,m$ and for $j=1,\ldots,c_i$ the s_i symbols in the jth column of the ith group can be mapped to a set of $r_{ij}\leq s_i$ symbols such that these mappings convert the array into an orthogonal array $OA[N,\prod_{i=1}^{m}\prod_{j=1}^{c_i}r_{ij}]$ of strength two.

In particular, if $s_i=s, c_i=c$ and $r_{ij}=r$ for every i and j, then a mappable nearly orthogonal array is denoted as $A=MNOA[N,(s^c)^m,(r^c)^m]$.

By property II, in a mappable nearly orthogonal array before mapping, each of the c_i columns in the i group is orthogonal to at least a proportion $\tilde{\pi}=(\tilde{c}-c_i)/(\tilde{c}-1)$ of the other columns. This leads to the following measures of the pre-mapping degree of orthogonality among the columns:

$$\overline{\pi}=\frac{\sum_{i=1}^{m}c_i\pi_i}{\sum_{i=1}^{m}c_i}=(\tilde{c}^2-\sum_{i=1}^{m}c_i^2)/\{\tilde{c}(\tilde{c}-1)\} \tag{1}$$

$$\pi_{min}=min_{1\leq i\leq m}\pi_i=(\tilde{c}-max_{1\leq i\leq m}c_i)/(\tilde{c}-1) \tag{2}$$

if $c_1=c_2=\cdots=c_m=c$, then $\tilde{c}=mc$, where m and c are the number of groups and number of columns respectively and by (1) and (2) we have

$$\overline{\pi}=\pi_{min}=(m-1)c/(mc-1) \tag{3}$$

In this paper, symmetric orthogonal arrays are constructed, so every columns c_i contains same set of symbols.

2.3. Projective geometry

The projective geometry $PG(r,s)$ over Galois field $GF(s)$ of order s, where s is a prime or a power of a prime number, consists of ordered set $(y_0, y_1, \ldots, y_r)$ called points where $y_i, i = 0, 1, \ldots, r$, are elements of $GF(s)$ and not all them are simultaneously zero. The point $(ay_0, ay_1, \ldots ay_r)$ represents the same point as $(y_0, y_1, \ldots, y_r)$, for any $a \in GF(s), (a \neq 0)$. The collection of all those points which satisfy a set of $(r-t)$ linearly independent homogeneous equation with coefficients from $GF(s)$, not all of them are simultaneously zero within the same equation, is said to represents a t-flat in $PG(r,s)$.

In particular a $0-$flat, a $1-$flat,$\ldots, a(r-1)-$flat respectively in $PG(r,s)$ are known as a point, a line, $\ldots$, a hyperplane of $PG(r,s)$. The number of points lying on a t-flats is $(t+1)$.

3. Method of construction

Projective Geometry is a direct representation of orthogonal arrays of strength two. Raghavarao (1971) obtained the orthogonal arrays $OA[s^{(r+1)}, (\frac{s^{(r+1)}-1}{(s-1)}), s, 2]$ using by $PG(r,s)$. Mukerjee *et al.* (2014) constructed mappable nearly orthogonal arrays of strength two using resolvable orthogonal arrays. Here, we give a method to construct new series of mappable nearly orthogonal arrays of strength two using projective geometry. The total number of points in $PG(r,s)$ is $|PG(r,s)| = [\frac{s^{(r+1)}-1}{(s-1)}]$. The $PG(r,s)$ has $p = [(\frac{s^{(r+1)}-1}{s^{t+1})-1}]$ disjoint t-flats if and only if $(t+1)|(r+1)$. An orthogonal array $OA[s^{(r+1)}, p, s^{(t+1)}, 2]$ can be constructed using the disjoint t-flats in $PG(r,s)$. The method of construction is described in the following steps and Theorem 2.

Step I: Consider the orthogonal array $D = OA[s^{(t+1)}, q, s, 2]$; $q = [(s^{(t+1)} - 1)/(s-1)]$, obtained from the collection of all points of $PG(t,s)$. The array D is of order $(s^{(t+1)} \times q)$ and each column of D has $s$symbols occurring equally often.

Step II: Replace the s^t occurrences of each of the s symbols in the kth column of D by $s^{(t+1)}$ symbols from the set $s_i = (0, 1, 2, \ldots, (s^{(t+1)} - 1))$ is as follows: For $k = 1, 2, \ldots q$, define

$$t_{kh} = \{hs, hs+1, \ldots, hs + (s^t - 1)\}, \qquad h = 0, 1, 2, \ldots, (s-1) \tag{4}$$

and replace the s^t occurrences of symbol h by the s^t members of t_{kh} in order as obtained in (4), that is, the first occurrence of h is replaced by hs and second occurrence by $hs+1$ and so on.

Step III: Let R denote the $(s^{(t+1)} \times q)$ array obtained from D after changing symbols of D according to (4), so that each column of R is a permutation of $\{0, 1, \ldots, (s^{(t+1)} - 1)\}$ symbols. Let $r(0), r(1), \ldots, r(s^{(t+1)} - 1)$ denote the $s^{(t+1)}$ rows of R.

Step IV: Consider an orthogonal array $A = OA[s^{(r+1)}, p, s^{(t+1)}, 2]$ obtained from $p = [(s^{(r+1)} - 1)/(s^{(t+1)} - 1)]$ disjoint t-flats of $PG(r,s)$. Let $0, 1, \ldots, (s^{(t+1)} - 1)$ denote the symbols in the ith column of orthogonal array $A = [a_{li}]; l = 1, 2, \ldots, s^{(r+1)}$ and $i = 1, 2, \ldots, p$.

Step V: Construct the following arrays using the array A and step III. Write

$$A = [A_1 : A_2 : \ldots : A_p],$$

where $A_i (i = 1, 2, \ldots, p)$ is of order $(s^{r+1} \times 1)$ with s^{t+1} symbols. Replace the s^{t+1} symbols $\{0, 1, 2, \ldots, (s^{t+1}-1)\}$ of A_i by the rows $r(0), r(1), r(2), \ldots\ldots\ldots, r(s^{(t+1)}-1)$ of R respectively and denote it by T_i. Then

1. T_i is of order $s^{(r+1)} \times q$, having rows $r(a_{1i}), r(a_{2i}), ..., r(a_{s^{(r+1)}i})$, for i=1, 2, . . . , p.
2. $T = [T_1 : T_2 : \cdots : T_p]$, of order $s^{(r+1)} \times (pq)$ with symbols 0,1,2,. . . ,$(s^{(t+1)} - 1)$, is the pre-mapping array.

Step VI: For the post mapping array, $s^{(t+1)}$ symbols are mapped to s symbols as follows: For i=1,2,. . . ,p, consider T_i and use reverse mapping of (4) as

$$\{hs, hs + 1, \ldots\ldots, hs + (s^t - 1)\} \to h = 0, 1, 2, \ldots, (s - 1)$$

in each of the q columns of T_i, to get B_i of order $(s^{(r+1)} \times q)$ with s symbols $\{0, 1, \ldots, (s-1)\}$ in every column. Write $B = [B_1 : B_2 : \cdots : B_p]$, then B is the post mapping array $B = OA[s^{(r+1)}, ((s)^q)^p]$. Thus, the mappable nearly orthogonal array

$$MNOA[s^{(r+1)}, \{(s^{(t+1)})^q\}^p, \{(s)^q\}^p]$$

is constructed and we have the following result.

Theorem 2: For given r, s and t where $(t+1)|(r+1)$, an $MNOA[s^{(r+1)}, \{(s^{(t+1)})^q\}^p, \{(s)^q\}^p]$ can always be constructed by using $PG(t, s)$ and p distinct t-flats of $PG(r, s)$.

The method of construction is illustrated through the following examples.

Example 1: Let $t = 1$, $r = 3$ and $s = 2$ in $PG(r, s)$. Using step I, we obtain the array $D = OA[4, 3, 2, 1]$ of order (4×3) as

$$D = \begin{bmatrix} 0 & 0 & 0 \\ 1 & 0 & 1 \\ 0 & 1 & 1 \\ 1 & 1 & 0 \end{bmatrix}$$

Replace two symbols of set $s = (0, 1)$ in each column of D by four symbols of set $s_i = (0, 1, 2, 3)$ as described in step II to obtain R, so that each column of R is a permutation of four symbols $0, 1, 2, 3$. The array R is

$$R = \begin{bmatrix} 0 & 0 & 0 \\ 2 & 1 & 2 \\ 1 & 2 & 3 \\ 3 & 3 & 1 \end{bmatrix}$$

Now, consider the orthogonal array $A = OA[16, 5, 4, 2]$ obtained by using the five disjoint 1-flat of $PG(3, 2)$ as given in step IV. The orthogonal array A is given below:

$$A^T = \begin{bmatrix} 0 & 1 & 2 & 2 & 0 & 3 & 3 & 1 & 0 & 2 & 2 & 1 & 3 & 0 & 3 & 1 \\ 0 & 0 & 1 & 2 & 2 & 1 & 2 & 2 & 3 & 3 & 0 & 3 & 3 & 1 & 0 & 1 \\ 0 & 2 & 2 & 1 & 2 & 0 & 3 & 0 & 3 & 0 & 3 & 1 & 2 & 1 & 1 & 3 \\ 0 & 2 & 0 & 2 & 1 & 2 & 0 & 3 & 2 & 1 & 3 & 0 & 3 & 3 & 1 & 1 \\ 0 & 1 & 3 & 0 & 2 & 2 & 1 & 3 & 3 & 1 & 2 & 2 & 0 & 1 & 3 & 0 \end{bmatrix}$$

Divide the columns of A into 5 groups, each group consisting of a single column denoted by A_i and replace the entries of A_i by rows of R as described in Step V to get T_i for $i = 1, 2, \ldots, 5$ each of order 16×3. The pre-mapping array $T = [T_1 : T_2 : T_3 : T_4 : T_5]$ is given in Table 1.

Table 1: Pre-mapping array using $PG(3, 2)$ and $t = 1$

Group 1	Group 2	Group 3	Group 4	Group 5
0 0 0	0 0 0	0 0 0	0 0 0	0 0 0
2 1 1	0 0 0	1 2 3	1 2 3	2 1 2
1 2 3	2 1 2	1 2 3	0 0 0	3 3 1
1 2 2	1 2 3	2 1 2	1 2 3	0 0 0
0 0 0	1 2 3	1 2 3	2 1 2	1 2 3
3 3 1	2 1 2	0 0 0	1 2 3	1 2 3
3 3 1	1 2 3	3 3 1	0 0 0	2 1 2
2 1 2	1 2 3	0 0 0	3 3 1	3 3 1
0 0 0	3 3 1	3 3 1	1 2 3	3 3 1
1 2 3	3 3 1	0 0 0	2 1 2	2 1 2
1 2 3	0 0 0	3 3 1	3 3 1	1 2 3
2 1 2	3 3 1	2 1 2	0 0 0	1 2 3
3 3 1	3 3 3	1 2 3	3 3 1	0 0 0
0 0 0	2 1 2	2 1 2	3 3 1	2 1 2
3 3 1	0 0 0	2 1 2	2 1 2	3 3 1
2 1 2	2 1 2	3 3 1	2 1 2	0 0 0

For post mapping, map the symbols $(0, 1, 2, 3)$ to $(0, 1)$ using the reverse mapping of (4). The post-mapping array is a symmetric tight orthogonal array $OA[16, (2^3)^5]$ of strength two and it is given in Table 2.

Thus, using projective geometry $PG(3, 2)$, the required mappable nearly orthogonal array $MNOA[16, (4^3)^5, (2^3)^5]$ is constructed, which is mappable to fully symmetric tight orthogonal array of strength two.

Example 2: Let $t = 1,\ r = 3$ and $s = 3$ in $PG(r, s)$. Using step I, we obtain the array $D = OA[9, 4, 3, 2]$ of order (9×4) as

$$D = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 0 & 1 & 1 & 2 \\ 0 & 2 & 2 & 1 \\ 1 & 0 & 1 & 1 \\ 1 & 1 & 2 & 0 \\ 1 & 2 & 0 & 2 \\ 2 & 0 & 2 & 2 \\ 2 & 1 & 0 & 1 \\ 2 & 2 & 1 & 0 \end{bmatrix}$$

Replace three symbols of set $s = (0, 1, 2)$ in each column of D by nine symbols of set $s_i = (0, 1, 2, 3, 4, 5, 6, 7, 8)$ as described in step II to obtain R, so that each column of R is a

Table 2: Post-mapping array using $PG(3,2)$ and $t=1$

Group 1	Group 2	Group 3	Group 4	Group 5
0 0 0	0 0 0	0 0 0	0 0 0	0 0 0
1 0 1	0 0 0	0 1 1	0 1 1	1 0 1
0 1 1	1 0 1	0 1 1	0 0 0	1 1 0
0 1 1	0 1 1	1 0 1	0 1 1	0 0 0
0 0 0	0 1 1	0 1 1	1 0 1	0 1 1
1 1 0	1 0 1	0 0 0	0 1 1	0 1 1
1 1 0	0 1 1	1 1 0	0 0 0	1 0 1
1 0 1	0 1 1	0 0 0	1 1 0	1 1 0
0 0 0	1 1 0	1 1 0	0 1 1	1 1 0
0 1 1	1 1 0	0 0 0	1 0 1	1 0 1
0 1 1	0 0 0	1 1 0	1 1 0	0 1 1
1 0 1	1 1 0	1 0 1	0 0 0	0 1 1
1 1 0	1 1 0	0 1 1	1 1 0	0 0 0
0 0 0	1 0 1	1 0 1	1 1 0	1 0 1
1 1 0	0 0 0	1 0 1	1 0 1	1 1 0
1 0 1	1 0 1	1 1 0	1 0 1	0 0 0

permutation of nine symbols $0, 1, 2, 3, 4, 5, 6, 7, 8$. The array R is

$$R = \begin{bmatrix} 0 & 0 & 0 & 0 \\ 1 & 3 & 3 & 6 \\ 2 & 6 & 6 & 3 \\ 3 & 1 & 4 & 4 \\ 4 & 4 & 6 & 1 \\ 5 & 7 & 1 & 7 \\ 6 & 3 & 8 & 8 \\ 7 & 5 & 2 & 5 \\ 8 & 8 & 5 & 2 \end{bmatrix}$$

Now, consider the orthogonal array $A = OA[81, 10, 9, 2]$ obtained by using the ten disjoint 1-flat of $PG(3, 3)$ as given in step IV. The orthogonal array A is given in Table 6 in Annexure.

Divide the columns of A into 10 groups, each group consisting of a single column denoted by A_i and replace the entries of A_i by rows of R as described in Step V to get T_i for $i = 1, 2, \ldots, 10$ each of order 81×4. The pre-mapping array $T = [T_1 : T_2 : T_3 : T_4 : T_5 : T_6 : T_7 : T_8 : T_9 : T_{10}]$ is given in Table 7 in Annexure.

For post mapping, map the symbols $(0, 1, 2, 3, 4, 5, 6, 7, 8,)$ to $(0, 1, 2)$ using the reverse mapping of (4). The post-mapping array is a symmetric tight orthogonal array $OA[81, (3^4)^{10}]$ of strength two and it is given in Table 8 in Annexure. Thus, using projective geometry $PG(3, 3)$, the required mappable nearly orthogonal array $MNOA[81, (9^4)^{10}, (3^4)^{10}]$ is constructed, which is mappable to fully symmetric tight orthogonal array of strength two.

Similarly, we can construct many more design using our proposed method, some of them are listed in the following tables along with the corresponding values of $\overline{\pi}$.

It may be noted here that all values in the last column of the above tables are obtained

Table 3: Some tight nearly orthogonal arrays based on t-flat and $PG(r, s)$ for $t = 1$

r	s	$D = OA[s^{t+1}, q, s, 2]$	$MNOA[s^{r+1}, \{(s^{t+1})^q)^p\}, \{(s)^q\}^p]$	$\overline{\pi}$
3	2	$D = OA[4, 3, 2, 2]$	$MNOA[16, \{(4)^3\}^5, \{(2)^3\}^5)]^*$	0.8571
5	2	$D = OA[4, 3, 2, 2]$	$MNOA[64, \{(4)^3\}^{21}, \{(2)^3\}^{21})]$	0.9677
7	2	$D = OA[4, 3, 2, 2]$	$MNOA[256, \{(4)^3\}^{85}, \{(2)^3\}^{85})]$	0.9921
3	3	$D = OA[9, 4, 3, 2]$	$MNOA[81, \{(9)^4\}^{10}, \{(3)^4\}^{10})]^*$	0.9230
5	3	$D = OA[9, 4, 3, 2]$	$MNOA[729, \{(9)^4\}^{91}, \{(3)^4\}^{91})]$	0.9890
7	3	$D = OA[9, 4, 3, 2]$	$MNOA[6561, \{(9)^4\}^{820}, \{(3)^4\}^{820})]$	0.9990
3	4	$D = OA[16, 5, 4, 2]$	$MNOA[256, \{(16)^5\}^{17}, \{(4)^5\}^{17})]^*$	0.9523
5	4	$D = OA[16, 5, 4, 2]$	$MNOA[4096, \{(16)^5\}^{273}, \{(4)^5\}^{273})]$	0.9970
7	4	$D = OA[16, 5, 4, 2]$	$MNOA[65536, \{(16)^5\}^{4369}, \{(4)^5\}^{4369})]$	0.9998
3	5	$D = OA[25, 6, 5, 2]$	$MNOA[625, \{(25)^6\}^{26}, \{(5)^6\}^{26})]$	0.9677
5	5	$D = OA[25, 6, 5, 2]$	$MNOA[15625, \{(25)^6\}^{651}, \{(5)^6\}^{651})]$	0.9987
3	9	$D = OA[81, 10, 9, 2]$	$MNOA[6561, \{(81)^{10}\}^{82}, \{(9)^{10}\}^{82})]$	0.9890
5	9	$D = OA[81, 10, 9, 2]$	$MNOA[531441, \{(81)^{10}\}^{6643}, \{(9)^{10}\}^{6643})]$	0.9998
3	25	$D = OA[125, 6, 5, 2]$	$MNOA[15625, \{(125)^6\}^{126}, \{(25)^6\}^{126})]$	0.9923

Table 4: Some tight nearly orthogonal arrays based on t-flat and $PG(r, s)$ for $t = 2$

r	s	$D = OA[s^{t+1}, q, s, 2]$	$MNOA[s^{r+1}, \{(s^{t+1})^q)^p\}, \{(s)^q\}^p]$	$\overline{\pi}$
5	2	$D = OA[8, 7, 2, 2]$	$MNOA[64, \{(8)^7\}^9, \{(2)^7\}^9)]^*$	0.9032
8	2	$D = OA[8, 7, 2, 2]$	$MNOA[512, \{(8)^7\}^{73}, \{(2)^7\}^{73})]$	0.9882
11	2	$D = OA[8, 7, 2, 2]$	$MNOA[4096, \{(8)^7\}^{585}, \{(2)^7\}^{585})]$	0.9985
5	3	$D = OA[27, 13, 3, 2]$	$MNOA[729, \{(27)^{13}\}^{28}, \{(3)^{13}\}^{28})]$	0.9669
8	3	$D = OA[27, 13, 3, 2]$	$MNOA[19683, \{(27)^{13}\}^{757}, \{(3)^{13}\}^{757})]$	0.9987
5	4	$D = OA[64, 21, 4, 2]$	$MNOA[4096, \{(64)^{21}\}^{65}, \{(4)^{21}\}^{65})]$	0.9853
8	4	$D = OA[64, 21, 4, 2]$	$MNOA[262144, \{(64)^{21}\}^{4161}, \{(4)^{21}\}^{4161})]$	0.9997
5	5	$D = OA[125, 31, 5, 2]$	$MNOA[15625, \{(125)^{31}\}^{126}, \{(5)^{31}\}^{126})]$	0.9920
5	9	$D = OA[729, 91, 9, 2]$	$MNOA[531441, \{(729)^{91}\}^{730}, \{(9)^{91}\}^{730})]$	0.9986

by using equation (3) and the MNOAs marked with $*$ are same as those obtained by Mukerjee *et al.* (2014).

4. Conclusion

In this paper, a method is proposed to construct, mappable nearly orthogonal arrays (MNOAs) using projective geometry. The constructed MNOAs are mappable to tight orthogonal arrays of strength two. It is observed that some new MNOAs are constructed with higher values of degree of orthogonality $\overline{\pi}$ and are therefore useful as better space filling designs.

Table 5: Some tight nearly orthogonal arrays based on t-flat and $PG(r,s)$ for $t=3$

r	s	$D=OA[s^{t+1},q,s,2]$	$MNOA[s^{r+1},\{(s^{t+1})^q)^p\},\{(s)^q\}^p]$	$\overline{\pi}$
7	2	$D=OA[16,15,2,2]$	$MNOA[256,\{(16)^{15}\}^{17},\{(2)^{15}\}^{17})]^*$	0.9448
11	2	$D=OA[16,15,2,2]$	$MNOA[4096,\{(16)^{15}\}^{273},\{(2)^{15}\}^{273})]$	0.9965
7	3	$D=OA[81,40,3,2]$	$MNOA[6561,\{(81)^{40}\}^{82},\{(3)^{40}\}^{82})]$	0.9881

Acknowledgements

We would like to thank the referee for valuable suggestions which has helped in the improvement of this paper.

References

Bose, R. C. and Bush, K. A. (1952). Orthogonal arrays of strength two and three. *Annals of Mathematical Statistics*, **23**, 508-524.

Hedayat, A. S., Sloane, N. J. A., and Stufken, J. (1999). *Orthogonal Arrays: Theory and Applications*. Springer, New York.

Li, W., Liu, M. Q., and Yang, J. F. (2023). Several new classes of space-filling designs. *Statistical Papers*, **65**, 357-379.

Mukerjee, R., Sun, F., and Tang, B. (2014). Nearly orthogonal arrays mappable into fully orthogonal arrays. *Biometrika*, **101**, 957-963.

Nguyen, N. K. (1996). A note on the construction of near-orthogonal arrays with mixed levels and economic run size. *Technometrics*, **38(3)**, 279-283.

Rao, C. R. (1946). Hypercubes of strength d, leading to confounded designs in factorial experiments. *Bulletin of the Calcutta Mathematical Society*, **38**, 67-78.

Rao, C. R. (1947). Factorial experiments derivable from combinatorial arrangements of arrays. *Supplement to the Journal of the Royal Statistical Society*, **9**, 128-139.

Raghavarao, D. (1971). *Constructions and Combinatorial Problems in Designs of Experiments*. John Wiley, New York.

Singh, P., Mazumder, M. D., and Babu, S. (2023a). Construction of nearly orthogonal arrays mappable into fully orthogonal arrays of strength two and three. *International Journal of Mathematics and Statistics*, **24**, 37- 50.

Singh, P., Mazumder, M. D., and Babu, S. (2023b). Mappable nearly orthogonal arrays using projective geometry. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, **14(03)**, 454-464.

Wang, J. C., and Wu, C. F. J. (1992). Nearly orthogonal arrays with mixed levels and small runs. *Technometrics*, **34(4)**, 409-422.

ANNEXURE

Table 6: Orthogonal array $A = [81, 10, 9, 2]$

Rows 1 to 27										Rows 28 to 54										Rows 55 to 81									
0	0	0	0	0	0	0	0	0	0	3	0	3	3	3	3	3	3	3	3	6	0	6	6	6	6	6	6	6	6
0	1	1	2	3	4	5	6	7	8	3	1	8	4	7	5	2	1	0	6	6	1	4	0	1	3	7	2	8	5
0	2	2	3	4	5	6	7	8	1	3	2	4	7	5	2	1	0	6	8	6	2	0	1	3	7	2	8	5	4
0	3	3	4	5	6	7	8	1	2	3	3	7	5	2	1	0	6	8	4	6	3	1	3	7	2	8	5	4	0
0	4	4	5	6	7	8	1	2	3	3	4	5	2	1	0	6	8	4	7	6	4	3	7	2	8	5	4	0	1
0	5	5	6	7	8	1	2	3	4	3	5	2	1	0	6	8	4	7	5	6	5	7	2	8	5	4	0	1	3
0	6	6	7	8	1	2	3	4	5	3	6	1	0	6	8	4	7	5	2	6	6	2	8	5	4	0	1	3	7
0	7	7	8	1	2	3	4	5	6	3	7	0	6	8	4	7	5	2	1	6	7	8	5	4	0	1	3	7	2
0	8	8	1	2	3	4	5	6	7	3	8	6	8	4	7	5	2	1	0	6	8	5	4	0	1	3	7	2	8
1	0	1	1	1	1	1	1	1	1	4	0	4	4	4	4	4	4	4	4	7	0	7	7	7	7	7	7	7	7
1	1	5	3	8	7	0	4	6	2	4	1	7	1	5	8	6	3	2	0	7	1	6	5	0	2	4	8	3	1
1	2	3	8	7	0	4	6	2	5	4	2	1	5	8	6	3	2	0	7	7	2	5	0	2	4	8	3	1	6
1	3	8	7	0	4	6	2	5	3	4	3	5	8	6	3	2	0	7	1	7	3	0	2	4	8	3	1	6	5
1	4	7	0	4	6	2	5	3	8	4	4	8	6	3	2	0	7	1	5	7	4	2	4	8	3	1	6	5	0
1	5	0	4	6	2	5	3	8	7	4	5	6	3	2	0	7	1	5	8	7	5	4	8	3	1	6	5	0	2
1	6	4	6	2	5	3	8	7	0	4	6	3	2	0	7	1	5	8	6	7	6	8	3	1	6	5	0	2	4
1	7	6	2	5	3	8	7	0	4	4	7	2	0	7	1	5	8	6	3	7	7	3	1	6	5	0	2	4	8
1	8	2	5	3	8	7	0	4	6	4	8	0	7	1	5	8	6	3	2	7	8	1	6	5	0	2	4	8	3
2	0	2	2	2	2	2	2	2	2	5	0	5	5	5	5	5	5	5	5	8	0	8	8	8	8	8	8	8	8
2	1	3	6	4	1	8	0	5	7	5	1	0	8	2	6	1	7	4	3	8	1	2	7	6	0	3	5	1	4
2	2	6	4	1	8	0	5	7	3	5	2	8	2	6	1	7	4	3	0	8	2	7	6	0	3	5	1	4	2
2	3	4	1	8	0	5	7	3	6	5	3	2	6	1	7	4	3	0	8	8	3	6	0	3	5	1	4	2	7
2	4	1	8	0	5	7	3	6	4	5	4	6	1	7	4	3	0	8	2	8	4	0	3	5	1	4	2	7	6
2	5	8	0	5	7	3	6	4	1	5	5	1	7	4	3	0	8	2	6	8	5	3	5	1	4	2	7	6	0
2	6	0	5	7	3	6	4	1	8	5	6	7	4	3	0	8	2	6	1	8	6	5	1	4	2	7	6	0	3
2	7	5	7	3	6	4	1	8	0	5	7	4	3	0	8	2	6	1	7	8	7	1	4	2	7	6	0	3	5
2	8	7	3	6	4	1	8	0	5	5	8	3	0	8	2	6	1	7	4	8	8	4	2	7	6	0	3	5	1

Table 7: Pre-mapping array using $PG(3,3)$and $t=1$

Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group7	Group 8	Group 9	Group 10
0000	0000	0000	0000	0000	0000	0000	0000	0000	0000
0000	1336	1336	2663	3144	4461	6388	6388	7525	8852
0000	2663	2663	3144	4461	5717	7525	7525	8852	1336
0000	3144	3144	4461	5717	6388	8852	8852	1336	2663
0000	4461	4461	5717	6388	7525	1336	1336	2663	3144
0000	5717	5717	6388	7525	8852	2663	2663	3144	4461
0000	6388	6388	7525	8852	1336	3144	3144	4461	5717
0000	7525	7525	8852	1336	2663	4461	4461	5717	6388
0000	8852	8852	1336	2663	3144	5717	5717	6388	7525
1336	0000	1336	1336	1336	1336	1336	1336	1336	1336
1336	1336	5763	3144	8852	7525	4461	4461	6388	2663
1336	2663	3144	8852	7525	0000	6388	6388	2663	5717
1336	3144	8852	7525	0000	4461	2663	2663	5717	3144
1336	4461	7525	0000	4461	6388	5717	5717	3144	8852
1336	5717	0000	4461	6388	2663	3144	3144	8852	7525
1336	6388	4461	6388	2663	5717	8852	8852	7525	0000
1336	7525	6388	2663	5717	3144	7525	7525	0000	4461
1336	8852	2663	5717	3144	8852	0000	0000	4461	6388
2663	0000	2663	2663	2663	2663	2663	2663	2663	2663
2663	1336	3144	6388	4461	1336	0000	0000	5717	7525
2663	2663	6388	4461	1336	8852	5717	5717	7525	3144
2663	3144	4461	1336	8852	0000	7525	7525	3144	6388
2663	4461	1336	8852	0000	5717	3144	3144	6388	4461
2663	5717	8852	0000	5717	7525	6388	6388	4461	1336
2663	6388	0000	5717	7525	3144	4461	4461	1336	8852
2663	7525	5717	7525	3144	6388	1336	1336	8852	0000
2663	8852	7525	3144	6388	4461	8852	8852	0000	5717

Table 7: Continued

Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group7	Group 8	Group 9	Group 10
3144	0000	3144	3144	3144	3144	3144	3144	3144	3144
3144	1336	8852	4461	7525	5717	2663	1336	0000	6388
3144	2663	4461	7525	5717	2663	1336	0000	6388	8852
3144	3144	7525	5717	2663	1336	0000	6388	8852	4461
3144	4461	5717	2663	1336	0000	6388	8852	4461	7525
3144	5717	2663	1336	0000	6388	8852	4461	7525	5717
3144	6388	1336	0000	6388	8852	4461	7525	5717	2663
3144	7525	0000	6388	8852	4461	7525	5717	2663	1336
3144	8852	6388	8852	4461	7525	5717	2663	1336	0000
4461	0000	4461	4461	4461	4461	4461	4461	4461	4461
4461	1336	7525	1336	5717	8852	6388	3144	2663	0000
4461	2663	1336	5717	8852	6388	3144	2663	0000	7525
4461	3144	5717	8852	6388	3144	2663	0000	7525	1336
4461	4461	8852	6388	3144	2663	0000	7525	1336	5717
4461	5717	6388	3144	2663	0000	7525	1336	5717	8852
4461	6388	3144	2663	0000	7525	1336	5717	8852	6388
4461	7525	2663	0000	7525	1336	5717	8852	6388	3144
4461	8852	0000	7525	1336	5717	8852	6388	3144	2663
5717	0000	5717	5717	5717	5717	5717	5717	5717	5717
5717	1336	0000	8852	2663	6388	1336	7525	4461	3144
5717	2663	8852	2663	6388	1336	7525	4461	3144	0000
5717	3144	2663	6388	1336	7525	4461	3144	0000	8852
5717	4461	6388	1336	7525	4461	3144	0000	8852	2663
5717	5717	1336	7525	4461	3144	0000	8852	2663	6388
5717	6388	7525	4461	3144	0000	8852	2663	6388	1336
5717	7525	4461	3144	0000	8852	2663	6388	1336	7525
5717	8852	3144	0000	8852	2663	6388	1336	7525	6388

Table 7: Continued

Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group7	Group 8	Group 9	Group 10
6388	0000	6388	6388	6388	6388	6388	6388	6388	6388
6388	1336	4461	0000	1336	3144	7525	2663	8852	5717
6388	2663	0000	1336	3144	7525	2663	8852	5717	4461
6388	3144	1336	3144	7525	2663	8852	5717	4461	0000
6388	4461	3144	7525	2663	8852	5717	4461	0000	1336
6388	5717	7525	2663	8852	5717	4461	0000	1336	3144
6388	6388	2663	8852	5717	4461	0000	1336	3144	7525
6388	7525	8852	5717	4461	0000	1336	3144	7525	2663
6388	8852	5717	4461	0000	1336	3144	7525	2663	8852
7525	0000	7525	7525	7525	7525	7525	7525	7525	7525
7525	1336	6388	5717	0000	2663	4461	8852	3144	1336
7525	2663	5717	0000	2663	4461	8852	3144	1336	6388
7525	3144	0000	2663	4461	8852	3144	1336	6388	5717
7525	4461	2663	4461	8852	3144	1336	6388	5717	0000
7525	5717	4461	8852	3144	1336	6388	5717	0000	2663
7525	6388	8852	3144	1336	6388	5717	0000	2663	4461
7525	7525	3144	1336	6388	5717	0000	2663	4461	8852
7525	8852	1336	6388	5717	0000	2663	4461	8852	3144
8852	0000	8852	8852	8852	8852	8852	8852	8852	8852
8852	1336	2663	7525	6388	0000	3144	5717	1336	4461
8852	2663	7525	6388	0000	3144	5717	1336	4461	2663
8852	3144	6388	0000	3144	5717	1336	4461	2663	7525
8852	4461	0000	3144	5717	1336	4461	2663	7525	6388
8852	5717	3144	5717	1336	4461	2663	7525	6388	0000
8852	6388	5717	1336	4461	2663	7525	6388	0000	3144
8852	7525	1336	4461	2663	7525	6388	0000	3144	5717
8852	8852	4461	2663	7525	6388	0000	3144	5717	1336

Table 8: Post-mapping array using $PG(3,3)$ and $t = 1$

Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group7	Group 8	Group 9	Group 10
0000	0000	0000	0000	0000	0000	0000	0000	0000	0000
0000	0112	0112	0221	1011	1120	2122	2122	2101	2210
0000	0221	0221	1011	1120	1202	2101	2101	2210	0112
0000	1011	1011	1120	1202	2122	2210	2210	0112	0221
0000	1120	1120	1202	2122	2101	0112	0112	0221	1011
0000	1202	1202	2112	2101	2210	0221	0221	1011	1120
0000	2122	2122	2101	2210	0112	1011	1011	1120	1202
0000	2101	2101	2210	0112	0221	1120	1120	1202	2122
0000	2210	2210	0112	0221	1011	1202	1202	2122	2101
0112	0000	0112	0112	0112	0112	0112	0112	0112	0112
0112	0112	1221	1011	2210	2101	1120	1120	2122	0221
0112	0221	1011	2210	2101	0000	2122	2122	0221	1202
0112	1011	2210	2101	0000	1120	0221	0221	1202	1011
0112	1120	2101	0000	1120	2122	1202	1202	1011	2210
0112	1202	0000	1120	2122	0221	1011	1011	2210	2101
0112	2122	1120	2122	0221	1202	2210	2210	2101	0000
0112	2101	2122	0221	1202	1011	2101	2101	0000	1120
0112	2210	0221	1202	1011	2210	0000	0000	1120	2122
0221	0000	0221	0221	0221	0221	0221	0221	0221	0221
0221	0112	1011	2122	1120	0112	0000	0000	1202	2101
0221	0221	2122	1120	0112	2210	1202	1202	2101	1011
0221	1011	1120	0112	2210	0000	2101	2101	1011	2122
0221	1120	0112	2210	0000	1202	1011	1011	2122	1120
0221	1202	2210	0000	1202	2101	2122	2122	1120	0112
0221	2122	0000	1202	2101	1011	1120	1120	0112	2101
0221	2101	1202	2101	1011	2122	0112	0112	2210	0000
0221	2210	2101	1011	2122	1120	2210	2210	0000	1202

Table 8: Continued

Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group7	Group 8	Group 9	Group 10
1011	0000	1011	1011	1011	1011	1011	1011	1011	1011
1011	0112	2210	1120	2101	1202	0221	0112	0000	2122
1011	0221	1120	2101	1202	0221	0112	0000	2122	2210
1011	1011	2101	1202	0221	0112	0000	2122	2210	1120
1011	1120	1202	0221	0112	0000	2122	2210	1120	2101
1011	1202	0221	0112	0000	2122	2210	1120	2101	1202
1011	2122	0112	0000	2122	2210	1120	2101	1202	0221
1011	2101	0000	2122	2210	1120	2101	1202	0221	0112
1011	2210	2122	2210	1120	2101	1202	0221	0112	0000
1120	0000	1121	1121	1121	1121	1121	1121	1121	1121
1120	0112	2101	0112	1202	2210	2122	1011	0221	0000
1120	0221	0112	1202	2210	2122	1011	0221	0000	2101
1120	1011	1202	2210	2122	1011	0221	0000	2101	0112
1120	1120	2210	2122	1011	0221	0000	2101	0112	1202
1120	1202	2122	1011	0221	0000	2101	0112	1202	2210
1120	2122	1011	0221	0000	2101	0112	1202	2210	2122
1120	2101	0221	0000	2101	0112	1202	2210	2122	1011
1120	2210	0000	2101	0112	1202	2210	2122	1011	0221
1202	0000	1202	1202	1202	1202	1202	1202	1202	1202
1202	0112	0000	2210	0221	2122	0112	2101	1120	1011
1202	0221	2210	0221	2122	0112	2101	1120	1011	0000
1202	1011	0221	2122	0112	2101	1120	1011	0000	2210
1202	1120	2122	0112	2101	1120	1011	0000	2210	0221
1202	1202	0112	2101	1120	1011	0000	2210	0221	2122
1202	2122	2101	1120	1011	0000	2210	0221	2122	0112
1202	2101	1120	1011	0000	2210	0221	2122	0112	2101
1202	2210	1011	0000	2210	0221	2122	0112	2101	2122

Table 8: Continued

Group 1	Group 2	Group 3	Group 4	Group 5	Group 6	Group7	Group 8	Group 9	Group 10
2122	0000	2122	2122	2122	2122	2122	2122	2122	2122
2122	0112	1120	0000	0112	1011	2101	0221	2210	1202
2122	0221	0000	0112	1011	2101	0221	2210	1202	1120
2122	1011	0112	1011	2101	0221	2210	1202	1120	0000
2122	1121	1011	2101	0221	2210	1202	1120	0000	0112
2122	1202	2101	0221	2210	1202	1120	0000	0112	1011
2122	2122	0221	2210	1202	1120	0000	0112	1011	2101
2122	2101	2210	1202	1120	0000	0112	1011	2101	0221
2122	2210	1202	1120	0000	0112	1011	2101	0221	2210
2101	0000	2101	2101	2101	2101	2101	2101	2101	2101
2101	0112	2122	1202	0000	0221	1120	2210	1011	0112
2101	0221	1202	0000	0221	1120	2210	1011	0112	2122
2101	1011	0000	0221	1120	2210	1011	0112	2122	1202
2101	1120	0221	1120	2210	1011	0112	2122	1202	0000
2101	1202	1120	2210	1011	0112	2122	1202	0000	0221
2101	2122	2210	1011	0112	2122	1202	0000	0221	1120
2101	2101	1011	0112	2122	1202	0000	0221	1120	2210
2101	2210	0112	2122	1202	0000	0221	1120	2210	1011
2210	0000	2210	2210	2210	2210	2210	2210	2210	2210
2210	0112	0221	2101	2122	0000	1011	1202	0112	1120
2210	0221	2101	2122	0000	1011	1202	0112	1120	0221
2210	1011	2122	0000	1011	1202	0112	1120	0221	2101
2210	1120	0000	1011	1202	0112	1120	0221	2101	2122
2210	1202	1011	1202	0112	1120	0221	2101	2122	0000
2210	2122	1202	0112	1120	0221	2101	2122	0000	1011
2210	2101	0112	1120	0221	2101	2122	0000	1011	1202
2210	2210	1120	0221	2101	2122	0000	1011	1202	0112

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 22, No. 2, 2024 (New Series), pp 323–359
http://www.ssca.org.in/journal

Power Generalized DUS Transformation of Inverse Kumaraswamy Distribution and Stress-Strength Analysis

Amrutha M. and V. M. Chacko
Department of Statistics
St. Thomas College (Autonomous) Thrissur, University of Calicut, Kerala, India-680001.

Received: 17 December 2023; Revised: 21 March 2024; Accepted: 30 March 2024

Abstract

Reliability analysis, including stress-strength analysis, for given data is more widely used in the reliability literature. A large number of new distributions are available, but many of them are not showing a good fit for the data under consideration. This inspires a researcher to introduce new lifetime distributions that demonstrate superior fitness in comparison to the existing distributions. So that more accurate reliability estimates can be obtained for the given data. The DUS transformation technique is widely used in reliability literature to create better models. Power generalized DUS(PGDUS) transformation to lifetime distributions, which is found to be useful to introduce more appropriate flexible distributions for the given data. Vinyl chloride data obtained from clean upgrading and monitoring wells in mg/L have been analyzed using DUS inverse Kumaraswamy (DUS IK), inverse Kumaraswamy (IK), and Weibull distributions. As a substitute for these distributions, this paper presents a new lifetime distribution employing PGDUS transformation, utilizing the inverse Kumaraswamy distribution as the baseline. The statistical properties of the proposed distribution are derived. The parameters of the proposed distribution are estimated using the maximum likelihood (ML) method, maximum product spacing (MPS), method of moment, and method of least squares. Additionally, Bayesian parameter estimates are acquired utilizing Lindley's approximation and the Metropolis-Hastings algorithm. The consistency of the model is verified using mean squared error (MSE) and biases, which are obtained based on simulated values. Then, the proposed distribution is compared with the DUS-IK, IK, and Weibull distributions. In this paper, single-component and multi-component stress-strength reliability analyses are also conducted.

Key words: PGDUS transformation; inverse-Kumaraswamy distribution; Stress-strength reliability.

1. Introduction

An appropriate lifetime distribution is essential to conducting reliability analysis with maximum accuracy. While using existing distributions, the fitness of the distributions for the

Corresponding Author: Dr. V. M. Chacko
Email: chackovm@gmail.com

given data is sometimes low. To overcome these problems, several researchers introduced new distributions with more fitness characteristics. Appropriate distributions are necessary for the stress-strength analysis in statistics and reliability engineering. Reliability distributions have different failure rate properties, like increasing failure rate, decreasing failure rate, bathtub and upside-down bathtub distributions, *etc.*

There are numerous ways to suggest new distributions in the statistical literature by using some baseline distributions without incorporating scale, shape, or location parameters, so that more appropriate statistical distributions can be made available in the statistical literature. DUS transformation is one of several methods (see Kumar *et al.* (2015)). Generalizing this DUS transformation will lead to the introduction of new distributions, which could be used while dealing with reliability analysis of parallel systems with components having DUS-transformed distributions.

Kumaraswamy (1980) introduced the Kumaraswamy distribution, which is also known as a beta-like distribution due to its similarity with the beta distribution in the sense that both have the same basic shape parameter. But the probability density function (pdf), cumulative distribution function (CDF), and quantile function are in closed form, which makes Kumaraswamy distribution a more practical choice for many applications, including modeling of biomedical data, reliability engineering, finance, hydrology, *etc.* (see Kumaraswamy (1976)) over Beta distribution.

Nowadays, many researchers focus on the inverse transformation of probability distributions and their applications, which proves the increase in model flexibility. Abd Al-Fattah *et al.* (2017) introduced the inverted Kumaraswamy (IK) distribution by introducing a transformation

$$U = \frac{1 - X}{X},$$

where $X \sim$ Kumaraswamy(α, β).
Iqbal (2017) generalized the IK distribution using a power transformation as

$$T = U^{\gamma},$$

where $U \sim$ IK distribution, called generalized IK distribution. All monotonic and non-monotonic failure rate patterns exhibits for this model. Jamal *et al.* (2019) proposed a new generator function based on the IK distribution and introduced a generalized IK-G family of distributions.

The DUS transformation approach was proposed by Kumar *et al.* (2015), utilizing a few baseline distributions that are sparse in computation and interpretation since they only ever contain the parameter(s) included in the baseline distribution. Let $h(u)$ and $H(u)$ be the pdf and CDF of the baseline distribution, then the pdf $g(u)$ and CDF $G(u)$ of the distribution obtained by the DUS transformation of the baseline distribution are given by

$$g(u) = \frac{1}{e-1} h(u) e^{H(u)}$$
$$G(u) = \frac{1}{e-1} \left(e^{H(u)} - 1 \right)$$

Maurya *et al.* (2016) proposed the DUS transformation of the Lindley distribution, and Tripati *et al.* (2019) introduced the DUS transformation of the exponential distribution. Deepthi and Chacko (2020) introduced the DUS transformation of the Lomax distribution, which is an upside-down bathtub-shaped failure rate model. Gauthami and Chacko (2021) proposed the DUS inverse-Weibull distribution, which is also an upside-down bathtub-shaped failure rate model. Anakha and Chacko (2022) introduced a non-monotonic hazard rate distribution using the DUS transformation with the IK distribution as the baseline distribution.

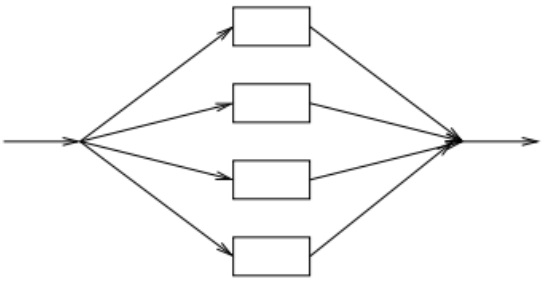


Figure 1: Parallel system

While considering a parallel system, for example, as shown in Figure 1, where each component is distributed to any DUS-transformed baseline distribution. Then the resulting distribution of parallel systems has to be investigated in detail. In order to address this problem, Thomas and Chacko (2021) introduced a method called exponentiation of DUS transformation, called PGDUS transformation, and introduced the PGDUS-Exponential distribution with exponential as the baseline distribution. Weibull and Lomax distributions are used by Thomas and Chacko (2023) to introduce new distributions using PGDUS transformation.

This paper introduces a new lifetime distribution for a system with components connected in parallel in which each of the components follows the DUS transformation of the IK distribution to study the distributions having monotone and non-monotone failure rate functions.

Consider a random variable U with pdf $h(u)$ and CDF $H(u)$. Then the pdf $q(u)$ and CDF $Q(u)$ of the PGDUS-IK(α, β, λ) distribution can be obtained as

$$q(u) = \frac{\lambda}{(e-1)^{\lambda}}(e^{H(u)} - 1)^{\lambda-1} e^{H(u)} h(u), \qquad \lambda > 0, \quad u > 0 \tag{1}$$

and

$$Q(u) = \left(\frac{e^{H(u)} - 1}{e - 1}\right)^{\lambda}, \qquad \lambda > 0, \quad u > 0. \tag{2}$$

respectively.

Similarly, the failure rate function of the PGDUS-IK distribution can be written as

$$r(u) = \frac{\lambda h(u) e^{H(u)} (e^{H(u)} - 1)^{\lambda - 1}}{(e-1)^{\lambda} - (e^{H(u)} - 1)^{\lambda}}, \qquad \lambda > 0, \quad u > 0. \tag{3}$$

This paper is divided into 10 sections: the PGDUS-IK distribution is proposed in Section 2. In Section 3, a detailed investigation into the properties of the PGDUS-IK distribution is undertaken. Section 4 discusses the mean residual life function of the PGDUS-IK distribution. In Section 5, the estimation of parameters for the proposed distribution has been done using the methods of maximum likelihood (ML), maximum product spacing (MPS), moments, and least squares. Also, bayesian estimators of α, β, and λ based on the squared error loss function, by taking gamma priors, are derived. The asymptotic confidence interval and bootstrap confidence interval for the unknown parameters of PGDUS-IK are derived in Section 6. The efficacy of the proposed estimators is investigated in terms of their bias and mean squared error (MSE) values in Section 7. Section 8 illustrates the applications of proposed estimators using the vinyl chloride data given in Bhaumik *et al.* (2009). In Section 9, stress-strength reliability for single components and for multi-components for the proposed distribution is investigated. A Simulation study to investigate and compare the performance of the reliability estimators is conducted, and data analysis for estimating single component and multi-component reliability is given, in the same section. Conclusions are provided in Section 10.

2. Power generalized DUS transformation of inverse-Kumaraswamy distribution

Kumaraswamy (1980) introduced the Kumaraswamy (K) distribution, which is empirically useful for a wide range of reliability applications. The pdf of the K distribution is given as

$$f(y; \alpha, \beta) = \alpha\beta y^{\alpha - 1} (1 - y^{\alpha})^{\beta - 1}, \qquad 0 < y < 1, \quad \alpha > 0, \quad \beta > 0. \tag{4}$$

IK distribution has the following pdf, CDF, and failure rate function

$$h(u) = \alpha\beta (1+u)^{-(\alpha+1)} (1 - (1+u)^{-\alpha})^{\beta - 1}, \qquad u > 0\ , \quad \alpha > 0, \quad \beta > 0, \tag{5}$$

$$H(u; \alpha, \beta) = (1 - (1+u)^{-\alpha})^{\beta}, \qquad u > 0\ , \quad \alpha > 0, \quad \beta > 0, \tag{6}$$

and

$$r(u) = \frac{\alpha\beta (1+u)^{-(\alpha+1)} (1 - (1+u)^{-\alpha})^{\beta}}{1 - (1 - (1+u)^{-\alpha})^{\beta}}, \qquad u > 0, \quad \alpha > 0, \quad \beta > 0 \tag{7}$$

respectively.

The DUS-IK distribution with pdf and CDF can be defined as

$$g(u) = \frac{\alpha\beta}{e-1} (1+u)^{-(\alpha+1)} \Big(1 - (1+u)^{-\alpha}\Big)^{\beta - 1} e^{(1 - (1+u)^{-\alpha})^{\beta}}, \quad u > 0, \quad \alpha > 0, \quad \beta > 0, \tag{8}$$

and

$$G(u) = \frac{e^{(1 - (1+u)^{-\alpha})^{\beta}} - 1}{e - 1}, \quad u > 0, \quad \alpha > 0, \quad \beta > 0 \tag{9}$$

respectively. The survival function will be

$$\bar{G}(u) = \frac{e - e^{(1-(1+u)^{-\alpha})^{\beta}}}{e-1}, \quad \text{for} \quad u > 0,\ \alpha > 0, \quad \beta > 0. \tag{10}$$

PGDUS-IK(α, β, λ)

By using the PGDUS transformation to IK distribution, the pdf, CDF, and failure rate functions can be written as

$$q(u) = \frac{\alpha\beta\lambda}{(e-1)^{\lambda}}(1+u)^{-(\alpha+1)}(1-(1+u)^{-\alpha})^{\beta-1}e^{(1-(1+u)^{-\alpha})^{\beta}}(e^{(1-(1+u)^{-\alpha})^{\beta}}-1)^{\lambda-1}, \tag{11}$$

$$Q(u) = \left(\frac{e^{(1-(1+u)^{-\alpha})^{\beta}}-1}{e-1}\right)^{\lambda}, \tag{12}$$

and

$$r(u) = \frac{\alpha\beta\lambda(1+u)^{-(\alpha+1)}(1-(1+u)^{-\alpha})^{\beta-1}e^{(1-(1+u)^{-\alpha})^{\beta}}(e^{(1-(1+u)^{-\alpha})^{\beta}}-1)^{\lambda-1}}{(e-1)^{\lambda}-(e^{(1-(1+u)^{-\alpha})^{\beta}}-1)^{\lambda}} \tag{13}$$

respectively, where, $u > 0, \quad \lambda > 0, \quad \alpha > 0, \quad \beta > 0.$

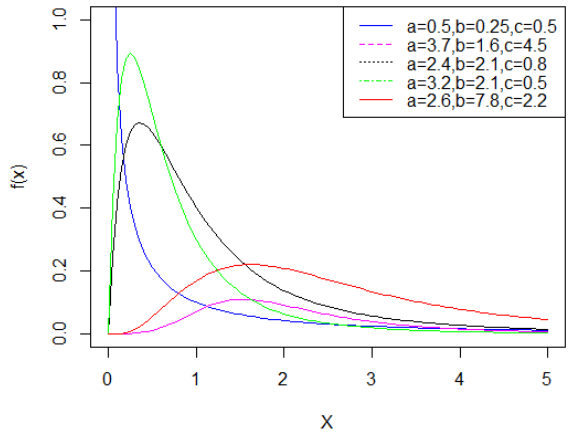


Figure 2: pdf plot for PGDUS-IK distribution

The PGDUS-IK(α, β, λ) distribution has both monotonic and non-monotonic hazard rates.

3. Statistical properties

Statistical properties are discussed in this section.

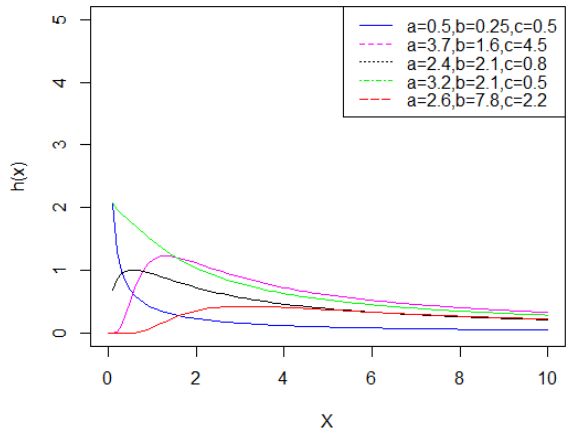


Figure 3: Failure rate plot of PGDUS-IK distribution

3.1. Moments

The rth raw moment of the PGDUS-IK(α, β, λ) distribution can be derived as follows

$$\mu_r^1 = E(U^r)$$
$$= \int_0^\infty u^r \frac{\alpha\beta\lambda}{(e-1)^\lambda}(1+u)^{-(\alpha+1)}(1-(1+u)^{-\alpha})^{\beta-1}e^{(1-(1+u)^{-\alpha})^\beta}(e^{(1-(1+u)^{-\alpha})^\beta}-1)^{\lambda-1}du$$

Substitute $a = (1-(1+u)^{-\alpha})^\beta$ in the above integral,

$$\mu_r^1 = \int_0^1 \frac{\lambda}{(e-1)^\lambda} e^a (e^a-1)^{\lambda-1}\left(\left(1-a^{\frac{1}{\beta}}\right)^{-\frac{1}{\alpha}}-1\right)^r da$$
$$= \frac{\lambda}{(e-1)^\lambda}\int_0^1 e^a \sum_{s=0}^{\lambda-1}(-1)^s \binom{\lambda-1}{s}(e^a)^{\lambda-s-1}\sum_{z=0}^{r}(-1)^{r-z}\binom{r}{z}\left(\left(1-a^{\frac{1}{\beta}}\right)^{-\frac{1}{\alpha}}\right)^s da$$
$$= \frac{\lambda}{(e-1)^\lambda}\sum_{s=0}^{\lambda-1}\sum_{z=0}^{r}(-1)^{s+r-z}\binom{\lambda-1}{s}\binom{r}{z}\int_0^1\left(1-a^{\frac{1}{\beta}}\right)^{-\frac{s}{\alpha}}e^{a(\lambda-s)}da$$
$$= \frac{\beta\lambda}{(e-1)^\lambda}\sum_{s=0}^{\lambda-1}\sum_{z=0}^{r}\sum_{l=0}^{\infty}\frac{(-1)^{s+r-z+l}}{l!}\binom{\lambda-1}{s}\binom{r}{z}(\lambda-s)^l\int_0^1 a^l\left(1-a^{\frac{1}{\beta}}\right)^{-\frac{s}{\alpha}}da$$
$$= \frac{\beta\lambda}{(e-1)^\lambda}\sum_{s=0}^{\lambda-1}\sum_{z=0}^{r}\sum_{l=0}^{\infty}\frac{(-1)^{s+r-z+l}}{l!}\binom{\lambda-1}{s}\binom{r}{z}(\lambda-s)^l\beta\left(1-\frac{s}{\alpha},\beta l+\beta\right). \quad (14)$$

By putting $r = 1, 2, ...,$ we get the corresponding raw moments as

$$\mu_1^1 = \frac{\beta\lambda}{(e-1)^\lambda}\sum_{s=0}^{\lambda-1}\sum_{l=0}^{\infty}\{(-1)^{s+l+1}+(-1)^{s+l}\}\binom{\lambda-1}{s}\frac{(\lambda-s)^l}{l!}\beta\left(1-\frac{s}{\alpha},\beta l+\beta\right). \quad (15)$$

$$\mu_2^1 = \frac{\beta\lambda}{(e-1)^\lambda}\sum_{s=0}^{\lambda-1}\sum_{l=0}^{\infty}\{(-1)^{s+l+2}+2.(-1)^{s+l+1}+(-1)^{s+l}\}\binom{\lambda-1}{s}\frac{(\lambda-s)^l}{l!}$$
$$\beta\left(1-\frac{s}{\alpha},\beta l+\beta\right). \quad (16)$$

3.2. Moment generating function

Let $U \sim \text{PGDUS-IK}(\alpha, \beta, \lambda)$, it's moments generating function is derived as

$$\begin{aligned} M_U(t) &= E(e^{tu}) \\ &= \int_0^\infty e^{tu} \frac{\alpha\beta\lambda}{(e-1)^\lambda}(1+u)^{-(\alpha+1)}(1-(1+u)^{-\alpha})^{\beta-1} e^{(1-(1+u)^{-\alpha})^\beta}(e^{(1-(1+u)^{-\alpha})^\beta}-1)^{\lambda-1} du \\ &= \frac{\alpha\beta\lambda}{(e-1)^\lambda}\sum_{j=0}^{\infty}\frac{t^j}{j!} \\ &\quad \int_0^\infty u^j (1+u)^{-(\alpha+1)}(1-(1+u)^{-\alpha})^{\beta-1} e^{(1-(1+u)^{-\alpha})^\beta}(e^{(1-(1+u)^{-\alpha})^\beta}-1)^{\lambda-1} du. \end{aligned} \tag{17}$$

Substituting $a = (1-(1+u)^{-\alpha})^\beta$ and by solving, we get

$$M_U(t) = \frac{\beta\lambda}{(e-1)^\lambda}\sum_{j=0}^{\infty}\frac{t^j}{j!}\sum_{s=0}^{\lambda-1}\sum_{z=0}^{j}\sum_{l=0}^{\infty}\frac{(-1)^{s+j-z+l}}{l!}\binom{\lambda-1}{s}\binom{j}{z}(\lambda-s)^l\beta\Big(1-\frac{s}{\alpha},\beta l+\beta\Big). \tag{18}$$

3.3. Characteristic function

The characteristic function of the distribution is derived as

$$\phi_U(t) = \frac{\beta\lambda}{(e-1)^\lambda}\sum_{j=0}^{\infty}\sum_{s=0}^{\lambda-1}\sum_{z=0}^{j}\sum_{l=0}^{\infty}\frac{(-1)^{s+j-z+l}}{j!l!}\binom{\lambda-1}{s}\binom{j}{z}(it)^j(\lambda-s)^l\beta\Big(1-\frac{s}{\alpha},\beta l+\beta\Big), \tag{19}$$

where $i = \sqrt{-1}$.

3.4. Cumulant generating function

The cumulant generating function of the distribution is derived as

$$\begin{aligned} K_U(t) &= \log\phi_U(t) \\ &= \log\left(\frac{\beta\lambda}{(e-1)^\lambda}\sum_{j=0}^{\infty}\sum_{s=0}^{\lambda-1}\sum_{z=0}^{j}\sum_{l=0}^{\infty}\frac{(-1)^{s+j-z+l}}{j!l!}\binom{\lambda-1}{s}\binom{j}{z}(it)^j(\lambda-s)^l \right. \\ &\qquad \left. \beta\Big(1-\frac{s}{\alpha},\beta l+\beta\Big)\right) \\ &= \log\frac{\beta\lambda}{(e-1)^\lambda} \\ &\quad + \log\left(\sum_{j=0}^{\infty}\sum_{s=0}^{\lambda-1}\sum_{z=0}^{j}\sum_{l=0}^{\infty}\frac{(-1)^{s+j-z+l}}{j!l!}\binom{\lambda-1}{s}\binom{j}{z}(it)^j(\lambda-s)^l\beta\Big(1-\frac{s}{\alpha},\beta l+\beta\Big)\right), \end{aligned} \tag{20}$$

where $i = \sqrt{-1}$.

3.5. Quantile function

The i^{th} quantile function, denoted by $D(i)$, of PGDUS-IK(α, β, λ) distribution is obtained by solving

$$Q(D(i)) = i, \qquad 0 < i < 1.$$

That is,

$$\left(\frac{e^{(1-(1+D)^{-\alpha})^{\beta}} - 1}{e-1}\right)^{\lambda} = i$$

By solving this, the quantile function of the distribution is obtained as

$$D(i) = \left(1 - \left(\log(1 + i^{\frac{1}{\lambda}}(e-1))\right)^{\frac{1}{\beta}}\right)^{-\frac{1}{\alpha}} - 1, \qquad i\epsilon(0,1). \tag{21}$$

Median of PGDUS-IK(α, β, λ) can be derived by substituting $i = \frac{1}{2}$ in $D(i)$. That is,

$$Median = \left(1 - \left(\log(1 + 0.5^{\frac{1}{\lambda}}(e-1))\right)^{\frac{1}{\beta}}\right)^{-\frac{1}{\alpha}} - 1. \tag{22}$$

Similarly, the inter-quartile range (IQR) of the distribution is,

$$IQR = \left(1 - \left[\log(1 + 0.75^{\frac{1}{\lambda}}(e-1))\right]^{\frac{1}{\beta}}\right)^{-\frac{1}{\alpha}} - \left(1 - \left[\log(1 + 0.25^{\frac{1}{\lambda}}(e-1))\right]^{\frac{1}{\beta}}\right)^{-\frac{1}{\alpha}}. \tag{23}$$

3.6. Order statistics

Let $U_{(1)}, U_{(2)}, ..., U_{(l)}$ be the order statistics for the random sample $U = (U_1, U_2, ..., U_l)$ taken from PGDUS-IK(α, β, λ). The pdf and CDF are given as

$$\begin{aligned} q_{(r)}(u) &= \frac{l!}{(r-1)!(l-r)!} q(u)(Q(u))^{r-1}(1-Q(u))^{l-r} \\ &= \frac{l!\alpha\beta\lambda}{(r-1)!(l-r)!}\left((e-1)^{\lambda} - (e^{(1-(1+u)^{-\alpha})^{\beta}} - 1)^{\lambda}\right)^{l-r} \\ &\quad \frac{(1+u)^{-(\alpha+1)}(1-(1+u)^{-\alpha})^{\beta-1} e^{(1-(1+u)^{-\alpha})^{\beta}} (e^{(1-(1+u)^{-\alpha})^{\beta}} - 1)^{\lambda r - 1}}{(e-1)^{l\lambda}} \end{aligned} \tag{24}$$

and

$$\begin{aligned} Q_{(r)}(u) &= \sum_{s=r}^{l} \binom{l}{s} (Q(u))^{s}(1-Q(u))^{l-s} \\ &= \sum_{l=r}^{l} \binom{l}{s} \left(\frac{e^{(1-(1+u)^{-\alpha})^{\beta}} - 1}{e-1}\right)^{\lambda s} \left(1 - \left(\frac{e^{(1-(1+u)^{-\alpha})^{\beta}} - 1}{e-1}\right)^{\lambda}\right)^{l-s}, \end{aligned} \tag{25}$$

respectively. Substituting $r = 1$ and $r = l$ into equations (24) and (25) allows us to derive the pdf and the CDF of the 1^{st} and l^{th} order statistics, respectively.

3.7. Entropy

Renyi entropy is derived as

$$\tau_R(\zeta) = \frac{1}{1-\zeta}\log\left(\int q^{\zeta}(u)du\right), \quad \zeta > 0, \quad \zeta \neq 1.$$

$$\begin{aligned}\int_0^{\infty} q^{\zeta}(u)du &= \left(\frac{\alpha\beta\lambda}{(e-1)^{\lambda}}\right)^{\zeta} \\ &\int_0^{\infty}(1+u)^{-\zeta(\alpha+1)}(1-(1+u)^{-\alpha})^{\zeta(\beta-1)}e^{\zeta(1-(1+u)^{-\alpha})^{\beta}}(e^{(1-(1+u)^{-\alpha})^{\beta}}-1)^{\zeta(\lambda-1)}du \\ &= \left(\frac{\alpha\beta\lambda}{(e-1)^{\lambda}}\right)^{\zeta}\sum_{s=0}^{\zeta(\lambda-1)}(-1)^{\zeta(\lambda-1)-s}\binom{\zeta(\lambda-1)}{s} \\ &\int_0^{\infty} e^{s(1-(1+u)^{-\alpha})^{\beta}}(1+u)^{-\zeta(\alpha+1)}(1-(1+u)^{-\alpha})^{\zeta(\beta-1)}e^{\zeta(1-(1+u)^{-\alpha})^{\beta}}du \\ &= \left(\frac{\alpha\beta\lambda}{(e-1)^{\lambda}}\right)^{\zeta}\sum_{z=0}^{\infty}\sum_{s=0}^{\zeta(\lambda-1)}\frac{(-1)^{\zeta(\lambda-1)-s}}{z!}\binom{\zeta(\lambda-1)}{s}(\zeta+s)^{z} \\ &\int_0^{\infty}(1+u)^{-\zeta(\alpha+1)}(1-(1+u)^{-\alpha})^{\zeta(\beta-1)+\beta z}du.\end{aligned}$$

Using the transformation $a = 1-(1+u)^{-\alpha}$,

$$\begin{aligned}\int_0^{\infty} q^{\zeta}(u)du = \left(\frac{\alpha\beta\lambda}{(e-1)^{\lambda}}\right)^{\zeta}\frac{1}{\alpha}\sum_{z=0}^{\infty}\sum_{s=0}^{\zeta(\lambda-1)}\frac{(-1)^{\zeta(\lambda-1)-s}}{z!}\binom{\zeta(\lambda-1)}{s}(\zeta+s)^{z} \\ \beta\Big(\zeta(\beta-1)+\beta z+1, \zeta\Big(1+\frac{1}{\alpha}\Big)-\frac{1}{\alpha}\Big).\end{aligned}$$

Then Renyi entropy form will be

$$\begin{aligned}\tau_R(\zeta) &= \frac{1}{1-\zeta}\log\left(\frac{\alpha\beta\lambda}{(e-1)^{\lambda}}\right)^{\zeta}\frac{1}{\alpha}\sum_{z=0}^{\infty}\sum_{s=0}^{\zeta(\lambda-1)}\frac{(-1)^{\zeta(\lambda-1)-s}}{z!}\binom{\zeta(\lambda-1)}{s}(\zeta+s)^{z} \\ &\beta\Big(\zeta(\beta-1)+\beta z+1, \zeta(1+\frac{1}{\alpha})-\frac{1}{\alpha}\Big) \\ &= \frac{1}{1-\zeta}\log\left(\frac{\alpha\beta\lambda}{(e-1)^{\lambda}}\right)^{\zeta}\frac{1}{\alpha} + \frac{1}{1-\zeta}\log\left(\sum_{z=0}^{\infty}\sum_{s=0}^{\zeta(\lambda-1)}\frac{(-1)^{\zeta(\lambda-1)-s}}{z!}\binom{\zeta(\lambda-1)}{s}(\zeta+s)^{z}\right. \\ &\left.\beta\Big(\zeta(\beta-1)+\beta z+1, \zeta(1+\frac{1}{\alpha})-\frac{1}{\alpha}\Big)\right) \qquad (26)\end{aligned}$$

where $\alpha > 0, \beta > 0, \lambda > 0, \zeta > 0, \zeta \neq 1$.

4. Mean residual life function

The mean residual life function at age ν is defined as the expected remaining life given survival at age ν and it is expressed as

$$\begin{aligned} MRL(\nu) &= \frac{1}{\bar{Q}(\nu)} \int_{\nu}^{\infty} u dQ(u) - \nu \\ &= \frac{\beta\lambda}{(e-1)^{\lambda} - (e^{(1-(1+u_i)^{-\alpha})^{\beta}} - 1)^{\lambda}} \sum_{s=0}^{\lambda-1} \sum_{l=0}^{\infty} \{(-1)^{s+l+1} + (-1)^{s+l}\} \binom{\lambda-1}{s} \frac{(\lambda-s)^l}{l!} \\ &\quad \beta\Big(1 - \frac{s}{\alpha}, \beta l + \beta\Big) - \nu. \end{aligned} \tag{27}$$

5. Estimation

To estimate the unknown parameters, methods of maximum likelihood, maximum product spacing, moments, and least squares are described below. Let $U = (U_1, U_2, ..., U_l)$ be a random sample of size l taken from PGDUS-IK(α, β, λ).

5.1. Maximum likelihood estimation

To obtain the maximum likelihood estimate (MLE) of unknown parameters α, β, and λ, consider

$$\begin{aligned} LF(u) &= \prod_{i=1}^{l} q(u) \\ &= \prod_{i=1}^{l} \frac{\alpha\beta\lambda}{(e-1)^{\lambda}} (1+u_i)^{-(\alpha+1)} (1-(1+u_i)^{-\alpha})^{\beta-1} e^{(1-(1+u_i)^{-\alpha})^{\beta}} (e^{(1-(1+u_i)^{-\alpha})^{\beta}} - 1)^{\lambda-1} \\ &= \left(\frac{\alpha\beta\lambda}{(e-1)^{\lambda}}\right)^{l} \prod_{i=1}^{l} (1+u_i)^{-(\alpha+1)} (1-(1+u_i)^{-\alpha})^{\beta-1} e^{(1-(1+u_i)^{-\alpha})^{\beta}} (e^{(1-(1+u_i)^{-\alpha})^{\beta}} - 1)^{\lambda-1}, \end{aligned} \tag{28}$$

the likelihood function and its logarithm will be

$$\begin{aligned} \log LF(u) = l\bigg(&\log\alpha + \log\beta + \log\lambda - \lambda\log(e-1)\bigg) - (\alpha+1)\sum_{i=1}^{l}\log(1+u_i) \\ &+ (\beta-1)\sum_{i=1}^{l}\log(1-(1+u_i)^{-\alpha}) + \sum_{i=1}^{l}(1-(1+u_i)^{-\alpha})^{\beta} \\ &+ (\lambda-1)\sum_{i=1}^{l}\log\Big(e^{(1-(1+u_i)^{-\alpha})^{\beta}} - 1\Big). \end{aligned}$$

To obtaine the MLEs, we find the first-order derivative of $\log L$ and equate it with zero.

$$\begin{aligned}\frac{\partial \log LF}{\partial \alpha} &= \frac{l}{\alpha} - \sum_{i=1}^{l} \log(1+u_i) + (\beta-1)\sum_{i=1}^{l} \frac{(1+u_i)^{-\alpha}\log(1+u_i)}{(1-(1+u_i)^{-\alpha})} \\ &\quad + \beta \sum_{i=1}^{l} (1+u_i)^{-\alpha}\log(1+u_i)(1-(1+u_i)^{-\alpha})^{\beta-1} \\ &\quad + \beta(\lambda-1)\sum_{i=1}^{l} \frac{(1+u_i)^{-\alpha}(1-(1+u_i)^{-\alpha})^{\beta-1}\log(1+u_i)e^{(1-(1+u_i)^{-\alpha})^{\beta}}}{e^{(1-(1+u_i)^{-\alpha})^{\beta}}-1} = 0 \end{aligned} \tag{29}$$

$$\begin{aligned}\frac{\partial \log LF}{\partial \beta} &= \frac{l}{\beta} + \sum_{i=1}^{l} \log(1-(1+u_i)^{-\alpha}) + \sum_{i=1}^{l} (1-(1+u_i)^{-\alpha})^{\beta}\log(1-(1+u_i)^{-\alpha}) \\ &\quad + (\lambda-1)\sum_{i=1}^{l} \frac{(1-(1+u_i)^{-\alpha})^{\beta}\log(1-(1+u_i)^{-\alpha})e^{(1-(1+u_i)^{-\alpha})^{\beta}}}{e^{(1-(1+u_i)^{-\alpha})^{\beta}}-1} = 0 \end{aligned} \tag{30}$$

$$\frac{\partial \log LF}{\partial \lambda} = \frac{l}{\lambda} - l\log(e-1) + \sum_{i=1}^{l} \log\left(e^{(1-(1+u_i)^{-\alpha})^{\beta}} - 1\right) = 0. \tag{31}$$

To solve equations (29), (30), (31) simultaneously, statistical software has to be used.

5.2. Maximum product spacing estimation

The maximum product spacing (MPS) estimation method was introduced by Cheng and Amin (1983) and explored in detailed by Ranneby (1984). The MPS estimation method ensures consistent estimators whether the MLE method exists or not.

To find the MPS estimators of $\alpha,\ \beta$, and λ, first define the spacings

$$D_i = Q(u_i, \alpha, \beta, \lambda) - Q(u_{i-1}, \alpha, \beta, \lambda); i = 1, 2, ..., l+1.$$

Hence, MPS estimators are nothing but parameter values that maximize the geometric mean of the spacings obtained from the observed samples. That is,

$$\begin{aligned} A &= \left(\prod_{i=1}^{l+1} D_i\right)^{1/l+1} \\ &= \left(\prod_{i=1}^{l+1}\left(\frac{e^{(1-(1+u_i)^{-\alpha})^{\beta}}-1}{e-1}\right)^{\lambda} - \left(\frac{e^{(1-(1+u_{i-1})^{-\alpha})^{\beta}}-1}{e-1}\right)^{\lambda}\right)^{1/l+1}. \end{aligned} \tag{32}$$

$$\log A = \frac{1}{l+1}\sum_{i=1}^{l+1} \log\left(\left(\frac{e^{(1-(1+u_i)^{-\alpha})^{\beta}}-1}{e-1}\right)^{\lambda} - \left(\frac{e^{(1-(1+u_{i-1})^{-\alpha})^{\beta}}-1}{e-1}\right)^{\lambda}\right).$$

$$\frac{\partial \log A}{\partial \alpha} = \frac{\beta\lambda}{l+1}\sum_{i=1}^{l+1}\Bigg(\frac{(1+u_i)^{-\alpha}(1-(1+u_i)^{-\alpha})^{\beta-1}e^{(1-(1+u_i)^{-\alpha})^{\beta}}\log(1+u_i)}{(e^{(1-(1+u_i)^{-\alpha})^{\beta}}-1)^{\lambda}-(e^{(1-(1+u_{i-1})^{-\alpha})^{\beta}}-1)^{\lambda}}$$
$$(e^{(1-(1+u_i)^{-\alpha})^{\beta}}-1)^{\lambda-1}-(e^{(1-(1+u_{i-1})^{-\alpha})^{\beta}}-1)^{\lambda-1}$$
$$\frac{(1+u_{i-1})^{-\alpha}(1-(1+u_{i-1})^{-\alpha})^{\beta-1}e^{(1-(1+u_{i-1})^{-\alpha})^{\beta}}\log(1+u_{i-1})}{(e^{(1-(1+u_i)^{-\alpha})^{\beta}}-1)^{\lambda}-(e^{(1-(1+u_{i-1})^{-\alpha})^{\beta}}-1)^{\lambda}}\Bigg), \tag{33}$$

$$\frac{\partial \log A}{\partial \beta} = \frac{\lambda}{l+1}\sum_{i=1}^{l+1}\Bigg(\frac{(1-(1+u_i)^{-\alpha})^{\beta}e^{(1-(1+u_i)^{-\alpha})^{\beta}}(e^{(1-(1+u_i)^{-\alpha})^{\beta}}-1)^{\lambda-1}\log(1-(1+u_i)^{-\alpha})}{(e^{(1-(1+u_i)^{-\alpha})^{\beta}}-1)^{\lambda}-(e^{(1-(1+u_{i-1})^{-\alpha})^{\beta}}-1)^{\lambda}}$$
$$-\frac{(1-(1+u_{i-1})^{-\alpha})^{\beta}e^{(1-(1+u_{i-1})^{-\alpha})^{\beta}}(e^{(1-(1+u_{i-1})^{-\alpha})^{\beta}}-1)^{\lambda-1}\log(1-(1+u_{i-1})^{-\alpha})}{(e^{(1-(1+u_i)^{-\alpha})^{\beta}}-1)^{\lambda}-(e^{(1-(1+u_{i-1})^{-\alpha})^{\beta}}-1)^{\lambda}}\Bigg), \tag{34}$$

and

$$\frac{\partial \log A}{\partial \lambda} = \frac{1}{l+1}\sum_{i=1}^{l+1}\Bigg(\frac{(e^{(1-(1+u_i)^{-\alpha})^{\beta}}-1)^{\lambda}\log\left(\frac{e^{(1-(1+u_i)^{-\alpha})^{\beta}}-1}{e-1}\right)}{(e^{(1-(1+u_i)^{-\alpha})^{\beta}}-1)^{\lambda}-(e^{(1-(1+u_{i-1})^{-\alpha})^{\beta}}-1)^{\lambda}}$$
$$-\frac{(e^{(1-(1+u_i)^{-\alpha})^{\beta}}-1)^{\lambda}\log\left(\frac{e^{(1-(1+u_{i-1})^{-\alpha})^{\beta}}-1}{e-1}\right)}{(e^{(1-(1+u_i)^{-\alpha})^{\beta}}-1)^{\lambda}-(e^{(1-(1+u_{i-1})^{-\alpha})^{\beta}}-1)^{\lambda}}\Bigg). \tag{35}$$

Setting the equations (33), (34) and (35) to zero, and solving simultaneously we get the MPS estimates of α, β, and λ. It is easy to obtain estimates using R software by numerical methods.

5.3. Method of moment estimation

The r^{th} order moment of PGDUS-IK(α, β, λ) is

$$\mu_r^1 = \frac{\beta\lambda}{(e-1)^{\lambda}}\sum_{s=0}^{\lambda-1}\sum_{z=0}^{r}\sum_{l=0}^{\infty}\frac{(-1)^{s+r-z+l}}{l!}\binom{\lambda-1}{s}\binom{r}{z}(\lambda-s)^{l}\beta\Big(1-\frac{s}{\alpha},\beta l+\beta\Big)$$

Taking $r = 1, 2$, and 3 we get first 3 raw moments of the PGDUS-IK distribution. Then, by equating these raw moments to corresponding sample moments, we get

$$\mu_1^1 = \frac{1}{l}\sum_{i=1}^{l}u_i \tag{36}$$

$$\mu_2^1 = \frac{1}{l}\sum_{i=1}^{l}u_i^2 \tag{37}$$

$$\mu_3^1 = \frac{1}{l}\sum_{i=1}^{l}u_i^3 \tag{38}$$

and solving these equations (36), (37), (38) simultaneously we get moment estimators. Statistical software can be used to solve these equations.

5.4. Method of least square estimation

The least-square estimators for the parameters in PGDUS-IK(α, β, λ) can be derived as follows:

$$LS = \sum_{i=1}^{l} \Big(Q(u_i) - \mathcal{Q}_i\Big)^2.$$

where, $Q(U_i)$ - theoretical CDF of the observation u_i
and $\mathcal{Q}_i$ - empirical CDF which is usually estimated by

$$\hat{\mathcal{Q}}_i = \frac{i}{l+1}.$$

There for,

$$LS = \sum_{i=1}^{l} \left(\left(\frac{e^{(1-(1+u_i)^{-\alpha})^{\beta}} - 1}{e-1} \right)^{\lambda} - \frac{i}{l+1} \right)^2.$$

$$\frac{\partial LS}{\partial \alpha} = 0 =>$$
$$\sum_{i=1}^{l} (1+u_i)^{-\alpha} \log(1+u_i) \Big(1-(1+u_i)^{-\alpha}\Big)^{\beta-1} e^{(1-(1+u_i)^{-\alpha})^{\beta}} \Big(e^{(1-(1+u_i)^{-\alpha})^{\beta}} - 1\Big)^{\lambda-1}$$
$$\left(\Big(\frac{e^{(1-(1+u_i)^{-\alpha})^{\beta}} - 1}{e-1} \Big)^{\lambda} - \frac{i}{l+1} \right) = 0. \tag{39}$$

$$\frac{\partial LS}{\partial \beta} = 0 =>$$
$$\sum_{i=1}^{l} \Big(1-(1+u_i)^{-\alpha}\Big)^{\beta} \log(1-(1+u_i)^{-\alpha}) e^{(1-(1+u_i)^{-\alpha})^{\beta}} \Big(e^{(1-(1+u_i)^{-\alpha})^{\beta}} - 1\Big)^{\lambda-1}$$
$$\left(\Big(\frac{e^{(1-(1+u_i)^{-\alpha})^{\beta}} - 1}{e-1} \Big)^{\lambda} - \frac{i}{l+1} \right) = 0. \tag{40}$$

$$\frac{\partial LS}{\partial \lambda} = 0 =>$$
$$\sum_{i=1}^{l} \Big(e^{(1-(1+u_i)^{-\alpha})^{\beta}} - 1\Big)^{\lambda} \log \left(\frac{e^{(1-(1+u_i)^{-\alpha})^{\beta}} - 1}{e-1} \right) \left(\Big(\frac{e^{(1-(1+u_i)^{-\alpha})^{\beta}} - 1}{e-1} \Big)^{\lambda} - \frac{i}{l+1} \right) = 0. \tag{41}$$

Solving (39), (40), and (41) simultaneosuly with respect to α, β and λ gives the least squares estimators. By using statistical softwares, we can find estimated values.

5.5. Bayesian analysis

The joint posterior density function of (α, β, λ) can be written as

$$\Phi(\alpha, \beta, \lambda|U) = \frac{L(\alpha, \beta, \lambda|U)w(\alpha, \beta, \lambda)}{\int_\alpha \int_\beta \int_\lambda L(\alpha, \beta, \lambda|U)w(\alpha, \beta, \lambda)d\alpha d\beta d\lambda},$$

where $w(\alpha, \beta, \lambda)$ is the joint prior density function of the parameters.

Then, the Bayes estimator under the squared error loss function is

$$I(U) = \hat{z}_B = E_{\underline{\theta}|U}(z(\underline{\theta})) = \frac{\int_{\underline{\theta}} z(\underline{\theta})L(\underline{\theta}|U)w(\underline{\theta})d\underline{\theta}}{\int_{\underline{\theta}} L(\underline{\theta}|U)w(\underline{\theta})d\underline{\theta}}.$$

There is no easy closed form for this estimated value since it involves an integral ration.

Lindley (1980) proposed the procedure to approximate the ratio of the two integrals. For a three-parameter distribution, Lindley's approximation can be written as (see Ali and Kanani (2021))

$$\begin{aligned} I(U) = v + (v_1\theta_1 + v_2\theta_2 + v_3\theta_3 + \theta_4 + \theta_5) + \frac{1}{2}\Big(B_1(v_1\sigma_{11} + v_2\sigma_{12} + v_3\sigma_{13})\Big) \\ + \frac{1}{2}\Big(B_2(v_1\sigma_{21} + v_2\sigma_{22} + v_3\sigma_{23})\Big) + \frac{1}{2}\Big(B_3(v_1\sigma_{31} + v_2\sigma_{32} + v_3\sigma_{33})\Big). \end{aligned} \tag{42}$$

where,

$$\begin{aligned} B_1 &= \sigma_{11}M_{111} + 2\sigma_{12}M_{121} + 2\sigma_{13}M_{131} + 2\sigma_{23}M_{231} + \sigma_{22}M_{221} + \sigma_{33}M_{331} \\ B_2 &= \sigma_{11}M_{112} + 2\sigma_{12}M_{122} + 2\sigma_{13}M_{132} + 2\sigma_{23}M_{232} + \sigma_{22}M_{222} + \sigma_{33}M_{332} \\ B_3 &= \sigma_{11}M_{113} + 2\sigma_{12}M_{123} + 2\sigma_{13}M_{133} + 2\sigma_{23}M_{233} + \sigma_{22}M_{223} + \sigma_{33}M_{333} \\ v_1 &= \frac{\partial v(\alpha, \beta, \lambda)}{\partial \alpha}, \qquad v_2 = \frac{\partial v(\alpha, \beta, \lambda)}{\partial \beta}, \qquad v_3 = \frac{\partial v(\alpha, \beta, \lambda)}{\partial \lambda} \\ v_{11} &= \frac{\partial^2 v(\alpha, \beta, \lambda)}{\partial^2 \alpha}, \qquad v_{22} = \frac{\partial^2 v(\alpha, \beta, \lambda)}{\partial^2 \beta}, \qquad v_{33} = \frac{\partial^2 v(\alpha, \beta, \lambda)}{\partial^2 \lambda} \end{aligned}$$

where M - the logarithm of the likelihood function. Then

$$\begin{aligned} M_1 &= \frac{\partial M}{\partial \alpha}, M_2 = \frac{\partial M}{\partial \beta}, \quad M_3 = \frac{\partial M}{\partial \lambda} \\ M_{ij} &= \frac{\partial^2 M}{\partial \tau_i \partial \tau_j}, \quad i, j = 1, 2, 3, \quad (\tau_i, \tau_j) = (\alpha, \beta, \lambda) \\ M_{ijk} &= \frac{\partial^3 M}{\partial \tau_i \partial \tau_j \partial \tau_k}, \quad (i, j, k) = 1, 2, 3, \quad (\tau_i, \tau_j, \tau_k) = (\alpha, \beta, \lambda) \\ \sigma_{ij} &= -\frac{1}{M_{ij}} \\ \theta_i &= \rho_1\sigma_{i1} + \rho_2\sigma_{i2} + \rho_3\sigma_{i3}, \qquad i = 1, 2, 3. \\ \theta_4 &= v_{12}\sigma_{12} + v_{13}\sigma_{13} + v_{23}\sigma_{23}, \quad \theta_5 = \frac{1}{2}(v_{11}\sigma_{11} + v_{22}\sigma_{22} + v_{33}\sigma_{33}). \\ \rho &= \log(w(\alpha, \beta, \lambda)), \qquad \rho_1 = \frac{\partial \rho}{\partial \alpha}, \qquad \rho_2 = \frac{\partial \rho}{\partial \beta}, \qquad \rho_3 = \frac{\partial \rho}{\partial \lambda} \end{aligned}$$

The detailed derivations of equation (42) are given in the appendix.

When information is unavailable for the parameters we use non-informative prior, like Uniform prior, where

$$w(\underline{\theta}) \propto 1.$$

Since the parameter ranges from 0 to ∞, we can choose the gamma distribution as the prior distribution. Therefore,

$$w(\underline{\theta}) \propto \alpha^{r-1}\beta^{p-1}\lambda^{t-1}e^{-(\alpha s+\beta q+\lambda v)}$$

Then the Bayes estimators of the parameters become

$$\hat{\alpha}_B = \hat{\alpha} + \theta_1 + \frac{1}{2}\Big(B_1\sigma_{11} + B_2\sigma_{21} + B_3\sigma_{31}\Big). \tag{43}$$

$$\hat{\beta}_B = \hat{\beta} + \theta_2 + \frac{1}{2}\Big(B_1\sigma_{12} + B_2\sigma_{22} + B_3\sigma_{32}\Big). \tag{44}$$

$$\hat{\lambda}_B = \hat{\lambda} + \theta_3 + \frac{1}{2}\Big(B_1\sigma_{13} + B_2\sigma_{23} + B_3\sigma_{33}\Big). \tag{45}$$

Metropolis-Hasting algorithm

The Metropolis-Hasting (MH) algorithm (see Tobias(2014)), a general Markov Chain Monte Carlo (MCMC) technique, is used to generate samples from models that are complicated. Metropolis *et al.* (1953) developed it initially, then Hastings (1970) developed it afterwards. The MH algorithm, for sampling from a target distribution, let it be π, and let $q(\theta_1^*|\theta_2, ..., \theta_k, \underline{x})$ denotes a proposal density that generates a candidate θ_1^*.

Algorithm:

The MH algorithm is used to simulate a probability distribution p from another probability distribution q, which is easier to simulate. Here p is called target distribution and q is the proposal. Let $\theta^{(t)}$ be the current draw from $p(\theta)$. The MH algorithm performs as follows:

1. Draw θ^* from $q(\theta|\theta^{(1)})$.
2. Accept $\theta^{(t+1)} = \theta^*$ with the probability $\min(1, p^*)$ where

$$p^* = \frac{p(\theta^*)q(\theta^{(t)}|\theta^*)}{p(\theta^{(t)})q(\theta^*|\theta^{(t)})}.$$

Otherwise, set $\theta^{(t+1)} = \theta^{(t)}$.

That is, accepting with the probability $\min(1, p^*)$ means that we will be drawing u according to a uniform distribution on (0,1), and if $u < \min(1, p^*)$, then accept θ^* is accepted; otherwise, it's not.

In the Bayesian context, the MH algorithm can be defined as follows: For that, the posterior distribution will be the form

$$p(\theta|y) \propto LF(y|\theta)w(\theta).$$

where LF is the likelihood function and w is the prior distribution. The MH algorithm can be used to simulate $p(\theta|y)$, by using $t(\theta|y) = LF(y|\theta) * w(\theta)$ and a proposal distribution $q(\theta_1|\theta_2, y)$, as follows.

1. Draw θ^* from $f(\theta|\theta^{(t)}, y)$.
2. Accept $\theta^{(t+1)} = \theta^*$ with the probability $\min(1, p^*)$, where

$$p^* = \frac{q(\theta^*|y)g(\theta^{(t)}|\theta^*, y)}{q(\theta^{(t)}|y)g(\theta^*|\theta^{(t)}, y)}.$$

Otherwise set $\theta^{(t+1)} = \theta^{(t)}$.

6. Confidence interval

In this section, we propose the asymptotic confidence interval and the bootstrap confidence interval, for the unknown parameters α, β, and λ of the PGDUS-IK distribution.

6.1. Asymptotic confidence interval

The asymptotic confidence intervals can be used when the MLEs are not in the closed form. Let us consider the Fisher information matrix I as

$$I = E\begin{bmatrix} \frac{-\partial^2 \log LF}{\partial\alpha^2} & \frac{-\partial^2 \log LF}{\partial\alpha\partial\beta} & \frac{-\partial^2 \log LF}{\partial\alpha\partial\lambda} \\ \frac{-\partial^2 \log LF}{\partial\alpha\partial\beta} & \frac{-\partial^2 \log LF}{\partial\beta^2} & \frac{-\partial^2 \log LF}{\partial\beta\partial\lambda} \\ \frac{-\partial^2 \log LF}{\partial\alpha\partial\lambda} & \frac{-\partial^2 \log LF}{\partial\beta\partial\lambda} & \frac{-\partial^2 \log LF}{\partial\lambda^2} \end{bmatrix}.$$

The second partial derivative of $\log LF$ is briefly given in the appendix.

The asymptotic distribution of MLEs $\tau = (\alpha, \beta, \lambda)$ is normal, with mean zero and variance-covariance matrix I^{-1}. That is,

$$l(\hat{\tau} - \tau) \to N(0, I^{-1}).$$

Hence, the asymptotic $100(1-\eta)\%$ confidence interval of α, β, and λ are

$$\hat{\alpha} \pm z_{\eta/2}\sqrt{Variance(\hat{\alpha})},$$
$$\hat{\beta} \pm z_{\eta/2}\sqrt{Variance(\hat{\beta})},$$

and

$$\hat{\lambda} \pm z_{\eta/2}\sqrt{Variance(\hat{\lambda})},$$

respectively.

6.2. Bootstrap confidence interval

The bootstrap method is a powerful statistical technique used for estimating the sampling distribution of a statistic by resampling with a replacement from the observed data. Let $\hat{\alpha}, \hat{\beta}$ and $\hat{\lambda}$ be the MLEs of parameters α, β and λ. Here we discussed the bootstrap percentile (Boot-p) confidence interval.

To do that, we need to generate a number (let B) of independent bootstrap samples from $u_1, u_2, ..., u_l$, and it is denoted as $u^*_{i1}, u^*_{i2}, ..., u^*_{il}$, for $i = 1, 2, ..., B$. Then, for each bootstrap sample, we calculated the MLEs of α, β, and λ, and the bootstrap MLEs are denoted as $\hat{\alpha}^*, \hat{\beta}^*$, and $\hat{\lambda}^*$, respectively.

Boot-p method

Let $\hat{Q}_1, \hat{Q}_2$, and $\hat{Q}_3$ be the CDF of $\hat{\alpha}^*, \hat{\beta}^*, \hat{\lambda}^*$ respectively. Then $(1-\eta)\%$ percentile confidence intervals are

$$\Big(\hat{Q}_1^{-1}(\eta/2), \hat{Q}_1^{-1}(1-\eta/2)\Big),$$
$$\Big(\hat{Q}_2^{-1}(\eta/2), \hat{Q}_2^{-1}(1-\eta/2)\Big),$$
and
$$\Big(\hat{Q}_3^{-1}(\eta/2), \hat{Q}_3^{-1}(1-\eta/2)\Big)$$

respectively.

7. Simulation study

In this section, the simulation study is used to examine the performance of estimators of PGDUS-IK(α, β, λ) distribution parameters.

By using the quantile function, a random sample of the PGDUS-IK(α, β, λ) distribution can be simulated by using

$$U = \left(1 - \Big(\log(1 + j^{\frac{1}{\lambda}}(e-1))\Big)^{\frac{1}{\beta}}\right)^{-\frac{1}{\alpha}} - 1, \qquad 0 < j < 1$$

where j from $U(0,1)$.

Here, differnt values of the sample size, $l = 50, 100, 200, 300,$ and 400 are considered and replicated 1000 times. The performance of MLE, MPS, and Bayes estimators of each parameter is examined using their biases and MSE values(see Table 1). Bayes estimators are obtained only by using informative prior gamma under the squared error loss function. It is observed that, biases and MSE values decreases to zero as sample size l increases.

8. Application

This section compares the PGDUS-IK distribution to DUS-IK distribution, IK distribution, and Weibull distribution. For that, we are using a vinyl chloride data obtained from clean upgrading monitoring wells in mg/L by Bhaumik *et al.* (2009) (Table 2).

A number of factors, including the p-value, log-likelihood value, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC), Kolmogrov-Smirnov (K-S) statistic can be applied to compare statistical models in order to evaluate which one has a better relative goodness-of-fit with the data. Lower K-S statistic, AIC, and BIC values indicate greater correspondence between the observed data and the model. Additionally, higher p-values and log-likelihood values indicate a stronger fit between the model and the observed data. If a single criterion consistently favors one model over another, that model is likely the better choice.

Based on the table values (see Table 3), compared to the other distributions described, PGDUS-IK(α, β, λ) possesses the lowest AIC, BIC and KS-statistic values moreover

Table 1: Simulation study for MLE, MPS and Bayes estimation for the values $\alpha = 0.4, \beta = 0.8, \lambda = 0.3$

method	l	bias($\hat{\alpha}$)	bias($\hat{\beta}$)	bias($\hat{\lambda}$)	MSE($\hat{\alpha}$)	MSE($\hat{\beta}$)	MSE($\hat{\lambda}$)
MLE	50	0.03124	0.26774	0.32454	0.01357	0.53269	4.10082
	100	0.01458	0.16442	0.14788	0.00618	0.31273	0.67138
	200	0.00782	0.06756	0.15049	0.00285	0.18506	0.43009
	300	0.00497	0.04771	0.08236	0.00187	0.12485	0.14573
	400	0.00307	0.03249	0.07471	0.00136	0.10573	0.13432
MPS	50	0.12090	1.08723	0.00376	0.03668	2.46970	0.47450
	100	0.06039	0.73472	-0.02447	0.01127	1.16509	0.13256
	200	0.03159	0.45171	-0.01279	0.00437	0.56686	0.07379
	300	0.01998	0.36702	-0.01325	0.00256	0.39957	0.04617
	400	0.01542	0.28923	-0.00465	0.00178	0.30120	0.03975
Bayesian	50	9.37736	2.68987	2.7458	0.00936	7.23541	7.53963
	100	0.03769	2.20796	0.49767	0.007187	4.87509	0.24768
	200	0.03386	1.03768	0.65245	0.00413	1.07678	0.42570
	300	0.03319	0.9113	0.6275	0.002381	0.83049	0.39383
	400	0.021983	0.2	-0.69736	0.00156	0.096756	0.48632

Table 2: Vinyl Chloride data

5.1	1.2	1.3	0.6	0.5	2.4
0.5	1.1	8.0	0.8	0.4	0.4
0.6	0.9	0.4	2.0	0.5	1.2
5.3	3.2	2.7	2.9	2.5	0.2
2.3	1.0	0.2	0.1	0.1	1.8
0.9	2.0	4.0	6.8		

a high log-likelihood value and p-value by the MLE, MPS, and Bayesian methods. We can therefore conclude that the PGDUS-IK distribution performs better than the given existing distribution for modeling a parallel system.

In Table 4, the estimated parameter values (based on the ML method) along with their 95% confidence interval, based on 1000 bootstrap samples for vinyl chloride data (see Table 2), are given.

9. Stress-Strength reliability(SSR)

Single-component SSR

Let U indicate the strength of a component or system that is subjected to a random stress, V. The system's functioning is then defined by stress-strength reliability. If U and V are distributed as PGDUS-IK(α, β, λ_1) and PGDUS-IK(α, β, λ_2), respectively, then stress-strength reliability is defined as

$$R = P(V < U) = \int_0^\infty q_U(u) Q_V(u) du$$

$$= \frac{\alpha\beta\lambda_1}{(e-1)^{\lambda_1+\lambda_2}} \int_0^\infty (1+u)^{-(\alpha+1)} \left(1-(1+u)^{-\alpha}\right)^{\beta-1} e^{(1-(1+u)^{-\alpha})^\beta} \left(e^{(1-(1+u)^{-\alpha})^\beta} - 1\right)^{\lambda_1+\lambda_2-1} du.$$

Table 3: Data Analysis

Distribution		Estimates	KS Statistic	Log(L)	p-value	AIC	BIC
PGDUS IK	MLE	2.0103 5.9354 0.3584	0.0884	-55.4280	0.953	114.856	117.9088
	MPS	1.7439 2.1148 0.9072	0.1229	-56.1242	0.6834	116.2484	119.3011
	Bayesian	2.0004 4.2964 0.5078	0.08707	-55.5094	0.9588	115.0187	118.0715
DUS-IK	MLE	1.9467 1.8296	0.0892	-55.5702	0.9497	115.1403	118.193
	MPS	1.7365 1.8928	0.1244	-56.5598	0.6692	117.1196	120.1723
	Bayesian	2.2306 2.8658	0.1569	-57.1064	0.3725	118.2127	121.2654
IK	MLE	1.7409 2.1059	0.0966	-55.7707	0.909	115.5414	118.5941
	MPS	1.5286 2.1388	0.1136	-59.4084	0.7729	122.8169	125.8696
	Bayesian	1.9060 2.9559	0.1409	-57.00978	0.5095	118.0194	121.0721
Weibull	MLE	1.0102 1.8879	0.0918	-55.4496	0.9366	114.8992	117.952
	MPS	1.1075 2.2840	0.1735	-105.4977	0.2577	214.9953	218.0481
	Bayesian	0.8033 1.5418	0.16938	-56.9383	0.2835	117.8766	120.9294

Take $a = \left(e^{(1-(1+u)^{-\alpha})^{\beta}} - 1\right)^{\lambda_1+\lambda_2}$, hence the stress-strength reliability becomes

$$R = \frac{\lambda_1}{\lambda_1 + \lambda_2}, \qquad \lambda_1 > 0, \quad \lambda_2 > 0. \tag{46}$$

To evaluate the reliability value, we need to estimate the parameters first.

Multi-component SSR

Let's consider a system comprising identical d components, which operates successfully if at least c $(1 \leq c \leq d)$ of these components survive a shared random stress. This

Table 4: Estimate value and 95 % bootstrap CI of Vinyl Chloride data

Method		α	β	λ
MLE	Estimate	2.0103428	5.9354142	0.3584253
	CI	(1.623967,2.791699)	(1.981512,23.06395)	(0.12504,1.286034)

situation is called multi-component systems. Bhattacharya and Johnson (1974) first studied the multi-component stress-strength reliability system and defined the reliability of a multi-component stress-strength model as

$$R_{c,d} = Pr\{\text{at least } \ c \ \text{ of } \ \text{ the } \ (U_1, U_2, ..., U_d) \ \text{ exceed } \ V\}. \tag{47}$$

Let $U = (U_1, U_2, ..., U_d)$ be random strength variables from PGDUS-IK$(\alpha, \beta, \lambda_1)$ with CDF $H(u)$, and V be the random stress variable from PGDUS-IK$(\alpha, \beta, \lambda_2)$ with CDF $Q(v)$. Then the reliability of a multi-component stress-strength model defined by Bhattacharya and Johnson (1974) is given as

$$\begin{aligned}
R_{c,d} &= \sum_{i=c}^{d} \binom{d}{i} \int_{-\infty}^{\infty} (1 - H(u))^i (H(u))^{d-i} dQ(u) \\
&= \sum_{i=c}^{d} \binom{d}{i} \int_{0}^{\infty} \frac{\alpha\beta\lambda_2}{(e-1)^{\lambda_2}} (1+u)^{-(\alpha+1)} \left(1 - (1+u)^{-\alpha}\right)^{\beta-1} e^{(1-(1+u)^{-\alpha})^{\beta}} \\
&\quad \left(e^{(1-(1+u)^{-\alpha})^{\beta}} - 1\right)^{\lambda_2 - 1} \left(\left(\frac{e^{(1-(1+u)^{-\alpha})^{\beta}} - 1}{e-1}\right)^{\lambda_1}\right)^{d-i} \left(1 - \left(\frac{e^{(1-(1+u)^{-\alpha})^{\beta}} - 1}{e-1}\right)^{\lambda_1}\right)^{i} du \\
&= \sum_{i=c}^{d} \binom{d}{i} \int_{0}^{\infty} \frac{\alpha\beta\lambda_2}{(e-1)^{\lambda_2}} (1+u)^{-(\alpha+1)} \left(1 - (1+u)^{-\alpha}\right)^{\beta-1} e^{(1-(1+u)^{-\alpha})^{\beta}} \\
&\quad \left(e^{(1-(1+u)^{-\alpha})^{\beta}} - 1\right)^{\lambda_1(d-i)+\lambda_2-1} \left((e-1)^{\lambda_1} - \left(e^{(1-(1+u)^{-\alpha})^{\beta}} - 1\right)^{\lambda_1}\right)^{i} du \\
&= \sum_{i=c}^{d} \sum_{p=0}^{i} \binom{d}{i} \binom{i}{p} \frac{(-1)^p \alpha\beta\lambda_2}{(e-1)^{\lambda_1(d+p-i)+\lambda_2}} \\
&\quad \int_{0}^{\infty} (1+u)^{-(\alpha+1)} \left(1 - (1+u)^{-\alpha}\right)^{\beta-1} e^{(1-(1+u)^{-\alpha})^{\beta}} \left(e^{(1-(1+u)^{-\alpha})^{\beta}} - 1\right)^{\lambda_1(d+p-i)+\lambda_2-1} du \\
&= \sum_{i=c}^{d} \sum_{p=0}^{i} \binom{d}{i} \binom{i}{p} \frac{(-1)^p \lambda_2}{\lambda_1(d+p-i)+\lambda_2}.
\end{aligned}$$

That is,

$$R_{c,d} = \sum_{i=c}^{d} \sum_{p=0}^{i} \binom{d}{i} \binom{i}{p} \frac{(-1)^p \lambda_2}{\lambda_1(d+p-i)+\lambda_2}, \qquad \lambda_1 > 0, \quad \lambda_2 > 0. \tag{48}$$

Suppose $U = (U_1, U_2, ..., U_d)$ are parallelly connected, then $c = 1$ and $R_{c,d}$ will become

$$R_{1,d} = \sum_{i=1}^{d} \sum_{p=0}^{i} \binom{d}{i} \binom{i}{p} \frac{(-1)^p \lambda_2}{\lambda_1(d+p-i)+\lambda_2}, \qquad \lambda_1 > 0, \quad \lambda_2 > 0.$$

Similarly, when $U = (U_1, U_2, ..., U_d)$ are connected in series, so $c = d$ and

$$R_{d,d} = \sum_{p=0}^{d} \binom{d}{p} \frac{(-1)^p \lambda_2}{\lambda_1 p + \lambda_2}, \qquad \lambda_1 > 0, \quad \lambda_2 > 0.$$

9.1. Estimation of reliability

To obtain the estimates of both single-component SSR and multi-component SSR, we need to get the respective parameter estimates. Hence, here we are using the ML method. Let $U = (U_1 < U_2 < ... < U_l)$ and $V = (V_1 < V_2 < ... < V_z)$ be the random samples from PGDUS-IK$(\alpha, \beta, \lambda_1)$ and PGDUS-IK$(\alpha, \beta, \lambda_2)$, respectively.

9.1.1. Estimation of R

The likelihood function for the observed samples for $\underline{\theta} = (\alpha, \beta, \lambda_1, \lambda_2)$ can be written as follows:

$$
\begin{aligned}
&LF(u, v, \underline{\theta}) \\
&= \prod_{i=1}^{l} \left(\frac{\alpha\beta\lambda_1}{(e-1)^{\lambda_1}} (1+u_i)^{-(\alpha+1)} (1-(1+u_i)^{-\alpha})^{\beta-1} e^{(1-(1+u_i)^{-\alpha})^{\beta}} (e^{(1-(1+u_i)^{-\alpha})^{\beta}} - 1)^{\lambda_1 - 1} \right) \\
&\quad \prod_{j=1}^{z} \left(\frac{\alpha\beta\lambda_2}{(e-1)^{\lambda_2}} (1+v_j)^{-(\alpha+1)} (1-(1+v_j)^{-\alpha})^{\beta-1} e^{(1-(1+v_j)^{-\alpha})^{\beta}} (e^{(1-(1+v_j)^{-\alpha})^{\beta}} - 1)^{\lambda_2 - 1} \right) \\
&= \left(\frac{\alpha\beta\lambda_1}{(e-1)^{\lambda_1}} \right)^{l} \prod_{i=1}^{l} (1+u_i)^{-(\alpha+1)} (1-(1+u_i)^{-\alpha})^{\beta-1} e^{(1-(1+u_i)^{-\alpha})^{\beta}} (e^{(1-(1+u_i)^{-\alpha})^{\beta}} - 1)^{\lambda_1 - 1} \\
&\quad \left(\frac{\alpha\beta\lambda_2}{(e-1)^{\lambda_2}} \right)^{z} \prod_{j=1}^{z} (1+v_j)^{-(\alpha+1)} (1-(1+v_j)^{-\alpha})^{\beta-1} e^{(1-(1+v_j)^{-\alpha})^{\beta}} (e^{(1-(1+v_j)^{-\alpha})^{\beta}} - 1)^{\lambda_2 - 1}.
\end{aligned}
$$

Then,

$$
\begin{aligned}
\log LF = {} & (l+z)\Big(\log\alpha + \log\beta\Big) + l\log\lambda_1 + z\log\lambda_2 - (l\lambda_1 + z\lambda_2)\log(e-1) \\
& - (\alpha+1)\left(\sum_{i=1}^{l} \log(1+u_i) + \sum_{j=1}^{z} \log(1+v_j) \right) + (\beta-1)\left(\sum_{i=1}^{l} \log(1-(1+u_i)^{-\alpha}) \right. \\
& \left. + \sum_{j=1}^{z} \log(1-(1+v_j)^{-\alpha}) \right) + \sum_{i=1}^{l} (1-(1+u_i)^{-\alpha})^{\beta} + \sum_{j=1}^{z} (1-(1+v_j)^{-\alpha})^{\beta} \\
& + (\lambda_1 - 1) \sum_{i=1}^{l} \log\left(e^{(1-(1+u_i)^{-\alpha})^{\beta}} - 1 \right) + (\lambda_2 - 1) \sum_{j=1}^{z} \log\left(e^{(1-(1+v_j)^{-\alpha})^{\beta}} - 1 \right).
\end{aligned}
$$

Compute the partial derivatives of the $logLF$ with respect to the parameters α, β, λ_1, and λ_2, respectively. That is,

$$
\begin{aligned}
\frac{\partial \log LF}{\partial\alpha} = {} & \frac{l+z}{\alpha} - \left(\sum_{i=1}^{l} \log(1+u_i) + \sum_{j=1}^{z} \log(1+v_j) \right) \\
& + (\beta-1)\left(\sum_{i=1}^{l} \frac{(1+u_i)^{-\alpha}\log(1+u_i)}{(1-(1+u_i)^{-\alpha})} + \sum_{j=1}^{z} \frac{(1+v_j)^{-\alpha}\log(1+v_j)}{(1-(1+v_j)^{-\alpha})} \right) \\
& + \beta\left(\sum_{i=1}^{l} (1+u_i)^{-\alpha} (1-(1+u_i)^{-\alpha})^{\beta-1} \log(1+u_i) \right.
\end{aligned}
\tag{49}
$$

$$+\sum_{j=1}^{z}(1+v_j)^{-\alpha}(1-(1+v_j)^{-\alpha})^{\beta-1}\log(1+v_j)$$
$$+(\lambda_1-1)\sum_{i=1}^{l}\frac{(1+u_i)^{-\alpha}(1-(1+u_i)^{-\alpha})^{\beta-1}\log(1+u_i)e^{(1-(1+u_i)^{-\alpha})^{\beta}}}{e^{(1-(1+u_i)^{-\alpha})^{\beta}}-1}$$
$$+(\lambda_2-1)\sum_{j=1}^{z}\frac{(1+v_j)^{-\alpha}(1-(1+v_j)^{-\alpha})^{\beta-1}\log(1+v_j)e^{(1-(1+v_j)^{-\alpha})^{\beta}}}{e^{(1-(1+v_j)^{-\alpha})^{\beta}}-1}\Bigg),$$

$$\frac{\partial\log LF}{\partial\beta}=\frac{l+z}{\beta}+\left(\sum_{i=1}^{l}\log(1-(1+u_i)^{-\alpha})+\sum_{j=1}^{z}\log(1-(1+v_j)^{-\alpha})\right)$$
$$+\textstyle\sum_{i=1}^{l}(1-(1+u_i)^{-\alpha})^{\beta}\log(1-(1+u_i)^{-\alpha})+\sum_{j=1}^{z}(1-(1+v_j)^{-\alpha})^{\beta}\log(1-(1+v_j)^{-\alpha})$$
$$+(\lambda_1-1)\left(\sum_{i=1}^{l}\frac{(1-(1+u_i)^{-\alpha})^{\beta}e^{(1-(1+u_i)^{-\alpha})^{\beta}}\log(1-(1+u_i)^{-\alpha})}{e^{(1-(1+u_i)^{-\alpha})^{\beta}}-1}\right)$$
$$+(\lambda_2-1)\left(\sum_{j=1}^{z}\frac{(1-(1+v_j)^{-\alpha})^{\beta}e^{(1-(1+v_j)^{-\alpha})^{\beta}}\log(1-(1+v_j)^{-\alpha})}{e^{(1-(1+v_j)^{-\alpha})^{\beta}}-1}\right),$$

$$\frac{\partial\log LF}{\partial\lambda_1}=\frac{l}{\lambda_1}-l\log(e-1)+\sum_{i=1}^{l}\log\left(e^{(1-(1+u_i)^{-\alpha})^{\beta}}-1\right),$$

$$\frac{\partial\log LF}{\partial\lambda_2}=\frac{z}{\lambda_2}-z\log(e-1)+\sum_{j=1}^{z}\log\left(e^{(1-(1+v_j)^{-\alpha})^{\beta}}-1\right).$$

Then, the MLEs of $\alpha,\ \beta,\ \lambda_1$, and λ_2 can be determined by solving the following equations::

$$\frac{\partial\log LF}{\partial\alpha}=0,\frac{\partial\log LF}{\partial\beta}=0,\frac{\partial\log LF}{\partial\lambda_1}=0,\frac{\partial\log LF}{\partial\lambda_2}=0.$$

Substituting the estimated values for α and β, we get the MLE of λ_1 and λ_2 as

$$\hat{\lambda_1}=\frac{l}{l\log(e-1)-\sum_{i=1}^{l}(\log(e^{(1-(1+u_i)^{-\hat{\alpha}})^{\hat{\beta}}})-1)},$$

$$\hat{\lambda_2}=\frac{z}{z\log(e-1)-\sum_{j=1}^{z}(\log(e^{(1-(1+v_j)^{-\hat{\alpha}})^{\hat{\beta}}})-1)}.$$

Hence, the MLE of R will be

$$\hat{R}=\frac{\hat{\lambda_1}}{\hat{\lambda_1}+\hat{\lambda_2}},\qquad \lambda_1>0\quad\lambda_2>0. \tag{50}$$

9.1.2. Estimation of $R_{c,d}$

To compute the MLE of multi-component reliability, $R_{c,d}$, assume that $U_{i1},U_{i2},...,U_{id}$ and $V_i,\ i=1,2,...,l$ denote the observed data obtained using PGDUS-IK(α,β,λ_1) with pdf

$h(u)$ and PGDUS-IK$(\alpha, \beta, \lambda_2)$ with pdf $q(v)$, respectively. The likelihood function can be defined as:

$$\begin{aligned} LF_{c,d}(u, v, \underline{\theta}) &= \prod_{i=1}^{l} \Big(\prod_{j=1}^{d} h(u_{ij}) \Big) q(v_i) \\ &= \frac{(\alpha\beta)^{l(d+1)} \lambda_1^{ld} \lambda_2^{l}}{(e-1)^{l(d\lambda_1+\lambda_2)}} \prod_{i=1}^{l} \Bigg(\prod_{j=1}^{d} (1+u_{ij})^{-(\alpha+1)} (1-(1+u_{ij})^{-\alpha})^{\beta-1} e^{(1-(1+u_{ij})^{-\alpha})^{\beta}} \\ &\quad \Big(e^{(1-(1+u_{ij})^{-\alpha})^{\beta}} - 1 \Big)^{\lambda_1 - 1} \Bigg) (1+v_i)^{-(\alpha+1)} (1-(1+v_i)^{-\alpha})^{\beta-1} e^{(1-(1+v_i)^{-\alpha})^{\beta}} \\ &\quad \Big(e^{(1-(1+v_i)^{-\alpha})^{\beta}} - 1 \Big)^{\lambda_2 - 1}. \end{aligned}$$

The logarithm likelihood function will be

$$\begin{aligned} &\log LF_{c,d} \\ &= l(d+1)\Big(\log\alpha + \log\beta \Big) + ld \log\lambda_1 + l\log\lambda_2 - l(d\lambda_1 + \lambda_2)\log(e-1) \\ &- (\alpha+1)\Bigg(\sum_{i=1}^{l}\sum_{j=1}^{d} \log(1+u_{ij}) + \sum_{i=1}^{l} \log(1+v_i) \Bigg) + (\beta-1)\Bigg(\sum_{i=1}^{l}\sum_{j=1}^{d} \log(1-(1+u_{ij})^{-\alpha}) \\ &+ \sum_{i=1}^{l} \log(1-(1+v_i)^{-\alpha}) \Bigg) + \sum_{i=1}^{l}\sum_{j=1}^{d} (1-(1+u_{ij})^{-\alpha})^{\beta} + \sum_{i=1}^{d} (1-(1+v_i)^{-\alpha})^{\beta} \\ &+ (\lambda_1 - 1)\sum_{i=1}^{l}\sum_{j=1}^{d} \log\Big(e^{(1-(1+u_{ij})^{-\alpha})^{\beta}} - 1 \Big) + (\lambda_2 - 1)\sum_{i=1}^{l} \log\Big(e^{(1-(1+v_i)^{-\alpha})^{\beta}} - 1 \Big). \end{aligned}$$

Consider the partial derivative of the $logLF_{c,d}$ with respect to the parameters and solving them by equating to zero, we can obtain the MLEs of the unknown parameters α, β, λ_1 and λ_2, respectively. That is,

$$\frac{\partial \log LF_{c,d}}{\partial\alpha} = 0, \frac{\partial \log LF_{c,d}}{\partial\beta} = 0, \frac{\partial \log LF_{c,d}}{\partial\lambda_1} = 0, \frac{\partial \log LF_{c,d}}{\partial\lambda_2} = 0$$

where

$$\begin{aligned} \frac{\partial \log LF_{c,d}}{\partial\alpha} &= \frac{l(d+1)}{\alpha} - \Bigg(\sum_{i=1}^{l}\sum_{j=1}^{d} \log(1+u_{ij}) + \sum_{i=1}^{l} \log(1+v_i) \Bigg) \\ &\quad + (\beta-1)\Bigg(\sum_{i=1}^{l}\sum_{j=1}^{d} \frac{(1+u_{ij})^{-\alpha}\log(1+u_{ij})}{(1-(1+u_{ij})^{-\alpha})} + \sum_{i=1}^{l} \frac{(1+v_i)^{-\alpha}\log(1+v_i)}{(1-(1+v_i)^{-\alpha})} \Bigg) \\ &\quad + \beta\Bigg(\sum_{i=1}^{l}\sum_{j=1}^{d} (1+u_{ij})^{-\alpha}(1-(1+u_{ij})^{-\alpha})^{\beta-1}\log(1+u_{ij}) \\ &\qquad + \sum_{i=1}^{l} (1+v_i)^{-\alpha}(1-(1+v_i)^{-\alpha})^{\beta-1}\log(1+v_i) \end{aligned} \tag{51}$$

$$+(\lambda_1-1)\sum_{i=1}^{l}\sum_{j=1}^{d}\frac{(1+u_{ij})^{-\alpha}(1-(1+u_{ij})^{-\alpha})^{\beta-1}\log(1+u_{ij})e^{(1-(1+u_{ij})^{-\alpha})^{\beta}}}{e^{(1-(1+u_{ij})^{-\alpha})^{\beta}}-1}$$
$$+(\lambda_2-1)\sum_{i=1}^{l}\frac{(1+v_i)^{-\alpha}(1-(1+v_i)^{-\alpha})^{\beta-1}\log(1+v_i)e^{(1-(1+v_i)^{-\alpha})^{\beta}}}{e^{(1-(1+v_i)^{-\alpha})^{\beta}}-1}\Bigg).$$

$$\frac{\partial \log LF_{c,d}}{\partial\beta}=\frac{l(d+1)}{\beta}+\left(\sum_{i=1}^{l}\sum_{j=1}^{d}\log(1-(1+u_{ij})^{-\alpha})+\sum_{i=1}^{l}\log(1-(1+v_i)^{-\alpha})\right)$$
$$+\sum_{i=1}^{l}\sum_{j=1}^{d}(1-(1+u_{ij})^{-\alpha})^{\beta}\log(1-(1+u_{ij})^{-\alpha})$$
$$+\sum_{i=1}^{l}(1-(1+v_i)^{-\alpha})^{\beta}\log(1-(1+v_i)^{-\alpha})$$
$$+(\lambda_1-1)\left(\sum_{i=1}^{l}\sum_{j=1}^{d}\frac{(1-(1+u_{ij})^{-\alpha})^{\beta}e^{(1-(1+u_{ij})^{-\alpha})^{\beta}}\log(1-(1+u_{ij})^{-\alpha})}{e^{(1-(1+u_{ij})^{-\alpha})^{\beta}}-1}\right)$$
$$+(\lambda_2-1)\left(\sum_{i=1}^{l}\frac{(1-(1+v_i)^{-\alpha})^{\beta}e^{(1-(1+v_i)^{-\alpha})^{\beta}}\log(1-(1+v_i)^{-\alpha})}{e^{(1-(1+v_i)^{-\alpha})^{\beta}}-1}\right).$$

$$\frac{\partial \log LF_{c,d}}{\partial\lambda_1}=\frac{ld}{\lambda_1}-ld\log(e-1)+\sum_{i=1}^{l}\sum_{j=1}^{d}\log\left(e^{(1-(1+u_{ij})^{-\alpha})^{\beta}}-1\right).$$

$$\frac{\partial \log LF_{c,d}}{\partial\lambda_2}=\frac{l}{\lambda_2}-l\log(e-1)+\sum_{i=1}^{l}\log\left(e^{(1-(1+v_i)^{-\alpha})^{\beta}}-1\right).$$

By substituting the MLEs of α and β, the MLEs of λ_1 and λ_2 will be in the form

$$\hat{\lambda_1}=\frac{ld}{ld\log(e-1)+\sum_{i=1}^{l}\sum_{j=1}^{d}\log\left(e^{(1-(1+u_{ij})^{-\hat{\alpha}})^{\hat{\beta}}-1}\right)},$$

and

$$\hat{\lambda_2}=\frac{l}{l\log(e-1)+\sum_{i=1}^{l}\log\left(e^{(1-(1+v_i)^{-\hat{\alpha}})^{\hat{\beta}}-1}\right)}.$$

Hence, the MLE of stress-strength reliability of the multi-component system will be

$$\hat{R_{c,d}}=\sum_{i=c}^{d}\sum_{p=0}^{i}\binom{d}{i}\binom{i}{p}\frac{(-1)^p\hat{\lambda_2}}{\hat{\lambda_1}(d+p-i)+\hat{\lambda_2}},\qquad \lambda_1>0,\quad \lambda_2>0.$$

9.2. Asymptotic distribution

This section discusses the asymptotic distribution of R and $R_{c,d}$ by using their MLEs.

9.2.1. Asymptotic distribution of R

The asymptotic distribution of MLE of R is normal with mean zero and variance-covariance matrix $I^{-1}(\underline{\theta})$. That is,

$$\sqrt{l+o}(\hat{R}-R) \to N(0, G^T I^{-1}(\underline{\theta})G) \tag{52}$$

where G^T - $\left(\frac{\partial R}{\partial \alpha}, \frac{\partial R}{\partial \beta}, \frac{\partial R}{\partial \lambda_1}, \frac{\partial R}{\partial \lambda_2}\right)$

$I^{-1}(\underline{\theta})$ - inverse of Fisher information matrix for unknown parameters

$$I(\underline{\theta}) = E\begin{bmatrix} \frac{-\partial^2 \log LF}{\partial \alpha^2} & \frac{-\partial^2 \log LF}{\partial \alpha \partial \beta} & \frac{-\partial^2 \log LF}{\partial \alpha \partial \lambda_1} & \frac{-\partial^2 \log LF}{\partial \alpha \partial \lambda_2} \\ \frac{-\partial^2 \log LF}{\partial \alpha \partial \beta} & \frac{-\partial^2 \log LF}{\partial \beta^2} & \frac{-\partial^2 \log LF}{\partial \beta \partial \lambda_1} & \frac{-\partial^2 \log LF}{\partial \beta \partial \lambda_2} \\ \frac{-\partial^2 \log LF}{\partial \alpha \partial \lambda_1} & \frac{-\partial^2 \log LF}{\partial \beta \partial \lambda_1} & \frac{-\partial^2 \log LF}{\partial \lambda_1^2} & \frac{-\partial^2 \log LF}{\partial \lambda_1 \partial \lambda_2} \\ \frac{-\partial^2 \log LF}{\partial \alpha \partial \lambda_2} & \frac{-\partial^2 \log LF}{\partial \beta \partial \lambda_2} & \frac{-\partial^2 \log LF}{\partial \lambda_1 \partial \lambda_2} & \frac{-\partial^2 \log LF}{\partial \lambda_2^2} \end{bmatrix}$$

Second order partial derivative of the log-likelihood function with respect to each parameters $\alpha,\ \beta,\ \lambda_1,$ and λ_2 are briefly derived and given in the appendix. Due to the complexity of the expectations, an approximate estimation of the variance-covariance matrix of $(\alpha, \beta, \lambda_1, \lambda_2)$ is $I^{-1}(\hat{\alpha}, \hat{\beta}, \hat{\lambda_1}, \hat{\lambda_2})$, where $\hat{\alpha},\ \hat{\beta},\ \hat{\lambda_1},$ and $\hat{\lambda_2}$ are the estimates of the respective parameters. From Eq.(45), we can obtain the approximate estimate of the variance of $\hat{R}$ as

$$\hat{Variance}(\hat{R}) \simeq G^T I^{-1} G.$$

Thus,

$$\frac{(\hat{R}-R)}{\sqrt{\hat{Variance}(\hat{R})}} \sim N(0,1).$$

This yields the asymptotic $100(1-\eta)\%$ confidence interval for R as

$$\hat{R} \pm Z_{\eta/2}\sqrt{\hat{Variance}(\hat{R})}$$

where $\hat{R}$ is the MLE of R and $Z_{\eta/2}$ is the upper $(\eta/2)^{th}$ quantile of the standard Normal distribution.

9.2.2. Asymptotic distribution of $R_{c,d}$

Similarly, for large sample size, the asymptotic distribution of MLE of $R_{c,d}$ is given by

$$\sqrt{l+ld}(\hat{R}_{c,d} - R_{c,d}) \to N(0, G^T I^{-1} G)$$

where G^T - $\left(\frac{\partial R_{c,d}}{\partial \alpha}, \frac{\partial R_{c,d}}{\partial \beta}, \frac{\partial R_{c,d}}{\partial \lambda_1}, \frac{\partial R_{c,d}}{\partial \lambda_2}\right)$

and I^{-1} - variance-covariance matrix or inverse of the Fisher information matrix and is given by

$$I^{-1} = E\begin{bmatrix} \frac{-\partial^2 \log LF_{c,d}}{\partial \alpha^2} & \frac{-\partial^2 \log LF_{c,d}}{\partial \alpha \partial \beta} & \frac{-\partial^2 \log LF_{c,d}}{\partial \alpha \partial \lambda_1} & \frac{-\partial^2 \log LF_{c,d}}{\partial \alpha \partial \lambda_2} \\ \frac{-\partial^2 \log LF_{c,d}}{\partial \alpha \partial \beta} & \frac{-\partial^2 \log LF_{c,d}}{\partial \beta^2} & \frac{-\partial^2 \log LF_{c,d}}{\partial \beta \partial \lambda_1} & \frac{-\partial^2 \log LF_{c,d}}{\partial \beta \partial \lambda_2} \\ \frac{-\partial^2 \log LF_{c,d}}{\partial \alpha \partial \lambda_1} & \frac{-\partial^2 \log LF_{c,d}}{\partial \beta \partial \lambda_1} & \frac{-\partial^2 \log LF_{c,d}}{\partial \lambda_1^2} & \frac{-\partial^2 \log LF_{c,d}}{\partial \lambda_1 \partial \lambda_2} \\ \frac{-\partial^2 \log LF_{c,d}}{\partial \alpha \partial \lambda_2} & \frac{-\partial^2 \log LF_{c,d}}{\partial \beta \partial \lambda_2} & \frac{-\partial^2 \log LF_{c,d}}{\partial \lambda_1 \partial \lambda_2} & \frac{-\partial^2 \log LF_{c,d}}{\partial \lambda_2^2} \end{bmatrix}^{-1}$$

The second order partial derivative of log-likelihood function with respect to the parameters $\alpha,\ \beta,\ \lambda_1,$ and λ_2 are derived and given in the appendix. An approximate estimation of the variance-covariance matrix, $I^{-1}(\hat{\alpha},\hat{\beta},\hat{\lambda_1},\hat{\lambda_2})$, of the parameters can be obtained by replacing the values with the estimate values of $\alpha,\ \beta,\ \lambda_1$ and λ_2, respectively. Hence we may estimate the variance of $\hat{R}_{c,d}$ as

$$\begin{aligned} \hat{Variance}(\hat{R}_{c,d}) &\simeq G^T I^{-1} G \\ &= Variance(\hat{\lambda_1})\Big(\frac{\partial R_{c,d}}{\partial \lambda_1}\Big)^2 + Variance(\hat{\lambda_2})\Big(\frac{\partial R_{c,d}}{\partial \lambda_2}\Big)^2 + 2\frac{\partial R_{c,d}}{\partial \lambda_1}\frac{\partial R_{c,d}}{\partial \lambda_2} I_{12}^{-1} \quad (53) \end{aligned}$$

where

$$Variance(\hat{\lambda_1}) = \left(E\Big(-\frac{\partial^2 \log LF_{c,d}}{\partial \hat{\lambda_1}^{\,2}}\Big)\right)^{-1}$$

$$Variance(\hat{\lambda_2}) = \left(E\Big(-\frac{\partial^2 \log LF_{c,d}}{\partial \hat{\lambda_2}^{\,2}}\Big)\right)^{-1}$$

and

$$I_{12}^{-1} = \left(E\Big(-\frac{\partial^2 \log LF_{c,d}}{\partial \hat{\lambda_1}\partial \hat{\lambda_2}}\Big)\right)^{-1}.$$

for large sample size,

$$\frac{(\hat{R}_{c,d} - R_{c,d})}{\sqrt{\hat{Variance}(\hat{R}_{c,d})}} \sim N(0,1)$$

and the asymptotic $100(1-\eta)\%$ confidence interval for $R_{c,d}$ is given by

$$\hat{R}_{c,d} \pm Z_{\eta/2}\sqrt{\hat{Variance}(\hat{R_{c,d}})}$$

where $\hat{R}_{c,d}$ is the MLE of $R_{c,d}$ and $Z_{\eta/2}$ is the upper $(\eta/2)^{th}$ quantile of the standard Normal distribution.

9.3. Simulation study

Here, a simulation study is carried out to compare the performance of MLEs of R and $R_{c,d}$ in terms of their biases and MSEs. Here we use the parameter values $(\alpha,\beta,\lambda_1,\lambda_2) = (2.5, 0.5, 6, 5)$, then the theoretical value of R is 0.5454545.Additionally, we calculate the confidence intervals using the ML method.The simulation results of R are given in Table 5. For $R_{c,d}$, take the values $(c,d) = \{(2,4),(3,6),(1,3)\}$ in each sample size (l,z). The simulation results of $R_{c,d}$ are reported in Table 6.

From the simulation results of both R and $R_{c,d}$, it is noted that as the sample size (l,z) increases, the biases and MSE values decrease. For the single-component stress-strength model, we considered $(l,z) = \{(10,10),(30,30),(50,50),(100,100)\}$. In the case $R_{c,d}$, we are considering another combination of sample size (l,z) as given in Table 6. The theoretical values of $R_{c,d}$ for different values of $(c,d) = \{(2,4),(3,6),(1,3)\}$ are 0.6476762, 0.6228523, and 0.7826087, respectively.

Table 5: Simulation study for Stress-Strength reliability for parameter values $\alpha = 2.5, \beta = 0.5, \lambda_1 = 6, \lambda_2 = 5, R = 0.5454545$

(l,z)	$\hat{R}$	Bias	MSE	95% ACI
(10,10)	0.5477	0.0022	0.0144	(0.54750, 0.54793)
(30,30)	0.5467	0.0012	0.0041	(0.54564, 0.54772)
(50,50)	0.5451	-0.0004	0.0026	(0.52439, 0.56579)
(100,100)	0.5453	-0.0002	0.0013	(0.48744, 0.60314)

Table 6: Simulation study for Multi-component Stress-Strength reliability for parameter values $\alpha = 2.5, \beta = 0.5, \lambda_1 = 6, \lambda_2 = 5$

$R_{2,4} = 0.6476762, R_{3,6} = 0.6228523, R_{1,3} = 0.7826087$

(c,d)	l	$\hat{R_{cd}}$	Bias	MSE	95% ACI
(2,4)	10	0.6824	0.0347	0.0012	(0.68195, 0.68270)
	30	0.6767	0.0289	0.0008	(0.66641,0.68694)
	40	0.6542	0.0065	4.1882e-05	(0.65342,0.65488)
	80	0.6296	-0.0179	0.0003	(0.62722,0.63228)
	500	0.6447	-0.0029	8.9723e-06	(0.61204,0.6773)
(3,6)	10	0.5876	0.0353	0.0013	(0.57622, 0.59892)
	20	0.6552	0.03231	0.0010	(0.58033,0.72999)
	50	0.6413	0.0184	0.0003	(0.63725,0.64531)
	100	0.6366	0.0138	0.0002	(0.53606,0.73716)
	500	0.6337	0.0108	0.0001	(0.62503,0.64230)
(1,3)	10	0.8313	0.0487	0.0024	(0.27759,1.38508)
	40	0.7972	0.01457	0.0002	(0.72687, 0.86749)
	80	0.7860	0.00342	1.1702e-05	(0.78464,0.78856)
	100	0.7729	0.0022	4.6651e-06	(0.75395,0.79199)
	500	0.7846	0.00191	3.7655e-06	(0.75172,0.81736)

9.4. Data analysis

In this section, we analyze two real datasets introduced by Badar and Priest (1982) to illustrate the use of our proposed estimation method. The first data set (denoted by U) is strength measured in GPA for single carbon fibers tested under tension at a gauge length of 20mm. The second one (denoted by V) is the strength measured in GPA for single carbon fiber tested under tension at a gauge of 10 mm.

The PGDUS-IK(α, β, λ) model fits both data sets. The estimated values of the parameters are obtained. Log-likelihood values, KS values with corresponding p-values, CVM values with corresponding p-values, AIC, and BIC values for both datasets are given in the table. The estimated value for reliability is obtained as 0.2127864.

In the case of the multi-component stress-strength model, the same data set fits with the model for each value of $(c, d) = \{(1, 3), (2, 4), (3, 6)\}$. The parameter estimators,reliability estimate value, K-S values with p-value, and CVM values with p-value are given in Table 8.

Table 7: Stress-Strength Data analysis

Estimates		K-S(p-value)	CVM(p-value)	Log-Likelihood	AIC	BIC
$\hat{\alpha} = 7.47$	X	0.13565	0.36097	-47.36073	98.72146	103.0077
$\hat{\beta}$ =2476.96		(0.1966)	(0.09158)			
$\hat{\lambda_1}$=1.36468	Y	0.087442	0.07484	-135.015	274.03	278.3136
$\hat{\lambda_2}$=5.04871		(0.7211)	(0.724)			

Table 8: Data Analysis of Multi-Component SSR Model

(c,d)	$\hat{R}_{c,d}$	Estimates	U		V	
			K-S (p)	CVM (p)	K-S (p)	CVM (p)
(1,3)	0.45163	7.3788 2124.06 1.4377 5.2369	0.13587 (0.1952)	0.35368 (0.09586)	0.08338 (0.7735)	0.07191 (0.7417)
(2,4)	0.23708	7.3625 2072.29 1.4474 5.2571	0.13651 (0.1909)	0.35254 (0.09655)	0.08263 (0.7829)	0.07148 (0.7443)
(3,6)	0.18989	7.3441 2017.05 1.4574 5.2762	0.13725 (0.1862)	0.35133 (0.09729)	0.08178 (0.7935)	0.07104 (0.747)

10. Summary

The present paper proposes a new lifetime distribution, called the PGDUS-IK distribution, with parameters α, β, and λ, respectively, by using the PGDUS transformation on the IK (α, β) distribution for modeling a parallel system. The statistical properties, including moments, moment generating function, characteristic function, cumulant generating function, quantile function, order statistics, and entropy, are derived. Also, the expected additional lifetime given that the system has survived until a time t is defined in terms of its mean residual life function. Then we move on to the topic estimation of unknown parameters α, β, and λ of the proposed distribution. In this paper, we consider different types of estimation methods, such as the MLE method, the method of maximum product spacing estimation, the method of moment estimation, the method of least squares estimation, and bayesian analysis, respectively. The confidence interval is a range of values that describes the uncertainty around an estimate. For PGDUS-IK(α, β, λ), asymptotic confidence interval and bootstrap confidence interval are obtained. Simulation of data from the proposed distribution is obtained by three different methods: MLE, MPS, and Bayesian. Table 1 shows that, biases and MSEs for the parameters α, β, and λ decrease with increasing sample size. A dataset of vinyl chloride data obtained from clean upgrading and monitoring wells is used for the data analysis. It can be concluded that the proposed PGDUS-IK is effective in providing a better fit of data when compared with other competing distributions, such as the DUS-IK, IK, and Weibull distributions. Stress-strength reliability for single-component and multi-component

models is discussed. Reliability estimates in both models are obtained from the parameter estimate values. The asymptotic distributions of single-component stress-strength reliability and multi-component stress-strength reliability are derived. As the sample size increases, the biases and MSEs of the simulated estimator of reliability in both models decrease. Both the single-component SSR model and the multi-component SSR model are applied to real data obtained from Badar and Priest (1982) and show that both models fit the data.

Acknowledgements

We are indeed grateful to the Editors for their guidance and counsel. We are very grateful to the reviewer for valuable comments and suggestions of generously listing many useful references.

References

Abd AL-Fattah, A. M., El-Helbawy, A. A., and Al-Dayian, G. R. (2017). Inverted Kumaraswamy distribution: properties and estimation. *Pakistan Journal of Statistics*, **33(1)**, 37–61.

Anakha, K. K. and Chacko, V. M. (2021). DUS-Kumaraswamy distribution: a Bathtub shaped failure rate model. *International Journal of Statistics and Reliability Engineering*, 359–367.

Ali, A. H. and Kanani, I. H. A. (2021). Bayesian methods to estimate the parameters of exponentiated Weibull distribution. *In Journal of Physics: Conference Series* (Vol. 1818, No. 1, p. 012143). IOP Publishing.

Bader, M. G. and Priest, A. M. (1982). Statistical aspects of fibre and bundle strength in hybrid composites. *Progress in Science and Engineering of Composites*, 1129–1136.

Basu, S. and Kundu, D. (2023). On three-parameter generalized exponential distribution. *Communication in Statistics- Simulation and Computation*. `https://doi.org/10.1080/03610918.2023.2226468`

Bhaumik, D.K., Kapur, K., and Gibbons, R.D. (2009). Testing parameters of a Gamma distribution for small samples. *Technometrics*, **51**, 326–334.

Cheng, R. C. H. and Amin, N. A. K. (1983). Estimating parameters in continuous univariate distributions with a shifted origin. *Journal of the Royal Statistical Society: Series B (Methodological)*, 394–403.

Deepthi, K. S. and Chacko, V. M.(2020). An upside-down bathtub-shaped failure rate model using a DUS transformation of Lomax distribution, Lirong Cui, Ilia Frenkel, Anatoly Lisnianski (Eds). *Stochastic Models in Reliability Engineering*, **6**, 81–100, Taylor and Francis Group, Boca Raton, CRC Press.

Dumonceaux, R. and Antle, C.E. (1973). Discrimination between the log-normal and the Weibull distributions. *Technometrics*, **15(4)**, 923–926.

Gauthami, P. and Chacko, V. M. (2021). Dus transformation of inverse Weibull distribution: an upside-down failure rate model. *Reliability Theory and Applications*, **16(2)**, 58–71.

Hastings, W.K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, **57**, 97–101.

Iqbal, Z., Tahir, M. M., Riaz, N., *et al.* (2017). Generalized inverted Kumaraswamy distribution: properties and application, *Open Journal of Statistics*, **7**, 645–662.

Jamal, F., Arslan Nasir, M., Ozel, G., *et al.* (2019). Generalized inverted Kumaraswamy generated family of distributions: theory and applications. *Journal of Applied Statistics*, **46(16)**, 2927–2944.

Kumar, D., Singh, U., and Singh, S. K. (2015). A method of proposing new distribution and its application to bladder cancer patients data. *Journal of Statistics Applications and Probability Letters*, **2(3)**, 235–245.

Kumaraswamy, P. (1976). Stochastic simulation of weekly hydrological processes (with computer programs), Part 1. *Institute of Hydraulics and Hydrology, Poondi*, 34–72.

Kumaraswamy, P. (1980). A generalized probability density function for double-bounded random processes. *Journal of Hydrology*, **46(1)**, 79–88.

Lindley, D. V. (1980). Approximate Bayesian methods. *Trabajos de estadística y de investigación operativa*, **31**, 223–245.

Maurya, S. K., Kaushik, A., Singh, S. K., and Singh, U. (2016). A new class of exponential transformed Lindley distribution and its application to Yarn data. *International Journal of Statistics and Economics*, **18(2)**.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, **21(6)**, 1087–1092.

Ranneby, B. (1984). The maximum spacing method: An estimation method related to the maximum likelihood method. *Scandinavian Journal of Statistics*, 93–112.

Thomas and Chacko, V. M. (2021). Power generalized DUS transformation of exponential distribution. *International Journal of Statistics and Reliability Engineering*, **8(3)**, 359–367.

Thomas and Chacko, V. M. (2023). Power generalized DUS transformation in Weibull and Lomax distributions. *Reliability: Theory and Applications*, **18(1)**, 368–384.

Tobias, J. L. (2014). Primer on the Use of Bayesian Methods in Health Economics, in *Encyclopedia of Health Economics*, pp 146–154.

Tripathi, A., Singh, U., and Singh, S. K. (2021). Inferences for the DUS-Exponential distribution based on upper record values. *Annals of Data Science*, **8**, 387–403.

Appendix

$$\log LF(x) = l\Bigg(\log\alpha + \log\beta + \log\lambda - \lambda\log(e-1)\Bigg) - (\alpha+1)\sum_{i=1}^{l}\log(1+u_i)$$
$$+ (\beta-1)\sum_{i=1}^{l}\log(1-(1+u_i)^{-\alpha}) + \sum_{i=1}^{l}(1-(1+u)^{-\alpha})^{\beta}$$
$$+ (\lambda-1)\sum_{i=1}^{l}\log\left(e^{(1-(1+u_i)^{-\alpha})^{\beta}}-1\right).$$

$$\frac{\partial^2\log LF}{\partial\alpha^2} = \frac{-l}{\alpha^2} + (\beta-1)\sum_{i=1}^{l}\left(\frac{(1+u_i)^{-\alpha-1}(1-(1+u_i)^{-\alpha}+(1+u_i)\log^2(1+u_i))}{(1-(1+u_i)^{-\alpha})^2}\right)$$
$$+ \quad \beta\sum_{i=1}^{l}(1+u_i)^{-\alpha}(1-(1+u_i)^{-\alpha})^{\beta-2}\log^2(1+u_i)(\beta(1+u_i)^{-\alpha}-1)$$
$$+ \quad \beta^2(\lambda-1)\sum_{i=1}^{l}\left(\frac{(1+u_i)^{-2\alpha}(1-(1+u_i)^{-\alpha})^{\beta-2}\log^2(1+u_i)e^{(1-(1+u_i)^{-\alpha})^{\beta}}}{(e^{(1-(1+u_i)^{-\alpha})^{\beta}}-1)^2}\right.$$
$$\left.\left(e^{(1-(1+u_i)^{-\alpha})^{\beta}}-(1-(1+u_i)^{-\alpha})^{\beta-1}-1\right)\right)$$
$$- \quad \beta(\lambda-1)\frac{\sum_{i=1}^{l}(1+u_i)^{-\alpha}(1-(1+u_i)^{-\alpha})^{\beta-2}e^{(1-(1+u_i)^{-\alpha})^{\beta}}\log^2(1+u_i)}{(e^{(1-(1+u_i)^{-\alpha})^{\beta}}-1)}$$

$$\frac{\partial^2\log LF}{\partial\beta^2} = \frac{-l}{\beta^2} + \sum_{i=1}^{l}(1-(1+u_i)^{-\alpha})^{\beta}\log^2(1-(1+u_i)^{-\alpha})$$
$$+ \quad \left((\lambda-1)\sum_{i=1}^{l}\frac{(1-(1+u_i)^{-\alpha})^{\beta}e^{(1-(1+u_i)^{-\alpha})^{\beta}}\log^2(1-(1+u_i)^{-\alpha})}{(e^{(1-(1+u_i)^{-\alpha})^{\beta}}-1)^2}\right.$$
$$\left.\left(e^{(1-(1+u_i)^{-\alpha})^{\beta}}-(1-(1+u_i)^{-\alpha})^{\beta}-1\right)\right)$$

$$\frac{\partial^2\log LF}{\partial\alpha\partial\beta} = \sum_{i=1}^{l}\frac{(1+u_i)^{-\alpha}\log(1+u_i)}{1-(1+u_i)^{-\alpha}}$$
$$+\beta\sum_{i=1}^{l}(1+u_i)^{-\alpha}(1-(1+u_i)^{-\alpha})^{\beta-1}\log(1+u_i)(\log(1-(1+u_i)^{-\alpha})+1)$$
$$-\beta(\lambda-1)\sum_{i=1}^{l}\left(\frac{(1+u_i)^{-\alpha}(1-(1+u_i)^{-\alpha})^{2\beta-1}e^{(1-(1+u_i)^{-\alpha})^{\beta}}\log(1+u_i)}{(e^{(1-(1+u_i)^{-\alpha})^{\beta}}-1)^2}\right.$$
$$\left.\log(1-(1+u_i)^{-\alpha})\right) + (\lambda-1)\sum_{i=1}^{l}\left(\frac{(1+u_i)^{-\alpha}(1-(1+u_i)^{-\alpha})^{\beta-1}e^{(1-(1+u_i)^{-\alpha})^{\beta}}}{e^{(1-(1+u_i)^{-\alpha})^{\beta}}-1}\right.$$
$$\left.(1+\beta(1-(1+u_i)^{-\alpha})\log(1+u_i)\log(1-(1+u_i)^{-\alpha})\right)$$

$$\frac{\partial^2\log LF}{\partial\lambda^2} = \frac{-l}{\lambda^2}$$

$$\frac{\partial^2 \log LF}{\partial\alpha\partial\lambda} = \beta\sum_{i=1}^{l}\frac{(1+u_i)^{-\alpha}(1-(1+u_i)^{-\alpha})^{\beta-1}e^{(1-(1+u_i)^{-\alpha})^{\beta}}\log(1+u_i)}{e^{(1-(1+u_i)^{-\alpha})^{\beta}}-1}$$

$$\frac{\partial^2 \log LF}{\partial\beta\partial\lambda} = \sum_{i=1}^{l}\frac{(1-(1+u_i)^{-\alpha})^{\beta}e^{(1-(1+u_i)^{-\alpha})^{\beta}}\log(1-(1+u_i)^{-\alpha})}{e^{(1-(1+u_i)^{-\alpha})^{\beta}}-1}$$

$$M_{11} = \frac{\partial M}{\partial^2\alpha}, \quad M_{22} = \frac{\partial M}{\partial^2\beta}, \quad M_{33} = \frac{\partial M}{\partial^2\lambda}$$

$$M_{12} = M_{21} = \frac{\partial^2 M}{\partial\alpha\partial\beta}, \quad M_{13} = M_{31} = \frac{\partial^2 M}{\partial\alpha\partial\lambda}, \quad M_{23} = M_{32} = \frac{\partial^2 M}{\partial\beta\partial\lambda}$$

$$M_{111} = \beta\sum\log^2(1+u_i)\Bigg(\frac{(1+u_i)^{-\alpha}(1-(1+u_i)^{-\alpha})^{\beta-2}e^{(1-(1+u_i)^{-\alpha})^{\beta}}\Big(\beta(1+u_i)^{-\alpha}-1\Big)}{e^{(1-(1+u_i)^{-\alpha})^{\beta}}-1}$$
$$-\frac{\beta(1+u_i)^{-2\alpha}(1-(1+u_i)^{-\alpha})^{2(\beta-1)}e^{(1-(1+u_i)^{-\alpha})^{\beta}}}{(e^{(1-(1+u_i)^{-\alpha})^{\beta}}-1)^2}\Bigg)$$

$$M_{122} = M_{221} = M_{212}$$
$$= \sum(1+u_i)^{-\alpha}\log(1+u_i)(1-(1+u_i)^{-\alpha})^{\beta-1}\log(1-(1+u_i)^{-\alpha})$$
$$\Big(2+\log(1-(1+u_i)^{-\alpha})\Big)+(\lambda-1)\sum(1+u_i)^{-\alpha}\log(1+u_i)\log(1-(1+u_i)^{-\alpha})$$
$$\left(\frac{(1-(1+u_i)^{-\alpha})^{\beta-1}e^{(1-(1+u_i)^{-\alpha})^{\beta}}(e^{(1-(1+u_i)^{-\alpha})^{\beta}}-1-(1-(1+u_i)^{-\alpha})^{\beta})}{\Big(e^{(1-(1+u_i)^{-\alpha})^{\beta}}-1\Big)^2}\right)$$
$$+(\lambda-1)\sum(1+u_i)^{-\alpha}\log(1+u_i)\log(1-(1+u_i)^{-\alpha})(1-(1+u_i)^{-\alpha})^{\beta-1}e^{(1-(1+u_i)^{-\alpha})^{\beta}}$$
$$\Bigg(\frac{\beta(1-(1+u_i)^{-\alpha})^{\beta}\log(1-(1+u_i)^{-\alpha})}{\Big(e^{(1-(1+u_i)^{-\alpha})^{\beta}}-1\Big)^4}+\frac{(1+(1-(1+u_i)^{-\alpha})^{\beta})}{\Big(e^{(1-(1+u_i)^{-\alpha})^{\beta}}-1\Big)^4}$$
$$\Big(1+\beta\log(1-(1+u_i)^{-\alpha})\Big(1+(1-(1+u_i)^{-\alpha})^{\beta}(1+e^{(1-(1+u_i)^{-\alpha})^{\beta}})\Big)\Big)$$
$$-\frac{2(1-(1+u_i)^{-\alpha})^{2\beta-1}e^{2(1-(1+u_i)^{-\alpha})^{\beta}}(1+(1-(1+u_i)^{-\alpha})^{\beta})log(1-(1+u_i)^{-\alpha})}{\Big(e^{(1-(1+u_i)^{-\alpha})^{\beta}}-1\Big)^2}\Bigg)$$

$$M_{112} = M_{121} = M_{211} = -\sum\frac{(1+u_i)^{-\alpha}\log^2(1+u_i)}{(1-(1+u_i)^{-\alpha})^2}+\sum(1+u_i)^{-\alpha}\log^2(1+u_i)$$
$$(1-(1+u_i)^{-\alpha})^{\beta-2}\Big((\beta(1+u_i)^{-\alpha}-1)(1+\beta\log(1-(1+u_i)^{-\alpha}))+\beta(1+u_i)^{-\alpha}\Big)$$
$$-(\lambda-1)\sum(1+u_i)^{-\alpha}\log^2(1+u_i)(1-(1+u_i)^{-\alpha})^{\beta-2}e^{(1-(1+u_i)^{-\alpha})^{\beta}}$$
$$\left(\frac{(e^{(1-(1+u_i)^{-\alpha})^{\beta}}-1)(1+\beta\log(1-(1+u_i)^{-\alpha}))-\beta(1-(1+u_i)^{-\alpha})^{\beta}\log(1-(1+u_i)^{-\alpha})}{(e^{(1-(1+u_i)^{-\alpha})^{\beta}}-1)^2}\right)$$
$$+\beta(\lambda-1)\sum(1+u_i)^{-2\alpha}\log^2(1+u_i)\Bigg((1-(1+u_i)^{-\alpha})^{\beta}e^{(1-(1+u_i)^{-\alpha})^{\beta}}$$
$$\left(\frac{(e^{(1-(1+u_i)^{-\alpha})^{\beta}}-(1-(1+u_i)^{-\alpha})^{\beta-1}-1)(2+\beta\log(1-(1+u_i)^{-\alpha})(1+(1-(1+u_i)^{-\alpha})^{\beta}))}{(e^{(1-(1+u_i)^{-\alpha})^{\beta}}-1)^2}\right.$$

$$+\frac{(1-(1+u_i)^{-\alpha})^{\beta-3}((1-(1+u_i)^{-\alpha})e^{(1-(1+u_i)^{-\alpha})^{\beta}}-1)\log(1-(1+u_i)^{-\alpha})}{(e^{(1-(1+u_i)^{-\alpha})^{\beta}}-1)^2}\Bigg)$$

$$-\frac{2\beta(1-(1+u_i)^{-\alpha})^{2\beta-2}e^{2(1-(1+u_i)^{-\alpha})^{\beta}}\log(1-(1+u_i)^{-\alpha})}{(e^{(1-(1+u_i)^{-\alpha})^{\beta}}-1)^3}\Bigg)$$

$$M_{123}=\sum(1+u_i)^{-\alpha}\log(1+u_i)(1-(1+u_i)^{-\alpha})^{\beta-1}e^{(1-(1+u_i)^{-\alpha})^{\beta}}$$

$$\left(\frac{1}{e^{(1-(1+u_i)^{-\alpha})^{\beta}}-1}+\beta\frac{(1-(1+u_i)^{-\alpha})^{\beta-1}\log(1-(1+u_i)^{-\alpha})e^{(1-(1+u_i)^{-\alpha})^{\beta}}}{(e^{(1-(1+u_i)^{-\alpha})^{\beta}}-1)^2}\right)$$

$$\left(e^{(1-(1+u_i)^{-\alpha})^{\beta}}-1-(1-(1+u_i)^{-\alpha})^{\beta}\right)$$

$$M_{231}=\sum\frac{(1+u_i)^{-\alpha}(1-(1+u_i)^{-\alpha})^{\beta-1}e^{(1-(1+u_i)^{-\alpha})^{\beta}}\log(1+u_i)}{e^{(1-(1+u_i)^{-\alpha})^{\beta}}-1}$$

$$+\beta\sum\log^2(1+u_i)(1+u_i)^{-\alpha}(1-(1+u_i)^{-\alpha})^{\beta-2}e^{(1-(1+u_i)^{-\alpha})^{\beta}}$$

$$\left(\frac{\left((\beta(1+u_i)^{-\alpha}-1)(e^{(1-(1+u_i)^{-\alpha})^{\beta}}-1)-\beta(1+u_i)^{-\alpha}(1-(1+u_i)^{-\alpha})^{\beta}\right)}{(e^{(1-(1+u_i)^{-\alpha})^{\beta}}-1)^2}\right)$$

$$M_{113}=M_{131}=M_{311}=\beta^2\sum(1+u_i)^{-2\alpha}(1-(1+u_i)^{-\alpha})^{\beta-1}\log^2(1+u_i)e^{(1-(1+u_i)^{-\alpha})^{\beta}}$$

$$\frac{e^{(1-(1+u_i)^{-\alpha})^{\beta}}-(1-(1+u_i)^{-\alpha})^{\beta-1}-1}{(e^{(1-(1+u_i)^{-\alpha})^{\beta}}-1)^2}$$

$$-\sum\frac{\beta(1+u_i)^{-\alpha}(1-(1+u_i)^{-\alpha})^{\beta-2}e^{(1-(1+u_i)^{-\alpha})^{\beta}}\log^2(1+u_i)}{e^{(1-(1+u_i)^{-\alpha})^{\beta}}-1}$$

$$M_{223}=M_{232}=M_{322}=\sum\Bigg(\log^2(1-(1+u_i)^{-\alpha})(1-(1+u_i)^{-\alpha})^{\beta}e^{(1-(1+u_i)^{-\alpha})^{\beta}}$$

$$\frac{\left(e^{(1-(1+u_i)^{-\alpha})^{\beta}}-1-(1-(1+u_i)^{-\alpha})^{\beta}\right)}{(e^{(1-(1+u_i)^{-\alpha})^{\beta}}-1)^2}\Bigg)$$

$$M_{331}=0=M_{332}$$

$$M_{222}=\frac{2l}{\beta^3}\sum(1-(1+u_i)^{(}-\alpha))^{\beta}\log^3(1-(1+u_i)^{-\alpha})$$

$$+(\lambda-1)\sum(1-(1+u_i)^{-\alpha})^{\beta}e^{(1-(1+u_i)^{-\alpha})^{\beta}}\log^3(1-(1+u_i)^{-\alpha})$$

$$\frac{e^{(1-(1+u_i)^{-\alpha})^{\beta}}(1+2(1-(1+u_i)^{-\alpha})^{\beta})-(1-(1+u_i)^{-\alpha})^{\beta}(3+(1-(1+u_i)^{-\alpha})^{\beta})-1}{(e^{(1-(1+u_i)^{-\alpha})^{\beta}}-1)^2}$$

$$-2\sum(1-(1+u_i)^{-\alpha})^{2\beta}e^{2(1-(1+u_i)^{-\alpha})^{\beta}}\log^3(1-(1+u_i)^{-\alpha})$$

$$\frac{e^{(1-(1+u_i)^{-\alpha})^{\beta}}-(1-(1+u_i)^{-\alpha})^{\beta}-1}{(e^{(1-(1+u_i)^{-\alpha})^{\beta}}-1)^3}$$

$$M_{333}=\frac{2l}{\lambda^3}$$

In the case stress-strength reliability,

$$\frac{\partial^2 \log LF}{\partial \lambda_1^2} = \frac{-l}{\lambda_1^2}$$

$$\frac{\partial^2 \log LF}{\partial \lambda_2^2} = \frac{-o}{\lambda_2^2}$$

$$\frac{\partial^2 \log LF}{\partial \lambda_1 \partial \lambda_2} = 0$$

$$\frac{\partial^2 \log LF}{\partial \beta \partial \lambda_1} = \sum_{i=1}^{l} \frac{(1-(1+u_i)^{-\alpha})^{\beta} e^{(1-(1+u_i)^{-\alpha})^{\beta}} \log(1-(1+u_i)^{-\alpha})}{e^{(1-(1+u_i)^{-\alpha})^{\beta}} - 1}$$

$$\frac{\partial^2 \log LF}{\partial \beta \partial \lambda_2} = \sum_{j=1}^{o} \frac{(1-(1+v_j)^{-\alpha})^{\beta} e^{(1-(1+v_j)^{-\alpha})^{\beta}} \log(1-(1+v_j)^{-\alpha})}{e^{(1-(1+v_j)^{-\alpha})^{\beta}} - 1}$$

$$\frac{\partial^2 \log LF}{\partial \alpha \partial \lambda_1} = \beta \sum_{i=1}^{l} \frac{(1+u_i)^{-\alpha}(1-(1+u_i)^{-\alpha})^{\beta-1} e^{(1-(1+u_i)^{-\alpha})^{\beta}} \log(1+u_i)}{e^{(1-(1+u_i)^{-\alpha})^{\beta}} - 1}$$

$$\frac{\partial^2 \log LF}{\partial \alpha \partial \lambda_2} = \beta \sum_{j=1}^{o} \frac{(1+v_j)^{-\alpha}(1-(1+v_j)^{-\alpha})^{\beta-1} e^{(1-(1+v_j)^{-\alpha})^{\beta}} \log(1+v_j)}{e^{(1-(1+v_j)^{-\alpha})^{\beta}} - 1}$$

$$\frac{\partial^2 \log LF}{\partial \alpha \partial \beta} = \sum_{i=1}^{l} \frac{(1+u_i)^{-\alpha} \log(1+u_i)}{(1-(1+u_i)^{-\alpha})} + \sum_{j=1}^{o} \frac{(1+v_j)^{-\alpha} \log(1+v_j)}{(1-(1+v_j)^{-\alpha})}$$

$$\begin{aligned}\frac{\partial^2 \log LF}{\partial \beta^2} = & -\frac{l+o}{\beta^2} + \sum_{i=1}^{l} (1-(1+u_i)^{-\alpha})^{\beta} \log^2(1-(1+u_i)^{-\alpha}) \\ & + \sum_{j=1}^{o} (1-(1+v_j)^{-\alpha})^{\beta} \log^2(1-(1+v_j)^{-\alpha}) \\ & + (\lambda_1 - 1) \sum_{i=1}^{l} \left(\frac{(1-(1+u_i)^{-\alpha})^{\beta} \log^2(1-(1+u_i)^{-\alpha}) e^{(1-(1+u_i)^{-\alpha})^{\beta}}}{e^{(1-(1+u_i)^{-\alpha})^{\beta}} - 1} \right. \\ & \left. - \frac{(1-(1+u_i)^{-\alpha})^{2\beta} \log^2(1-(1+u_i)^{-\alpha}) e^{(1-(1+u_i)^{-\alpha})^{\beta}}}{(e^{(1-(1+u_i)^{-\alpha})^{\beta}} - 1)^2} \right) \\ & + (\lambda_2 - 1) \sum_{j=1}^{o} \left(\frac{(1-(1+v_j)^{-\alpha})^{\beta} \log^2(1-(1+v_j)^{-\alpha}) e^{(1-(1+v_j)^{-\alpha})^{\beta}}}{e^{(1-(1+v_j)^{-\alpha})^{\beta}} - 1} \right.\end{aligned}$$

$$- \frac{(1-(1+v_j)^{-\alpha})^{2\beta}\log^2(1-(1+v_j)^{-\alpha})e^{(1-(1+v_j)^{-\alpha})^{\beta}}}{(e^{(1-(1+v_j)^{-\alpha})^{\beta}}-1)^2}\Bigg)$$

$$\frac{\partial^2 \log LF}{\partial\alpha^2} = -\frac{l+o}{\alpha^2} - (\beta-1)\Bigg(\sum_{i=1}^{l}\frac{(1+u_i)^{-\alpha}\log^2(1+u_i)}{(1-(1+u_i)^{-\alpha})^2} + \sum_{j=1}^{o}\frac{(1+v_j)^{-\alpha}\log^2(1+v_j)}{(1-(1+v_j)^{-\alpha})^2}\Bigg)$$

$$+ \quad \beta\Bigg(\sum_{i=1}^{l}(1+u_i)^{-\alpha}(1-(1+u_i)^{-\alpha})^{\beta-2}(\beta(1+u_i)^{-\alpha}-1)\log^2(1+u_i)$$

$$+ \quad \sum_{j=1}^{o}(1+v_j)^{-\alpha}(1-(1+v_j)^{-\alpha})^{\beta-2}(\beta(1+v_i)^{-\alpha}-1)\log^2(1+v_j)\Bigg)$$

$$+ \quad \beta(\lambda_1-1)\sum_{i=1}^{l}\Bigg(-\beta\frac{(1+u_i)^{-2\alpha}(1-(1+u_i)^{-\alpha})^{2(\beta-1)}e^{(1-(1+u_i)^{-\alpha})^{\beta}}\log^2(1+u_i)}{(e^{(1-(1+u_i)^{-\alpha})^{\beta}}-1)^2}$$

$$+ \quad \frac{(1+u_i)^{-\alpha}(1-(1+u_i)^{-\alpha})^{(\beta-2)}e^{(1-(1+u_i)^{-\alpha})^{\beta}}(\beta(1+u_i)^{-\alpha}-1)\log^2(1+u_i)}{e^{(1-(1+u_i)^{-\alpha})^{\beta}}-1}\Bigg)$$

$$+ \quad \beta(\lambda_2-1)\sum_{j=1}^{o}\Bigg(-\beta\frac{(1+v_j)^{-2\alpha}(1-(1+v_j)^{-\alpha})^{2(\beta-1)}e^{(1-(1+v_j)^{-\alpha})^{\beta}}\log^2(1+v_j)}{(e^{(1-(1+v_j)^{-\alpha})^{\beta}}-1)^2}$$

$$+ \quad \frac{(1+v_j)^{-\alpha}(1-(1+v_j)^{-\alpha})^{(\beta-2)}e^{(1-(1+v_j)^{-\alpha})^{\beta}}(\beta(1+v_j)^{-\alpha}-1)\log^2(1+v_j)}{e^{(1-(1+v_j)^{-\alpha})^{\beta}}-1}\Bigg)$$

In the case of multi-component stress-strength reliability, the log-likelihood function is given as

$$\log LF_{c,d} = l(d+1)\Big(\log\alpha+\log\beta\Big) + ld\log\lambda_1 + l\log\lambda_2 - l(d\lambda_1+\lambda_2)\log(e-1)$$
$$-(\alpha+1)\Big(\textstyle\sum_{i=1}^{l}\sum_{j=1}^{d}\log(1+u_{ij}) + \sum_{i=1}^{l}\log(1+v_i)\Big) + (\beta-1)\Big(\sum_{i=1}^{l}\sum_{j=1}^{d}\log(1-(1+u_{ij})^{-\alpha})$$
$$+\sum_{i=1}^{l}\log(1-(1+v_i)^{-\alpha})\Big) + \sum_{i=1}^{l}\sum_{j=1}^{d}(1-(1+u_{ij})^{-\alpha})^{\beta} + \sum_{i=1}^{d}(1-(1+v_i)^{-\alpha})^{\beta}$$
$$+(\lambda_1-1)\textstyle\sum_{i=1}^{l}\sum_{j=1}^{d}\log\Big(e^{(1-(1+u_{ij})^{-\alpha})^{\beta}}-1\Big) + (\lambda_2-1)\sum_{i=1}^{l}\log\Big(e^{(1-(1+v_i)^{-\alpha})^{\beta}}-1\Big).$$

$$\frac{\partial^2 \log LF_{c,d}}{\partial\lambda_1^2} = \frac{-ld}{\lambda_1^2}$$

$$\frac{\partial^2 \log LF_{c,d}}{\partial\lambda_1\partial\lambda_2} = 0 = \frac{\partial^2 \log LF_{c,d}}{\partial\lambda_2\partial\lambda_1}$$

$$\frac{\partial^2 \log LF_{c,d}}{\partial\alpha\partial\lambda_1} = \beta\sum_{i=1}^{l}\sum_{j=1}^{c}\frac{(1+u_{ij})^{-\alpha}(1-(1+u_{ij})^{-\alpha})^{\beta-1}e^{(1-(1+u_{ij})^{-\alpha})^{\beta}}\log(1+u_{ij})}{e^{(1-(1+u_{ij})^{-\alpha})^{\beta}}-1}$$

$$\frac{\partial^2 \log LF_{c,d}}{\partial\beta\partial\lambda_1} = \sum_{i=1}^{l}\sum_{j=1}^{c}\frac{(1-(1+u_{ij})^{-\alpha})^{\beta}e^{(1-(1+u_{ij})^{-\alpha})^{\beta}}\log(1-(1+u_{ij})^{-\alpha})}{e^{(1-(1+u_{ij})^{-\alpha})^{\beta}}-1}$$

$$\frac{\partial^2 \log LF_{c,d}}{\partial\lambda_2^2} = \frac{-l}{\lambda_2^2}$$

$$\frac{\partial^2 \log LF_{c,d}}{\partial\alpha\partial\lambda_2} = \beta\sum_{i=1}^{l}\frac{(1+v_i)^{-\alpha}(1-(1+v_i)^{-\alpha})^{\beta-1}e^{(1-(1+v_i)^{-\alpha})^{\beta}}\log(1+v_i)}{e^{(1-(1+v_i)^{-\alpha})^{\beta}}-1}$$

$$\frac{\partial^2 \log LF_{c,d}}{\partial\beta\partial\lambda_2} = \sum_{i=1}^{l}\frac{(1-(1+v_i)^{-\alpha})^{\beta}e^{(1-(1+v_i)^{-\alpha})^{\beta}}\log(1-(1+v_i)^{-\alpha})}{e^{(1-(1+v_i)^{-\alpha})^{\beta}}-1}$$

$$\begin{aligned}\frac{\partial^2 \log LF_{c,d}}{\partial\alpha\partial\beta} &= \sum_{i=1}^{l}\sum_{j=1}^{c}\frac{(1+u_{ij})^{-\alpha}\log(1+u_{ij})}{1-(1+u_{ij})^{-\alpha}} + \sum_{i=1}^{l}\frac{(1+v_i)^{-\alpha}\log(1+v_i)}{1-(1+v_i)^{-\alpha}}\\ &+ \beta\Bigg(\sum_{i=1}^{l}\sum_{j=1}^{c}(1-(1+u_{ij})^{-\alpha})^{\beta-1}\log(1-(1+u_{ij})^{-\alpha})(1+u_{ij})^{-\alpha}\log(1+u_{ij})\\ &+ \sum_{i=1}^{l}(1-(1+v_i)^{-\alpha})^{\beta-1}\log(1-(1+v_i)^{-\alpha})(1+v_i)^{-\alpha}\log(1+v_i)\Bigg)\\ &+ \sum_{i=1}^{l}\sum_{j=1}^{c}(1-(1+u_{ij})^{-\alpha})^{\beta-1}(1+u_{ij})^{-\alpha}\log(1+u_{ij})\\ &+ \sum_{i=1}^{l}(1-(1+v_i)^{-\alpha})^{\beta-1}(1+v_i)^{-\alpha}\log(1+v_i)\\ &+ (\lambda_1-1)\sum_{i=1}^{l}\sum_{j=1}^{c}(1-(1+u_{ij})^{-\alpha})^{\beta-1}(1+u_{ij})^{-\alpha}\log(1+u_{ij})e^{(1-(1+u_{ij})^{-\alpha})^{\beta}}\\ &\left(\frac{(\beta\log(1-(1+u_{ij})^{-\alpha})+1)(e^{(1-(1+u_{ij})^{-\alpha})^{\beta}}-1)-\beta(1-(1+u_{ij})^{-\alpha})^{\beta}\log(1-(1+u_{ij})^{-\alpha})}{(e^{(1-(1+u_{ij})^{-\alpha})^{\beta}}-1)^2}\right)\\ &+ (\lambda_2-1)\sum_{i=1}^{l}(1-(1+v_i)^{-\alpha})^{\beta-1}(1+v_i)^{-\alpha}\log(1+v_i)e^{(1-(1+v_i)^{-\alpha})^{\beta}}\\ &\left(\frac{(\beta\log(1-(1+v_i)^{-\alpha})+1)(e^{(1-(1+v_i)^{-\alpha})^{\beta}}-1)-\beta(1-(1+v_i)^{-\alpha})^{\beta}\log(1-(1+v_i)^{-\alpha})}{(e^{(1-(1+v_i)^{-\alpha})^{\beta}}-1)^2}\right)\end{aligned}$$

$$\begin{aligned}\frac{\partial^2 \log LF_{c,d}}{\partial^2\beta} &= -\frac{l(d+1)}{\beta^2} + \sum_{i=1}^{l}\sum_{j=1}^{c}(1-(1+u_{ij})^{-\alpha})^{\beta}\log^2(1-(1+u_{ij})^{-\alpha})\\ &+ \sum_{i=1}^{l}(1-(1+v_i)^{-\alpha})^{\beta}\log^2(1-(1+v_i)^{-\alpha}) + (\lambda_1-1)\sum_{i=1}^{l}\sum_{j=1}^{c}e^{(1-(1+u_{ij})^{-\alpha})^{\beta}}\end{aligned}$$

$$\frac{(1-(1+u_{ij})^{-\alpha})^{\beta}\log^2(1-(1+u_{ij})^{-\alpha})\Big(e^{(1-(1+u_{ij})^{-\alpha})^{\beta}}-1-(1-(1+u_{ij})^{-\alpha})^{\beta}\Big)}{(e^{(1-(1+u_{ij})^{-\alpha})^{\beta}}-1)^2}$$

$$+(\lambda_2-1)\sum_{i=1}^{l}e^{(1-(1+v_i)^{-\alpha})^{\beta}}(1-(1+v_i)^{-\alpha})^{\beta}\log^2(1-(1+v_i)^{-\alpha})$$

$$\left(\frac{e^{(1-(1+v_i)^{-\alpha})^{\beta}}-1-(1-(1+v_i)^{-\alpha})^{\beta}}{(e^{(1-(1+v_i)^{-\alpha})^{\beta}}-1)^2}\right)$$

$$\frac{\partial^2\log LF_{c,d}}{\partial^2\alpha}=-\frac{l(d+1)}{\alpha^2}-(\beta-1)\left(\sum_{i=1}^{l}\sum_{j=1}^{c}\frac{(1+u_{ij})^{-\alpha}\log^2(1+u_{ij})}{(1-(1+u_{ij})^{-\alpha})^2}+\sum_{i=1}^{l}\frac{(1+v_i)^{-\alpha}\log^2(1+v_i)}{(1-(1+v_i)^{-\alpha})^2}\right)$$

$$+\beta\left(\sum_{i=1}^{l}\sum_{j=1}^{c}(1+u_{ij})^{-\alpha}\log^2(1+u_{ij})(1-(1+u_{ij})^{-\alpha})^{\beta-2}\left((\beta(1+u_{ij})^{-\alpha}-1)\right.\right.$$

$$\left.+\frac{e^{(1-(1+u_{ij})^{-\alpha})^{\beta}}\Big((\beta(1+u_{ij})^{-\alpha}-1)(e^{(1-(1+u_{ij})^{-\alpha})^{\beta}}-1)-\beta(1+u_{ij})^{-\alpha}(1-(1+u_{ij})^{-\alpha})^{\beta}\Big)}{(e^{(1-(1+v_i)^{-\alpha})^{\beta}}-1)^2}\right)$$

$$+\sum_{i=1}^{l}(1+v_i)^{-\alpha}\log^2(1+v_i)(1-(1+v_i)^{-\alpha})^{\beta-2}\left((\beta(1+u_{ij})^{-\alpha}-1)\right.$$

$$\left.\left.+\frac{e^{(1-(1+v_i)^{-\alpha})^{\beta}}\Big((\beta(1+v_i)^{-\alpha}-1)(e^{(1-(1+v_i)^{-\alpha})^{\beta}}-1)-\beta(1+v_i)^{-\alpha}(1-(1+v_i)^{-\alpha})^{\beta}\Big)}{(e^{(1-(1+v_i)^{-\alpha})^{\beta}}-1)^2}\right)\right)$$

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 22, No. 2, 2024 (New Series), pp 361–362
http://www.ssca.org.in/journal

Resolvability of a BIB Design of Takeuchi (1962)

Shyam Saurabh
Department of Mathematics, Tata College, Kolhan University, Chaibasa, India

Received: 15 December 2023; Revised: 11 May 2024; Accepted: 11 June 2024

Abstract

A $(3, 3, 9)$-resolvable solution of a BIB design with parameters: $v = 21, b = 35, r = 15, k = 9, \lambda = 6$, and listed as T47 in the Table of Takeuchi (1962), is obtained. The resolvable solution is obtained by decomposing the incidence matrix into incidence matrices of smaller BIB designs.

Key words: Balanced incomplete block design; Resolvable solution; Circulant matrix.

AMS Subject Classifications: 62K10; 05B05

1. The solution

Let the incidence matrix $\mathbf{N}$ of a balanced incomplete block (BIB) design may be decomposed into submatrices as $\mathbf{N} = [\mathbf{N}_1|\mathbf{N}_2|\dots|\mathbf{N}_t]$ such that each row sum of $\mathbf{N}_i$ $(1 \le i \le t)$ is α_i. Then the design is $(\alpha_1, \alpha_2, \dots, \alpha_t)$-resolvable [see Kageyama (1976)]. If $\alpha_1 = \alpha_2 = \dots = \alpha_t = \alpha$ then the design is α-resolvable.

The following solution of a BIB design with parameters: $v = 21, b = 35, r = 15, k = 9, \lambda = 6$ using the method of differences may be found in Takeuchi (1962):
$[0_0, 1_0, 2_0, 4_0, 0_1, 1_1, 2_1, 4_1, 2_2]; [0_0, 6_0, 5_0, 3_0, 6_2, 4_2, 3_2, 2_2, 0_1]; [0_1, 6_1, 5_1, 3_1, 6_2, 4_2, 3_2, 2_2, 0_0]$;
$[0_0, 2_0, 6_0, 1_1, 3_1, 4_1, 1_2, 2_2, 4_2]; [1_0, 3_0, 4_0, 0_1, 2_1, 6_1, 1_2, 2_2, 4_2]$ mod 7.

The incidence matrix $\mathbf{N}$ of the design may be decomposed into block submatrices as follows:

$$\mathbf{N} = (\mathbf{N}_1|\mathbf{N}_2|\mathbf{N}_3) =$$

$$\left(\begin{array}{c|c|cc|c} \beta+\beta^3+\beta^4 & \mathbf{I}_7+\beta^2+\beta^6 & \mathbf{I}_7+\beta+\beta^2+\beta^4 & \mathbf{I}_7+\beta^3+\beta^5+\beta^6 & \mathbf{I}_7 \\ \mathbf{I}_7+\beta^2+\beta^6 & \beta+\beta^3+\beta^4 & \mathbf{I}_7+\beta+\beta^2+\beta^4 & \mathbf{I}_7 & \mathbf{I}_7+\beta^3+\beta^5+\beta^6 \\ \beta+\beta^2+\beta^4 & \beta+\beta^2+\beta^4 & \beta^2 & \beta^2+\beta^3+\beta^4+\beta^6 & \beta^2+\beta^3+\beta^4+\beta^6 \end{array}\right)$$

where $\mathbf{I}_7$ is the identity matrix of order 7 and $\beta = circ(0\ 1\ 0 \dots 0)$ is a permutation circulant matrix of order 7 such that $\beta^7 = \mathbf{I}_7$. Since each row sum of the block matrices $\mathbf{N}_1, \mathbf{N}_2$ and $\mathbf{N}_3$ are 3, 3 and 9 respectively, the BIB design is $(3, 3, 9)$-resolvable. The resolvable solution is given below in Table 1.

Further repeating above solution of the BIB design three times, we obtain a 9-resolvable solution of BIB design with parameters: $v = 21, b = 105, r = 45, k = 9, \lambda = 18$.

Corresponding Author: Shyam Saurabh
Email: shyamsaurabh785@gmail.com

This solution may be considered new as this is not reported in the tables of Kageyama (1973), Kageyama and Mohan (1983) and Subramani (1990).

Table 1: (3,3,9)-resolvable solution of the BIB design T47

Replication I	Replication II
(4, 5, 7, 8, 9, 13, 18, 20, 21)	(1, 2, 6, 11, 12, 14, 18, 20, 21)
(1, 5, 6, 9, 10, 14, 15, 19, 21)	(2, 3, 7, 8, 12, 13, 15, 19, 21)
(2, 6, 7, 8, 10, 11, 15, 16, 20)	(1, 3, 4, 9, 13, 14, 15, 16, 20)
(1, 3, 7, 9, 11, 12, 16, 17, 21)	(2, 4, 5, 8, 10, 14, 16, 17, 21)
(1, 2, 4, 10, 12, 13, 15, 17, 18)	(3, 5, 6, 8, 9, 11, 15, 17, 18)
(2, 3, 5, 11, 13, 14, 16, 18, 19)	(4, 6, 7, 9, 10, 12, 16, 18, 19)
(3, 4, 6, 8, 12, 14, 17, 19, 20)	(1, 5, 7, 10, 11, 13, 17, 19, 20)

Replication III		
(1, 4, 6, 7, 8, 11, 13, 14, 20)	(1, 8, 9, 10, 12, 16, 18, 19, 20)	(1, 2, 3, 5, 8, 16, 18, 19, 20)
(1, 2, 5, 7, 8, 9, 12, 14, 21)	(2, 9, 10, 11, 13, 17, 19, 20, 21)	(2, 3, 4, 6, 9, 17, 19, 20, 21)
(1, 2, 3, 6, 8, 9, 10, 13, 15)	(3, 10, 11, 12, 14, 15, 18, 20, 21)	(3, 4, 5, 7, 10, 15, 18, 20, 21)
(2, 3, 4, 7, 9, 10, 11, 14, 16)	(4, 8, 11, 12, 13, 15, 16, 19, 21)	(1, 4, 5, 6, 11, 15, 16, 19, 21)
(1, 3, 4, 5, 8, 10, 11, 12, 17)	(5, 9, 12, 13, 14, 15, 16, 17, 20)	(2, 5, 6, 7, 12, 15, 16, 17, 20)
(2, 4, 5, 6, 9, 11, 12, 13, 18)	(6, 8, 10, 13, 14, 16, 17, 18, 21)	(1, 3, 6, 7, 13, 16, 17, 18, 21)
(3, 5, 6, 7, 10, 12, 13, 14, 19)	(7, 8, 9, 11, 14, 15, 17, 18, 19)	(1, 2, 4, 7, 14, 15, 17, 18, 19)

Acknowledgements

The author is thankful to anonymous reviewers and Editor-in-Chief for their valuable suggestions in improving the presentation of the note.

References

Kageyama, S. (1973). On μ-resolvable and affine μ-resolvable balanced incomplete block designs. *Annals of Statistics*, **1**, 195–203.

Kageyama, S. (1976). Resolvability of block designs. *The Annals of Statistics*, **4**, 655–661.

Kageyama, S. and Mohan, R. N. (1983). On μ-resolvable BIB designs. *Discrete Mathematics*, **45**, 113–122.

Subramani, J. (1990). A note on μ-resolvable BIB designs. *Journal of the Indian Society of Agricultural Statistics*, **XLII**, 226–233.

Takeuchi, K. (1962). A table of difference sets generating balanced incomplete block designs. *Review of International Statistical Institute*, **30**, 361–366.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 22, No. 2, 2024 (New Series), pp 363–368
http://www.ssca.org.in/journal

Improving Data Validation

A. K. Nigam
Consultant Advisor, IASDS, Lucknow

Received: 24 March 2024; Revised: 12 June 2024; Accepted: 16 June 2024

Abstract

Data validation in Official System is usually taken as reporting at different levels, identifying the inconsistencies, and taking remedial measures. The idea of present write-up is to overhaul the concept of Validation and to increase its scope in a multi-pronged way to increase its visibility and utility to the policy makers. This would be of great help to the Official Reporting System and would revolutionize the whole approach to validation.

Key words: Official reporting system; Data validation; Improving data validation.

1. Introduction

Good quality data is a fundamental requirement for framing efficient policies. The COVID-19 experience has taught the world the importance of timely availability of reliable and relevant data for making informed decisions.

In India, data quality and reliability has long been the center of debate. Acknowledging poor data quality, several steps are now being taken at state and central levels to overcome the data quality issues. Policy framing and implementation is hugely dependent on the data. Thus, it becomes extremely important that good quality data is produced to have informed decisions on policy issues. Quality data includes factors such as accuracy, consistency, and reliability which is often lacking in the National/ State level data. One way to address this problem could be data triangulation and validation at different levels of data production and compilation through robust statistical techniques.

Data validation in Official System is perceived in a very conservative way. Validation is usually taken as reporting at different levels, identifying the inconsistencies, and taking remedial measures. For instance, for immunization coverage in children, it is usually taken as validation of reporting at different levels *viz.*, session site/ village, PHC/UPHC, CHC and district. Validation of routine immunization coverage also requires identifying errors which take place in reporting at these levels during the roll up process.

The idea of present write-up is to overhaul the concept of validation and to increase its scope in a multi-pronged way to increase its visibility and utility to the policy makers. This would help the official reporting system and would virtually revolutionize the whole approach to validation.

Corresponding Author: Arun K. Nigam
Email: dr_aknigam@yahoo.com

2. The revised validation

Good quality data is a pre-requisite to draw sound and meaningful inference from any research study involving data analysis. Poor quality data may emanate from (i) use of poor methodology, (ii) lack of sound data scrutiny and (iii) poor reporting of collected/analysed data at different stages of reporting in-built in the system. Data validation focuses on identification of the causes of poor-quality data and taking/suggesting remedial measures thereof.

2.1. Methodology

Choice of proper methodology and adequate sample size is crucial for any research study. Use of poor methodology mostly includes choosing a poor study design and inadequate sample size. Both these are widespread in not only in Indian context but are also prevalent worldwide. The book by Nigam (2016) deals these issues in detail. The discussion which follows derives heavily from this book.

Among matters of concern are choice of proper survey design including adequate sample size, clarity and coverage of questionnaires, data cleaning/handling and choice of analytical techniques for obtaining valid and efficient estimates as departure from these result in wasting precious funds employed for research programs. It also often ends up with invalid and misleading estimates, which may have strong policy implications.

2.2. Questionnaire

Besides a proper and efficient sampling design required for obtaining efficient and valid estimates, the type and coverage of questionnaires is a crucial deciding factor in obtaining quality data. Any ill-conceived questionnaire leads to substantive non-response, incorrect and evasive responses. In many surveys, questionnaires are unduly lengthy having questions not relevant to the study. On the other hand, sometimes, these are too short to provide a satisfactory coverage. A lengthy questionnaire escalates the cost of the survey and makes management and supervision work cumbersome and time consuming. It also creates problems in editing and cleaning of data and in a decrease in efficiency. A questionnaire with insufficient coverage is likely to be less efficient because of the failure to collect some vital information.

To refine the questionnaire, it is necessary to train interviewers, data editors/cleaners, and through test data analysis. Adequate time should be allotted for field practice and the training should be evaluated. There should be effective and quality monitoring during the field work and this allows for making amends for the ambiguity and inconsistencies. Proper and effective training and pre-testing allow both the project handlers and the interviewers gain insight into the spirits underlying different questions. At data entry level also, there should be data validation employing range check, valid value check as well as internal consistency checks. The follow-up checks and corrective measures improve not only the quality of data gathered but also making the resulting estimates much more relevant and consistent. This aspect, however, is usually taken rather casually in many surveys conducted in our country.

2.3. Sample size and related issues

A close look into the research studies reveals that sample size is often arbitrarily decided, without considering the extent and nature of the variability of the character being studied and even when adequate sample size is taken, there is an attempt to present analysis by sub-groups in terms of related socio-economic, demographic, housing, or household characteristics. This practice leads to decomposition of sample size according to these sub-groups. While a smaller sample size leads to invalid estimates with unduly large standard errors, a larger sample involves avoidable wasteful expenditure. In view of this, it is worthwhile to highlight some of the observations on these issues by Nigam (2006) and Nigam and Singh (2011).

2.4. Poor reporting of collected / analyzed data

In most of the large-scale surveys reporting of indicators is usually done by sub-groups like caste, religion, gender, age group, grades of nutritional status, grades of anemia *etc.* In many situations, sample size for some of these sub-groups is grossly inadequate. Examples of this can be found in the reporting of National Family Health Survey (NFHS), NNMB, Reproductive Child Health (RCH), and District Level Household Surveys (DLHS) and others. The sample size is usually ascertained for all the groups keeping in mind the precision, complexity of the design and expected non-response. Any attempt to the reporting by sub-groups makes such estimates highly imprecise. In view of this, it may be better to go for interval estimates (confidence interval) instead of point estimates. The best alternative, however, is to develop small area estimates for the sub-groups (Chapter 13 in Nigam 2016). For examples of these types of dis-aggregated reporting one may refer to Nigam (2006) and Nigam and Singh (2011). For example, in NFHS-2, nutritional status was reported only for 77 children in Hill Region, for 57 children of ST and for 65 children of Self-employed parents. The reporting has further categorization according to grades of nutritional status. The prevalence of undernutrition ranged from 40-60 percent for below-2sd and 16-30 percent for below-3sd in these groups. Any anaemia among children has been reported for 72 children in Hills, 73 in Bundelkhand and for 33 children of ST, with further division according to grades of anaemia (severe, mild *etc.*). The reported prevalence of any anaemia ranged 73-80 percent and 5-13 percent for severe anaemia. One can easily notice that sample sizes were not adequate for any of these sub-group estimates

Poor reporting of data in different stages are also widely prevalent and can be controlled through proper monitoring. Ways for controlling errors, bridging data gaps, data reduction and improving the quality of data are being discussed now. These can be applied at different stages, *viz.*, at handling of data, sample selection and estimation. Every survey, without exception, encounters the problem of missing data or data with inconsistencies. The main reasons of missing data (i) non-collection of the responses of a sample element, (ii) deletion of some responses as they fail to satisfy certain edit checks. Inconsistencies in data may be attributed to various reasons, such as errors in tabulation, data entry or even in copying from a secondary source. In all such events, it is a norm to treat it as missing data and handle it accordingly. Whereas total non-response, *i.e.*, when all of the responses on a unit are not available, can be handled by some form of weighting adjustment techniques, item non-responses are taken care of by imputation. In the sequel, we add another dimension to data validation.

Framework for data monitoring and quality assessment

Several data quality frameworks have been developed in the health sciences, most of them related to administrative data, registries and electronic health records (Mariño *et al.*, 2022, Schmidt *et al.*,2021). Although they may differ in functionality, their fundamental data quality parameters are similar. The four broad dimensions of such frameworks are - Data Integrity, Completeness, Consistency, and Accuracy (Mariño *et al.*, 2022, Schmidt *et al.*, 2021).

Data Integrity

Data integrity is the degree to which the data conforms to structural and technical requirements. This dimension includes the domains of correct structural representation of data, correct matching of records and elements across multiple datasets, and use of appropriate data value formats.

Completeness

This dimension evaluates the degree to which the required data values are present. It is evaluated with respect to two domains - crude missingness and qualified missingness. Crude missingness is the metrics of missing data values that ignore the underlying reasons for missing data. In qualified missingness assessment, metrics are developed for underlying reasons for missing data such as non-response rate, refusal rate, drop-out rate, *etc.*

Consistency

This data quality dimension encompasses the aspects of range and value violations (compliance with admissible data values or value ranges), and contradictions (presence of improbable combinations).

Accuracy

It is assessed in terms of three domains- unexpected distributions, unexpected associations, and disagreement of repeated measurements. Unexpected distribution entails presence of outliers, unexpected location, shape, scale, among other distributional discrepancies. Unexpected association may include unexpected direction or strength of association between variables. Disagreement of repeated measurements is defined in terms of intra-class reliability, inter-class reliability, or disagreement with gold standard.

Various software packages, especially on R platform, have been developed in recent years to implement data quality monitoring on the four dimensions discussed above (Mariño *et al.*, 2022). These packages offer user-friendly platform to examine data properties in an automated and efficient fashion. However, the performance of such packages depends largely on the quality of the metadata file, which is required for creating the metrics for data quality assessments. So, developing standard framework for metadata is another important aspect for data quality assessment.

2.5. Some useful statistical techniques to address data quality

As stated earlier, use of poor methodology mostly includes choosing a poor study design and inadequate sample size. We now describe some useful statistical techniques

which may go a long way in strengthening the methodology.

Imputation

Almost every large survey suffers from missing values which may also include non-response and/or outliers. Imputation technique consists of handling non-responses by replacing each missing value with a real value. Several imputation procedures are now available for assigning values for missing responses and these are deductive imputation, overall and class-mean imputations, random imputation, hot-deck imputation, and imputation based upon regression.

Principal components technique

Large scale surveys mostly have large data sets requiring reduction. Principal components technique can be used to reduce data. This technique uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables called principal components. The number of principal components is less than or equal to the number of original variables.

Randomized response technique

In many surveys, the intention is to seek information on sensitive characteristics to which response is either false or evasive. It is, therefore, useful to employ the randomized response technique in such cases. The technique ensures confidentiality to the respondent and has become popular in recent years. Sensitive characteristics include unsafe sexual behaviour, child abuse, drugs, *etc.* Detailed applications of this technique are discussed in Nigam (2016).

Small area estimation

One of the areas of data gaps is related to micro-level planning, which requires estimates of different activities for 'smaller' areas having inadequate sample size. This can be achieved by using small area estimation technique which is discussed in detail in Nigam (2016).

Some other useful techniques

Two other techniques need special, though brief, mention. Re-sampling inference is a technique which aims at finding the standard error of variance estimates of non-linear statistics, such as ratios, regression coefficient, index numbers, *etc.* Some other applications are the standard errors of statistics such as median (height or weight), inflation rate, wholesale price index number and the like.

Another useful technique is Snowball Sampling. This can be used in situations where large sample size is required. One such example is estimation of Maternal Mortality Ratio *etc.*

Acknowledgements

Part of this work was done exclusively by the author and included in the Draft Report- Development of Data Validation Protocol Manual, (2022) by the Institute of Applied Statistics and Development Studies (IASDS), Lucknow. An initial draft was discussed with Prof. Arvind Pandey who made several useful suggestions. The reviewer also provided additional material adding another dimension to the write-up.

References

Draft Report (2022). *Development of Data Validation Protocol Manual*. Institute of Applied Statistics and Development Studies, Lucknow.

Mariño, J., Kasbohm, E., Struckmann, S., Kapsner, L. A., and Schmidt, C. O. (2022). R packages for data quality assessments and data monitoring: a software scoping review with recommendations for future developments. *Applied Sciences*, **12**, 4238.

Nigam, A. K. (2006). *Strengthening of NNMB Surveys*. In; Arvind Pandey (Ed) Biostatistical Aspects of Health and Population, Indian Society of Medical Statistics,155-60.

Nigam, A. K. (2016). *Statistical Aspects of Community Health and Nutrition*. Woodhead Publishing India in Food Science.

Nigam, A. K. and Singh, Padam (2011). *Research Methods in Public Health Nutrition: Common Critical Factors: Chapter 11*. In Shiela C. Vir (Ed) Public Health Nutrition in Developing Countries, WPI Publication.

Schmidt, C. O., Struckmann, S., Enzenbach, C., Reineke, A., *et al*. (2021). Facilitating harmonized data quality assessments. A data quality framework for observational health research data collections with software implementations in R. *BMC Medical Research Methodology*, **21**, 1-15.

Publisher
Society of Statistics, Computer and Applications
Registered Office: I-1703, Chittaranjan Park, New Delhi- 110019, INDIA
Mailing Address: B-133, Ground Floor, Chittaranjan Park, New Delhi-110019, INDIA
Tele: 011 - 40517662
https://ssca.org.in/
statapp1999@gmail.com
2024

Printed by : Galaxy Studio & Graphics
Mob: +91 9818 35 2203, +91 9582 94 1203