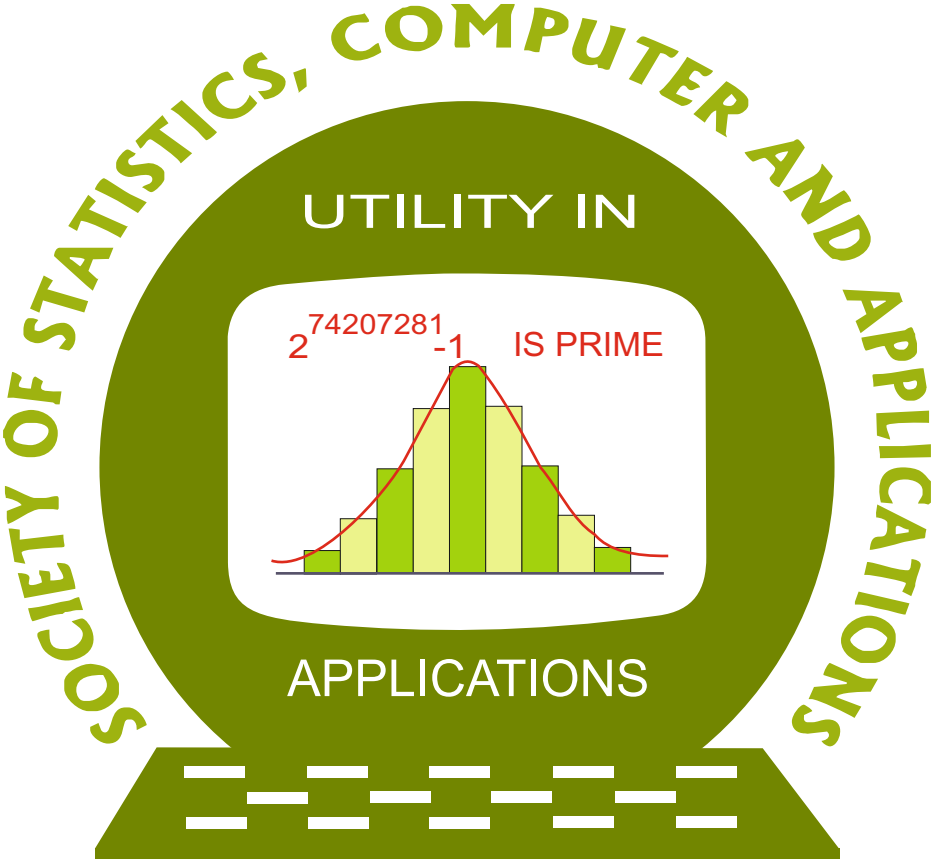
ISSN 2454-7395(online)

STATISTICS AND APPLICATIONS



FOUNDED 1998

Journal of the Society of Statistics, Computer and Applications

https://ssca.org.in/journal.html
Volume 22, No. 1, 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.

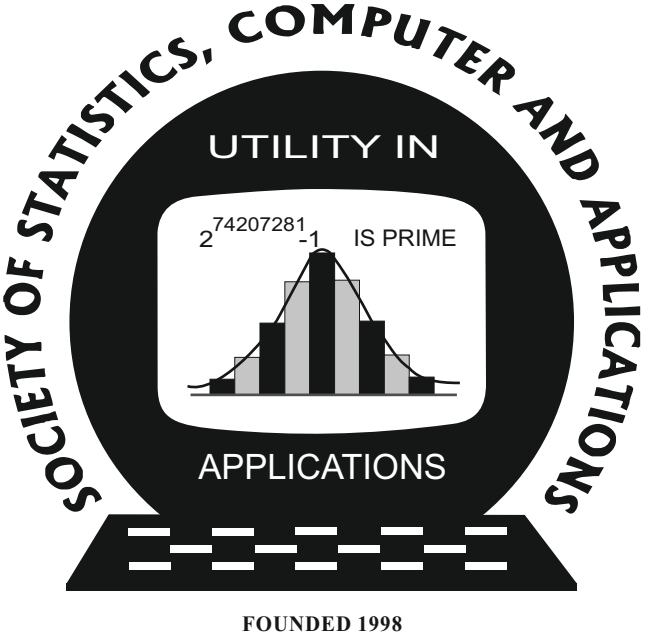
Ex-Officio Members (By Designation)

Director General, Central Statistics Office, Government of India, New Delhi
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. 1, 2024 (New Series)

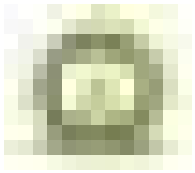
Statistics and Applications

Volume 22, No. 1, 2024 (New Series)

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

CONTENTS

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 22, No. 1, 2024 (New Series), pp 1–20
http://www.ssca.org.in/journal



Controllability and Observability of Fuzzy Matrix Lyapunov Discrete Dynamical System

L. N. Charyulu Rompicharla[1*], Venkata Sundaranand Putcha[2], and G. V. S. R. Deekshitulu[3]

[1] *Department of Mathematics, V.R.Siddhartha Engineering College, Kanuru, Vijayawada, A.P., India.*

[*] *Research Scholar, Jawaharlal Nehru Technological University Kakinada, A.P., India.*

[2] *Department of Mathematics, Rayalaseema University, Kurnool, A.P., India.*

[3] *Department of Mathematics, JNTU College of Engineering, Kakinada, A.P., India.*

Received: 15 December 2021; Revised: 19 July 2022; Accepted: 26 July 2022

Abstract

This paper deals with the controllability and observability of the fuzzy matrix Lyapunov discrete dynamical system. The considered fuzzy system is vectorised by using Kronecker product. The resulting vector system is converted to matrix Lyapunov difference inclusion. For the considered fuzzy system, a symmetric controllability matrix is constructed and derived fuzzy control. A sufficient condition for complete controllability of the fuzzy matrix Lyapunov discrete dynamical system is established by fuzzy rule based approach. Center average defuzzifier approach is used to establish the sufficient conditions for the complete observability of the fuzzy matrix Lyapunov discrete dynamical system. A numerical example is presented to illustrate the theories established, results proved and formulae derived.

Key words: Lyapunov systems; Fuzzy discrete dynamical systems; Fuzzy rule; Controllability; Observability; Defuzzifier.

AMS Subject Classifications: 93B05, 93C55, 93C42, 93B07

1. Introduction

Real world systems represented by mathematical models require the knowledge of exact parameter model values. Many mathematical models do exhibit some degree of uncertainty because of the limitations in obtaining the exact values of the model parameters. This will naturally inspire scientists and engineers to construct models with uncertain parameters and uncertain initial conditions. This uncertainty cannot be ignored or neglected because of its influence on the model predictions. One of the important ways of incorporating the uncertainty or vagueness is by fuzzy dynamical modeling. The fundamental prerequisites for the design process are the controllability and the observability. The controllability conditions guarantee for the existence of control which will steer the state from the initial point to the

Corresponding Author: L.N.Charyulu Rompicharla
Email: narayanarompicharla@gmail.com

desired final point. So these two metrics are mandatory to test the possibility and feasibility of achieving the design requirements for the system of consideration. A simple criteria for the controllability and observability for the fuzzy dynamical systems similar to that of deterministic dynamical systems cannot be found because of the vagueness and uncertainty involved in the systems as well as initial condition. So the controllability and observability in fuzzy sense are to be explored. In the fuzzy case, the controllability cannot be characterized by finding a suitable control which can transfer the system from the initial state to any desired final state in a finite time interval since finite number of options emerge because coefficients, variables in the system and initial conditions are fuzzy, not deterministic. Cai and Tang (2000), Ding and Kandel (2000a), Farinwata and Vachtsevanos (1993) have studied the controllability of fuzzy systems.

Mastiani and Effati (2018) have investigated the controllability and the observability property of two systems that one of them has fuzzy variables and the other one has fuzzy coefficients and fuzzy variables(fully fuzzy system) by normalizing the fuzzy matrices. Gabr (2015) studied impact of propagation of fuzziness in the coefficients of dynamical systems in modeling, analysis, and design of automatic control systems. Difference equations describe the observed evolution phenomena in a better manner when compared to that of differential equations. Lyapunov matrix systems appear in determining the stability of the autonomous systems by the second method of Lyapunov without finding the solution of the system and also in the minimization of quadratic cost functionals in optimal control problems. Matrix Lyapunov systems have extensive applications in control theory, digital computers, optimal filters, population dynamics, differential games, power systems, signal processing and boundary value problems. Putcha *et al.* (2012) established variation of parameters formula for the matrix fuzzy dynamical systems and studied the controllability and observability of the fuzzy discrete dynamical systems by the defuzzifier approach. Anand and Murty (2005), Murty *et al.* (1997) studied the controllability and observability of the continuous and discrete dynamical systems. It is very important to study the controllability and observability of the mathematical models represented by fuzzy difference equations governing the ambiguity in dynamics which is not probabilistic. In general the problem of steering an initial state of a system to a desired final state in R^n become a problem of steering a fuzzy state to another fuzzy-state in E^s. Many of the physical applications may not have the exact information about their deterministic dynamics which is prerequisite to construct a dynamical system. The importance of control theory in mathematics and its occurrence in several problems such as mechanics, electromagnetic theory, thermodynamics, and artificial satellites are well known. In general, fuzzy systems are classified into 3 categories, (i) Pure fuzzy systems (ii) T-S fuzzy systems, and (iii) Fuzzy logic systems, using fuzzifiers and defuzzifiers. In this paper, we use fuzzy matrix Lyapunov discrete dynamical system to describe fuzzy logic system and establish sufficient conditions for controllability and observability of first order fuzzy matrix Lyapunov discrete dynamical system S_1 modeled by

$$\Delta T(n) = A(n)T(n) + T(n)A^T(n) + A(n)T(n)A^T(n) + F(n)U(n), T(n_0) = T_0, n > 0 \quad (1)$$

$$Y(n) = C(n)T(n) + D(n)U(n) \quad (2)$$

where $U(n)$ is an $m \times s$ fuzzy input matrix called fuzzy control and $Y(n)$ is an $r \times s$ fuzzy output matrix. Here $T(n), A(n), F(n), C(n)$ and $D(n)$ are matrices of order $s \times s, s \times s, s \times m, r \times s$ and $r \times m$ whose elements are continuous functions of n on $J = [0, N] \subset R(N > 0)$. Barnett (1975) studied the problem of controllability and observability for a system of ordinary

differential equations. Anand and Murty (2005) established necessary and sufficient conditions for the controllability and observability of continuous matrix Lyapunov systems. Using fuzzy control, a complex system can be decomposed into several subsystems according to the expertise of human ability to understand the system and using the human control strategy represented by a simple control law. The popular fuzzy controllers in the literature are Mamdani fuzzy controllers and Takagi-Sugeno(TS) fuzzy controllers . The main difference between them is that the Mamdani fuzzy controllers use fuzzy sets whereas the (TS) fuzzy controllers use linear functions, to represent the fuzzy rules. The accessibility and the controllability properties of TS fuzzy logic control systems are studied by Biglarbegian *et al.* (2012) by using differential geometric and Lie-algebraic techniques. Ding and Kandel (2000b,a), established that the observability is a characteristic of the system to estimate the range of the fuzzy initial state with to the knowledge of the fuzzy input and the fuzzy output in a finite time interval for the fuzzy dynamical system with the fuzzy initial state. In the works of Takagi and Sugeno (1985), Johansen *et al.* (2000), Sugeno (1999) a crisp analytical function is used instead of a membership function in a fuzzy model. In recent years many authors Alwadie *et al.* (2003); Ying (1999, 2006); Ding *et al.* (2003, 1999) are studying TS fuzzy controllers, because of their ability to model real world problems. Anand and Murty (2005); Murty *et al.* (1995) established conditions for the controllability and observability of Liapunov type matrix difference system. Murty *et al.* (2008) presented criteria for the existence and uniqueness of solution to Kronecker product initial value problem associated with general first order matrix difference system. Murty *et al.* (2009) studied qualitative properties of general first order matrix difference systems. We obtain a unique solution of the system (1), when $U(n)$ is a crisp continuous matrix. We use fuzzy matrix discrete system to describe fuzzy logic system and extend some of the results of Ding and Kandel (2000a,b) developed for continuous case to that of discrete case by vectorizing the fuzzy matrix discrete system. We obtain sufficient conditions for controllability and observability of the system (1) satisfying the initial condition. The fundamental results established in Murty *et al.* (1995), Rompicharla *et al.* (2019, 2020) have in-fact motivated us to develop our results on fuzzy matrix Lyapunov discrete dynamical systems.

The paper is organized as follows. Section 2 presents basic definitions and results required to understand the paper. Section 3 is concerned with the formation of fuzzy matrix Lyapunov discrete dynamical systems. Sufficient conditions for the controllability and observability of fuzzy matrix Lyapunov discrete dynamical systems are presented in Section 4 and Section 5 respectively. Section 6 presents a numerical example.

2. Preliminaries

In this section basic definitions of Kronecker product, properties of vectorization, α-level set, fundamental matrix solutions of homogeneous and non-homogeneous matrix Lyapunov discrete dynamical systems and corresponding initial value problems are presented.
Let $\left(N_{n_0}^{+}\right) = \{n_0, n_0 \pm 1, ..., n_0 \pm k, ...\}$ where n_0 is an integer number.
Let $P_k\left(N_{n_0}^{+}\right)^s$ denotes the family of all nonempty compact convex subsets of $\left(N_{n_0}^{+}\right)^{s\times s}$.
Define the addition and scalar multiplication in $P_k\left(N_{n_0}^{+}\right)^s$ as usual. Rådström (1952) states that $P_k\left(N_{n_0}^{+}\right)^s$ is a commutative semi group under addition, which satisfies the cancellation law. Moreover, if $\alpha, \beta \in (N_{n_0}^{+})$ and $A, B \in P_k\left(N_{n_0}^{+}\right)^s$, then $\alpha(A+B) = \alpha A + \alpha B, \alpha(\beta A) =$

$(\alpha\beta)A, 1.A = A$, and if $\alpha, \beta \geq 0$, then $(\alpha + \beta)A = \alpha A + \beta A$. The distance between A and B is defined by Hausdroff metric $d(A, B) = inf\{\epsilon : A \subset N(B, \epsilon), B \subset N(A, \epsilon)\}$, where $N(A, \epsilon) = \{x \in \left(N_{n_0}^+\right)^s : \|x - y\| < \epsilon, \text{ for some } y \in A\}$.

Definition 1: A set valued function $F : J \to P_k\left(N_{n_0}^+\right)^s$, where $J = [0, N] \subset R(N > 0)$ is said to be measurable if it satisfies any one of the following equivalent conditions:

(1) for all $u \in (N_{n_0}^+)^s, n \to d_{F(n)}(u) = inf_{v \in F(n)} \parallel u - v \parallel$ is measurable,

(2) Gr $F = \{(t, u) \in J \times (N_{n_0}^+)^s : u \in F(n)\} \in \sum \times \beta(N_{n_0}^+)^s$, where $\sum, \beta(N_{n_0}^+)^s$ are Borel σ - field of J and $(N_{n_0}^+)^s$, respectively (Graph measurability),

(3) there exists a sequence $\{f_n(.)\}_{n \geq 1}$ of measurable functions such that $F(n) = \overline{\{f_n(.)\}_{n \geq 1}}$, for all $n \in J$ (Castaing's representation).

We denote by S_F^1 the set of all selections of $F(.)$ that belong to the Lebesgue Bochner space $L^1_{(N_{n_0}^+)^s}(J)$, that is, $S_F^1 = \left\{f(.) \in L^1_{(N_{n_0}^+)^s}(J) : f(n) \in F(n) \text{ almost every where (a.e)}\right\}$. We present the Aumann's integral as follows: $\int_J F(t)dt = \left\{\int_J f(t)dt, f(.) \in S_F^1\right\}$. We say that $F : J \to P_k\left(N_{n_0}^+\right)^s$ is integrably bounded if it is measurable and there exists a function $h : J \to (N_{n_0}^+), h \in L^1_{(N_{n_0}^+)^s}(J)$, such that $\parallel u \parallel \leq h(t), u \in F(t)$. We know that if F is a closed valued measurable multifunction, then $\int_J F(t)dt$ is convex in $(N_{n_0}^+)^s$. Furthermore, if F is integrably bounded, then $\int_J F(t)dt$ is compact in $(N_{n_0}^+)^s$.
Let $E^s = \{u : \left(N_{n_0}^+\right)^s \to [0, 1]/u$ satisfies the following $\}$;

(1) u is normal, that is, there exists an $n_0 \in \left(N_{n_0}^+\right)^{s \times s}$ such that $u(n_0) = 1$;

(2) u is fuzzy convex, that is, for $x, y \in \left(N_{n_0}^+\right)^s$ and $0 \leq \lambda \leq 1$, $u(\lambda x + (1 - \lambda)y) \geq min[u(x), u(y)]$;

(3) u is upper semicontinuous;

(4) $[u]^0 = \overline{\{x \in \left(N_{n_0}^+\right)^s / u(x) > 0\}}$ is compact.

For $0 < \alpha \leq 1$, the $0 < \alpha \leq 1$, the α- level set is denoted and defined by $[u]^\alpha = \{x \in (N_{n_0}^+ \; u(x) \geq \alpha\}$. Then, from $(1) - (4)$ above, it follows that $[u]^\alpha \in P_k(N_{n_0}^+)$ for all $0 \leq \alpha \leq 1$. Define $D : E^n \times E^n \to [0, \infty]$ by $D(u, v) = sup\{d([u]^\alpha, [v]^\alpha)/\alpha \in [0, 1]\}$, where d is the Hausdroff metric defined in $P_k\left(N_{n_0}^+\right)^s$. It is easy to show that D is a metric in E^s and using results of Rådström (1952), we see that (E^s, D) is a complete metric space, but not locally compact. Moreover, the distance D verifies that $D(u + w, v + w) = D(u, v), u, v \in E^s, D(\lambda u, \lambda v) = |\lambda| D(u, v), u, v \in E^s, \lambda \in R$, $D(u+w, v+z) \leq D(u, v)+D(w, z), u, v, w, z \in E^s$. We note that (E^s, D) is not a vector space. But it can be embedded isomorphically as a cone in Banach space (Rådström (1952)). Regarding fundamentals of differentiability and integrability of fuzzy functions, we refer to Kaleva (1987) and Lakshmikantham and Mohapatra (2003).

Definition 2: Let $A \in \mathbb{C}^{r\times s}(\mathbb{R}^{r\times s})$ and $B \in \mathbb{C}^{p\times q}(\mathbb{R}^{p\times q})$. Then Kronecker product of A and B is written as $A \otimes B$ and is defined as a partitioned matrix $\begin{bmatrix} a_{11}B & a_{12}B & ...a_{1s}B \\ a_{21}B & a_{22}B & ...a_{2s}B \\ ... & ... & ... \\ a_{r1}B & a_{r2}B & a_{rs}B \end{bmatrix}$ which is an $rp \times sq$ matrix and is in $\mathbb{C}^{rp\times sq}(\mathbb{R}^{rp\times sq})$.
The Kronecker product has the following properties.

(1) $(A \otimes B)^* = A^* \otimes B^*$

(2) $(A \otimes B)^{-1} = A^{-1} \otimes B^{-1}$

(3) $(A \otimes B)(C \otimes D) = (AC \otimes BD)$.
This rule holds, provided the dimensions of the matrices are such that expressions are defined.

(4) $|| A \otimes B || = \|A\| \, \|B\|$, where $(\|A\| = max_{i,j} |a_{ij}|)$

(5) $(A + B) \otimes C = (A \otimes C) + (B \otimes C)$.
Vectorization of matrix A is denoted by VEC $(A) = \widehat{A}$ and defined as follows.

Definition 3: Let $A = [a_{ij}] \in \mathbb{C}^{r\times s}(\mathbb{R}^{r\times s})$, we denote VEC$(A) = \widehat{A} = [A_{.1}, A_{.2}, ..., A_{.s}]^T$ where $A_{.j} = [a_{1j}, a_{2j}, ..., a_{rj}]^T, (1 \le j \le s)$, where X is a matrix of size $s \times s$.
Vectorization has the following properties.
1. VEC $(AXB) = (B^* \otimes A)$ VEC X.
2. If A and B are square matrices of order s, then
VEC $(AX) = (I_s \otimes A)$ VEC X;
VEC $(XB) = (B^* \otimes I_s)$ VEC X.

Theorem 1: [Ralescu (1979); Murty and Kumar (2008)]
If $u \in E^s$ then

(1) $[u]^\alpha \in P_k \left(N_{n_0}^+\right)^{s\times s}$ for all $0 \le \alpha \le 1$;

(2) $[u]^{\alpha_2} \subset [u]^{\alpha_1}$ for all $0 \le \alpha_1 \le \alpha_2 \le 1$;

(3) α_k is non decreasing sequence converging to $\alpha > 0$, then $[u]^\alpha = \cap_{k\ge1}[u]^{\alpha_k}$. Conversely, if $\{A^\alpha : 0 \le \alpha \le 1\}$ is a family of subsets of $\left(N_{n_0}^+\right)^{s\times s}$ satisfying $(1) - (3)$, then there exists a $u \in E^s$ such that $[u]^\alpha = A^\alpha$ for $0 < \alpha \le 1$ and $[u]^0 = \overline{U_{0\le\alpha\le1}A^\alpha} \subset A^0$.

Theorem 2: [Sundaranand Putcha (2014); Rompicharla *et al.* (2019)]
Let $\phi(n, 0)$ and $\phi^*(n, 0)$ be the fundamental matrix solutions of
$\Delta T(n) = A(n)T(n)$ and $\Delta T(n) = T(n)A^T(n)$.
Then the matrix $\phi(n, n_0)C\phi^*(n, n_0)$ (where C is a constant square matrix of size s) be the fundamental matrix for the system

$$\Delta T(n) = A(n)T(n) + T(n)A^T(n) + A(n)T(n)A^T(n), T(n_0) = I_s. \tag{3}$$

The matrix $(\phi^*(n,n_0)\otimes\phi(n,n_0))$ is the fundamental matrix of

$$\Delta\widehat{T}(n)=((A(n)\otimes A(n))+(A(n)\otimes I_s)+(I_s\otimes A(n)))\widehat{T}(n),\widehat{T}(n_0)=\widehat{T_0} \tag{4}$$

and the solution of (4) is $\widehat{T}(n)=(\phi^*(n,0)\otimes\phi(n,0)\widehat{T_0}$.

Theorem 3: [Putcha and Prathyusha (2019)]
Let $\phi(n,n_0)C\phi^*(n,n_0)$ be the fundamental matrix for the system (3). Then the unique solution of the initial value problem

$$\Delta\widehat{T}(n)=[(A(n)\otimes A(n))+(A(n)\otimes I_s)+(I_s\otimes A(n))]\widehat{T}(n)+(I_s\otimes F(n))\widehat{U}(n),\widehat{T}(n_0)=\widehat{T}_0 \tag{5}$$

is given by

$$\widehat{T}(n)=(\phi^*(n,0)\otimes\phi(n,0))\widehat{T}_0+\sum_{j=0}^{n-1}(\phi^*(n,j+1)\otimes\phi(n,j+1))(I_s\otimes F(j))\widehat{U}(j).$$

3. Inclusion approach to Fuzzy Matrix Lyapunov discrete dynamical system

This section presents a method of the conversion of fuzzy matrix Lyapunov discrete dynamical system to a matrix Lyapunov difference inclusion. Thus the solution of a fuzzy matrix Lyapunov discrete dynamical system can be expressed as the solution set of the corresponding matrix Lyapunov difference inclusion.
Let $u_i(n)\in E^1, n\in J, i=1,2,..,s^2$, and define

$$\begin{aligned}\widehat{U}(n)=(u_1(n),u_2(n),...,u_{s^2}(n)) &= u_1(n)\times u_2(n)\times ...\times u_{s^2}(n)\\ &=\{(u_1^\alpha(n),u_2^\alpha(n),...,u_{s^2}^\alpha(n):\alpha\in[0,1]\}\\ &=\{(\widetilde{u}_1(n),\widetilde{u}_2(n),...,\widetilde{u}_{s^2}(n):\widetilde{u}_i(n)\in u_i^\alpha(n),\alpha\in[0,1]\},\end{aligned} \tag{6}$$

where $u_i^\alpha(n)$ is the α-level set of $u_i(n)$. From the above definition of $\widehat{U}(n)$ and Theorem 1, it can be easily seen that $\widehat{U}(n)\in E^{s^2}$. We now show that the following system S_2 defined by

$$\Delta\widehat{T}(n)=((A(n)\otimes A(n))+(A(n)\otimes I_s)+(I_s\otimes A(n)))\widehat{T}(n)+(I_s\otimes F(n))\widehat{U}(n),\widehat{T}(0)=\widehat{T}_0 \tag{7}$$

and

$$\widehat{Y}(n)=((I_s\otimes C(n)))\widehat{T}(n)+(I_s\otimes D(n))\widehat{U}(n) \tag{8}$$

determines a fuzzy system by using the fuzzy control $\widehat{U}(n)$. Assume that $\widehat{U}(n)$ is continuous in E^{s^2}. Then the set $\widehat{U}^\alpha=u_1(n)\times u_2(n)\times,...,\times u_{s^2}(n)$ is a convex and compact set in $\left(N_{n_0}^+\right)^{s^2}$. For any positive number N, consider the following inclusions

$$\Delta\widehat{T}(n)\in[(A(n)\otimes A(n))+(A(n)\otimes I_s)+(I_s\otimes A(n))]\widehat{T}(n)+(I_s\otimes F(n))\widehat{U}^\alpha(n),n\in[0,N], \tag{9}$$

$$\widehat{T}(n_0)\in\widehat{T}_0. \tag{10}$$

Let $\widehat{T}^\alpha$ be the solution of (9) satisfying (10)

Lemma 1: $[\widehat{T}(n)]^{\alpha} \in P_k(N_{n_0}^{+})^{s^2}$, for every $0 \leq \alpha \leq 1, n \in [0, N]$.

Proof: We can observe that $\widehat{T}^{\alpha}$ is non empty since $\widehat{U}^{\alpha}(n)$ has measurable selection.
By choosing

$$K = max_{n\in[0,N]} \parallel \phi(n, n_0) \parallel, L = max_{n\in[0,N]} \parallel \phi^*(n, n_0)) \parallel,$$

$$M = max\{\parallel u(n) \parallel : u(n) \in \widehat{U}^{\alpha}(n), n \in [0, N]\},$$
$$T = max_{n\in[0,N]} \parallel F(n) \parallel, J = max_{n\in[0,N]} \parallel I_s \parallel = 1.$$

If for any $\widehat{T} \in \widehat{T}^{\alpha}$, there exists a control $u(n) \in \widehat{U}^{\alpha}(n)$ such that

$$\widehat{T}(n) = (\phi^*(n, n_0) \otimes \phi(n, n_0))\widehat{T}_0 + \sum_{j=n_0}^{n-1} (\phi^*(n, j+1) \otimes \phi(n, j+1))(I_s \otimes F(j))U(j). \quad (11)$$

By taking norm on both sides of the equation (11), we get

$$\parallel \widehat{T}(n) \parallel \leq KL \parallel \widehat{T}_0 \parallel + KLTMN.$$

Hence $\widehat{T}^{\alpha}$ is bounded.
For any $n_1, n_2 \in [0, N]$, consider,

$$\widehat{T}(n_1) - \widehat{T}(n_2) = (\phi^*(n, n_1) \otimes \phi(n, n_1))\widehat{T}_0 + \sum_{j=n_0}^{n_1-1} (\phi^*(n_1, j+1) \otimes \phi(n_1, j+1))(I_s \otimes F(j))u(j) -$$

$$(\phi^*(n, n_2) \otimes \phi(n, n_2))\widehat{T}_0 - \sum_{j=n_0}^{n_2-1} (\phi^*(n_2, j+1) \otimes \phi(n_2, j+1))(I_s \otimes F(j))u(j)$$

Therefore

$$\parallel \widehat{T}(n_1) - \widehat{T}(n_2) \parallel \leq \parallel (\phi^*(n, n_1) \otimes \phi(n, n_1)) - (\phi^*(n, n_2) \otimes \phi(n, n_2)) \parallel \parallel \widehat{T}_0 \parallel +$$

$$\sum_{j=n_2-1}^{n_1-1} \parallel (\phi^*(n_1, j+1) \otimes \phi(n_1, j+1))(I_s \otimes F(j))u(j) \parallel +$$

$$\sum_{j=n_0}^{n_2-1} \parallel [(\phi^*(n_1, j+1) \otimes \phi(n_1, j+1)) - (\phi^*(n_2, j+1) \otimes \phi(n_2, j+1))](I_s \otimes F(j))u(j) \parallel$$

$$\leq \parallel (\phi^*(n, n_1) \otimes \phi(n, n_1)) - (\phi^*(n, n_2) \otimes \phi(n, n_2)) \parallel \parallel \widehat{T}_0 \parallel + KLTM \mid n_1 - n_2 \mid +$$

$$MT \sum_{j=n_0}^{N-1} \parallel (\phi^*(n_1, j+1) \otimes \phi(n_1, j+1) - (\phi^*(n_2, j+1) \otimes \phi(n_2, j+1)) \parallel .$$

Since $(\phi^*(n, n_0))$ and $(\phi(n, n_0))$ are both uniformly continuous on $[0, N]$,
$\widehat{T}$ is equicontinous. Thus, $\widehat{T}^{\alpha}$ is relatively compact.
Let $\widehat{T}_k \in \widehat{T}^{\alpha}$ and $\widehat{T}_k \to \widehat{T}$. For each $\widehat{T}_k$, there is a $u_k \in \widehat{U}^{\alpha}(n)$ such that

$$\widehat{T}_k(n) = ((\phi^*(n, n_0) \otimes \phi(n, n_0))\widehat{T}_0 + \sum_{j=n_0}^{n-1} (\phi^*(n, j+1) \otimes \phi(n, j+1))(I_s \otimes F(j))U_k(j). \quad (12)$$

Since $u_k \in \widehat{U}^{\alpha}(n)$ is closed, then there is a subsequence $< u_{k_i} >$ of $< u_k >$ converging weakly to $u \in \widehat{U}^{\alpha}(n)$. From Mazur's theorem Conway and Voglmeir (2016), there exists a sequence of numbers $\lambda_i > 0$, $\Sigma\lambda_i = 1$ such that $\Sigma\lambda_i u_{k_i}$ converges strongly to u. Thus from (12) we have

$$\Sigma\lambda_i \widehat{T}_{K_i}(n) = \sum \lambda_i((\phi^*(n,n_0)\otimes\phi(n,n_0))\widehat{T}_0 + \sum_{j=n_0}^{n-1}(\phi^*(n,j+1)\otimes\phi(n,j+1))(I_s\otimes F(j))\Sigma\lambda_i u_{k_i}(j). \tag{13}$$

As $i \to \infty$ from equation (13) and Fatuou's lemma, it follows that

$$\widehat{T}(n) = (\phi^*(n,n_0)\otimes\phi(n,n_0))\widehat{T}_0 + \sum(\phi^*(n,j+1)\otimes\phi(n,j+1))(I_s\otimes F(j))u(j).$$

Thus $\widehat{T}(n) \in \widehat{T}^{\alpha}$, and hence $\widehat{T}^{\alpha}$ is closed.
Let $\widehat{T}_1, \widehat{T}_2 \in \widehat{T}^{\alpha}$, then there exists $u_1, u_2 \in \widehat{U}^{\alpha}(n)$ such that

$$\Delta\widehat{T}_1(n) = [(A(n)\otimes A(n)) + (A(n)\otimes I_s) + (I_s\otimes A(n))]\widehat{T}_1(n) + (I_s\otimes F(n))u_1(n),$$

$$\Delta\widehat{T}_2(n) = [(A(n)\otimes A(n)) + (A(n)\otimes I_s) + (I_s\otimes A(n))]\widehat{T}_2(n) + (I_s\otimes F(n))u_2(n).$$

Let $\widehat{T}(n) = \lambda\widehat{T}_1(n) + (1-\lambda)\widehat{T}_2(n), 0 \le \lambda \le 1$ then

$$\begin{aligned}
&\Delta\widehat{T}(n)\\
&\quad= \lambda\Delta\widehat{T}_1(n) + (1-\lambda)\Delta\widehat{T}_2(n)\\
&\quad= \lambda((A(n)\otimes A(n)) + (A(n)\otimes I_s) + (I_s\otimes A(n)))\widehat{T}_1(n)\\
&\qquad + (I_s\otimes F(n))u_1(n) + (1-\lambda)(((A(n)\otimes A(n)) + (A(n)\otimes I_s)\\
&\qquad + (I_s\otimes A(n)))\widehat{T}_2(n) + (I_s\otimes F(n))u_2(n)\\
&\quad= [(I_s\otimes A(n)) + (A(n)\otimes I_s) + (A(n)\otimes A(n))][\lambda\widehat{T}_1(n)\\
&\quad+ (1-\lambda)\widehat{T}_2(n)] + (I_s\otimes F(n))[\lambda u_1(n) + (1-\lambda)u_2(n)]
\end{aligned}$$

Since $\widehat{U}^{\alpha}(n)$ is convex, $\lambda u_1(n) + (1-\lambda)u_2(n) \in \widehat{U}^{\alpha}(n)$, we have

$$\Delta\widehat{T}(n) \in (I_s\otimes A(n) + A(n)\otimes I_s + A(n)\otimes A(n))\widehat{T}(n) + (I_s\otimes F(n))\widehat{U}^{\alpha}(n),$$

i.e.,$\widehat{T} \in \widehat{T}^{\alpha}$. Thus $\widehat{T}^{\alpha}$ is convex. Therefore $\widehat{T}^{\alpha}$ is non empty, compact and convex in $\mathbb{C}[[0,N],(N_{n_0}^+)^{s^2}]$. Thus, from Arzela-Ascoli theorem, it follows that $[\widehat{T}(n)]^{\alpha}$ is convex in $(N_{n_0}^+)^{s^2}$, for every $n \in [0,N]$. Therefore $[\widehat{T}(n)]^{\alpha} \in P_k((N_{n_0}^+)^{s^2})$ for every $0 \le \alpha \le 1, n \in [0,N]$. □

Lemma 2: $[\widehat{T}(n)]^{\alpha_2} \subset [\widehat{T}(n)]^{\alpha_1}$, for all $0 \le \alpha_1 \le \alpha_2 \le 1$.

Proof: Let $0 \le \alpha_1 \le \alpha_2 \le 1$. Since $\widehat{U}^{\alpha_2}(n)$ is contained in $\widehat{U}^{\alpha_1}(n)$, it follows that

$$\widehat{U}^{\alpha_2}(n) = u_1^{\alpha_2}(n)\times u_2^{\alpha_2}(n)\times ...\times u_{s^2}^{\alpha_2}(n) \subset u_1^{\alpha_1}(n)\times u_2^{\alpha_1}(n)\times ...\times u_{s^2}^{\alpha_1}(n) = \widehat{U}^{\alpha_1}(n)$$

and also the following inclusions:

$$\begin{aligned}
\Delta\widehat{T}(n) &\in [(A(n)\otimes A(n)) + (A(n)\otimes I_s) + (I_s\otimes A(n))]\widehat{T}(n) + (I_s\otimes F(n))\widehat{U}^{\alpha_2}(n)\\
&\subset [(A(n)\otimes A(n)) + (A(n)\otimes I_s) + (I_s\otimes A(n))]\widehat{T}(n) + (I_s\otimes F(n))\widehat{U}^{\alpha_1}(n)
\end{aligned} \tag{14}$$

Consider the following inclusions:

$$\Delta\widehat{T}(n) \in [(A(n)\otimes A(n))+(A(n)\otimes I_s)+(I_s\otimes A(n))]\widehat{T}(n)+(I_s\otimes F(n))\widehat{U}^{\alpha_2}(n), n \in [0,N] \quad (15)$$

$$\Delta\widehat{T}(n) \in [(A(n)\otimes A(n))+(A(n)\otimes I_s)+(I_s\otimes A(n))]\widehat{T}(n)+(I_s\otimes F(n))\widehat{U}^{\alpha_1}(n), n \in [0,N] \quad (16)$$

Let $\widehat{T}^{\alpha_2}$ and $\widehat{T}^{\alpha_1}$ be the solution sets of (15) and (16) respectively. Clearly the solution of (15) satisfies the following inclusion:

$$\widehat{T}(n) \in (\phi^*(n,n_0)\otimes\phi(n,n_0))\widehat{T}_0 + \sum_{j=n_0}^{n-1}(\phi^*(n,j+1)\otimes\phi(n,j+1))(I_s\otimes F(j))S^1_{\widehat{U}^{\alpha_2}(j+1)}$$

$$\subset (\phi^*(n,n_0)\otimes\phi(n,n_0))\widehat{T}_0 + \sum_{j=n_0}^{n-1}(\phi^*(n,j+1)\otimes\phi(n,j+1))(I_s\otimes F(j))S^1_{\widehat{U}^{\alpha_1}(j+1)}. \quad (17)$$

Thus $\widehat{T}^{\alpha_2} \subset \widehat{T}^{\alpha_1}$. And hence $\widehat{T}^{\alpha_2}(n) \subset \widehat{T}^{\alpha_1}(n)$ □

Lemma 3: If $<\alpha_k>$ is nondecreasing sequence converging to $\alpha > 0$ then $\widehat{T}^{\alpha}(n) = \cap_{k\geqslant 1}\widehat{T}^{\alpha_k}(n)$.

Proof: Let

$$\widehat{U}^{\alpha_k}(n) = u_1^{\alpha_k} \times u_2^{\alpha_k}\times, ..., \times u_{s^2}^{\alpha_k}, \widehat{U}^{\alpha}(n) = u_1^{\alpha} \times u_2^{\alpha}\times, ..., u_{s^2}^{\alpha}$$

and consider the inclusions

$$\Delta\widehat{T}(n) \in [(A(n)\otimes A(n)) + (A(n)\otimes I_s) + (I_s\otimes A(n))]\widehat{T}(n) + (I_s\otimes F(n))\widehat{U}^{\alpha_k}(n) \quad (18)$$

$$\Delta\widehat{T}(n) \in [(A(n)\otimes A(n)) + (A(n)\otimes I_s) + (I_s\otimes A(n))]\widehat{T}(n) + (I_s\otimes F(n))\widehat{U}^{\alpha}(n) \quad (19)$$

Let $\widehat{T}^{\alpha_k}$ and $\widehat{T}^{\alpha}$ be the solution sets of 18 and 19 respectively. Since $u_i(n)$ is a fuzzy set and from Theorem 1, we have

$$u_i^{\alpha} = \cap_{k\geq 1} u_i^{\alpha_k}, \quad (20)$$

we consider

$$\widehat{U}^{\alpha}(n) = u_1^{\alpha} \times u_2^{\alpha}\times, ..., \times u_{s^2}^{\alpha} = \cap_{k\geq 1}u_1^{\alpha_k} \times \cap_{k\geq 1}u_2^{\alpha_k}\times, ..., \cap_{k\geq 1}u_{s^2}^{\alpha_k} = \cap_{k\geq 1}\widehat{U}^{\alpha_k}(n) \quad (21)$$

and then $S^1_{\widehat{U}^{\alpha}(n)} = S^1_{\cap_{k\geqslant 1}\widehat{U}^{\alpha_k}(n)}$.
Therefore

$$\Delta\widehat{T}(n) \in [(A(n)\otimes A(n)) + (A(n)\otimes I_s) + (I_s\otimes A(n))]\widehat{T}(n) + (I_s\otimes F(n))\widehat{U}^{\alpha}(n)$$

$$= [(A(n)\otimes A(n)) + (A(n)\otimes I_s) + (I_s\otimes A(n))]\widehat{T}(n) + (I_s\otimes F(n))\cap_{k\geq 1}\widehat{U}^{\alpha_k}(n)$$

$$\subset [(A(n)\otimes A(n)) + (A(n)\otimes I_s) + (I_s\otimes A(n))]\widehat{T}(n) + (I_s\otimes F(n))\widehat{U}^{\alpha_k}(n), k = 1,2,3,... \quad (22)$$

Thus we have $\widehat{T}^{\alpha} \subset \widehat{T}^{\alpha_k}, k = 1,2,3...$, which implies that

$$\widehat{T}^{\alpha} \subset \cap_{k\geq 1}\widehat{T}^{\alpha_k}. \quad (23)$$

Let $\widehat{T}$ be the solution set of the inclusion (18) for $k \geq 1$. Then

$$\widehat{T}(n) \in (\phi^*(n,n_0) \otimes \phi(n,n_0))\widehat{T}_0 + \sum_{j=n_0}^{n-1} (\phi^*(n,j+1) \otimes \phi(n,j+1))(I_s \otimes F(j))S^1_{\widehat{U}^{\alpha_k}}(n). \quad (24)$$

It follows that

$$\begin{aligned}\widehat{T}(n) &\in (\phi^*(n,n_0) \otimes \phi(n,n_0))\widehat{T}_0 + \cap_{k\geq 1} \sum_{j=n_0}^{n-1} (\phi^*(n,j+1) \otimes \phi(n,j+1))(I_s \otimes F(j))S^1_{\widehat{U}^{\alpha k}(n)} \\ &\subset (\phi^*(n,n_0) \otimes \phi(n,n_0))\widehat{T}_0 + \sum_{j=n_0}^{n-1} (\phi^*(n,j+1) \otimes \phi(n,j+1))(I_s \otimes F(j))S^I_{\cap k\geq 1 \widehat{U}^{\alpha_k}} \\ &= (\phi^*(n,n_0) \otimes \phi(n,n_0))\widehat{T}_0 + \sum_{j=n_0}^{n-1} (\phi^*(n,j+1) \otimes \phi(n,j+1))(I_s \otimes F(j))S^1_{\widehat{U}^{\alpha_k}}.\end{aligned}$$

This implies that $\widehat{T} \in \widehat{T}^{\alpha}$. Therefore,

$$\cap_{k\geq 1}\widehat{T}^{\alpha_k} \subset \widehat{T}^{\alpha}. \quad (25)$$

From (23) and (25), we have $\widehat{T}^{\alpha} = \cap_{k\geq 1}\widehat{T}^{\alpha_k}$ and hence, $\widehat{T}^{\alpha}(n) = \cap_{k\geq 1}\widehat{T}^{\alpha_k}(n)$.

□

The following theorem establishes the equivalence of fuzzy matrix Lyapunov discrete dynamical system with that of matrix Lyapunov difference inclusion and presents the solution set.

Theorem 4: The system (7) and (8) is a fuzzy matrix Lyapunov discrete dynamical system, and it can be expressed as

$$\Delta\widehat{T}(n) = [(A(n) \otimes A(n)) + (A(n) \otimes I_s) + (I_s \otimes A(n))]\widehat{T}(n) + (I_s \otimes F(n))\widehat{U}(n), \widehat{T}(n_0) = \{\widehat{T}_0\}; \quad (26)$$

$$\widehat{Y}(n) = (I_s \otimes C(n))\widehat{T}(n) + (I_s \otimes D(n))\widehat{U}(n). \quad (27)$$

The solution set of fuzzy matrix Lyapunov discrete dynamical system (26) and (27) is given by

$$\widehat{T}(n) \in (\phi^*(n,n_0) \otimes \phi(n,n_0))\widehat{T}_0 + \sum_{j=n_0}^{n-1} (\phi^*(n,j+1) \otimes \phi(n,j+1))(I_s \otimes F(j))\widehat{U}(j). \quad (28)$$

Proof: Proof follows from the Lemmas 1,2,3 and Theorem 1 since there exists $\widehat{T}(n) \in E^{s^2}$ on $[0,N]$ such that $\widehat{T}^{\alpha}(n)$ is a solution set to the difference inclusions (9) and (10).

□

The following corollary is the input characterization of the solution set of the initial value problem associated with the non homogeneous matrix Lyapunov discrete dynamical system.corollary

Corollary 1: If the input is in the form $\widehat{U}(n) = \widetilde{u}_1(n) \times \widetilde{u}_2(n) \times ... \times \widetilde{u}_i(n) \times ... \times \widetilde{u}_{s^2}(n)$ where $\widetilde{u}_k(n) \in R^1, k \neq i$ are crisp numbers, then the ith component of the solution set of (5) is a fuzzy set in E^1.

The following definitions fuzzy controllability, fuzzy observability, α-level sets and product of fuzzy matrix with α-level sets are essential for exploring the controllability and observability of the fuzzy matrix Lyapunov discrete dynamical system.

Definition 4: The fuzzy system given by equations (26)-(27) is said to be completely controllable if for any initial state $\widehat{T}(n_0) = \widehat{T}_0$ and any given final state $\widehat{T}_f$ there exists a finite time $n_1 > 0$ and a control $\widehat{U}(n)$, $0 \leq n \leq n_1$ such that $\widehat{T}(n_1) = \widehat{T}_f$.

Definition 5: The fuzzy system given by equations (26)-(27) is said to be completely observable over the interval $[0, N]$ if the knowledge of rule base of input $\widehat{U}$ and output $\widehat{Y}$ over $[0, N]$ suffices to determine a rule base of initial state $\widehat{T}_0$. Let $u_i^l, y_i^l, i = 1, 2, .., s^2, l = 1, 2, ..., m$, be fuzzy sets in E^l. We assume that the rule base for the input and output is given by

$$R^l : \text{If } \widetilde{u}_1(n) \text{ is in } u_1^l(n), \widetilde{u}_2(n) \text{ is in } u_2^l(n), ..., \widetilde{u}_{s^2}(n) \text{ is in } u_{s^2}^l(n),$$

$$\text{Then } \widetilde{y}_1(n) \text{ is in } y_1^l(n), \widetilde{y}_2(n) \text{ is in } y_2^l(n), ..., \widetilde{y}_{s^2}(n) \text{ is in } y_{s^2}^l(n), l = 1, 2, ..., m \tag{29}$$

and the output can be expressed as a function of input by the equation

$$\widehat{Y}(n) = (I_s \otimes C(n))\widehat{T}(n) + (I_s \otimes D(n))\widehat{U}(n).$$

Definition 6: Let $x, y \in E^{s^2}$ and $x = x_1 \times x_2 \times ... \times x_{s^2}$ and $y = y_1 \times y_2 \times ... \times y_{s^2}$, $x_i, y_i \in E^1, i = 1, 2, ..., s^2$.

If $y = z + x$, then $z = y - x$ which is defined by $[z]^\alpha = [y - x]^\alpha = [y]^\alpha - [x]^\alpha = \begin{bmatrix} [y_1]^\alpha - [x_1]^\alpha \\ ... \\ [y_{s^2}]^\alpha - [x_{s^2}]^\alpha \end{bmatrix}$

If $y = w - x$, then $w = y + x$ which is defined by
$[w]^\alpha = [y + x]^\alpha = [y]^\alpha + [x]^\alpha = \begin{bmatrix} [y_1]^\alpha + [x_1]^\alpha \\ ... \\ [y_{s^2}]^\alpha + [x_{s^2}]^\alpha \end{bmatrix}$.

Definition 7: Let $C = \begin{bmatrix} c_{11} & c_{12} & ... & c_{1s^2} \\ c_{21} & c_{22} & ... & c_{2s^2} \\ ... & ... & ... & ... \\ c_{s^2 1} & c_{s^2 2} & ... & c_{s^2 s^2} \end{bmatrix}$ be an $s^2 \times s^2$ matrix, $p = p_1 \times p_2 \times ... \times p_{s^2}$, let $p_i \in E^1, i = 1, 2, ..., s^2$, be a fuzzy set in E^{s^2}, and let $[p_i]^\alpha$ be α- level sets of p_i, define the product Cp of C and p as

$$[Cp]^\alpha = C[p]^\alpha = \begin{bmatrix} c_{11} & c_{12} & ... & c_{1s^2} \\ c_{21} & c_{22} & ... & c_{2s^2} \\ ... & ... & ... & ... \\ c_{s^2 1} & c_{s^2 2} & ... & c_{s^2 s^2} \end{bmatrix} \begin{bmatrix} [p_1]^\alpha \\ [p_2]^\alpha \\ ... \\ [p_{s^2}]^\alpha \end{bmatrix} = \begin{bmatrix} c_{11}[p_1]^\alpha + ... + c_{1s^2}[p_{s^2}]^\alpha \\ c_{21}[p_1]^\alpha + ... + c_{2s^2}[p_{s^2}]^\alpha \\ ... \\ c_{s^2 1}[p_1]^\alpha + ... + c_{s^2 s^2}[p_{s^2}]^\alpha \end{bmatrix}$$

4. Controllability of Fuzzy Matrix Lyapunov discrete dynamical system

A sufficient condition for controllability of fuzzy matrix Lyapunov discrete dynamical system is derived by fuzzy rule based approach via corresponding Lyapunov difference inclusion.

Theorem 5: The fuzzy system (26)-(27) is completely controllable if the $s^2 \times s^2$ symmetric

controllable matrix

$$W(n_0,N)=\sum_{j=n_0}^{N-1}[(\phi^*(N,j+1)\otimes\phi(N,j+1))(I_s\otimes F(j))(I_s\otimes F(j))^*(\phi^*(N,j+1)\otimes\phi(N,j+1))^*] \tag{30}$$

(Where $*$ represents the conjugate transpose) is nonsingular. Furthermore, the fuzzy control $\widehat{U}(n)$ which transfers the state of the system from $\widehat{T}(0)=\widehat{T}_0$ to a fuzzy state

$$\widehat{T}(N)=\widehat{T}_f=(t_{f_1},t_{f_2},...,t_{f_{s^2}}) \tag{31}$$

can be modified by the following fuzzy rule base:

$$R:\text{IF}\quad \widetilde{t_1}\quad\text{is in}\quad t_{f_1},t_{f_2},...,\widetilde{t}_{f_{s^2}}\quad\text{is in}\quad t_{f_{s^2}}\text{THEN}\quad \widetilde{u}_1\quad\text{is in}\quad u_1...\widetilde{u}_{s^2}\quad\text{is in}\quad u_{s^2} \tag{32}$$

where

$$\begin{aligned}(\widetilde{u}_1(n),\widetilde{u}_2(n),...,\widetilde{u}_{s^2}(n))=&\\ &\frac{1}{N}(I_s\otimes F(n))^{-1}(\phi^*(N,j+1)\otimes\phi(N,j+1)))^{-1}\\ &\times(\widetilde{t_1}(N),\widetilde{t_2}(N),...,t_{f_i},..,\widetilde{t}_{s^2}(N))\\ &-(I_s\otimes F(n))^*(\phi^*(N,j+1)\otimes\phi(N,j+1))^*\\ &W^{-1}(n_0,N)(\phi^*(N,N_0)\otimes\phi(N,N_0))\widehat{T}(0),i=1,2,..,s^2.\end{aligned}$$

Proof: Suppose that the symmetric controllability matrix $W(n_0,N)$ is nonsingular. Therefore $W^{-1}(n_0,N)$ exists. By multiplying equation (30) on both sides by $W^{-1}(n_0,N)(\phi^*(N,N_0)\otimes\phi(N,N_0))\widehat{T}_0$, we get

$$(\phi^*(N,N_0)\otimes\phi(N,N_0))\widehat{T}_0=\sum_{j=0}^{N-1}(\phi^*(N,j+1)\otimes\phi(N,j+1))(I_s\otimes F(j))$$

$$\times(I_s\otimes F(j))^*(\phi^*(N,j+1)\otimes\phi(N,j+1)^*)W^{-1}(0,N)(\phi^*(N,N_0)\otimes\phi(N,N_0))\widehat{T}_0. \tag{33}$$

Now our problem is to find the control $\widehat{U}(n)$ such that

$$\widehat{T}(N)=\widehat{T}_f=(\phi^*(N,N_0)\otimes\phi(N,N_0))\widehat{T}_0+\sum_{j=n_0}^{N-1}(\phi^*(N,j+1)\otimes\phi(N,j+1))(I_s\otimes F(j))\widehat{U}(j). \tag{34}$$

Since $\widehat{T}$ is fuzzy and $\widehat{U}(n)$ must be fuzzy, otherwise the left side of equation (34) cannot be equal to the crisp right side. Now $\widehat{T}_f$ can be written as

$$\begin{aligned}\widehat{T}_f=\frac{1}{N}\sum_{j=n_0}^{N-1}\widehat{T}_f=&\frac{1}{N}\sum_{j=n_0}^{N-1}(\phi^*(N,j+1)\otimes\phi(N,j+1))(I_s\otimes F(j))\\ &\times(I_s\otimes F(j))^{-1}(\phi^*(N,j+1)\otimes\phi(N,j+1))^{-1}\widehat{T}_f.\end{aligned} \tag{35}$$

From (34) and (35) we have

$$\frac{1}{N}\sum_{j=n_0}^{N-1}(\phi^*(N,j+1)\otimes\phi(N,j+1))(I_s\otimes F(j))\times(I_s\otimes F(j))^{-1}(\phi^*(N,j+1)\otimes\phi(N,j+1))^{-1}\widehat{T}_f$$

$$= (\phi^*(N, N_0) \otimes \phi(N, N_0))\widehat{T}_0 + \sum_{j=0}^{N-1} (\phi^*(N, j+1) \otimes \phi(N, j+1))(I_s \otimes F(j))\widehat{U}(j). \quad (36)$$

From (33) and (35) it follows that

$$\frac{1}{N} \sum_{j=n_0}^{N-1} (\phi^*(N, j+1) \otimes \phi(N, j+1))(I_s \otimes F(j))(I_s \otimes F(j))^{-1}$$

$$(\phi^*(N, j+1) \otimes \phi(N, j+1))^{-1}\widehat{T}_f = \sum_{j=n_0}^{N-1} (\phi^*(N, j+1) \otimes \phi(N, j+1))(I_s \otimes F(j))$$

$$(I_s \otimes F(j))^*(\phi^*(N, j+1) \otimes \phi(N, j+1))^* \times W^{-1}(n_0, N)(\phi^*(N, N_0) \otimes \phi(N, N_0))\widehat{T}_0$$

$$+ \sum_{j=n_0}^{N-1} (\phi^*(N, j+1) \otimes \phi(N, j+1))(I_s \otimes F(j))\widehat{U}(j) \quad (37)$$

i.e.,

$$\sum_{j=n_0}^{N-1} (\phi^*(N, j+1) \otimes \phi(N, j+1))(I_s \otimes F(j))\widehat{U}(j) = \sum_{j=n_0}^{N-1} (\phi^*(N, j+1) \otimes \phi(N, j+1))$$

$$(I_s \otimes F(j))\{\frac{1}{N}(I_s \otimes F(j))^{-1}(\phi^*(N, j+1) \otimes \phi(N, j+1))^{-1}\widehat{T}_f -$$

$$(I_s \otimes F(j))^*(\phi^*(N, j+1) \otimes \phi(N, j+1))^* W^{-1}(n_0, N)(\phi^*(N, N_0) \otimes \phi(N, N_0))\widehat{T}_0\}. \quad (38)$$

Now $\widehat{U}(N)$ can be expressed as

$$\widehat{U}(N) = \frac{1}{N}(I_s \otimes F(n))^{-1}(\phi^*(N, n+1) \otimes \phi(N, n+1))^{-1}\widehat{T}_f -$$

$$(I_s \otimes F(n))^*(\phi^*(N, n+1) \otimes \phi(N, n+1))^* \times W^{-1}(n_0, N)(\phi^*(N, N_0) \otimes \phi(N, N_0))\widehat{T}_0\}. \quad (39)$$

Now we have the following two possible cases for (39)

Case(i)
When $\widehat{T}(N) = \widehat{T}_f = (\widetilde{t}_1(N), \widetilde{t}_2(N), ..., \widetilde{t}_{s^2}(N))$ is a crisp point, equation (39) gives corresponding control $\widehat{U}(n)$ and is given by $\widehat{U}(n) = (\widetilde{u}_1, \widetilde{u}_2, ..., \widetilde{u}_{s^2})$.

Case(ii)
When $\widehat{T}(N) = (\widetilde{t}_1(N), \widetilde{t}_2(N), ..., t_{f_i}, ..., \widetilde{t}_{s^2}(N))$, equation (39) gives the corresponding control $\widehat{U}(n)$ and is given by $\widehat{U}(n) = (\widetilde{u}_1, \widetilde{u}_2, ..., u_i, ..., \widetilde{u}_{s^2})$ in which the component of $\widehat{U}(n)$ is a fuzzy set in E^1.
Clearly $\widetilde{u}_i(n)$ is in $u_i(n)$, $\mu_{t_{f_i}}(\widetilde{t}_i(N))$ gives the grade of the membership of $\widetilde{t}_i(N)$ in t_{f_i}.
Hence fuzzy rule base for the control $\widehat{U}$ given by equations (31) and (32) follows. □

Note: The converse of the above theorem need not be true. Since fuzzy rule base cannot imply the non singularity of the controllability matrix $W(n_0, N)$ given by (30). It follows that the condition in the above theorem is only sufficient condition but not necessary.

5. Observability of Fuzzy Matrix Lyapunov discrete dynamical system

A sufficient condition for observability of fuzzy matrix Lyapunov discrete dynamical system is constructed by center average defuzzifier approach via corresponding Lyapunov difference inclusion.

Theorem 6: Assume that the fuzzy rule base (29) holds, then the fuzzy system (26)-(27) is completely observable over the interval $[0, N]$ and $(I_s \otimes C(N))(\phi^*(N, N_0) \otimes \phi(N, N_0))$ is nonsingular. Furthermore, if

$$\widehat{T_0} = (\widetilde{t_0^1}, \widetilde{t_0^2}, ..., \widetilde{t_0^{s^2}}) \tag{40}$$

then one has the following rule base for the initial value $\widehat{T_0}$,

$$R^l : \text{If } \ \widetilde{u}_1(N) \in u_1^l(N), ..., \widetilde{u}_{s^2}(N) \in u_{s^2}^l(N) \ \text{ and } \ \widetilde{y}_1(N) \in y_1^l(N), ..., \widetilde{y}_{s^2}(N) \in y_{s^2}^l(N)$$

$$\text{Then } \ \widetilde{t_0^1} \ \text{ is in } \ t_0^l(1), ..., \widetilde{t_0^{s^2}}(n) \ \text{ is in } \ t_0^l(S^2), l = 1, 2, .., m. \tag{41}$$

where

$$t_0^l(i) = [(I_S \otimes C(N))(\phi^*(N, N_0) \otimes \phi(N, N_0))]^{-1}\{V_i^l(N) - (I_S \otimes D(N))\widehat{U}(N) -$$

$$(I_s \otimes C(N)) \sum_{j=n_0}^{N-1} (\phi^*(N, j+1) \otimes \phi(N, j+1))(I_s \otimes F(j))H_i^l(j)\}, \tag{42}$$

$$\widehat{T_0} = ((I_s \otimes C(N))(\phi^*(N, N_0) \otimes \phi(N, N_0))^{-1}\{\widetilde{y}(N) - (I_s \otimes D(N))\widetilde{U}(N) -$$

$$(I_s \otimes C(N)) \times \sum_{j=n_0}^{N-1} (I_s \otimes \phi(N-j-1))(I_s \otimes F(j))\widetilde{U}(j), \tag{43}$$

$$H_i^l(n) = \widetilde{u}_1(n) \times \widetilde{u}_2(n) \times, ..., u_i^l(n) ... \times \widetilde{u}_{s^2}(n), \tag{44}$$

$$V_i^l(n) = \widetilde{y}_1(n) \times \widetilde{y}_2(n) \times, ..., y_i^l(n) ... \times \widetilde{y}_{s^2}(n), i = 1, 2, .., s^2; l = 1, 2, .., m. \tag{45}$$

Proof: Consider the case when $l = 1$. Let

$$\widetilde{u}(n) = (\widetilde{u}_1(n), \widetilde{u}_2(n), ..., \widetilde{u}_{s^2}(n)), \tag{46}$$

$$\widetilde{y}(n) = (\widetilde{y}_1(n), \widetilde{y}_2(n), ..., \widetilde{y}_{s^2}(n)) \tag{47}$$

Let $\mu_{u_i^1(n)}(\widetilde{u}_i(n))$ be the grade of the membership of $\widetilde{u}_i(n)$ in $u_i^1(n)$, and let $\mu_{y_i^1(n)}(\widetilde{y}_i(n))$ be the grade of membership of $\widetilde{y}_i(n)$ in $y_i^1(n)$. Since $(I_s \otimes C(N))(\phi^*(N, N_0) \otimes \phi(N, N_0))$ is nonsingular and from (28) we have

$$\widehat{T_0} = [(I_s \otimes C(N))(\phi^*(N, N_0) \otimes \phi(N, N_0))]^{-1}\{\widetilde{y}(N) - (I_s \otimes D(N))\widetilde{u}(N) -$$

$$(I_s \otimes C(N)) \sum_{j=n_0}^{N-1} (\phi^*(n, j+1) \otimes \phi(n, j+1))(I_s \otimes F(j))\widetilde{u}(j)\} \tag{48}$$

When the input and output are both fuzzy sets it follows from equation 8 that

$$(I_s \otimes C(N))\widehat{T}(n) = \widehat{Y}(n) - (I_s \otimes D(N))\widetilde{u}(N) \tag{49}$$

is a fuzzy set. From equation (28), we get

$$(I_s \otimes C(N))(\phi^*(N,N_0)\otimes\phi(N,N_0))\widehat{T}_0 + \sum_{j=n_0}^{N-1}(\phi^*(n,j+1)\otimes\phi(n,j+1))(I_s\otimes F(j))\widehat{U}(j)$$
$$= \widehat{Y}(n) - ((I_s\otimes D(N)))\widehat{U}(n). \tag{50}$$

Using Definition 6, it follows that

$$(I_s \otimes C(N))(\phi^*(N,N_0)\otimes\phi(N,N_0))\widehat{T}_0 = \{\widehat{Y}(n) - (I_s\otimes D(N))\widehat{U}(n) -$$
$$(I_s\otimes C(N))\}\sum_{j=n_0}^{N-1}(\phi^*(n,j+1)\otimes\phi(n,j+1))(I_s\otimes F(j))\widehat{U}(j). \tag{51}$$

Since $(I_s \otimes C(N))(\phi^*(N,N_0)\otimes\phi(N,N_0))$ is nonsingular, we have

$$\widehat{T}_0 = [(I_s \otimes C(N))(\phi^*(N,N_0)\otimes\phi(N,N_0)]^{-1}\{\widehat{Y}(N) - ((I_s\otimes D(N))\widehat{U}(N)) -$$
$$(I_s\otimes C(N))\times\sum_{j=n_0}^{N-1}(\phi^*(n,j+1)\otimes\phi(n,j+1))(I_s\otimes F(j))\widehat{U}(j)\} \tag{52}$$

Now, the initial value $\widehat{T}_0$ should be a fuzzy set but not a crisp value. The following assumptions will enable us to determine each component of $\widehat{T}_0$

$$H_i^1(n) = \widetilde{u}_1(n)\times u_i(n+1)\times ...\times \widetilde{u}_{s^2}(n)$$
$$V_i^1(n) = \widetilde{y}_1(n)\times y_i(n+1)\times ...\times \widetilde{y}_{s^2}(n) \text{ where } \ i=1,2,...,s^2 \tag{53}$$

From the Corollary 3, we know that the ith component of the set

$$(\phi^*(N,N_0)\otimes\phi(N,N_0))\widehat{T}_0 + \sum_{j=n_0}^{n-1}(\phi^*(n,j+1)\otimes\phi(n,j+1))(I_s\otimes F(j))H_i^1(n) \tag{54}$$

is a fuzzy set in E^1. From the fact that the product of a square matrix of size s^2 and column vector whose elements are α- level sets defined on fuzzy set in E^{s^2} is again a fuzzy set in E^{s^2}, it follows that the product

$$(I_s\otimes C(N))\times\sum_{j=n_0}^{n-1}(\phi^*(n,j+1)\otimes\phi(n,j+1))(I_s\otimes F(j))H_i^1(n) \tag{55}$$

is a fuzzy set in E^{s^2}. Hence $\widehat{T}_o$ is a fuzzy set in E^{s^2} and the i^{th} component of it denoted by $t_0^1(i)$ is a fuzzy set in E^1. The grade of membership of $\widetilde{t}_0^i$ in $t_0^1(i)$ is defined by $\mu_{t_0^1(i)}(\widetilde{t}_0^i) = min\{\mu_{u_i^1(n)}(\widetilde{u}_i(n)), \mu_{y_i^1}(n)(\widetilde{y}_i(n))\}$. Now the initial value is determined by using the equations (41) to (45). In general, computation of $t_0^l(i)$ is very difficult, but to solve the real value problem the following approximation is chosen. Now we take the point $(\widetilde{t}_0^i, \mu_{t_0^l}(i)(\widetilde{t_0}^i))$ and the zero level set $[t_0^l(i)]^0$ to determine a triangle as the new fuzzy set $t_0^l(i)$. We can use the centre average defuzzifier

$$\widetilde{t}_0^i = \frac{\sum_{l=1}^m (\widetilde{t}_0^i)^l \mu_{t_0^l(i)}(\widetilde{t}_0^i)^l}{\sum_{l=1}^m \mu_{t_0^l(i)}(\widetilde{t}_0^i)^l} \tag{56}$$

to determine the initial value $\widehat{T}_0 = (\widetilde{t}_0^1, \widetilde{t}_0^2, ..., \widetilde{t}_0^{s^2})$. To obtain more accurate value for the initial state, more rule bases may be provided. □

6. Numerical example

In this section, a numerical example which verify and validate the established conditions of controllability and observability of fuzzy matrix Lyapunov discrete dynamical system is presented. Consider the fuzzy matrix Lyapunov discrete dynamical system (27) satisfying (28) with $A(n) = \begin{bmatrix} 1 & 0 \\ 0 & -2 \end{bmatrix}$, $F(n) = \begin{bmatrix} 2^n & 0 \\ 0 & 3^n \end{bmatrix}$, $C(n) = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ and $D(n) = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}$, $N = 2$, $T(0) = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$. Let the final state $\widehat{t_f} = (t_{f_1}, t_{f_2}, t_{f_3}, t_{f_4})$ in E^4, where $[\widehat{T_f}]^\alpha = ([t_{f_1}]^\alpha, [t_{f_2}]^\alpha, [t_{f_3}]^\alpha, [t_{f_4}]^\alpha)^T = [[\alpha - 1, 1 - \alpha], [\alpha - 1, 1 - \alpha], [0.1(\alpha - 1), 0.1(1 - \alpha)], [0.1(\alpha - 1), 0.1(1 - \alpha)]]^T$. Choose the points $\widetilde{t}_{f_1} = 0.5, \widetilde{t}_{f_2} = 0.25, \widetilde{t}_{f_3} = 0.05,$ and $\widetilde{t}_{f_4} = 0.025$, which are in $t_{f_1}, t_{f_2}, t_{f_3}$, and t_{f_4} whose membership function values are $0.5, 0.75, 0.5$ and 0.75 respectively. The fundamental matrix of homogeneous discrete dynamical system $\Delta T(n) = A(n)T(n)$ is given by $\phi(n, n_0) = \begin{bmatrix} 1^{n-n_0} & 0 \\ 0 & (-2)^{n-n_0} \end{bmatrix}$. The $2^2 \times 2^2$ symmetric controllable matrix $W(0, 2)$ obtained by equation (30) of Theorem 5 we get $W(0, 2) = \begin{bmatrix} 5 & 0 & 0 & 0 \\ 0 & 13 & 0 & 0 \\ 0 & 0 & 5 & 0 \\ 0 & 0 & 0 & 13 \end{bmatrix}$ is nonsingular. Thus from Theorem 5 the α- level fuzzy control $\widehat{U}(n)$ is computed by

$$\widehat{U}^\alpha(n) = \begin{bmatrix} (2)^{-n-1}[\alpha - 1, 1 - \alpha] \\ 3^{-n}(-2)^n[\alpha - 1, 1 - \alpha] \\ (2)^{-n-1}[0.1(\alpha - 1), 0.1(1 - \alpha)] \\ 3^{-n}(-2)^n[0.1(\alpha - 1), 0.1(1 - \alpha)] \end{bmatrix} - \begin{bmatrix} (2)^n 0.2 \\ (0.304)3^n(-2)^{-n+1} \\ (0.8)2^n \\ (1.1216)3^n(-2)^{-n+1} \end{bmatrix}.$$

The α- level sets of fuzzy input $\widehat{U}(n)$ and fuzzy output $\widehat{Y}(n)$ by Rule Base 1 are denoted by $[\widehat{U}^{(1)}]^\alpha$, $[\widehat{Y}^{(1)}]^\alpha$ and are given by

Rule Base 1:

$$[\widehat{U}^{(1)}]^\alpha = \begin{bmatrix} [0, -0.75(\alpha - 1)] \\ [0.75(\alpha - 1) + 1, 1] \\ [0, -0.5(\alpha - 1)] \\ [0.5(\alpha - 1) + 1, 1] \end{bmatrix} [\widehat{Y}^{(1)}]^\alpha = \begin{bmatrix} [0, -2(\alpha + 1)] \\ [0.5\alpha + 2.5, 3] \\ [0, -1.5(\alpha - 1)] \\ [0.5(\alpha - 1) + 3, 3] \end{bmatrix}$$

The α- level sets of fuzzy input $\widehat{U}(n)$ and fuzzy output $\widehat{Y}(n)$ by Rule Base 2 are denoted by $[\widehat{U}^{(2)}]^\alpha$, $[\widehat{Y}^{(21)}]^\alpha$ and are given by.

Rule Base 2:

$$[\widehat{U}^{(2)}]^\alpha = \begin{bmatrix} [0, -0.8(\alpha - 1)] \\ [0.8\alpha + 0.2, 1] \\ [0, -0.5(\alpha - 1)] \\ [0.5\alpha + 0.5, 1] \end{bmatrix} [\widehat{Y}^{(2)}]^\alpha = \begin{bmatrix} [0, -1.5(\alpha - 1)] \\ [\alpha + 1, 2] \\ [0, -2.5(\alpha - 1)] \\ [(2\alpha + 1), 3] \end{bmatrix}.$$

From Rule Base 1, select

$$\widetilde{u}^1 = (\widetilde{u}_1, \widetilde{u}_2, \widetilde{u}_3, \widetilde{u}_4) = (0.5, 0.85, 0.4, 0.75)$$

the values of the membership function of $\widetilde{u}_1, \widetilde{u}_2, \widetilde{u}_3$ and $\widetilde{u}_4$ are $\frac{1}{3}, 0.8, 0.2$, and $\frac{1}{2}$, respectively. Also

$$\widetilde{y}^1 = (\widetilde{y}^1, \widetilde{y}^2, \widetilde{y}^3, \widetilde{y}^4) = (1, 2.8, 0.5, 2.9)$$

the values of the membership function of the output $\widetilde{y}_1, \widetilde{y}_2, \widetilde{y}_3,$ and $\widetilde{y}_4$, are $\frac{1}{2}, 0.6, \frac{2}{3}$ and 0.8 respectively.
From Rule Base 2, we select

$$\widetilde{u}^2 = (\widetilde{u}_1, \widetilde{u}_2, \widetilde{u}_3, \widetilde{u}_4) = (0.5, 0.8, 0.25, 0.75),$$

the values of the membership function of $\widetilde{u}_1, \widetilde{u}_2, \widetilde{u}_3$ and $\widetilde{u}_4$ respectively are $\frac{1}{3}, 0.8, 0.2$ and $\frac{1}{2}$ respectively. Also

$$\widetilde{y}^2 = (\widetilde{y}^1, \widetilde{y}^2, \widetilde{y}^3, \widetilde{y}^4) = (1, 1.75, 2, 1.5)$$

the values of the membership function of $\widetilde{y}^1, \widetilde{y}^2, \widetilde{y}^3,$ and $\widetilde{y}^4$ are $\frac{1}{3}, \frac{3}{4}, 0.2$ and 0.25 respectively.
From Rule Base 1 and equation (43) we have $\widehat{T}_0 = \begin{bmatrix} [1.3] \\ [0.0375] \\ [1.7] \\ [-0.0625] \end{bmatrix}$. From Rule Base 1 and equation 42 we have $t_0^1(1) = \begin{bmatrix} [2.8; 2.2.5\alpha + 0.55] \\ [-0.2125; -0.5\alpha - 0.7125] \\ [2.9] \\ [-1.1875] \end{bmatrix}$. When $\alpha = 0$, we observed that $\widetilde{t}_0^1 = 1.3$ belong to the interval $[2.8, 0.55]$. We choose its function in t_0^1 as
$\mu_{t_0^1}(1) = min\{\mu_{u_1^1}(\widetilde{u}_1(n)), \mu_{y_1^1}(n)(\widetilde{y}_1(n))\} = min(\frac{1}{3}, \frac{1}{2}) = \frac{1}{3}$.
$t_0^1(2) = \begin{bmatrix} [0.5\alpha + 1; 1.5] \\ [0.1875(1-\alpha); 0] \\ [2.9] \\ [-1.175] \end{bmatrix}$. When $\alpha = 0$ we observed that $\widetilde{t}_0^2 = 0.0375$ belong to the interval $[0.1875; 0]$. We choose its membership grade in $t_0^1(2)$ as $\mu_{t_0^1(2)}\widetilde{t}_0^1 = min(0.8, 0.6) = 0.6$.
$t_0^1(3) = \begin{bmatrix} [1.3] \\ [0.0375] \\ [2.9] \\ [-13125, -0.375\alpha - 1.5625] \end{bmatrix}$. When $\alpha = 0$, we observed that $\widetilde{t}_0^3 = 1.7$ belong to the interval $[2.9; 0]$. We choose its membership grade in $t_0^1(3)$ as $\mu_{t_0^1}(3)\widetilde{t}_0^1 = min(0.2, \frac{2}{3}) = 0.2$.
$t_0^1(4) = \begin{bmatrix} [1.3] \\ [0.0375] \\ [0.5\alpha + 2.5; 3] \\ [-0.875\alpha - 0.75; -1.625] \end{bmatrix}$. When $\alpha = 0$, we observed that $\widetilde{t}_0^4 = -0.0625$ belong to the interval $[-0.75, -1.625]$. We choose its membership grade in $t_0^1(4)$ as $\mu_{t_0^1}(4)\widetilde{t}_0^1 = min(\frac{1}{2}, 0.8) = \frac{1}{2}$.
Similarly for Rule Base 2 using equation (43) we get $\widehat{T}_0 = \begin{bmatrix} [0.95] \\ [-0.125] \\ [0.75] \\ [0.3125] \end{bmatrix}$.
By using Rule Base 2 and equation (42) we get

$$t_0^2(1) = \begin{bmatrix} [1.75; 2.4\alpha - 0.25] \\ [-0.2; -0.375\alpha + 0.05] \\ [1.5] \\ [-0.8125] \end{bmatrix}, \mu_{t_0^2}(1)\widetilde{t}_0^2 = min(\tfrac{3}{8}, \tfrac{1}{3}) = \tfrac{1}{3},$$

$$t_0^2(2) = \begin{bmatrix} [\alpha - 0.5; 0.5] \\ [0.2 - 0.2\alpha; 0] \\ [1.5] \\ [-0.2] \end{bmatrix}, \mu_{t_0^2}(2)\widetilde{t}_0^2 = min(\tfrac{3}{4}, \tfrac{3}{4}) = \tfrac{3}{4},$$

$$t_0^2(3) = \begin{bmatrix} [0.25] \\ [0.05] \\ [1.5] \\ [-1.3125; -0.625\alpha - 0.6875] \end{bmatrix}, \mu_{t_0^2}(3)\widetilde{t}_0^2 = min(\tfrac{1}{2}, 0.2) = 0.2,$$

$$t_0^2(4) = \begin{bmatrix} [0.25] \\ [0.5] \\ [2\alpha + 1; 3] \\ [-0.875\alpha - 0.375; -1.25] \end{bmatrix}, \mu_{t_0^2}(4)\widetilde{t}_0^2 = min(\tfrac{1}{2}, 0.25) = 0.25.$$

By using the center average defuzzifier given by equation (56) the initial value $\widehat{T}_0 = (\widetilde{t}_0^1, \widetilde{t}_0^2, \widetilde{t}_0^3, \widetilde{t}_0^4)$ is given by

$$\widetilde{t}_0^1 = \frac{[1.3 \times \frac{1}{3} + (0.95) \times \frac{1}{3}]}{\frac{1}{3} + \frac{1}{3}} = 2.55142, \widetilde{t}_0^2 = \frac{[0.0375 \times (0.6) + (-0.125) \times 0.75]}{0.6 + 0.75} = -0.0527,$$

$$\widetilde{t}_0^3 = \tfrac{[1.7\times(0.2)+(0.75)\times(0.2)]}{0.2+0.2} = 1.225, \widetilde{t}_0^4 = \tfrac{[-0.0625\times(0.5)+(0.3125)\times(0.25)]}{0.5+0.25} = 0.0625.$$

By considering more rule bases the accuracy of the initial state can be improved.

7. Conclusion

In this paper, by visualizing fuzzy matrix Lyapunov discrete dynamical system as a Lyapunov difference inclusion, sufficient conditions for the controllability and observability of the fuzzy matrix Lyapunov discrete dynamical system are constructed by following the fuzzy rule base. We have constructed the rule base for the initial value without the knowledge of the solution of the system. This approach is new for the Lyapunov discrete dynamical systems. The constructed example clearly demonstrates the established results.

Acknowledgements

One of the authors L.N.Charyulu Rompicharla would like to thank the management and Principal of V.R.Sidhhartha Engineering College, Vijayawada for the support.
The authors are thankful to the referee for the constructive comments with resulted in the improvement of the paper.

References

Alwadie, A., Ying, H., and Shah, H. (2003). A practical two-input two-output takagi-sugeno fuzzy controller. *International Journal of Fuzzy Systems*, **5**, 123–130.

Anand, P. and Murty, K. (2005). Controllability and observability of liapunov type matrix difference system. In *Proceedings of 50th Congress of ISTAM (An International Meet) IIT Kharagpur*, pages 125–132.

Barnett, S. (1975). *Introduction to Mathematical Control Theory.* Clarendon press.

Biglarbegian, M., Sadeghian, A., and Melek, W. (2012). On the accessibility/controllability of fuzzy control systems. *Information Sciences*, **202**, 58–72.

Cai, Z. and Tang, S. (2000). Controllability and robustness of t-fuzzy control systems under directional disturbance. *Fuzzy Sets and Systems*, **115**, 279–285.

Chen, Y., Yang, B., Abraham, A., and Peng, L. (2007). Automatic design of hierarchical takagi–sugeno type fuzzy systems using evolutionary algorithms. *IEEE Transactions on Fuzzy Systems*, **15**, 385–397.

Conway, L. P. and Voglmeir, J. (2016). Functional analysis of anomeric sugar kinases. *Carbohydrate Research*, **432**, 23–30.

Ding, Y., Ying, H., and Shao, S. (1999). Structure and stability analysis of a takagi–sugeno fuzzy pi controller with application to tissue hyperthermia therapy. *Soft Computing*, **2**, 183–190.

Ding, Y., Ying, H., and Shao, S. (2003). Typical takagi–sugeno pi and pd fuzzy controllers: analytical structures and stability analysis. *Information Sciences*, **151**, 245–262.

Ding, Z. and Kandel, A. (2000a). On the controllability of fuzzy dynamical systems (ii). *Journal of Fuzzy Mathematics*, **8**, 295–306.

Ding, Z. and Kandel, A. (2000b). On the observability of fuzzy dynamical control systems (ii). *Fuzzy Sets and Systems*, **115**, 261–277.

Farinwata, S. S. and Vachtsevanos, G. (1993). A survey on the controllability of fuzzy logic systems. In *Proceedings of 32nd IEEE Conference on Decision and Control*, pages 1749–1750. IEEE.

Gabr, W. I. (2015). A new approach for automatic control modeling, analysis and design in fully fuzzy environment. *Ain Shams Engineering Journal*, **6**, 835–850.

Johansen, T. A., Shorten, R., and Murray-Smith, R. (2000). On the interpretation and identification of dynamic takagi-sugeno fuzzy models. *IEEE Transactions on Fuzzy Systems*, **8**, 297–313.

Kaleva, O. (1987). Fuzzy differential equations. *Fuzzy Sets and Systems*, **24**, 301–317.

Lakshmikantham, V. and Mohapatra, R. (2003). *Theory of Fuzzy Differential Equations and Inclusions*. Taylor and Francis Publishers, London.

Mastiani, R. and Effati, S. (2018). On controllability and observability of fuzzy control systems. *Iranian Journal of Fuzzy Systems*, **15**, 41–64.

Murty, K., Anand, P., and Prasannam, V. (1997). First order difference system-existence and uniqueness. *Proceedings of the American Mathematical Society*, **125**, 3533–3539.

Murty, K., Andreou, S., and Viswanadh, K. (2009). Qualitative properties of general first order matrix difference systems. *Nonlinear Studies*, **16**, 1–17.

Murty, K., Balaram, V., and Viswanadh, K. (2008). Solution of kronecker product initial value problems associated with first order difference system via tensor—based hardness of the shortest vector problem. *Electronic Modelling*, **30**, 1–15.

Murty, K., Prasad, K., and Anand, P. (1995). Two-point boundary value problems associated with liapunov type matrix difference system. *Dynamic Systems Applications*, **4**, 205–213.

Murty, M. and Kumar, G. S. (2008). On controllability and observability of fuzzy dynamical matrix lyapunov systems. *Advances in Fuzzy Systems*, **1**, 1–16.

Putcha, V. S. and Katakol, S. (2022). Qualitative and analytical treatment of nonlinear dynamical systems in neurological diseases. In *Role of Nutrients in Neurological Disorders*, pages 85–114. Springer.

Putcha, V. S. and Prathyusha, M. (2019). Dichotomy and well-conditioning of discrete matrix sylvester systems. *International Journal of Differential Equations and Applications*, **18**, 63–85.

Putcha, V. S., Rao, D., and Malladi, R. (2019). Existence of ψ-bounded solutions for first order matrix difference system on z. *International Journal of Embedded Systems and Emerging Technologies*, **5**, 6–16.

Putcha, V. S., Rompicharla, C., and Deekshitulu, G. (2012). A note on fuzzy discrete dynamical systems. *International Journal of Contemporary Mathematical Sciences*, **7**, 1931–1939.

Rådström, H. (1952). An embedding theorem for spaces of convex sets. *Proceedings of the American Mathematical Society*, **3**, 165–169.

Ralescu, D. (1979). Applications of fuzzy sets to systems analysis. *Journal of Symbolic Logic*, **44**.

Rompicharla, C. L., Putcha, S. V., and Deekshitulu, G. (2020). Existence of $(\phi \otimes \psi)$ bounded solutions for linear first order kronecker product systems. *International Journal of Recent Scientific Research*, **11**, 39047–39053.

Rompicharla, C. L., Putcha, V. S., and Deekshithulu, G. (2019). Controllability and observability of fuzzy matrix discrete dynamical systems. *Journal of Nonlinear Sciences & Applications (JNSA)*, **12**.

Sugeno, M. (1999). On stability of fuzzy systems expressed by fuzzy rules with singleton consequents. *IEEE Transactions on Fuzzy Systems*, **7**, 201–224.

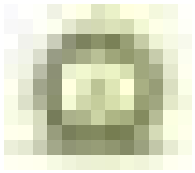
Sundaranand Putcha, V. (2014). Discrete linear sylvester repetitive process. *Nonlinear Studies*, **21**.

Takagi, T. and Sugeno, M. (1985). Fuzzy identification of systems and its applications to modeling and control. *IEEE Transactions on Systems, Man, and Cybernetics*, **1**, 116–132.

Ying, H. (1999). Analytical analysis and feedback linearization tracking control of the general takagi-sugeno fuzzy dynamic systems. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, **29**, 290–298.

Ying, H. (2006). Deriving analytical input–output relationship for fuzzy controllers using arbitrary input fuzzy sets and zadeh fuzzy and operator. *IEEE Transactions on Fuzzy Systems*, **14**, 654–662.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 22, No. 1, 2024 (New Series), pp 21–37
http://www.ssca.org.in/journal



Some Improved Separate Estimators of Population Mean in Stratified Ranked Set Sampling

Rajesh Singh and Anamika Kumari
Department of Statistics
Banaras Hindu University, Varanasi, India

Received: 11 October 2022; Revised: 13 April 2023; Accepted: 21 May 2023

Abstract

This paper presents improved population mean estimators using auxiliary variable in Stratified Ranked Set Sampling. We have derived the expressions for bias and mean square errors up to the first order of approximation and shown that the proposed estimators under optimum conditions are more efficient than other estimators taken in this paper. In an attempt to verify the efficiencies of proposed estimators, theoretical results are supported by empirical study and simulation study for which we have considered two populations.

Key words: Study variable; Auxiliary variable; Bias; Mean square error; Ranked set sampling.

AMS Subject Classifications: 62K05, 05B05

1. Introduction

In theory of sampling it is evident that suitable use of auxiliary information improves the efficiency of the estimator. These auxiliary information may be used either at the design phase or the estimation phase or at both phases. Cochran (1940) was the first to introduce a ratio estimator of Population Mean using auxiliary information. Shabbir and Gupta (2007), Koyuncu and Kadilar (2009) and Chaudhary *et al.* (2009) have considered the problem of estimating population mean taking into consideration information on auxiliary variable.

When population is heterogenous stratified random sampling (SSRS) is used for better accuracy. Several authors like Kadilar and Cingi (2003), Shabbir and Gupta (2006) and Haq and Shabbir (2013) have proposed estimators in stratified random sampling using information on a single auxiliary variable. Singh and Kumar (2012) have proposed improved estimators of population mean using two auxiliary variables in stratified random sampling. Recently, Muneer *et al.* (2020) have proposed family of chain exponential estimators in SSRS.

Ranked set sampling (RSS) is an improved sampling method over Simple Random Set Sampling (SRS). McIntyre (1952) was the first to explain RSS for estimating the population means. Takahasi and Wakimoto (1968) gave the necessary mathematical theory of RSS.

Corresponding Author: Anamika Kumari
Email: anamikatiwari1410@gmail.com

Samawi and Muttlak (1996) suggested ratio estimators of population mean in RSS and showed that the RSS estimators gave improved results over their SRS counterparts. Shiva (2006) compared RSS with SRS for estimation of the unknown mean of study variable and the ratio of study variable to auxiliary variable. He concluded that RSS gives a better estimate for both the mean and the ratio. Singh *et al.* (2014) suggested a general procedure for estimating the population mean using RSS. Bouza (2014) and Bouza *et al.* (2018) provided a review of RSS, its modification, and its application.

Stratified ranked set sampling (SRSS) was first introduced by Samawi (1996) for increasing the efficiency of estimator of population mean. Samawi and Siam (2003) have proposed the combined and the separate ratio estimators in SRSS.

2. Sampling methodology

In ranked set sampling (RSS), we rank randomly selected units from the population merely by observation or prior experience after which only a few of these sampled units are measured. In RSS, k independent random sets each of size k are selected from the population and each unit in the set is being selected with equal probability. The members of each random set are ranked with respect to the characteristic of the auxiliary variable. Then the smallest unit is selected from the first ordered set and the second smallest unit is selected from the second ordered set. By this way, this procedure is continued until the largest rank is chosen from the k^{th} set. This cycle may be repeated r times, so rk $(=n)$ units have been measured during this process.

SRSS takes the following steps.

- Step 1: Select k_h^2 bivariate sample units randomly from the h^{th} stratum of the population.
- Step 2: Arrange these selected units randomly into k_h sets, each of size k_h.
- Step 3: The procedure of ranked set sampling (RSS) is then applied, on each of the sets to obtain the k_h sets of ranked set sample units. Here ranking is done with respect to the auxiliary variable X_h.
- Step 4: Repeat the above steps r times for each stratum to get the desired sample of size $n_h = k_h r$.

Consider a finite population $U = (U_1, U_2, ..., U_N)$ based on N identifiable units with a study variable Y and auxiliary variables X associated with each unit U_i, i = 1,2,...,N of the population. Let the population be divided into L disjoint strata with stratum h based on N_h, h = 1,2,...,L units.

Let $(Y_{h[1]j}, X_{h(1)j}), (Y_{h[2]j}, X_{h(2)j})$, ..., $(Y_{h[k_h]j}, X_{h(k_h)j})$ be the stratified ranked set sample for j^{th}, j=1,2,...,r cycle in h^{th} stratum.

Let $\overline{y}_{[SRSS]} = \sum_{h=1}^{L} W_h \overline{y}_{h[rss]}$ and $\overline{x}_{[SRSS]} = \sum_{h=1}^{L} W_h \overline{x}_{h[rss]}$

respectively be the stratified ranked set sample means corresponding to the population means

$\overline{Y} = \sum_{h=1}^{L} W_h \overline{Y}_h$ and $\overline{X} = \sum_{h=1}^{L} W_h \overline{X}_h$

of variables Y and X, where $W_h = \frac{N_h}{N}$ is the weight in stratum h.

Let $\overline{y}_{h[rss]} = \sum_{i=1}^{k_h} \sum_{j=1}^{r} \frac{Y_{h[i]j}}{k_h r}$ and $\overline{x}_{h[rss]} = \sum_{i=1}^{k_h} \sum_{j=1}^{r} \frac{X_{h(i)j}}{k_h r}$

be the stratified ranked set sample means corresponding to the population means

$\overline{Y}_h = \sum_{j=1}^{N_h} \frac{Y_{h[i]j}}{N_h}$ and $\overline{X}_h = \sum_{j=1}^{N_h} \frac{X_{h(i)j}}{N_h}$

of variables Y and X in stratum h.

Let $s_{yh}^2 = \frac{1}{n_h - 1} \sum_{h=1}^{L} (Y_{h[i]} - \overline{y}_{h[rss]})^2$, $s_{xh}^2 = \frac{1}{n_h - 1} \sum_{h=1}^{L} (X_{h(i)} - \overline{x}_{h[rss]})^2$ and

$s_{xyh} = \frac{1}{n_h - 1} \sum_{h=1}^{L} (Y_{h[i]} - \overline{y}_{h[rss]})(X_{h(i)} - \overline{x}_{h[rss]})$

respectively be the sample variances and covariances corresponding to the population variances and covariances.

$S_{yh}^2 = \frac{1}{N_h - 1} \sum_{h=1}^{L} (Y_{h[i]} - \overline{Y}_h)^2$, $S_{xh}^2 = \frac{1}{N_h - 1} \sum_{h=1}^{L} (X_{h(i)} - \overline{X}_h)^2$

and $S_{xyh} = \frac{1}{N_h - 1} \sum_{h=1}^{L} (Y_{h[i]} - \overline{Y}_h)(X_{h(i)} - \overline{X}_h)$ in the stratum h.

Let C_{yh} and C_{xh} respectively be the population coefficient of variation of variables Y and X.

3. Existing estimators

The conventional separate estimator of the population mean $\overline{Y}$ under SRSS is given by

$$t^s = \sum_{h=1}^{L} W_h \overline{y}_{h[rss]} \tag{1}$$

The variance of the estimator t^s is given by

$$Var(t^s) = \sum_{h=1}^{L} W_h^2 \overline{Y}_h^2 U_{20h} \tag{2}$$

The classical separate ratio estimator of the population mean $\overline{Y}$ under SRSS is defined as

$$t_r^s = \sum_{h=1}^{L} W_h \overline{y}_{h[rss]} \frac{\overline{X}}{\overline{x}_{h[rss]}} \tag{3}$$

The Mean Squared Error (MSE) of the estimator t_r^c is given by

$$MSE(t_r^s) = \sum_{h=1}^{L} W_h^2 \overline{Y}_h^2 [U_{20h} + U_{02h} - 2U_{11h}] \tag{4}$$

The classical separate regression estimator of the population mean $\overline{Y}$ under SRSS is

given as

$$t_{lr}^{s} = \sum_{h=1}^{L} W_h \overline{y}_{h[rss]} + \beta(\overline{X} - \overline{x}_{h[rss]}) \tag{5}$$

The Mean Squared Error (MSE) of the estimator t_{lr}^{c} is given by

$$MSE(t_{lr}^{s}) = \sum_{h=1}^{L} W_h^2 [\overline{Y}_h^2 U_{20h} + \beta_h^2 \overline{X}_h^2 U_{02h} - 2\beta_h \overline{Y}_h \overline{X}_h U_{11h}] \tag{6}$$

where β_h is the regression coefficient of Y_h on X_h.

4. Proposed estimators

Motivated by Bhushan *et al.* (2020), we suggest some estimators of the population mean $\overline{Y}$ using SRSS as

$$t_{p1}^{s} = \sum_{h=1}^{L} W_h \overline{y}_{h[rss]} \exp\left(\alpha_{1h}\left(\frac{\overline{x}_{h[rss]}}{\overline{X}_h} - 1\right)\right) \tag{7}$$

$$t_{p2}^{s} = \sum_{h=1}^{L} W_h \overline{y}_{h[rss]} \exp\left(\alpha_{2h} \log \frac{\overline{x}_{h[rss]}}{\overline{X}_h}\right) \tag{8}$$

where α_{1h} and α_{2h} are constants such that MSE of the estimators is minimum.

The biases of the proposed estimators are

$$Bias(t_{p1}^{s}) = \sum_{h=1}^{L} W_h \overline{Y}_h \left(\frac{\alpha_{1h}^2}{2} U_{02h} + \alpha_{1h} U_{11h}\right) \tag{9}$$

$$Bias(t_{p2}^{s}) = \sum_{h=1}^{L} W_h \overline{Y}_h \left(\frac{(\alpha_{2h}^2 - \alpha_{2h})}{2} U_{02h} + \alpha_{2h} U_{11h}\right) \tag{10}$$

The mean square errors of the proposed estimators are

$$MSE(t_{p1}^{s}) = \sum_{h=1}^{L} W_h^2 \overline{Y}_h^2 \left(U_{20h} + \alpha_{1h}^2 U_{02h} + 2\alpha_{1h} U_{11h}\right) \tag{11}$$

$$MSE(t_{p2}^{s}) = \sum_{h=1}^{L} W_h^2 \overline{Y}_h^2 \left(U_{20h} + \alpha_{2h}^2 U_{02h} + 2\alpha_{2h} U_{11h}\right) \tag{12}$$

The minimum mean square errors at the optimum values are

$$MinMSE(t_{p1}^{s}) = \sum_{h=1}^{L} W_h^2 \overline{Y}_h^2 \left(U_{20h} - \frac{U_{11h}^2}{U_{02h}}\right) \tag{13}$$

$$MinMSE(t_{p2}^{s}) = \sum_{h=1}^{L} W_h^2 \overline{Y}_h^2 \left(U_{20h} - \frac{U_{11h}^2}{U_{02h}}\right) \tag{14}$$

Outline of the derivations are given in Appendix

5. Some other proposed estimators

We propose modified estimators of population mean by $\overline{Y}$ under SRSS as

$$t_{p3}^{s} = \sum_{h=1}^{L} W_h[(1+w_{1h})\overline{y}_{h[rss]} + w_{2h}(\overline{X}_h - \overline{x}_{h[rss]})]\frac{\overline{X}_h}{\overline{x}_{h[rss]}} \tag{15}$$

$$t_{p4}^{s} = \sum_{h=1}^{L} W_h[(1+w_{3h})\overline{y}_{h[rss]} + w_{4h}(\overline{X}_h - \overline{x}_{h[rss]})]\exp\left(\frac{\overline{X}_h - \overline{x}_{h[rss]}}{\overline{X}_h + \overline{x}_{h[rss]}}\right) \tag{16}$$

$$t_{p5}^{s} = \sum_{h=1}^{L} W_h\left[w_{5h}\overline{y}_{h[rss]} + w_{6h}\exp\left(\frac{\overline{X}_h - \overline{x}_{h[rss]}}{\overline{X}_h + \overline{x}_{h[rss]}}\right)\left(1 + log\frac{\overline{x}_{h[rss]}}{\overline{X}_h}\right)\right] \tag{17}$$

$$t_{p6}^{s} = \sum_{h=1}^{L} W_h\left[w_{7h}\overline{y}_{h[rss]} + w_{8h}\left(\frac{\overline{X}_h}{\overline{x}_{h[rss]}}\right)\exp\left(\frac{\overline{X}_h - \overline{x}_{h[rss]}}{\overline{X}_h + \overline{x}_{h[rss]}}\right)\right] \tag{18}$$

The biases of the proposed estimators are

$$bias(t_{p3}^{s}) = \sum_{h=1}^{L} W_h[\overline{Y}_h w_{1h} + \overline{Y}_h(U_{02h} + w_{1h}U_{02h} + w_{2h}\delta U_{02h} - U_{11} - w_{1h}U_{11h})] \tag{19}$$

$$bias(t_{p4}^{s}) = \sum_{h=1}^{L} W_h\left[\overline{Y}_h w_{3h} + \overline{Y}_h\left(\frac{3}{8}U_{02h} + \frac{3}{8}w_{3h}U_{02h} + \frac{1}{2}w_{4h}\delta_h U_{02h} - \frac{1}{2}U_{11h} - \frac{1}{2}w_{3h}U_{11h}\right)\right] \tag{20}$$

$$Bias(t_{p5}^{s}) = \sum_{h=1}^{L} W_h\left[(w_{5h}-1)\overline{Y}_h + w_{6h}\left(1 - \frac{5}{8}U_{02h}\right)\right] \tag{21}$$

$$Bias(t_{p6}^{s}) = \sum_{h=1}^{L} W_h\left[(w_{7h}-1)\overline{Y}_h + w_{8h}\left(1 + \frac{15}{8}U_{02h}\right)\right] \tag{22}$$

The mean square errors of the proposed estimators are

$$MSE(t_{p3}^{s}) = \sum_{h=1}^{L} W_h^2\overline{Y}_h^2(A_{1h} + w_{1h}^2B_{1h} + w_{2h}^2C_{1h} + 2w_{1h}D_{1h} - 2w_{2h}E_{1h} - 2w_{1h}w_{2h}F_{1h}) \tag{23}$$

$$MSE(t_{p4}^{s}) = \sum_{h=1}^{L} W_h^2\overline{Y}_h^2(A_{2h} + w_{3h}^2B_{2h} + w_{4h}^2C_{2h} + 2w_{3h}D_{2h} - 2w_{4h}E_{2h} - 2w_{3h}w_{4h}F_{2h}) \tag{24}$$

The minimum mean square errors at the optimum values are

$$MinMSE(t_{p3}^{s}) = \sum_{h=1}^{L} W_h^2\overline{Y}_h^2\left(A_{1h} + \frac{C_{1h}D_{1h}^2 + B_{1h}E_{1h}^2 - 2D_{1h}E_{1h}F_{1h}}{F_{1h}^2 - B_{1h}C_{1h}}\right) \tag{25}$$

$$MinMSE(t_{p4}^{s}) = \sum_{h=1}^{L} W_h^2\overline{Y}_h^2\left(A_{2h} + \frac{C_{2h}D_{2h}^2 + B_{2h}E_{2h}^2 - 2D_{2h}E_{2h}F_{2h}}{F_{2h}^2 - B_{2h}C_{2h}}\right) \tag{26}$$

Outline of the derivations are given in Appendix

5.1. Case 1: Sum of weights is unity ($w_5 + w_6 = 1$ and $w_7 + w_8 = 1$)

The mean square errors of the proposed estimators are

$$MSE(t_{p5}^s) = \sum_{h=1}^{L} W_h^2 \overline{Y}_h^2 (U_{20h} + w_{6h}^2 U_{02h} - 2w_{6h} V_{11h}) \tag{27}$$

$$MSE(t_{p6}^c) = \sum_{h=1}^{L} W_h^2 \overline{Y}_h^2 (U_{20h} + w_{8h}^2 U_{02h} - 2w_{8h} U_{11h}) \tag{28}$$

The minimum mean square errors at the optimum values are

$$MinMSE(t_{p5}^s) = \sum_{h=1}^{L} W_h^2 \overline{Y}_h^2 \left(U_{20h} - \frac{U_{11h}^2}{U_{02h}} \right) \tag{29}$$

$$MinMSE(t_{p6}^s) = \sum_{h=1}^{L} W_h^2 \overline{Y}_h^2 \left(U_{20h} - \frac{U_{11h}^2}{U_{02h}} \right) \tag{30}$$

Outline of the derivations are given in Appendix

5.2. Case 2: Sum of weights is flexible ($w_5 + w_6 \neq 1$ and $w_7 + w_8 \neq 1$)

The mean square errors of the proposed estimators are

$$MSE(t_{p5}^s) = \sum_{h=1}^{L} W_h^2 [C_{3h} + w_{5h}^2 A_{3h} + w_{6h}^2 B_{3h} - 2w_{5h} C_{3h} - 2w_{6h} D_{3h} + 2w_{5h} w_{6h} E_{3h}] \tag{31}$$

$$MSE(t_{p6}^s) = \sum_{h=1}^{L} W_h^2 [C_{4h} + w_{7h}^2 A_{4h} + w_{8h}^2 B_{4h} - 2w_{7h} C_{4h} - 2w_{8h} D_{4h} + 2w_{7h} w_{8h} E_{4h}] \tag{32}$$

The minimum mean square errors at the optimum values are

$$MinMSE(t_{p5}^s) = \sum_{h=1}^{L} W_h^2 \left[C_{3h} + \frac{B_{3h} C_{3h}^2 + A_{3h} D_{3h}^2 - 2C_{3h} D_{3h} E_{3h}}{E_{3h}^2 - A_{3h} B_{3h}} \right] \tag{33}$$

$$MinMSE(t_{p6}^s) = \sum_{h=1}^{L} W_h^2 \left[C_{4h} + \frac{B_{4h} C_{4h}^2 + A_{4h} D_{4h}^2 - 2C_{4h} D_{4h} E_{4h}}{E_{4h}^2 - A_{4h} B_{4h}} \right] \tag{34}$$

Outline of the derivations are given in Appendix

6. Empirical study

In this section, we compare the performance of the proposed estimators with the other estimators considered in this paper. For comparison, we have taken a stratified population with 3 strata of sizes 20, 30, 17 respectively from the Singh (2003) (page no. 1119 (Appendix)). Where y is production (study variable) in metric tons and x is area (auxiliary variable) in hectares. For the above population, the parameters are given as below: For total population, N=67, $\overline{Y}$=72247.6, $\overline{X}$=26438

Table 1

Stratum 1	Stratum 2	Stratum 3
N_1=20	N_2=30	N_3=17
n_1=12	n_2=18	n_3=9
W_1=0.29851	W_2=0.44776	W_3=0.25373
$\overline{X}_1$=6801.25	$\overline{X}_2$=11025.3	$\overline{X}_3$=82464.1
$\overline{Y}_1$=17511.7	$\overline{Y}_2$=18937.4	$\overline{Y}_3$=377960.5
S^2_{x1}=175539558	S^2_{x2}=595679198.4	S^2_{x3}=20255478994
S^2_{y1}=1366895911	S^2_{y2}=2421559069	S^2_{y3}=687956456787
S_{y1x1}=489224338	S_{y2x2}=1174423304	S_{y3x3}=46735680920
C_{x1}=1.94804	C_{x2}=2.21368	C_{x3}=1.72586
$D^2_{yh1[i]}$=0.322701311	$D^2_{yh2[i]}$= 0.284750439	$D^2_{yh3[i]}$= 0.352112122
$D^2_{xh1[i]}$=0.277106302	$D^2_{xh2[i]}$= 0.191404888	$D^2_{xh3[i]}$= 0.201142044
$D_{yxh1[i]}$=0.298636371	$D_{yxh2[i]}$= 0.227030958	$D_{yxh3[i]}$= 0.01248969
R_1=2.57477	R_2=1.71763	R_3=4.58333

From this population we took ranked set samples of sizes k_1=4 , k_2=6 and k_3=3 from the stratum 1^{st} , 2^{nd} and 3^{rd} respectively. Further each ranked set sample from each stratum were repeated with number of cycles r=3. Hence sample size of stratifird ranked set sample is equivalent to $n_h = k_h r$.

Table 2: The MSE and PRE of the estimators

Estimators	MSE	Bias	PRE
t^s	1759632517	0.0000	100.0000
t^s_r	1204001473	17677.2090	146.1400
t^s_{lr}	11702271788	0.0000	150.3600
t^s_{p1}	11702271788	-2020.0767	150.3600
t^s_{p2}	11702271788	321.8933	150.3600
t^s_{p3}	811711525	-18442.3400	216.7800
t^s_{p4}	545563651	-27933.6290	281.5500
t^s_{p5}	425689034	11761.6920	413.3600
t^s_{p6}	315596791	-8835.3558	557.5500

The formula for Percent Relative Efficiency (PRE) is
PRE (estimators) = $\frac{MSE(t^s)}{MSE(estimator)} \times 100$

From Table 2, it is observed that

- The estimators t^s_{p1} and t^s_{p2} are almost equally efficient estimators as separate linear regression estimators under SRSS as these estimators show the MSE almost equal to the MSE of the combined linear regression estimator (t^s_{lr}). These two estimators t^s_{p1} and t^s_{p2} are more efficient estimators than that the other competitive estimators.

- t^s_{p3} ,t^s_{p4} , t^s_{p5} and t^s_{p6} are more efficient than other estimators used in this paper. It is observed that t^s_{p3} ,t^s_{p4} , t^s_{p5} and t^s_{p6} are more efficient than convention, ratio estimator and linear regression estimator under SRSS.

- From Table 2, we can conclude that the proposed estimators perform better than existing estimators as our proposed estimators have greater PRE.

7. Simulation study

To generalize the results of the numerical study, we have conducted simulation study over two hypothetically generated normal populations. The simulation procedure is explained in the following points:

- We generated bivariate random observations of size N=600 units from a bivariate normal distribution with parameters μ_y =20, σ_y =15, and μ_x=15, σ_x=10 and possibly chosen values of ρ_{yx}= 0.6, 0.7, 0.8, 0.9.
- Similarly, generate the population-2 with the parameters μ_y =120, σ_y =25, and μ_x=100, σ_x=20 .
- The population generated above is divided into 3 equal strata and a stratified ranked set sample of size 12 units with number of cycles 4 and set size 3 is drawn from each stratum.
- Compute the required statistics.
- Iterate the above steps 10,000 times to calculate the MSE and PRE of various combined estimators using the following expression.

$$MSE(T) = \frac{1}{10000} \sum_{i=1}^{10000} (T_i - \overline{Y})^2 \tag{35}$$

$$PRE = \frac{Var(t^c)}{MSE(T)} \times 100 \tag{36}$$

The MSE and PRE of the separate estimators are calculated using (35) and (36) and the results are reported for various values of correlation coefficients in Table 3.

Table 3 also shows that our proposed estimators perform better than the existing estimators. The MSE of the estimators decreases when the correlation and sample size increases for the population 1 and 2.

8. Conclusions

In this article we have proposed estimators for the population mean in stratified Ranked set sampling using the information of auxiliary variable. The expressions for Bias and MSE of the suggested estimators have been derived up to the first order of approximation. Empirical approach and simulation study for comparing the efficiency of the proposed estimators with other estimators have been used. The results have been shown the Tables 2 and 3. The Tables show that the proposed estimators turn out to be more efficient as compared to the other estimators for both populations. The proposed estimators are found to be rather improved in terms of lesser MSE and greater PRE as compared to the existing

Table 3: The MSE and PRE of the estimators

ρ_{yx}	Estimators	Population1			Population2		
		MSE	Bias	PRE	MSE	Bias	PRE
0.9	t^s	0.007284	0.000000	100.000000	0.066100	0.000000	100.000000
	t^s_r	0.006384	-0.000207	114.095495	0.043496	0.002922	151.969606
	t^s_{lr}	0.004961	0.000000	146.827826	0.042869	0.000000	154.189069
	t^s_{p1}	0.004945	-0.000239	147.285914	0.042656	-0.001315	154.960149
	t^s_{p2}	0.004943	-0.000279	147.352653	0.042767	-0.002903	154.558070
	t^s_{p3}	0.003387	-0.000311	215.050896	0.034651	-0.001423	190.761305
	t^s_{p4}	0.003339	0.000190	218.094473	0.024678	0.001060	267.848401
	t^s_{p5}	0.003245	-0.000178	224.414416	0.020015	-0.001060	330.255808
	t^s_{p6}	0.003090	-0.000426	235.710841	0.019309	-0.002208	342.320400
0.8	t^s	0.006984	0.000000	100.000000	0.090219	0.000000	100.000000
	t^s_r	0.006809	0.000672	102.560246	0.070926	0.001426	127.201411
	t^s_{lr}	0.004670	0.000000	149.540410	0.059455	0.000000	151.741804
	t^s_{p1}	0.004634	-0.000181	150.699420	0.059354	-0.004426	152.000269
	t^s_{p2}	0.004687	-0.000102	148.992987	0.059456	0.004611	151.739251
	t^s_{p3}	0.004174	-0.000314	167.327264	0.043999	-0.008287	205.046909
	t^s_{p4}	0.003587	0.000145	194.697844	0.030173	0.001628	299.004742
	t^s_{p5}	0.002991	-0.000314	233.454669	0.026272	-0.001067	343.397088
	t^s_{p6}	0.002657	-0.000537	262.768911	0.024469	-0.002098	368.707344
0.7	t^s	0.009693	0.000000	100.000000	0.074859	0.000000	100.000000
	t^s_r	0.006455	0.000136	150.158004	0.061038	0.002001	122.642758
	t^s_{lr}	0.005928	0.000000	163.507324	0.058294	0.000000	128.416597
	t^s_{p1}	0.005934	-0.000124	163.345039	0.058345	0.002856	128.303578
	t^s_{p2}	0.005976	0.000362	162.191669	0.058568	0.001072	127.814362
	t^s_{p3}	0.005562	-0.000330	174.267715	0.040771	-0.007203	183.608732
	t^s_{p4}	0.005109	0.000123	189.710173	0.038156	0.001765	196.189650
	t^s_{p5}	0.003267	-0.000429	296.645336	0.036670	-0.001165	204.142055
	t^s_{p6}	0.002752	-0.000625	352.113929	0.020043	-0.002084	373.482308
0.6	t^s	0.008782	0.000000	100.000000	0.091577	0.000000	100.000000
	t^s_r	0.008134	0.000191	107.954650	0.086847	0.002652	105.447165
	t^s_{lr}	0.007273	0.000000	120.745030	0.078933	0.000000	116.018800
	t^s_{p1}	0.007145	-0.000832	122.898593	0.078557	0.001270	116.573953
	t^s_{p2}	0.007108	0.000521	123.537736	0.078345	0.001939	116.889576
	t^s_{p3}	0.005597	-0.000356	156.880083	0.030133	-0.004797	303.904598
	t^s_{p4}	0.003695	0.000945	237.630218	0.023471	0.002048	390.162195
	t^s_{p5}	0.002241	-0.000552	391.795672	0.016980	-0.009593	539.310974
	t^s_{p6}	0.001342	-0.000719	654.023741	0.013250	-0.001763	691.136804

estimators in both real and simulated data sets. It is also observed from the simulation that the MSE of the proposed estimators decreases as the values of the correlation coefficient increase whereas the PRE of the suggested estimators increases as the values of the correlation coefficients increase. Based on our empirical study and simulation study, we can conclude that our proposed estimators can be preferred over the other estimators taken in this paper in several real situations.

Acknowledgements

We are very grateful to the anonymous referee for his valuable comments that improved the quality of the paper.

References

Al-Omari Ibrahim, A. and Bouza, C. (2014). Review of ranked set sampling: modifications and applications. *Investigaci´on Operacional*, **35**, 215–235.

Bhushan, S., Gupta, R., Singh, S., and Kumar, A. (2020). A new efficient log type class of estimators using auxiliary variable. *International Journal of Statistics and Systems*, **15**, 19–28.

Bouza, C., Singh, P., and Singh, R. (2018). Ranked set sampling and optional scrambling randomized response modeling. *Investigaci´on Operacional*, **39**, 100–107.

Chaudhary, M., Singh, R., Shukla, R., Kumar, M., and Smarandache, F. (2009). A family of estimators for estimating population mean in stratified sampling under non-response. *Pakistan Journal of Statistics and Operation Research*, **5**, 47–54.

Cochran, W. (1977). *Sampling Techniques* Wiley Eastern Limited. New Delhi.

Ganeslingam, S. and Ganesh, S. (2006). Ranked set sampling versus simple random sampling in the estimation of the mean and the ratio. *Journal of Statistics and Management Systems*, **9**, 459–472.

Haq, A. and Shabbir, J. (2013). Improved family of ratio estimators in simple and stratified random sampling. *Communications in Statistics-Theory and Methods*, **42**, 782–799.

Kadilar, C. and Cingi, H. (2003). Ratio estimators in stratified random sampling. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, **45**, 218–225.

Koyuncu, N. and Kadilar, C. (2009). Family of estimators of population mean using two auxiliary variables in stratified random sampling. *Communications in Statistics-Theory and Methods*, **38**, 2398–2417.

McIntyre, G. (1952). A method for unbiased selective sampling, using ranked sets.*Australian Journal of Agricultural Research*, **3**, 385–390.

Muneer, S., Shabbir, J., and Khalil, A. (2017). Estimation of finite population mean in simple random sampling and stratified random sampling using two auxiliary variables. *Communications in Statistics-Theory and Methods*, **46**, 2181–2192.

Samawi, H. and Muttlak, H. (1996). Estimation of ratio using rank set sampling. *Biometrical Journal*, **38**, 753–764.

Samawi, H. and Siam, M. (2003). Ratio estimation using stratified ranked set sample. *Metron*, **61**, 75–90.

Shabbir, J. and Gupta, S. (2006). A new estimator of population mean in stratified sampling. *Communications in Statistics-Theory and Methods*, **35**, 1201–1209.

Shabbir, J. and Gupta, S. (2007). On estimating the finite population mean with known population proportion of an auxiliary variable. *Pakistan Journal of Statistics*, **23**, 1-9.

Singh, H.P., Tailor, R., and Singh, S. (2014). General procedure for estimating the population mean using ranked set sampling. *Journal of Statistical Computation and Simulation*, **84**, 931–945.

Singh, R. and Kumar, M. (2012). Improved estimators of population mean using two auxiliary variables in stratified random sampling. *Pakistan Journal of Statistics and Operation Research*, **8**, 65–72.

Singh, S. (2003). *Advanced Sampling Theory With Applications: How Michael ""Selected""Amy*. Springer Science & Business Media.

Takahasi, K. and Wakimoto, K. (1968). On unbiased estimates of the population mean based on the sample stratified by means of ordering. *Annals of the Institute of Statistical Mathematics*, **20**, 1–31.

APPENDIX

This section consider the proof of the Theorems of Section 4 & 5.
To derive the MSE of the proposed estimators, the following notations will be used throughout the paper.

$$\overline{y}_{h[srss]} = \overline{Y_h}(1+\epsilon_{0h})$$
$$\overline{x}_{h[srss]} = \overline{X_h}(1+\epsilon_{1h})$$

such that $\mathrm{E}(\epsilon_{0h})= \mathrm{E}(\epsilon_{1h})= 0$

$$E(\epsilon_{0h}^2) = (\eta_h C_{yh}^2 - D_{yh[i]}^2) = U_{20h}$$
$$E(\epsilon_{1h}^2) = (\eta_h C_{xh}^2 - D_{xh[i]}^2) = U_{02h}$$
$$E(\epsilon_{0h}\epsilon_{1h}) = (\eta_h C_{xyh} - D_{xyh[i]}) = U_{11h}$$

where $\eta_h = \frac{1}{k_h r}$, $C_{xh} = \frac{S_{xh}}{\overline{X}}$,$C_{yh} = \frac{S_{yh}}{\overline{Y}}$, $D_{xh[i]}^2 = \frac{1}{k_h^2 r \overline{X}^2} \sum_{i=1}^{k_h} (\overline{X}_{h(i)} - \overline{X}_h)^2$,

$D_{yh[i]}^2 = \frac{1}{k_h^2 r \overline{Y}^2} \sum_{i=1}^{k_h} (\overline{Y}_{h(i)} - \overline{Y}_h)^2$ and $D_{xyh[i]} = \frac{1}{k_h^2 r \overline{Y}\overline{X}} \sum_{i=1}^{k_h} (\overline{Y}_{h(i)} - \overline{Y}_h)(\overline{X}_{h(i)} - \overline{X}_h)$

where $\overline{Y}_{h[i]}$ and $\overline{X}_{h(i)}$ are the means of the i^{th} is ranked set and are given by

$$\overline{Y}_{h[i]} = \frac{1}{r}\sum_{j=1}^{r} Y_{h[i]j}, \overline{X}_{h(i)} = \frac{1}{r}\sum_{j=1}^{r} X_{h(i)j}$$

Now, consider the estimator

$$t_{p1}^s = \sum_{h=1}^{L} W_h \overline{y}_{h[rss]} \exp\left(\alpha_{1h}\left(\frac{\overline{x}_{h[rss]}}{\overline{X_h}} - 1\right)\right)$$

Using the above notations we have

$$t_{p1}^s = \sum_{h=1}^{L} W_h \overline{Y}_h (1+\epsilon_{0h}) \exp\left(\alpha_{1h}\left(\frac{\overline{X}_h(1+\epsilon_1)}{\overline{X}_h} - 1\right)\right) \tag{37}$$

The bias of the estimator t_{p1}^s is given by

$$Bias(t_{p1}^s) = \sum_{h=1}^{L} W_h \overline{Y}_h \left(\frac{\alpha_{1h}^2}{2} U_{02h} + \alpha_{1h} U_{11h}\right) \tag{38}$$

The MSE of the estimator t_{p1}^s is given by

$$MSE(t_{p1}^s) = \sum_{h=1}^{L} W_h^2 \overline{Y}_h^2 \left(U_{20h} + \alpha_{1h}^2 U_{02h} + 2\alpha_{1h} U_{11h}\right) \tag{39}$$

To find out the minimum MSE for t_{p1}^s, we partially differentiate equation (39) *w.r.t.* α_{1h} and equating to zero we get

$$\alpha_{1h}^* = -\frac{U_{11h}}{U_{02h}} \tag{40}$$

Putting the optimum value of α_{1h} in the equation (39), we get a minimum MSE of t_{p1}^s as

$$MinMSE(t_{p1}^s) = \sum_{h=1}^{L} W_h^2 \overline{Y}_h^2 \left(U_{20h} - \frac{U_{11h}^2}{U_{02h}}\right) \tag{41}$$

Similarly, we can obtain the optimum values of constants and minimum MSEs of other proposed estimators which are given as

$$t_{p2}^s = \sum_{h=1}^{L} W_h \overline{Y}_h (1+\epsilon_{0h}) \exp\left(\alpha_{2h} \log \frac{\overline{X}_h (1+\epsilon_{1h})}{\overline{X}_h}\right) \tag{42}$$

The bias of the estimator t_{p2}^s is given by

$$Bias(t_{p2}^s) = \sum_{h=1}^{L} W_h \overline{Y}_h \left(\frac{(\alpha_{2h}^2 - \alpha_{2h})}{2} U_{02h} + \alpha_{2h} U_{11h}\right) \tag{43}$$

The MSE of the estimator t_{p2}^s is given by

$$MSE(t_{p2}^s) = \sum_{h=1}^{L} W_h^2 \overline{Y}_h^2 \left(U_{20h} + \alpha_{2h}^2 U_{02h} + 2\alpha_{2h} U_{11h}\right) \tag{44}$$

To find out the minimum MSE for t_{p2}^s, we partially differentiate equation (44) *w.r.t.* α_{2h} and equating to zero we get

$$\alpha_{2h}^* = -\frac{U_{11h}}{U_{02h}} \tag{45}$$

Putting the optimum value of α_{2h} in the equation (44), we get a minimum MSE of t_{p2}^s as

$$MinMSE(t_{p2}^s) = \sum_{h=1}^{L} W_h^2 \overline{Y}_h^2 \left(U_{20h} - \frac{U_{11h}^2}{U_{02h}}\right) \tag{46}$$

$$t_{p3}^s = \sum_{h=1}^{L} W_h [(1+w_{1h})\overline{Y}_h (1+\epsilon_{0h}) + w_{2h}\epsilon_{1h}](1-\epsilon_{1h}+\epsilon_{1h}^2) \tag{47}$$

$$t_{p3}^s - \overline{Y} = \sum_{h=1}^{L} W_h \overline{Y}_h [(\epsilon_{0h} + w_{1h} + \epsilon_{0h}w_{1h} - \epsilon_{1h} - \epsilon_{1h}w_{1h} - \epsilon_{0h}\epsilon_{1h} - \epsilon_{0h}\epsilon_{1h}w_{1h} + \epsilon_{1h}^2 + w_{1h}\epsilon_{1h}^2) - w_{2h}\delta(\epsilon_{1h} - \epsilon_{1h}^2)] \tag{48}$$

The bias of the estimator t_{p3}^s is given by

$$Bias(t_{p3}^s) = \sum_{h=1}^{L} W_h[\overline{Y}_h w_{1h} + \overline{Y}_h(U_{02h} + w_{1h}U_{02h} + w_{2h}\delta U_{02h} - U_{11} - w_{1h}U_{11h})] \tag{49}$$

The MSE of the estimator t_{p3}^s is given by

$$MSE(t_{p3}^s) = \sum_{h=1}^{L} W_h^2\overline{Y}_h^2[U_{20h} + U_{02h} - 2U_{11h} + w_{1h}^2(1 + U_{20h} + 3U_{02h} - 4U_{11h}) + w_{2h}^2\delta_h^2 U_{02h} + 2w_{1h}(U_{20h} + 2U_{02h} - 3U_{11h}) - 2w_{2h}\delta(U_{11h} - U_{02h}) - 2w_{1h}w_{2h}\delta(U_{11h} - 2U_{02h})] \tag{50}$$

$$MSE(t_{p3}^s) = \sum_{h=1}^{L} W_h^2\overline{Y}_h^2(A_{1h} + w_{1h}^2 B_{1h} + w_{2h}^2 C_{1h} + 2w_{1h}D_{1h} - 2w_{2h}E_{1h} - 2w_{1h}w_{2h}F_{1h}) \tag{51}$$

where

$$A_{1h} = U_{20h} + U_{02h} - 2U_{11h}$$
$$B_{1h} = 1 + U_{20h} + 3U_{02h} - 4U_{11h}$$
$$C_{1h} = \delta^2 U_{02h}, \delta_h = \frac{\overline{X}_h}{\overline{Y}_h}$$
$$D_{1h} = U_{20h} + 2U_{02h} - 3U_{11h}$$
$$E_{1h} = \delta_h(U_{02h} - U_{11h})$$
$$F_{1h} = \delta_h(U_{02h} - 2U_{11h})$$

To find out the minimum MSE for t_{p3}^s, we partially differentiate equation (51) *w.r.t.* w_{1h} and w_{2h} and equating to zero we get

$$w_{1h}^* = \frac{C_{1h}D_{1h} - E_{1h}F_{1h}}{F_{1h}^2 - B_{1h}C_{1h}} \tag{52}$$

$$w_{2h}^* = \frac{D_{1h}F_{1h} - B_{1h}C_{1h}}{F_{1h}^2 - B_{1h}C_{1h}} \tag{53}$$

Putting the optimum values of w_{1h} and w_{2h} in the equation (51), we get a minimum MSE of t_{p3}^s as

$$MinMSE(t_{p3}^s) = \sum_{h=1}^{L} W_h^2\overline{Y}_h^2\left(A_{1h} + \frac{C_{1h}D_{1h}^2 + B_{1h}E_{1h}^2 - 2D_{1h}E_{1h}F_{1h}}{F_{1h}^2 - B_{1h}C_{1h}}\right) \tag{54}$$

$$t_{p4}^s = \sum_{h=1}^{L} W_h[(1 + w_{3h})\overline{Y}_h(1 + \epsilon_{0h}) + w_{4h}\epsilon_{1h}]\left(1 - \frac{3}{2}\epsilon_{1h} + \frac{15}{8}\epsilon_{1h}^2\right) \tag{55}$$

$$t_{p4}^s - \overline{Y} = \sum_{h=1}^{L} W_h\overline{Y}_h[(\epsilon_{0h} + W_{3h} + \epsilon_{0h}w_{3h} - \frac{1}{2}\epsilon_{1h} - \frac{1}{2}\epsilon_{1h}w_{3h} - \frac{1}{2}\epsilon_{0h}\epsilon_{1h} - \frac{1}{2}\epsilon_{0h}\epsilon_{1h}w_{3h} + \frac{3}{8}\epsilon_{1h}^2 + \frac{3}{8}w_{3h}\epsilon_1^2) - w_{4h}\delta_h(\epsilon_{1h} - \epsilon_{1h}^2)] \tag{56}$$

The bias of the estimator t_{p4}^s is given by

$$Bias(t_{p4}^s) = \sum_{h=1}^{L} W_h \left[\overline{Y}_h w_{3h} + \overline{Y}_h \left(\frac{3}{8} U_{02h} + \frac{3}{8} w_{3h} U_{02h} + \frac{1}{2} w_{4h} \delta_h U_{02h} - \frac{1}{2} U_{11h} - \frac{1}{2} w_{3h} U_{11h} \right) \right] \tag{57}$$

The MSE of the estimator t_{p4}^s is given by

$$MSE(t_{p4}^s) = \sum_{h=1}^{L} W_h^2 \overline{Y}_h^2 (U_{20h} + \frac{1}{4} U_{02h} - U_{11} + w_{3h}^2 (1 + U_{20h} + U_{02h} - 2U_{11h}) + w_{4h}^2 \delta_h^2 U_{02h} + 2w_{3h}$$
$$(U_{20h} + \frac{5}{4} U_{02h} - \frac{3}{2} U_{11h}) - 2w_{4h} \delta_h (U_{11h} - \frac{1}{2} U_{02h}) - 2w_{3h} w_{4h} \delta (U_{11h} - U_{02h})) \tag{58}$$

$$MSE(t_{p4}^s) = \sum_{h=1}^{L} W_h^2 \overline{Y}_h^2 (A_{2h} + w_{3h}^2 B_{2h} + w_{4h}^2 C_{2h} + 2w_{3h} D_{2h} - 2w_{4h} E_{2h} - 2w_{3h} w_{4h} F_{2h}) \tag{59}$$

where

$$A_{2h} = U_{20h} + \frac{1}{4} U_{02h} - U_{11h}$$
$$B_{2h} = 1 + U_{20h} + U_{02h} - 2U_{11h}$$
$$C_{2h} = \delta^2 U_{02h}, \delta_h = \frac{\overline{X}_h}{\overline{Y}_h}$$
$$D_{2h} = U_{20h} + \frac{5}{4} U_{02h} - \frac{3}{2} U_{11h}$$
$$E_{2h} = \delta \left(U_{02h} - \frac{1}{2} U_{11h} \right)$$
$$F_{2h} = \delta (U_{02h} - U_{11h})$$

To find out the minimum MSE for t_{p4}^s, we partially differentiate equation (59) *w.r.t.* w_{3h} and w_{4h} and equating to zero we get

$$w_{3h}^* = \frac{C_{2h} D_{2h} - E_{2h} F_{2h}}{F_{2h}^2 - B_{2h} C_{2h}} \tag{60}$$

$$w_{4h}^* = \frac{D_{2h} F_{2h} - B_{2h} C_{2h}}{F_{2h}^2 - B_{2h} C_{2h}} \tag{61}$$

Putting the optimum values of w_{3h} and w_{4h} in the equation (59), we get a minimum MSE of t_{p4}^s as

$$MinMSE(t_{p4}^s) = \sum_{h=1}^{L} W_h^2 \overline{Y}_h^2 \left(A_{2h} + \frac{C_{2h} D_{2h}^2 + B_{2h} E_{2h}^2 - 2D_{2h} E_{2h} F_{2h}}{F_{2h}^2 - B_{2h} C_{2h}} \right) \tag{62}$$

$$t_{p5}^s = \sum_{h=1}^{L} W_h \left[w_{5h} \overline{Y}_h (1 + \epsilon_{0h}) + w_{6h} \exp \left(\frac{-\epsilon_{1h}}{2 + \epsilon_{1h}} \right) (1 + log(1 + \epsilon_{1h})) \right] \tag{63}$$

$$t_{p5}^s - \overline{Y} = \sum_{h=1}^{L} W_h \left[(w_{5h} - 1) \overline{Y}_h + w_{5h} \overline{Y}_h \epsilon_{0h} + w_{6h} \left(1 + \frac{\epsilon_{1h}}{2} - \frac{5}{8} \epsilon_{1h}^2 \right) \right] \tag{64}$$

$$Bias(t_{p5}^{s}) = \sum_{h=1}^{L} W_h \left[(w_{5h} - 1)\overline{Y}_h + w_{6h} \left(1 - \frac{5}{8} U_{02h} \right) \right] \tag{65}$$

$$t_{p6}^{s} = \sum_{h=1}^{L} W_h \left[w_{7h}\overline{Y}_h(1 + \epsilon_{0h}) + w_{8h} \exp\left(\frac{-\epsilon_{1h}}{2 + \epsilon_{1h}} \right) (1 + \epsilon_{1h})^{-1} \right] \tag{66}$$

$$t_{p6}^{s} - \overline{Y} = \sum_{h=1}^{L} W_h \left[(w_{7h} - 1)\overline{Y}_h + w_{7h}\overline{Y}_h\epsilon_{0h} + w_{8h} \left(1 - \frac{3}{2}\epsilon_{1h} - \frac{15}{8}\epsilon_{1h}^{2} \right) \right] \tag{67}$$

$$Bias(t_{p6}^{s}) = \sum_{h=1}^{L} W_h \left[(w_{7h} - 1)\overline{Y}_h + w_{8h} \left(1 + \frac{15}{8} U_{02h} \right) \right] \tag{68}$$

CASE 1: SUM OF WEIGHTS IS UNITY ($w_5 + w_6 = 1$ and $w_7 + w_8 = 1$)

The MSE of the estimator t_{p5}^{s} is given by

$$MSE(t_{p5}^{s}) = \sum_{h=1}^{L} W_h^2 \overline{Y}_h^2 (U_{20h} + w_{6h}^2 U_{02h} - 2w_{6h} V_{11h}) \tag{69}$$

To find out the minimum MSE for t_{p5}^{s} , we partially differentiate equation (69) *w.r.t.* w_{6h}, and equating to zero we get

$$w_{6h}^{*} = \frac{V_{11h}}{V_{02h}} \tag{70}$$

Putting the optimum value of w_{6h} in the equation (69), we get a minimum MSE of t_{p5}^{s} as

$$MinMSE(t_{p5}^{s}) = \sum_{h=1}^{L} W_h^2 \overline{Y}_h^2 \left(U_{20h} - \frac{U_{11h}^2}{U_{02h}} \right) \tag{71}$$

The MSE of the estimator t_{p6}^{s} is given by

$$MSE(t_{p6}^{c}) = \sum_{h=1}^{L} W_h^2 \overline{Y}_h^2 (U_{20h} + w_{8h}^2 U_{02h} - 2w_{8h} U_{11h}) \tag{72}$$

To find out the minimum MSE for t_{p6}^{s} , we partially differentiate equation (72) *w.r.t.* w_{8h}, and equating to zero we get

$$w_{8h}^{*} = \frac{U_{11h}}{U_{02h}} \tag{73}$$

Putting the optimum value of w_{8h} in the equation (72), we get a minimum MSE of t_{p6}^{s} as

$$MinMSE(t_{p6}^{s}) = \sum_{h=1}^{L} W_h^2 \overline{Y}_h^2 \left(U_{20h} - \frac{U_{11h}^2}{U_{02h}} \right) \tag{74}$$

CASE 2: SUM OF WEIGHTS IS FLEXIBLE ($w_5 + w_6 \neq 1$ and $w_7 + w_8 \neq 1$)

$$t_{p5}^{s} - \overline{Y} = \sum_{h=1}^{L} W_h \left[(w_{5h} - 1)\overline{Y}_h + w_{5h}\overline{Y}_h\epsilon_{0h} + w_{6h} \left(1 + \frac{\epsilon_{1h}}{2} - \frac{5}{8}\epsilon_{1h}^{2} \right) \right] \tag{75}$$

Squaring on both sides we get

$$(t_{p5}^{s}-\overline{Y})^2=\sum_{h=1}^{L}W_h^2\Big[\overline{Y}_h^2+\overline{Y}_h^2w_{5h}^2(1+\epsilon_{0h}^2)+w_{6h}^2(1-\epsilon_{1h}^2)-2w_{5h}\overline{Y}_h^2-2w_{6h}\overline{Y}_h\Big(1-\frac{5}{8}\epsilon_{1h}^2\Big)+2w_{5h}w_{6h}\Big(1-\frac{5}{8}\epsilon_{1h}^2+\frac{1}{2}\epsilon_{0h}\epsilon_{1h}\Big)\Big] \quad (76)$$

Taking expectations on both sides we get

$$MSE(t_{p5}^{s})=\sum_{h=1}^{L}W_h^2\Big[\overline{Y}_h^2+\overline{Y}_h^2w_{5h}^2(1+U_{20h})+w_{6h}^2(1-U_{02h})-2w_{5h}\overline{Y}_h^2-2w_{6h}\overline{Y}_h\Big(1-\frac{5}{8}U_{02h}\Big)+2w_{5h}w_{6h}\Big(1-\frac{5}{8}U_{02h}+\frac{1}{2}U_{11h}\Big)\Big] \quad (77)$$

$$MSE(t_{p5}^{s})=\sum_{h=1}^{L}W_h^2[C_{3h}+w_{5h}^2A_{3h}+w_{6h}^2B_{3h}-2w_{5h}C_{3h}-2w_{6h}D_{3h}+2w_{5h}w_{6h}E_{3h}] \quad (78)$$

where

$$A_{3h}=\overline{Y}_h^2(1+U_{20h})$$
$$B_{3h}=1-U_{02h}$$
$$C_{3h}=\overline{Y}_h^2$$
$$D_{3h}=\overline{Y}_h\left(1-\frac{5}{8}U_{02h}\right)$$
$$E_{3h}=\overline{Y}_h\left(1-\frac{5}{8}U_{02h}+\frac{1}{2}U_{11h}\right)$$

To find out the minimum MSE for the estimator t_{p5}^{s} , we partially differentiate equation (78) *w.r.t.* w_{5h} and w_{6h} and equating to zero we get

$$w_{5h}^{*}=\frac{B_{3h}C_{3h}-D_{3h}E_{3h}}{A_{3h}B_{3h}-E_{3h}^2} \quad (79)$$

$$w_{6h}^{*}=\frac{A_{3h}D_{3h}-C_{3h}E_{3h}}{A_{3h}B_{3h}-E_{3h}^2} \quad (80)$$

Putting the optimum values of w_{5h} and w_{6h} in the equation (78), we get a minimum MSE of t_{p5}^{s} as

$$MinMSE(t_{p5}^{s})=\sum_{h=1}^{L}W_h^2\left[C_{3h}+\frac{B_{3h}C_{3h}^2+A_{3h}D_{3h}^2-2C_{3h}D_{3h}E_{3h}}{E_{3h}^2-A_{3h}B_{3h}}\right] \quad (81)$$

$$t_{p6}^{c}-\overline{Y}=\sum_{h=1}^{L}W_h\left[(w_{7h}-1)\overline{Y}_h+w_{7h}\overline{Y}_h\epsilon_{0h}+w_{8h}\left(1-\frac{3}{2}\epsilon_{1h}+\frac{15}{8}\epsilon_{1h}^2\right)\right] \quad (82)$$

Squaring on both sides we get

$$(t_{p6}^s - \overline{Y})^2 = \sum_{h=1}^{L} W_h^2 \Big[\overline{Y}_h^2 + \overline{Y}_h^2 w_{7h}^2(1+\epsilon_{0h}^2) + w_{8h}^2(1+6\epsilon_{1h}^2) - 2w_{7h}\overline{Y}_h^2 - 2w_{8h}\overline{Y}_h\Big(1 - \frac{15}{8}\epsilon_{1h}^2\Big) + 2w_{7h}w_{8h}\Big(1 + \frac{15}{8}\epsilon_{1h}^2 - \frac{3}{2}\epsilon_{0h}\epsilon_{1h}\Big)\Big] \quad (83)$$

Taking expectations on both sides we get

$$MSE(t_{p6}^s) = \sum_{h=1}^{L} W_h^2 \Big[\overline{Y}_h^2 + \overline{Y}_h^2 w_{7h}^2(1+U_{20h}) + w_{8h}^2(1+6U_{02h}) - 2w_{7h}\overline{Y}_h^2 - 2w_{8h}\overline{Y}_h\Big(1 + \frac{15}{8}U_{02h}\Big) + 2w_{7h}w_{8h}\Big(1 + \frac{15}{8}U_{02h} - \frac{3}{2}U_{11h}\Big)\Big] \quad (84)$$

$$MSE(t_{p6}^s) = \sum_{h=1}^{L} W_h^2 [C_{4h} + w_{7h}^2 A_{4h} + w_{8h}^2 B_{4h} - 2w_{7h}C_{4h} - 2w_{8h}D_{4h} + 2w_{7h}w_{8h}E_{4h}] \quad (85)$$

where

$$A_{4h} = \overline{Y}_h^2(1 + U_{20h})$$
$$B_{4h} = 1 + 6U_{02h}$$
$$C_{4h} = \overline{Y}_h^2$$
$$D_{4h} = \overline{Y}_h\left(1 + \frac{15}{8}U_{02h}\right)$$
$$E_{4h} = \overline{Y}_h\left(1 + \frac{15}{8}U_{02h} - \frac{3}{2}U_{11h}\right)$$

To find out the minimum MSE for the estimator t_{p6}^s, we partially differentiate equation (85) *w.r.t.* w_{7h} and w_{8h} and equating to zero we get

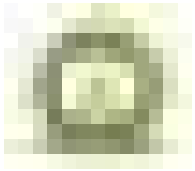
$$w_{7h}^* = \frac{B_{4h}C_{4h} - D_{4h}E_{4h}}{A_{4h}B_{4h} - E_{4h}^2} \quad (86)$$

$$w_{8h}^* = \frac{A_{4h}D_{4h} - C_{4h}E_{4h}}{A_{4h}B_{4h} - E_{4h}^2} \quad (87)$$

Putting the optimum values of w_{7h} and w_{8h} in the equation (85), we get a minimum MSE of t_{p6}^s as

$$MinMSE(t_{p6}^s) = \sum_{h=1}^{L} W_h^2 \left[C_{4h} + \frac{B_{4h}C_{4h}^2 + A_{4h}D_{4h}^2 - 2C_{4h}D_{4h}E_{4h}}{E_{4h}^2 - A_{4h}B_{4h}}\right] \quad (88)$$

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 22, No. 1, 2024 (New Series), pp 39–56
http://www.ssca.org.in/journal



E-Bayesian and Hierarchical Bayesian Estimation for Inverse Rayleigh Distribution Based on Left Censoring Scheme

R. B. Athirakrishnan and E. I. Abdul Sathar
Department of Statistics, University of Kerala,
Thiruvananthapuram - 695 581, India.

Received: 01 September 2022; Revised: 15 April 2023; Accepted: 24 May 2023

Abstract

This study is concerned with estimating the scale parameter and the reversed hazard rate of the Inverse Rayleigh distribution based on left censoring, one of the most noticeable distributions in lifetime studies. Even though different estimation methods are employed, each method suffers from its problems such as complexity of calculations, high risk, *etc.* Results derived under squared error, entropy, and precautionary loss functions. E-Bayesian and H-Bayesian estimations are obtained based on different priors of the hyper parameters to investigate the influence on these estimations. We investigated the asymptotic behaviors of E-Bayesian estimates and relations among them. Finally, a comparison among the Bayes, H-Bayes, and E-Bayes estimates in different sample sizes made using real and the simulated data. Numerical study shows that the newly presented method is more efficient than previous methods and is also easy to operate.

Key words: Inverse Rayleigh distribution; Left censoring; Bayesian estimation; E-Bayesian estimation; H-Bayesian estimation.

AMS Subject Classifications: 62F15; 62N05

1. Introduction

Several authors used Inverse Rayleigh (IR) distribution to model applications in the area of reliability. Voda (1972) used this distribution to model the lifetimes of several experimental units. Several works related to inference using complete samples based on parameters of inverse Rayleigh (IR) distribution are available in the literature. El-Helbawy and Abd-El-Monem (2005) developed Bayes estimators for the parameters of the IR distribution using different loss functions. For more works related to inference using IR distribution, one can refer to Soliman *et al.* (2010), Dey (2012), Feroze and Aslam (2012) and Shawky and Badr (2012). In the context of reliability and survival analysis, censoring is unavoidable, and there

Corresponding Author: Dr. E. I. Abdul Sathar
Email: sathare@gmail.com

are different censoring schemes available. One of the practical censoring schemes is the left censoring, and it occurs when we cannot identify the exact time the event occurred.

Considering the advantage of using the E-Bayesian estimation method recently, many papers are published in the literature using this approach. Han (2009) proposed the E-Bayesian estimate of the failure rate of exponential distribution using type-1 censoring. E-Bayesian estimates of Burr type XII distribution parameters using type-2 censoring had proposed by Jaheen and Okasha (2011). Okasha and Wang (2016) derived E-Bayesian estimators of the geometric distribution parameters when samples are available only in the form of records. Kızılaslan (2017) discusses the E-Bayesian estimation of the proportional hazard rate model. E-bayesian and hierarchical bayesian estimates of the power function distribution parameters had proposed by Abdul-Sathar and Athirakrishnan (2019). This paper aims to propose E-Bayesian and H-Bayesian estimates of the inverse Rayleigh distribution parameters when left-censored data are available. We additionally provide estimates of the reversed hazard rate using three different loss functions. The asymptotic performance of the proposed estimators for different priors is also studied.

The organization of the rest of the works is as follows. We discuss Bayesian estimation of the scale parameter and reversed hazard rate of the IR distribution using left-censored data in Section 2. In Section 3, we discuss the H-Bayesian estimation of the scale parameter and the reversed hazard rate. E-Bayesian estimators of the scale parameter and reversed hazard rate are discussed in Section 4. The properties exhibited by all these estimators discusses in Section 5. The estimator's performance using simulated and real data sets discuss respectively in Sections 6 and 7. Finally, concluding remarks about the proposed study are given in Section 8.

2. Bayesian estimation

In this section, we derive the Bayesian estimators of the parameter λ of IR distribution using left-censored data under the squared error loss function (SELF), the entropy loss function (ELF), and the precautionary loss function (PLF). The pdf, cdf, and reversed hazard rate of the one-parameter IR distribution are respectively given by

$$f(x;\lambda) = \frac{2\lambda}{x^3} e^{\frac{-\lambda}{x^2}}, \quad x > 0, \quad \lambda > 0, \tag{1}$$

$$F(x;\lambda) = e^{\frac{-\lambda}{x^2}}, \quad x > 0, \quad \lambda > 0, \tag{2}$$

and

$$h(\bar{t}) = \frac{2\lambda}{t^3}, t > 0. \tag{3}$$

Let $\underline{X} = X_{(r+1)}, ..., X_{(n)}$ be the last $(n-r)$ order statistics using a random sample of size n from IR distribution. Likelihood function in this context is given as

$$L(X_{(r+1)}, ..., X_{(n)}|\lambda) \propto \lambda^{n-r} e^{-\lambda\tau_{(ir)}}, \tag{4}$$

where $\tau_{(ir)} = rx_{(r+1)}^{-2} + \sum_{i=r+1}^{n} x_{(i)}^{-2}$. The prior for the parameter λ assumes Gamma distribution with density function

$$\pi(\lambda|a,b) = \frac{b^a}{\Gamma(a)}\lambda^{a-1}e^{-b\lambda}, \quad \lambda > 0, \quad a,b > 0,$$

where a and b are the hyper parameters. Here we only consider the case of $a = 1$, then the density function $\pi(\lambda|a,b)$ reduces to

$$\pi(\lambda|b) = be^{-b\lambda}, \quad b > 0. \tag{5}$$

Hence the posterior distribution using (4) and (5) simplifies to

$$f(r|\lambda) = \frac{(\tau_{(ir)}+b)^{n-r+1}}{\Gamma(n-r+1)}\lambda^{(n-r)}e^{-\lambda(\tau_{(ir)}+b)}, \quad \lambda > 0 \tag{6}$$

Now we derive the Bayes estimators of λ and reversed hazard rate of left censored IR distribution under three different loss functions.
Using SELF, the Bayes estimators of λ and reversed hazard rate simplify to

$$\hat{\lambda}_{B1} = E(\lambda|\underline{x}) = \frac{n-r+1}{\tau_{(ir)}+b}, \tag{7}$$

$$\hat{h(t)}_{B1} = E\left(\frac{2\lambda}{t^3}\middle|\underline{x}\right) = \frac{2(n-r+1)}{t^3(\tau_{(ir)}+b)}. \tag{8}$$

The Bayes estimators of λ and reversed hazard rate using ELF simplifies to

$$\hat{\lambda}_{B2} = \left[E\left(\frac{1}{\lambda}\middle|\underline{x}\right)\right]^{-1} = \frac{n-r}{\tau_{(ir)}+b}. \tag{9}$$

$$\hat{h(t)}_{B2} = \left[E\left(\left(\frac{2\lambda}{t^3}\right)^{-1}\middle|\underline{x}\right)\right]^{-1} = \frac{2(n-r)}{t^3(\tau_{(ir)}+b)}. \tag{10}$$

The Bayes estimators of λ and reversed hazard rate using PLF simplifies to

$$\hat{\lambda}_{B3} = \sqrt{E(\lambda^2|\underline{x})} = \sqrt{\frac{(n-r+1)(n-r+2)}{(\tau_{(ir)}+b)^2}}. \tag{11}$$

$$\hat{h(t)}_{B3} = \sqrt{E\left(\left(\frac{2\lambda}{t^3}\right)^2\middle|\underline{x}\right)} = \frac{2}{t^3}\sqrt{\frac{(n-r+1)(n-r+2)}{(\tau_{(ir)}+b)^2}}. \tag{12}$$

3. Hierarchical Bayesian estimation

Lindley and Smith (1972) first introduced the idea of hierarchical prior distribution. For the parameter λ, the hierarchical prior density function is defined as

$$\pi(\lambda) = \int_0^c \pi(\lambda|b)\pi(b)db.$$

Hierarchical Bayesian (H-Bayesian) estimation of λ is obtained based on three different distributions of the hyper parameter b. The influence of the different prior distributions on the H-Bayesian estimation of λ is studied by using these distributions. The following distributions of b may be used

$$\pi_1(b) = \frac{2(c-b)}{c^2}, \quad 0 < b < c, \tag{13}$$

$$\pi_2(b) = \frac{1}{c}, \quad 0 < b < c, \tag{14}$$

$$\pi_3(b) = \frac{2b}{c^2}, \quad 0 < b < c, \tag{15}$$

3.1. Hierarchical Bayesian estimation of λ

For $\pi_1(b)$, the hierarchical prior density function simplifies to

$$\pi_4(\lambda) = \frac{2}{c^2}\int_0^c b(c-b)e^{-b\lambda}db, \lambda > 0. \tag{16}$$

Using Bayesian theorem, the hierarchical posterior density for λ can be defined as

$$\begin{aligned} H_1(\lambda|\underline{x}) &= \frac{\pi_4(\lambda)L(r|\lambda)}{\int_0^\infty \pi_4(\lambda)L(r|\lambda)d\lambda} \\ &= \frac{\int_0^c b(c-b)\lambda^{n-r}e^{-\lambda(\tau_{(ir)}+2b)}(\tau_{(ir)}+b)^{n-r+1}db}{\int_0^c b(c-b)\frac{\Gamma(n-r+1)}{(\tau_{(ir)}+2b)^{n-r+1}}(\tau_{(ir)}+b)^{n-r+1}db}. \end{aligned} \tag{17}$$

The H-Bayesian estimators of λ under SELF is given as

$$\hat{\lambda}_{HS1} = \frac{\int_0^c b(c-b)\frac{\Gamma(n-r+2)}{(\tau_{(ir)}+2b)^{(n-r+2)}}(\tau_{(ir)}+b)^{(n-r+1)}db}{\int_0^c b(c-b)\frac{\Gamma(n-r+1)}{(\tau_{(ir)}+2b)^{(n-r+1)}}(\tau_{(ir)}+b)^{(n-r+1)}db}, \tag{18}$$

Similarly, the H-Bayesian estimators of λ under ELF and PLF are given respectively as

$$\hat{\lambda}_{HE1} = \frac{\int_0^c b(c-b)\frac{\Gamma(n-r+1)}{(\tau_{(ir)}+2b)^{(n-r+1)}}(\tau_{(ir)}+b)^{(n-r+1)}db}{\int_0^c b(c-b)\frac{\Gamma(n-r)}{(\tau_{(ir)}+2b)^{(n-r)}}(\tau_{(ir)}+b)^{(n-r+1)}db}, \tag{19}$$

and

$$\hat{\lambda}_{HP1} = \sqrt{\frac{\int_0^c b(c-b)\frac{\Gamma(n-r+3)}{(\tau_{(ir)}+2b)^{(n-r+3)}}(\tau_{(ir)}+b)^{(n-r+1)}db}{\int_0^c b(c-b)\frac{\Gamma(n-r+1)}{(\tau_{(ir)}+2b)^{(n-r+1)}}(\tau_{(ir)}+b)^{(n-r+1)}db}}, \tag{20}$$

For $\pi_2(b)$, the hierarchical prior density function simplifies to

$$\pi_5(\lambda) = \frac{1}{c}\int_0^c be^{-b\lambda}db, \lambda > 0. \tag{21}$$

Using Bayesian theorem, the hierarchical posterior density for λ can be defined as

$$\begin{aligned} H_2(\lambda|\underline{x}) &= \frac{\pi_5(\lambda)L(r|\lambda)}{\int_0^\infty \pi_5(\lambda)L(r|\lambda)d\lambda} \\ &= \frac{\int_0^c b\lambda^{n-r}e^{-\lambda(\tau_{(ir)}+2b)}(\tau_{(ir)}+b)^{n-r+1}db}{\int_0^c b\frac{\Gamma(n-r+1)}{(\tau_{(ir)}+2b)^{n-r+1}}(\tau_{(ir)}+b)^{n-r+1}db}. \end{aligned} \tag{22}$$

The H-Bayesian estimator of λ under SELF is given as

$$\hat{\lambda}_{HS2} = \frac{\int_0^c b\frac{\Gamma(n-r+2)}{(\tau_{(ir)}+2b)^{(n-r+2)}}(\tau_{(ir)}+b)^{(n-r+1)}db}{\int_0^c b\frac{\Gamma(n-r+1)}{(\tau_{(ir)}+2b)^{(n-r+1)}}(\tau_{(ir)}+b)^{(n-r+1)}db}, \tag{23}$$

Similarly, the H-Bayesian estimators of λ under ELF and PLF are given respectively as

$$\hat{\lambda}_{HE2} = \frac{\int_0^c b\frac{\Gamma(n-r+1)}{(\tau_{(ir)}+2b)^{(n-r+1)}}(\tau_{(ir)}+b)^{(n-r+1)}db}{\int_0^c b\frac{\Gamma(n-r)}{(\tau_{(ir)}+2b)^{(n-r)}}(\tau_{(ir)}+b)^{(n-r+1)}db}, \tag{24}$$

and

$$\hat{\lambda}_{HP2} = \sqrt{\frac{\int_0^c b\frac{\Gamma(n-r+3)}{(\tau_{(ir)}+2b)^{(n-r+3)}}(\tau_{(ir)}+b)^{(n-r+1)}db}{\int_0^c b\frac{\Gamma(n-r+1)}{(\tau_{(ir)}+2b)^{(n-r+1)}}(\tau_{(ir)}+b)^{(n-r+1)}db}}, \tag{25}$$

For $\pi_3(b)$, the hierarchical prior density function simplifies to

$$\pi_6(\lambda) = \frac{2}{c^2}\int_0^c b^2e^{-b\lambda}db, \lambda > 0. \tag{26}$$

Using Bayesian theorem, the hierarchical posterior density for λ can be defined as

$$\begin{aligned} H_3(\lambda|\underline{x}) &= \frac{\pi_6(\lambda)L(r|\lambda)}{\int_0^\infty \pi_6(\lambda)L(r|\lambda)d\lambda} \\ &= \frac{\int_0^c b^2\lambda^{n-r}e^{-\lambda(\tau_{(ir)}+2b)}(\tau_{(ir)}+b)^{n-r+1}db}{\int_0^c b^2\frac{\Gamma(n-r+1)}{(\tau_{(ir)}+2b)^{n-r+1}}(\tau_{(ir)}+b)^{n-r+1}db}. \end{aligned} \tag{27}$$

The H-Bayesian estimator of λ under SELF is given as

$$\hat{\lambda}_{HS3} = E(\lambda|\underline{x}) = \frac{\int_0^c b^2\frac{\Gamma(n-r+2)}{(\tau_{(ir)}+2b)^{(n-r+2)}}(\tau_{(ir)}+b)^{(n-r+1)}db}{\int_0^c b^2\frac{\Gamma(n-r+1)}{(\tau_{(ir)}+2b)^{(n-r+1)}}(\tau_{(ir)}+b)^{(n-r+1)}db}. \tag{28}$$

Similarly, the H-Bayesian estimators of λ under ELF and PLF are given respectively as

$$\hat{\lambda}_{HE3} = \left[E\left(\lambda^{-1}|\underline{x}\right)\right]^{-1} = \frac{\int_0^c b^2\frac{\Gamma(n-r+1)}{(\tau_{(ir)}+2b)^{(n-r+1)}}(\tau_{(ir)}+b)^{(n-r+1)}db}{\int_0^c b^2\frac{\Gamma(n-r)}{(\tau_{(ir)}+2b)^{(n-r)}}(\tau_{(ir)}+b)^{(n-r+1)}db}, \tag{29}$$

and

$$\hat{\lambda}_{HP3} = \sqrt{E(\lambda^2|\underline{x})} = \sqrt{\frac{\int_0^c b^2\frac{\Gamma(n-r+3)}{(\tau_{(ir)}+2b)^{(n-r+3)}}(\tau_{(ir)}+b)^{(n-r+1)}db}{\int_0^c b^2\frac{\Gamma(n-r+1)}{(\tau_{(ir)}+2b)^{(n-r+1)}}(\tau_{(ir)}+b)^{(n-r+1)}db}}. \tag{30}$$

3.2. Hierarchical Bayesian estimation of reversed hazard rate

Based on SELF, ELF and PLF, the H-Bayesian estimators of the reversed hazard rate is computed for the three different distributions of the hyperparameter b given by (13), (14) and (15). For $\pi_1(b)$, the H-Bayesian estimator of the reversed hazard rate is obtained from (17). Under SELF, the H-Bayesian estimator of reversed hazard rate is given as

$$\hat{h(t)}_{HS1} = \frac{\frac{2}{t^3}\int_0^c b(c-b)\frac{\Gamma(n-r+2)}{(\tau_{(ir)}+2b)^{(n-r+2)}}(\tau_{(ir)}+b)^{(n-r+1)}db}{\int_0^c b(c-b)\frac{\Gamma(n-r+1)}{(\tau_{(ir)}+2b)^{(n-r+1)}}(\tau_{(ir)}+b)^{(n-r+1)}db}. \tag{31}$$

The H-Bayesian estimators of reversed hazard rate under ELF and PLF are given as

$$\hat{h(t)}_{HE1} = \frac{\frac{2}{t^3}\int_0^c b(c-b)\frac{\Gamma(n-r+1)}{(\tau_{(ir)}+2b)^{(n-r+1)}}(\tau_{(ir)}+b)^{(n-r+1)}db}{\int_0^c b(c-b)\frac{\Gamma(n-r)}{(\tau_{(ir)}+2b)^{(n-r)}}(\tau_{(ir)}+b)^{(n-r+1)}db}. \tag{32}$$

and

$$\hat{h(t)}_{HP1} = \frac{2}{t^3}\sqrt{\frac{\int_0^c b(c-b)\frac{\Gamma(n-r+3)}{(\tau_{(ir)}+2b)^{(n-r+3)}}(\tau_{(ir)}+b)^{(n-r+1)}db}{\int_0^c b(c-b)\frac{\Gamma(n-r+1)}{(\tau_{(ir)}+2b)^{(n-r+1)}}(\tau_{(ir)}+b)^{(n-r+1)}db}}. \tag{33}$$

For $\pi_2(b)$, the H-Bayesian estimator of the reversed hazard rate is obtained from (22). Under SELF, the H-Bayesian estimator of reversed hazard rate is given as

$$\hat{h(t)}_{HS2} = \frac{\frac{2}{t^3}\int_0^c b\frac{\Gamma(n-r+2)}{(\tau_{(ir)}+2b)^{(n-r+2)}}(\tau_{(ir)}+b)^{(n-r+1)}db}{\int_0^c b\frac{\Gamma(n-r+1)}{(\tau_{(ir)}+2b)^{(n-r+1)}}(\tau_{(ir)}+b)^{(n-r+1)}db}. \tag{34}$$

The H-Bayesian estimators of reversed hazard rate under ELF and PLF are given as

$$\hat{h(t)}_{HE2} = \frac{\frac{2}{t^3}\int_0^c b\frac{\Gamma(n-r+1)}{(\tau_{(ir)}+2b)^{(n-r+1)}}(\tau_{(ir)}+b)^{(n-r+1)}db}{\int_0^c b\frac{\Gamma(n-r)}{(\tau_{(ir)}+2b)^{(n-r)}}(\tau_{(ir)}+b)^{(n-r+1)}db}. \tag{35}$$

and

$$\hat{h(t)}_{HP2} = \frac{2}{t^3}\sqrt{\frac{\int_0^c b\frac{\Gamma(n-r+3)}{(\tau_{(ir)}+2b)^{(n-r+3)}}(\tau_{(ir)}+b)^{(n-r+1)}db}{\int_0^c b\frac{\Gamma(n-r+1)}{(\tau_{(ir)}+2b)^{(n-r+1)}}(\tau_{(ir)}+b)^{(n-r+1)}db}}. \tag{36}$$

For $\pi_3(b)$, the H-Bayesian estimator of the reversed hazard rate is obtained from (27). Under SELF, the H-Bayesian estimator of reversed hazard rate is given as

$$\hat{h(t)}_{HS3} = \frac{\frac{2}{t^3}\int_0^c b^2\frac{\Gamma(n-r+2)}{(\tau_{(ir)}+2b)^{(n-r+2)}}(\tau_{(ir)}+b)^{(n-r+1)}db}{\int_0^c b^2\frac{\Gamma(n-r+1)}{(\tau_{(ir)}+2b)^{(n-r+1)}}(\tau_{(ir)}+b)^{(n-r+1)}db}. \tag{37}$$

The H-Bayesian estimators of reversed hazard rate under ELF and PLF are given as

$$\hat{h(t)}_{HE3} = \frac{\frac{2}{t^3}\int_0^c b^2\frac{\Gamma(n-r+1)}{(\tau_{(ir)}+2b)^{(n-r+1)}}(\tau_{(ir)}+b)^{(n-r+1)}db}{\int_0^c b^2\frac{\Gamma(n-r)}{(\tau_{(ir)}+2b)^{(n-r)}}(\tau_{(ir)}+b)^{(n-r+1)}db}. \tag{38}$$

and

$$\hat{h(t)}_{HP3} = \frac{2}{t^3}\sqrt{\frac{\int_0^c b^2 \frac{\Gamma(n-r+3)}{(\tau_{(ir)}+2b)^{(n-r+3)}}(\tau_{(ir)}+b)^{(n-r+1)}db}{\int_0^c b^2 \frac{\Gamma(n-r+1)}{(\tau_{(ir)}+2b)^{(n-r+1)}}(\tau_{(ir)}+b)^{(n-r+1)}db}}. \tag{39}$$

4. E-Bayesian estimation

According to Han (1997) the E-Bayesian estimate of λ is defined as

$$\hat{\lambda}_E = \int_b \hat{\lambda}_B(b)\pi(b)db. \tag{40}$$

where $\hat{\lambda}_B(b)$ is the Bayesian estimator of λ with prior density $\pi(b)$. From (40), we can see that E-Bayesian estimation is the expectation of Bayesian estimator of the parameters for the hyper parameter. E-Bayesian estimation based on three different prior distributions of the hyper parameter (13), (14) and (15) are used to investigate the influence of different prior distributions on the E-Bayesian estimation of λ and reversed hazard rate.

4.1. E-Bayesian estimation for λ

Based on SELF, ELF and PLF, the E-Bayesian estimators of λ is computed for the three different distributions of the hyperparameter b given by (13), (14) and (15). For $\pi_1(b)$, the E-Bayesian estimate of λ under SELF is obtained from (7) and (13) as

$$\hat{\lambda}_{ES1} = \int_0^c \hat{\lambda}_{B1}(b)\pi_1(b)db = \frac{2(n-r+1)}{c^2}\left\{(\tau_{(ir)}+c)\ln\left(\frac{\tau_{(ir)}+c}{\tau_{(ir)}}\right) - c\right\}. \tag{41}$$

Similarly, the E-Bayesian estimates of λ under ELF and PLF are computed from (9), (11) and (13) and are given respectively, by

$$\hat{\lambda}_{EE1} = \frac{2(n-r)}{c^2}\left\{(\tau_{(ir)}+c)\ln\left(\frac{\tau_{(ir)}+c}{\tau_{(ir)}}\right) - c\right\}, \tag{42}$$

and

$$\hat{\lambda}_{EP1} = 2\sqrt{\frac{(n-r+1)(n-r+2)}{c}}\left\{(\tau_{(ir)}+c)\ln\left(\frac{\tau_{(ir)}+c}{\tau_{(ir)}}\right) - c\right\}. \tag{43}$$

For $\pi_2(b)$, the E-Bayesian estimate of λ under SELF is obtained from (7) and (14) as

$$\hat{\lambda}_{ES2} = \frac{n-r+1}{c}\ln\left(\frac{\tau_{(ir)}+c}{\tau_{(ir)}}\right), \tag{44}$$

Similarly, the E-Bayesian estimates of λ under ELF and PLF are computed from (9), (11) and (14) and are given respectively, by

$$\hat{\lambda}_{EE2} = \frac{n-r}{c}\ln\left(\frac{\tau_{(ir)}+c}{\tau_{(ir)}}\right), \tag{45}$$

and

$$\hat{\lambda}_{EP2} = \sqrt{\frac{(n-r+1)(n-r+2)}{c}} \ln\left(\frac{\tau_{(ir)}+c}{\tau_{(ir)}}\right), \tag{46}$$

For $\pi_3(b)$, the E-Bayesian estimate of λ under SELF is obtained from (7) and (15) as

$$\hat{\lambda}_{ES3} = \frac{2(n-r+1)}{c^2}\left\{c - \tau_{(ir)} \ln\left(\frac{\tau_{(ir)}+c}{\tau_{(ir)}}\right)\right\}, \tag{47}$$

Similarly, the E-Bayesian estimates of λ under ELF and PLF are computed from (9), (11) and (14) and are given respectively, by

$$\hat{\lambda}_{EE3} = \frac{2(n-r)}{c^2}\left\{c - \tau_{(ir)} \ln\left(\frac{\tau_{(ir)}+c}{\tau_{(ir)}}\right)\right\}, \tag{48}$$

and

$$\hat{\lambda}_{EP3} = 2\sqrt{\frac{(n-r+1)(n-r+2)}{c}}\left\{c - \tau_{(ir)} \ln\left(\frac{\tau_{(ir)}+c}{\tau_{(ir)}}\right)\right\}. \tag{49}$$

4.2. E-Bayesian estimation for reversed hazard rate

Based on SELF, ELF and PLF, the E-Bayesian estimators of reversed hazard rate is computed for the three different distributions of the hyperparameter b given by (13), (14) and (15). For $\pi_1(b)$, the E-Bayesian estimate of reversed hazard rate under SELF is obtained from (8) and (13) as

$$\hat{h(t)}_{ES1} = \frac{4(n-r+1)}{c^2t^3}\left\{(\tau_{(ir)}+c)\ln\left(\frac{\tau_{(ir)}+c}{\tau_{(ir)}}\right) - c\right\}. \tag{50}$$

Similarly, the E-Bayesian estimates of reversed hazrd rate under ELF and PLF are computed from (10), (12) and (13) and are given respectively, by

$$\hat{h(t)}_{EE1} = \frac{4(n-r)}{c^2t^3}\left\{(\tau_{(ir)}+c)\ln\left(\frac{\tau_{(ir)}+c}{\tau_{(ir)}}\right) - c\right\}, \tag{51}$$

and

$$\hat{h(t)}_{EP1} = \frac{4}{t^3}\sqrt{\frac{(n-r+1)(n-r+2)}{c}}\left\{(\tau_{(ir)}+c)\ln\left(\frac{\tau_{(ir)}+c}{\tau_{(ir)}}\right) - c\right\}. \tag{52}$$

For $\pi_2(b)$, the E-Bayesian estimate of reversed hazard rate under SELF is obtained from (8) and (14) as

$$\hat{h(t)}_{ES2} = \frac{2(n-r+1)}{ct^3}\ln\left(\frac{\tau_{(ir)}+c}{\tau_{(ir)}}\right). \tag{53}$$

Similarly, the E-Bayesian estimates of reversed hazard rate under ELF and PLF are computed from (10), (12) and (14) and are given respectively, by

$$\hat{h(t)}_{EE2} = \frac{2(n-r)}{ct^3}\ln\left(\frac{\tau_{(ir)}+c}{\tau_{(ir)}}\right), \tag{54}$$

and

$$\hat{h(t)}_{EP2} = \frac{2}{t^3}\sqrt{\frac{(n-r+1)(n-r+2)}{c}}\ln\left(\frac{\tau_{(ir)}+c}{\tau_{(ir)}}\right). \tag{55}$$

For $\pi_3(b)$, the E-Bayesian estimate of reversed hazard rate under SELF is obtained from (8) and (15) as

$$\hat{h(t)}_{ES3} = \frac{4(n-r+1)}{c^2t^3}\left\{c - \tau_{(ir)}\ln\left(\frac{\tau_{(ir)}+c}{\tau_{(ir)}}\right)\right\}. \tag{56}$$

Similarly, the E-Bayesian estimates of reversed hazard rate under ELF and PLF are computed from (10), (12) and (15) and are given respectively, by

$$\hat{h(t)}_{EE3} = \frac{4(n-r)}{c^2t^3}\left\{c - \tau_{(ir)}\ln\left(\frac{\tau_{(ir)}+c}{\tau_{(ir)}}\right)\right\}, \tag{57}$$

and

$$\hat{h(t)}_{EP3} = \frac{4}{t^3}\sqrt{\frac{(n-r+1)(n-r+2)}{c}}\left\{c - \tau_{(ir)}\ln\left(\frac{\tau_{(ir)}+c}{\tau_{(ir)}}\right)\right\}. \tag{58}$$

5. Properties

In this section, we discussed the important properties of E-Bayesian estimators including the relation of this estimators with the hierarchical Bayesian estimators. In the following theorem, we gives the relationship of E-Bayes estimators of λ under different loss functions.

Theorem 1: The relationship of E-Bayes estimators of λ using respectively the SELF, ELF and PLF are given as

i) $\hat{\lambda}_{EEi} < \hat{\lambda}_{ESi} < \hat{\lambda}_{EPi}, i = 1, 2, 3$

ii) $\lim_{\tau_{(ir)}\to\infty}\hat{\lambda}_{ESi} = \lim_{\tau_{(ir)}\to\infty}\hat{\lambda}_{EEi} = \lim_{\tau_{(ir)}\to\infty}\hat{\lambda}_{EPi} = 0.$

Proof:

i) The relationship $\hat{\lambda}_{EE1} < \hat{\lambda}_{ES1} < \hat{\lambda}_{EP1}$ is a particular case of $\hat{\lambda}_{EEi} < \hat{\lambda}_{ESi} < \hat{\lambda}_{EPi}$ and it is same as

$$n-r < n-r+1 < \sqrt{(n-r+1)(n-r+2)}. \tag{59}$$

We use the concept of mathematical induction for proving the relation. For n=1, we have $1-r < (2-r) < \sqrt{(2-r)(3-r)}$. Hence the result is true for n=1. Squaring the above equation, we get

$$(n-r)^2 < (n-r+1)^2 < (n-r+1)(n-r+2). \tag{60}$$

Now assume that the result hold for n=k. That is

$$(k-r)^2 < (k-r+1)^2 < (k-r+1)(k-r+2). \tag{61}$$

Now, we prove the result for *n=k+1*, so we have

$$\begin{aligned}((k+1)+r+1)((k+1)+r+2) &= (k-r+2)(k-r+3)\\ &= (k-r+1)(k-r+2)\\ &\quad +2(k-r+2). \end{aligned} \tag{62}$$

Using (33), we get

$$\begin{aligned}(k-r)^2+2(k-r+2) &< (k-r+1)^2+2(k-r+2)\\ &< (k-r+1)(k-r+2)+2(k-r+2). \end{aligned} \tag{63}$$

we have

$$(k-r)^2+2(k-r+2) = ((k+1)-r)^2+3 > ((k+1)-r)^2. \tag{64}$$

Also, we have

$$(k-r+1)^2+2(k-r+2) = ((k+1)-r+1)^2+1 > ((k+1)-r+1)^2. \tag{65}$$

Using (33) to (36) ,we have

$$((k+1)-r)^2 < ((k+1)-r+1)^2 < ((k+1)-r+1)((k+1)-r+2). \tag{66}$$

Hence the result.

ii) From the derivation of $\hat{\lambda}_{ES1}$, we have

$$\hat{\lambda}_{ES1} = \frac{2(n-r+1)}{c^2}\int_0^c \frac{c-b}{\tau_{(ir)}+b}db.$$

Using the generalized mean value theorem, we can find atleast one number $b_1 \in (0,c)$ such that

$$\hat{\lambda}_{ES1} = \frac{2(n-r+1)}{c^2}\frac{1}{\tau_{(ir)}+b_1}\int_0^c (c-b)db.$$

Taking the limit as $\tau_{(ir)} \to \infty$

$$\lim_{\tau_{(ir)}\to\infty} \hat{\lambda}_{ES1} = 0. \tag{67}$$

Using the generalized mean value theorem, we can find atleast one number $b_2 \in (0,c)$ such that

$$\hat{\lambda}_{EE1} = \frac{2(n-r)}{c^2}\frac{1}{\tau_{(ir)}+b_2}\int_0^c (c-b)db.$$

Taking the limit as $\tau_{(ir)} \to \infty$

$$\lim_{\tau_{(ir)}\to\infty} \hat{\lambda}_{EE1} = 0. \tag{68}$$

Using the generalized mean value theorem, we can find atleast one number $b_3 \in (0,c)$ such that, we have

$$\hat{\lambda}_{EP1} = \frac{2\sqrt{(n-r+2)(n-r+1)}}{c^2(\tau_{(ir)}+b_3)}\int_0^c (c-b)db.$$

Taking the limit as $\tau_{(ir)} \to \infty$

$$\lim_{\tau_{(ir)}\to\infty} \hat{\lambda}_{EP1} = 0. \tag{69}$$

Using (38) to (40), we have the proof. From the above theorem, we can see that, E-Bayesian estimators for λ are different for different loss functions. It can also be noted that the estimators are asymptotically equal or close to each other when $\tau_{(ir)}$ is sufficiently large. The rest of the proof is same as the above. In the following theorem we provide the relationship of E-Bayes estimators of reversed hazard rate for different loss functions. The proof is similar to the above theorem and hence omitted. □

Theorem 2: The relationship of E-Bayes estimators of reversed hazard rate using respectively the SELF, ELF and PLF are given as

i) $\hat{h(t)}_{EE1} < \hat{h(t)}_{ES1} < \hat{h(t)}_{EP1}$

ii) $\lim_{\tau_{(ir)}\to\infty} \hat{h(t)}_{ES1} = \lim_{\tau_{(ir)}\to\infty} \hat{h(t)}_{EE1} = \lim_{\tau_{(ir)}\to\infty} \hat{h(t)}_{EP1} = 0.$

In the following theorem, we gives the relationship between E-Bayes and hierarchical Bayes estimators of λ under the same loss function.

Theorem 3: The relation between E-Bayes and hierarchical Bayes estimators of λ for SELF, ELF and PLF are respectively given as

i) $\lim_{\tau_{(ir)}\longrightarrow ß\infty} \hat{\lambda}_{ESi} = \lim_{\tau_{(ir)}\longrightarrow ß\infty} \hat{\lambda}_{HSi} = 0, i = 1, 2, 3.$

ii) $\lim_{\tau_{(ir)}\longrightarrow ß\infty} \hat{\lambda}_{EEi} = \lim_{\tau_{(ir)}\longrightarrow ß\infty} \hat{\lambda}_{HEi} = 0, i = 1, 2, 3.$

iii) $\lim_{\tau_{(ir)}\longrightarrow ß\infty} \hat{\lambda}_{EPi} = \lim_{\tau_{(ir)}\longrightarrow ß\infty} \hat{\lambda}_{HPi} = 0, i = 1, 2, 3.$

Proof:

i) Under SELF, from the above theorem, using (22), we get

$$\lim_{\tau_{(ir)}\to\infty} \hat{\lambda}_{ES1} = 0. \tag{70}$$

Using the result $\Gamma(n+r+2) = (n+r+1)\Gamma(n+r+1)$ and by using the generalized mean value theorem, we can find atleast one number $b_4 \in (0, c)$

$$\int_0^c b(c-b)(\tau_{(ir)}+b)^{(n-r+1)} \frac{(n-r+1)\Gamma(n-r+1)}{(\tau_{(ir)}+2b)(\tau_{(ir)}+2b)^{(n-r+1)}} db =$$
$$\frac{n-r+1}{(\tau_{(ir)}+2b_4)} \int_0^c b(c-b)(\tau_{(ir)}+b)^{(n-r+1)} \frac{\Gamma(n-r+1)}{(\tau_{(ir)}+2b)^{(n-r+1)}} db.$$

$$\begin{aligned} \therefore \hat{\lambda}_{HS1} &= \frac{\int_0^c b(c-b)(\tau_{(ir)}+b)^{(n-r+1)} \frac{\Gamma(n-r+2)}{(\tau_{(ir)}+2b)^{(n-r+2)}} db}{\int_0^c b(c-b)(\tau_{(ir)}+b)^{(n-r+1)} \frac{\Gamma(n-r+1)}{(\tau_{(ir)}+2b)^{(n-r+1)}} db} \\ &= \frac{n-r+1}{(\tau_{(ir)}+2b_4)} \end{aligned} \tag{71}$$

Taking limit as $\tau_{(ir)} \to \infty$

$$\lim_{\tau_{(ir)}\to\infty} \hat{\lambda}_{HS1} = 0. \tag{72}$$

Hence using (41) and (43), we have

$$\lim_{\tau_{(ir)}\to\infty} \hat{\lambda}_{ES1} = \lim_{\tau_{(ir)}\to\infty} \hat{\lambda}_{HS1} = 0. \tag{73}$$

ii) Under ELF, from the above theorem, using (22), we get

$$\lim_{\tau_{(ir)}\to\infty} \hat{\lambda}_{EE1} = 0. \tag{74}$$

Using the result $\Gamma(n-r+1) = (n-r)\Gamma(n-r)$ and by sing the generalized mean value theorem, we can find atleast one number $b_5 \in (0, c)$ such that

$$\int_0^c b(c-b)(\tau_{(ir)}+b)^{(n-r+1)} \frac{(n-r)\Gamma(n-r)}{(\tau_{(ir)}+2b)(\tau_{(ir)}+2b)^{(n-r)}} db =$$
$$\frac{(n-r)}{(\tau_{(ir)}+2b_5)} \int_0^c b(c-b)(\tau_{(ir)}+b)^{(n-r+1)} \frac{\Gamma(n-r)}{(\tau_{(ir)}+2b)^{(n-r)}} db. \tag{75}$$

Using (18) we have

$$\begin{aligned} \hat{\lambda}_{HE1} &= \frac{\int_0^c b(c-b)(\tau_{(ir)}+b)^{(n-r+1)} \frac{\Gamma(n-r+1)}{(\tau_{(ir)}+2b)^{(n-r+1)}} db}{\int_0^c b(c-b)(\tau_{(ir)}+b)^{(n-r+1)} \frac{\Gamma(n-r)}{(\tau_{(ir)}+2b)^{(n-r)}} db} \\ &= \frac{(n-r)}{(\tau_{(ir)}+2b_5)}. \end{aligned} \tag{76}$$

Taking limit as $\tau_{(ir)} \to \infty$

$$\lim_{\tau_{(ir)}\to\infty} \hat{\lambda}_{HE1} = 0. \tag{77}$$

Hence using (45) and (47), we have

$$\lim_{\tau_{(ir)}\to\infty} \hat{\lambda}_{EE1} = \lim_{\tau_{(ir)}\to\infty} \hat{\lambda}_{HE1} = 0. \tag{78}$$

iii) Under PLF, from the above theorem, using (22), we get

$$\lim_{\tau_{(ir)}\to\infty} \hat{\lambda}_{EP1} = 0. \tag{79}$$

Using the result $\Gamma(n+a+2) = (n+a+1)\Gamma(n+a+1)$ and by using the generalized mean value theorem, we can find atleast one number $b_6 \in (0, c)$ such that

$$\int_0^c b(c-b) \frac{\Gamma(n-r+3)}{(\tau_{(ir)}+2b)^{(n-r+3)}} (\tau_{(ir)}+b)^{(n-r+1)} db =$$
$$\int_0^c b(c-b) \frac{\Gamma(n-r+3)}{(\tau_{(ir)}+2b)^{(n-r+3)}} (\tau_{(ir)}+b)^{(n-r+1)} db. \tag{80}$$

Using (20) we have

$$\hat{\lambda}_{HP1} = \sqrt{\frac{\int_0^c b(c-b)\frac{\Gamma(n-r+3)}{(\tau_{(ir)}+2b)^{(n-r+3)}}(\tau_{(ir)}+b)^{(n-r+1)}db}{\int_0^c b(c-b)\frac{\Gamma(n-r+1)}{(\tau_{(ir)}+2b)^{(n-r+1)}}(\tau_{(ir)}+b)^{(n-r+1)}db}}$$
$$= \frac{\sqrt{(n-r+2)(n-r+1)}}{(\tau_{(ir)}+2b_6)}. \quad (81)$$

Taking limit as $\tau_{(ir)} \to \infty$

$$\lim_{\tau_{(ir)}\to\infty} \hat{\lambda}_{HP1} = 0. \quad (82)$$

Hence using (49) and (51), we have

$$\lim_{\tau_{(ir)}\to\infty} \hat{\lambda}_{EP1} = \lim_{\tau_{(ir)}\to\infty} \hat{\lambda}_{HP1} = 0. \quad (83)$$

□

The rest of the proof can be proved in the similar way and omitted. In the following theorem, we gives the relationship between E-Bayes and hierarchical Bayes estimators of reversed hazard rate under the same loss function. The proof is similar to the above theorem and hence omitted.

Theorem 4: The relation between E-Bayes and hierarchical Bayes estimators of reversed hazard rate for SELF, ELF and PLF are respectively given as

i) $\lim_{\tau_{(ir)}\longrightarrow ß\infty} \hat{h(t)}_{ESi} = \lim_{\tau_{(ir)}\longrightarrow ß\infty} \hat{h(t)}_{HSi} = 0, i = 1, 2, 3.$

ii) $\lim_{\tau_{(ir)}\longrightarrow ß\infty} \hat{h(t)}_{EEi} = \lim_{\tau_{(ir)}\longrightarrow ß\infty} \hat{h(t)}_{HEi} = 0, i = 1, 2, 3.$

iii) $\lim_{\tau_{(ir)}\longrightarrow ß\infty} \hat{h(t)}_{EPi} = \lim_{\tau_{(ir)}\longrightarrow ß\infty} \hat{h(t)}_{HPi} = 0, i = 1, 2, 3.$

6. Monte Carlo Simulation

In this section, we inspect the performance of the proposed estimators using a simulation study. We use the following steps for performing the study.

Step 1: Generate samples of sizes n=500,1000 and 1500 from the inverse Rayleigh distribution with pdf (1) for $\lambda = 13$.

Step 2: Fix the value of $c = 1$.

Step 3: For computing the Bayesian estimators, use (7), (8), (9), (10), (11) and (12), for E-Bayesian estimators, use (40), (41), (42), (43), (44), (45), (46), (47) and (48) and for calculating hierarchical Bayesian estimators, use (70), (75) and (80).

Step 4: Repeat steps 1-3, 10000 times and compute the MSE.

Table 1: MSE for Bayesian, E-Bayesian and H-Bayesian estimates of λ for simulated data

	$n=500$			$n=1000$			$n=1500$			CP	ACI
	$r=50$	$r=100$	$r=150$	$r=100$	$r=200$	$r=300$	$r=200$	$r=300$	$r=400$		
$\hat{\lambda}_{B1}$	0.6168	0.4591	0.4501	0.2479	0.2384	0.1945	0.1653	0.1473	0.1414	92.3 %	(11.5621, 13.7678)
$\hat{\lambda}_{B2}$	0.5334	0.4832	0.4403	0.2521	0.2431	0.1991	0.1701	0.1500	0.1439	94.3 %	(11.4505, 13.8129)
$\hat{\lambda}_{B3}$	0.4918	0.4543	0.4260	0.2459	0.2363	0.1924	0.1631	0.1460	0.1403	99.1 %	(11.0848, 14.2832)
$\hat{\lambda}_{ES1}$	0.3851	0.3791	0.3727	0.2376	0.2113	0.1752	0.1378	0.1338	0.1312	90.6 %	(11.8765, 14.0324)
$\hat{\lambda}_{ES2}$	0.3970	0.3844	0.3758	0.2365	0.2082	0.1771	0.1428	0.1358	0.1321	95.8 %	(11.5952, 14.1703)
$\hat{\lambda}_{ES3}$	0.4212	0.4036	0.3822	0.2378	0.2084	0.1810	0.1491	0.1387	0.1342	95.3 %	(11.5601, 14.0589)
$\hat{\lambda}_{EE1}$	0.3912	0.3788	0.3782	0.2374	0.2100	0.1762	0.1402	0.1348	0.1317	98.2 %	(11.4222, 14.4218)
$\hat{\lambda}_{EE2}$	0.4088	0.3940	0.3794	0.2374	0.2085	0.1791	0.1458	0.1372	0.1331	97.4 %	(11.4457, 14.2521)
$\hat{\lambda}_{EE3}$	0.4387	0.4167	0.3903	0.2399	0.2102	0.1839	0.1527	0.1406	0.1357	98.7 %	(11.2301, 14.3229)
$\hat{\lambda}_{EP1}$	0.3830	0.3804	0.3702	0.2379	0.2121	0.1748	0.1367	0.1333	0.1310	93.1 %	(11.8072, 14.1385)
$\hat{\lambda}_{EP2}$	0.3920	0.3802	0.3747	0.2362	0.2083	0.1763	0.1414	0.1351	0.1317	96.9 %	(11.5367, 14.2634)
$\hat{\lambda}_{EP3}$	0.4135	0.3977	0.3789	0.2370	0.2078	0.1798	0.1474	0.1379	0.1335	94.5 %	(11.6169, 14.0357)
$\hat{\lambda}_{HS1}$	0.6005	0.4286	0.4392	0.3148	0.2663	0.1816	0.1879	0.1819	0.1651	90.1 %	(11.7812, 13.9065)
$\hat{\lambda}_{HS2}$	0.6861	0.4473	0.4210	0.3345	0.2789	0.1912	0.1983	0.1904	0.1734	96.3 %	(11.4715, 14.1431)
$\hat{\lambda}_{HS3}$	0.6861	0.4473	0.4310	0.3345	0.2789	0.1912	0.1992	0.1940	0.1743	96.0 %	(11.4920, 14.1226)
$\hat{\lambda}_{HE1}$	0.5910	0.4334	0.4138	0.2987	0.2568	0.2044	0.1770	0.1752	0.1704	96.1 %	(11.6000, 14.3119)
$\hat{\lambda}_{HE2}$	0.4987	0.4332	0.4138	0.2855	0.2503	0.1945	0.1756	0.1703	0.1686	98.9 %	(11.2829, 14.6215)
$\hat{\lambda}_{HE3}$	0.5535	0.4331	0.4139	0.2929	0.2537	0.2002	0.1743	0.1725	0.1658	90.9 %	(11.8395, 14.0574)
$\hat{\lambda}_{HP1}$	0.6668	0.4322	0.4173	0.3299	0.2759	0.1890	0.1968	0.1914	0.1723	90.7 %	(11.7994, 13.9757)
$\hat{\lambda}_{HP2}$	0.6200	0.4330	0.4187	0.2794	0.2505	0.1885	0.1760	0.1741	0.1739	94.9 %	(11.6174, 14.1429)
$\hat{\lambda}_{HP3}$	0.5110	0.4339	0.4202	0.2863	0.2502	0.1955	0.1710	0.1673	0.1604	94.1 %	(11.6524, 14.0932)

Step 5: For creating the credible intervals, we first order $\lambda_1, \lambda_2, ..., \lambda_N$ as $\lambda_{(1)} < \lambda_{(2)} < ... < \lambda_{(N)}$ and $h_1, h_2, ..., h_N$ as $h_{(1)} < h_{(2)} < ... < h_{(N)}$. The $100(1-\gamma)$ symmetric credible intervals of λ and reversed hazard rate are obtained respectively as $(\lambda_{(N\gamma/2)}, \lambda_{(N(1-\gamma/2))})$ and $(h_{(N\gamma/2)}, h_{(N(1-\gamma/2))})$.

The MSE, average credible intervals (ACI) and coverage probabilities (CP) of the estimators computed using the simulated data are reported in Tables 1 and 2.

From Tables 1 and 2, we have the following conclusions.

- For a fixed value of n and r the MSE is less for E-Bayesian estimators as compared to Bayesian and Hierarchical Bayesian estimators.
- The performance of the proposed estimators are better than Bayesian and Hierarchical Bayesian estimators in terms of MSE.

7. Real data set

To study the performance of the estimators derived in this article, for real life situations, we considered the real data set reported by Ma and Gui (2020) representing 23 deep-groove ball bearing failure times. We fit inverse Rayleigh distribution to the data and the corresponding p-value and test statistic value for the Kolmogorov-Smirnov test are 0.6942 and 0.1415 respectively. Using MLE we estimated $\hat{\lambda} = 0.2244$. Using the bootstrapping concept, we computed the MSE, average credible interval (ACI) and coverage probability (CP) of the estimators and are given in Tables 3 and 4.

Table 2: MSE for Bayesian, E-Bayesian H-Bayesian estimates of $\hat{h(t)}$ for simulated data

	$n=500$			$n=1000$			$n=1500$			CP	ACI
	$r=50$	$r=100$	$r=150$	$r=100$	$r=200$	$r=300$	$r=200$	$r=300$	$r=400$		
$\hat{h}_{B1}$	0.0152	0.0038	0.0011	0.0059	0.0018	0.0006	0.0022	0.0008	0.0006	92.1 %	(0.6645, 2.0725)
$\hat{h}_{B2}$	0.0157	0.0039	0.0012	0.0060	0.0018	0.0006	0.0023	0.0008	0.0006	97.5 %	(0.6519, 2.1202)
$\hat{h}_{B3}$	0.0150	0.0037	0.0010	0.0058	0.0018	0.0006	0.0022	0.0008	0.0005	96.8 %	(0.6577, 2.1129)
$\hat{h}_{ES1}$	0.0125	0.0031	0.0005	0.0053	0.0016	0.0004	0.0019	0.0007	0.0004	99.0 %	(0.6636, 2.2009)
$\hat{h}_{ES2}$	0.0130	0.0032	0.0006	0.0054	0.0016	0.0004	0.0020	0.0007	0.0004	91.4 %	(0.6783, 2.0970)
$\hat{h}_{ES3}$	0.0135	0.0033	0.0007	0.0055	0.0017	0.0005	0.0021	0.0007	0.0005	94.1 %	(0.6701, 2.1060)
$\hat{h}_{EE1}$	0.0127	0.0031	0.0005	0.0053	0.0016	0.0004	0.0020	0.0007	0.0004	95.8 %	(0.6737, 2.1367)
$\hat{h}_{EE2}$	0.0132	0.0032	0.0006	0.0054	0.0016	0.0005	0.0020	0.0007	0.0005	97.2 %	(0.6657, 2.1447)
$\hat{h}_{EE3}$	0.0139	0.0034	0.0008	0.0056	0.0017	0.0005	0.0021	0.0007	0.0005	92.4 %	(0.6707, 2.0890)
$\hat{h}_{EP1}$	0.0125	0.0031	0.0005	0.0053	0.0016	0.0004	0.0019	0.0007	0.0004	98.4 %	(0.6683, 2.1850)
$\hat{h}_{EP2}$	0.0128	0.0032	0.0006	0.0054	0.0016	0.0004	0.0020	0.0007	0.0004	97.4 %	(0.6680, 2.1550)
$\hat{h}_{EP3}$	0.0134	0.0033	0.0007	0.0055	0.0017	0.0005	0.0020	0.0007	0.0005	99.3 %	(0.6532, 2.1964)
$\hat{h}_{HS1}$	0.0110	0.0035	0.0020	0.0072	0.0021	0.0009	0.0032	0.0011	0.0007	95.3 %	(1.6906, 2.0910)
$\hat{h}_{HS2}$	0.0113	0.0040	0.0024	0.0076	0.0023	0.0008	0.0034	0.0012	0.0008	95.9 %	(1.6799, 2.0910)
$\hat{h}_{HS3}$	0.0106	0.0043	0.0028	0.0080	0.0024	0.0007	0.0037	0.0013	0.0009	94.1 %	(1.7096, 2.0915)
$\hat{h}_{HE1}$	0.0104	0.0035	0.0016	0.0069	0.0020	0.0008	0.0031	0.0010	0.0008	95.6 %	(1.7032, 2.1112)
$\hat{h}_{HE2}$	0.0104	0.0037	0.0013	0.0066	0.0019	0.0007	0.0028	0.0009	0.0007	98.0 %	(1.6711, 2.1423)
$\hat{h}_{HE3}$	0.0104	0.0042	0.0015	0.0068	0.0020	0.0009	0.0030	0.0010	0.0008	92.7 %	(1.7246, 2.0876)
$\hat{h}_{HP1}$	0.0107	0.0041	0.0023	0.0075	0.0023	0.0008	0.0034	0.0012	0.0008	99.7 %	(1.5973, 2.1971)
$\hat{h}_{HP2}$	0.0107	0.0037	0.0009	0.0064	0.0019	0.0006	0.0026	0.0010	0.0007	97.7 %	(1.6665, 2.1258)
$\hat{h}_{HP3}$	0.0108	0.0037	0.0013	0.0067	0.0019	0.0008	0.0029	0.0009	0.0008	98.4 %	(1.6518, 2.1383)

It can also be noted that the estimators are satisfying the inequalities mentioned in Theorems 1 and 2. From the Tables, we can conclude that E-Bayesian estimators perform better than Bayesian and H-Bayesian estimators in terms of MSE.

8. Conclusion

The Bayesian, E-Bayesian and H-Bayesian techniques are used for estimating the parameter and reversed hazard rate of the inverse Rayleigh distribution based on left censoring. A real data and the Monte Carlo simulation are used for computing the estimates and the comparisons of these estimation methods are also carried out.Using E-Bayesian method we can see that the complex integrals involved in the calculation of hierarchical estimation methods are reduced to some extent. One of the important finding of the study is the close dependency of the proposed method with existing method and are established in Theorems 3 and 4. Another finding of the present study is the superiority of the proposed estimators with existing estimators. We also study the effect of various loss functions theoretically and are presented in Theorems 1 and 2. Important concluding remarks from our study are listed below:

1. Results showed that the MSE of the estimates decreases as the sample size increases.
2. The MSE of the E-Bayesian estimates of λ is less than the MSE of the Bayesian and H-Bayesian estimates, so E-Bayesian estimators perform better than the other two existing estimation methods.
3. The MSE of Bayesian, H-Bayesian and E-Bayesian estimates decrease when r increases.

Table 3: Comparison of MSE of the proposed estimators of λ with Bayesian estimates for real data

	$n = 23$				CP	ACI
	$r = 2$	$r = 4$	$r = 8$	$r = 12$		
$\hat{\lambda}_{B1}$	0.0106	0.0083	0.0054	0.0052	97.4 %	(0.0839, 0.4024)
$\hat{\lambda}_{B2}$	0.0115	0.0082	0.0048	0.0047	98.8 %	(0.0606, 0.4036)
$\hat{\lambda}_{B3}$	0.0103	0.0084	0.0059	0.0056	95.9 %	(0.0992, 0.3980)
$\hat{\lambda}_{ES1}$	0.0106	0.0085	0.0057	0.0054	93.0 %	(0.1134, 0.3768)
$\hat{\lambda}_{ES2}$	0.0106	0.0084	0.0056	0.0054	96.0 %	(0.0959, 0.3933)
$\hat{\lambda}_{ES3}$	0.0106	0.0084	0.0055	0.0053	98.8 %	(0.0630, 0.4252)
$\hat{\lambda}_{EE1}$	0.0115	0.0084	0.0049	0.0049	95.1 %	(0.0974, 0.3705)
$\hat{\lambda}_{EE2}$	0.0115	0.0083	0.0049	0.0048	97.2 %	(0.0816, 0.3853)
$\hat{\lambda}_{EE3}$	0.0115	0.0083	0.0049	0.0048	96.9 %	(0.0845, 0.3815)
$\hat{\lambda}_{EP1}$	0.0104	0.0087	0.0062	0.0058	97.8 %	(0.0804, 0.4208)
$\hat{\lambda}_{EP2}$	0.0103	0.0086	0.0061	0.0057	98.9 %	(0.0619, 0.4383)
$\hat{\lambda}_{EP3}$	0.0103	0.0086	0.0060	0.0057	97.5 %	(0.0844, 0.4148)
$\hat{\lambda}_{HS1}$	0.0106	0.0082	0.0053	0.0052	91.8 %	(0.1188, 0.3663)
$\hat{\lambda}_{HS2}$	0.0106	0.0081	0.0053	0.0051	92.4 %	(0.1163, 0.3677)
$\hat{\lambda}_{HS3}$	0.0106	0.0081	0.0053	0.0051	98.3 %	(0.0729, 0.4110)
$\hat{\lambda}_{HE1}$	0.0115	0.0085	0.0050	0.0050	91.6 %	(0.1140, 0.3558)
$\hat{\lambda}_{HE2}$	0.0115	0.0083	0.0049	0.0048	99.8 %	(0.0202, 0.4460)
$\hat{\lambda}_{HE3}$	0.0115	0.0083	0.0048	0.0047	93.7%	(0.1051, 0.3602)
$\hat{\lambda}_{HP1}$	0.0102	0.0083	0.0057	0.0054	91.3 %	(0.1234, 0.3714)
$\hat{\lambda}_{HP2}$	0.0104	0.0087	0.0061	0.0058	98.7 %	(0.0661, 0.4347)
$\hat{\lambda}_{HP3}$	0.0103	0.0085	0.0060	0.0056	96.7 %	(0.0925, 0.4059)

4. The MSE of E-Bayesian estimates under ELF is less than the MSE of E-Bayesian estimates under SELF and PLF, so E-Bayesian estimators under ELF perform better than the E-Bayesian estimator SELF and PLF.

5. We can conclude that the E-Bayesian estimators perform better than Bayesian and H-Bayesian estimators in terms of MSE.

Acknowledgements

The authors would like to thank the associate editor, the editor and the referees for their constructive comments and suggestions that improved the content and the style of this article.

Table 4: Comparison of MSE of the proposed estimators of $h(\bar{t})$ with Bayesian estimates for real data

	$n = 23$				CP	ACI
	$r = 2$	$r = 4$	$r = 8$	$r = 12$		
$\hat{h}_{B1}$	60.6478	46.0936	42.6285	41.8922	91.9 %	(-10.9944, 19.0189)
$\hat{h}_{B2}$	55.1956	41.6762	37.7979	36.1012	98.4 %	(-15.9456, 23.6054)
$\hat{h}_{B3}$	63.4686	48.3812	45.1682	44.9786	95.9 %	(-13.8673, 22.0723)
$\hat{h}_{ES1}$	61.5741	46.7373	43.315	42.7344	99.0 %	(-18.2402, 26.3238)
$\hat{h}_{ES2}$	61.3408	46.5753	43.1418	42.5214	98.8 %	(-17.6651, 25.7339)
$\hat{h}_{ES3}$	61.1081	46.4137	42.9692	42.3093	99.3 %	(-19.234, 27.2879)
$\hat{h}_{EE1}$	56.0256	42.2482	38.3836	36.7883	91.1 %	(-10.185, 17.9011)
$\hat{h}_{EE2}$	55.8165	42.1042	38.2359	36.6144	97.1 %	(-14.1523, 21.8543)
$\hat{h}_{EE3}$	55.6081	41.9606	38.0886	36.4414	95.0 %	(-12.2928, 19.9807)
$\hat{h}_{EP1}$	64.4444	49.062	45.9071	45.9024	92.8 %	(-11.7803, 20.0455)
$\hat{h}_{EP2}$	64.1986	48.8906	45.7207	45.6688	99.4 %	(-20.1433, 28.3934)
$\hat{h}_{EP3}$	63.9536	48.7197	45.5350	45.4362	90.9 %	(-10.7881, 19.0231)
$\hat{h}_{HS1}$	60.3752	45.9039	42.4273	41.6472	99.9 %	(-24.2466, 32.2537)
$\hat{h}_{HS2}$	60.105	45.7157	42.228	41.4053	96.1 %	(-13.6962, 21.6858)
$\hat{h}_{HS3}$	59.8369	45.5289	42.0308	41.1667	97.0 %	(-14.5804, 22.5527)
$\hat{h}_{HE1}$	56.448	42.539	38.6829	37.1418	98.4 %	(-16.0768, 23.8213)
$\hat{h}_{HE2}$	55.6528	41.9915	38.1201	36.478	97.7 %	(-14.8781, 22.569)
$\hat{h}_{HE3}$	55.4419	41.8461	37.9714	36.3037	99.7 %	(-20.5672, 28.2438)
$\hat{h}_{HP1}$	62.8967	47.9816	44.7371	44.4445	92.7 %	(-11.6261, 19.7953)
$\hat{h}_{HP2}$	64.3447	48.9925	45.8313	45.8067	98.0 %	(-16.4345, 24.6936)
$\hat{h}_{HP3}$	63.7582	48.5834	45.3870	45.251	95.6 %	(-13.6303, 21.8532)

References

Abdul-Sathar, E. I. and Athirakrishnan, R. B. (2019). E-Bayesian and hierarchical Bayesian estimation for the shape parameter and reversed hazard rate of power function distribution under different loss functions. *Journal of the Indian Society for Probability and Statistics*, **20**, 227–253.

Dey, S. (2012). Bayesian estimation of the parameter and reliability function of an inverse Rayleigh distribution. *Malaysian Journal of Mathematical Sciences*, **6**, 113–124.

El-Helbawy, A. and Abd-El-Monem (2005). Bayesian estimation and prediction for the inverse rayleigh lifetime distribution. In *Proceeding of the 40th annual conference of Statistics, Computer sciences and Operation Research*, pages 45–59, ISSR, Cairo University.

Feroze, N. and Aslam, M. (2012). On posterior analysis of inverse Rayleigh distribution under singly and doubly type II censored data. *International Journal of Probability and Statistics*, **1**, 145–152.

Han, M. (1997). The structure of hierarchical prior distribution and its applications. *Chinese Operations Research and Management Science*, **6**, 31–40.

Han, M. (2009). E-Bayesian estimation and hierarchical Bayesian estimation of failure rate. *Applied Mathematical Modelling*, **33**, 1915–1922.

Jaheen, Z. F. and Okasha, H. M. (2011). E-Bayesian estimation for the Burr type XII model based on type-2 censoring. *Applied Mathematical Modelling*, **35**, 4730–4737.

Kızılaslan, F. (2017). The E-Bayesian and hierarchical Bayesian estimations for the proportional reversed hazard rate model based on record values. *Journal of Statistical Computation and Simulation*, **87**, 2253–2273.

Lindley, D. V. and Smith, A. F. (1972). Bayes estimates for the linear model. *Journal of the Royal Statistical Society. Series B (Methodological)*, **34**, 1–41.

Ma, Y. and Gui, W. (2020). Entropy-based and non-entropy-based goodness of fit test for the inverse Rayleigh distribution with progressively type-II censored data. *Probability in the Engineering and Informational Sciences*, **35**, 1–19.

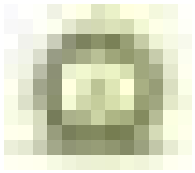
Okasha, H. M. and Wang, J. (2016). E-Bayesian estimation for the geometric model based on record statistics. *Applied Mathematical Modelling*, **40**, 658–670.

Shawky, A. and Badr, M. (2012). Estimations and prediction from the inverse Rayleigh model based on lower record statistics. *Life Science Journal*, **9**, 985–990.

Soliman, A., Amin, E. A., and Abd-El Aziz, A. A. (2010). Estimation and prediction from inverse Rayleigh distribution based on lower record values. *Applied Mathematical Sciences*, **4**, 3057–3066.

Voda, V. G. (1972). On the inverse Rayleigh distributed random varaible. *Reports of Statistical Application Research*, **19**, 13–21.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 22, No. 1, 2024 (New Series), pp 57-71
http://www.ssca.org.in/journal



Eigenspace Based Online Path Planner for Autonomous Mobile Robots

Shyba Zaheer[1], Imthias Ahamed T.P.[1], Tauseef Gulrez[2] and Zoheb Zaheer[3]
[1]*Department of EEE, TKM College of Engineering, Kerala, India*
[2] *Department of Computing and Research, Syscon Private Limited, Melbourne, Australia*
[3]*Mitsogo Inc, Kerala, India*

Received: 03 August 2022; Revised: 10 December 2022; Accepted: 25 May 2023

Abstract

This paper proposes a target oriented online path planning algorithm which is capable of navigating a mobile robot autonomously in unknown environments. The proposed technique called Free Configuation Eigenspace (FCE) finds collision free path from laser sensor data by computing its eigenvectors. The paper describes an online 2D simulation method of FCE with static obstacles and start and goal positions. The proposed method is benchmarked aginst the well known online path planner Vector Field Histogram (VFH). In this 2D simulation, the robot model used is a differential drive robot and it is assumed that the robot is equipped with a laser scanner. Simulation experiments are done with start and goal positions on simulated 2D maps in MATLAB with different obstacle courses. The respective trajectories for different start and goal positions were generated on the map and path lengths analyzed

Key words: Online path planning; Online obstacle avoidance; Eigenspace; Eigenvector.

1. Introduction

An autonomous robot's ability to plan its motion in real-time has become a crucial part of modern intelligent robotics. Applications of path planning in online environments include the mining industry, planet exploration, reconnaissance, *etc.* This is also known as local path planning. Online path planning deals with the assessment of the dynamic conditions of the environment and identifying the positional relationships among various elements in the environment. In online navigation, the robot can autonomously decide its motion using equipped sensors such as laser sensors, ultrasonic range finders, sharp infrared range sensors, vision (camera) sensors, *etc.*

Pioneering work in online obstacle avoidance and path planning was initiated by Khatib (1986), konown as Artificialal Potential Fields method (APF) and is popular in mobile robotics. The idea of (APF) comes from the concept of the potential field in physics,

Corresponding Author: Shyba Zaheer
Email: s.shyba@gmail.com

which regards the movement of objects as the result of two kinds of forces. The robot is subjected to attractive forces from the target and repulsive forces from the obstacle. Under the action of the two forces, the robot moves toward the target point due to the resultant force and during the moving process, it can effectively avoid the obstacles and reach the target. Bug algorithms by Lumelsky and Stepanov (1987) are also used for online path planning which uses two-dimensional scenes filled with unknown obstacles. Bug algorithm assumes the robot as a point operating in the plane with a contact sensor or range sensor to detect obstacles. Bug algorithm uses a straightforward path planning approach to move towards the goal unless an obstacle is encountered, in which case it circumnavigates the obstacle until motion towards the goal is once again allowable. Another online planner in literature is the Dynamic Windows Approach (DWA), by Fox *et al.* (1997). This approach is derived directly from the motion dynamics of the robot and is therefore particularly well suited for robots operating at high speed. The dynamic window contains the feasible linear and angular velocities taking into consideration the acceleration capability of the robot.The collision cone concept based online path planning was proposed by Chakravarthy and Ghose (1998). The collision cone can be used to predict the possibility of collisions between two objects and to design collision avoidance strategies. In this method, a collision of a robot can be averted if the relative velocity of a robot with respect to a particular obstacle falls exterior to the collision cone.

One of the widely used sensor-based online path planning algorithms is Vector Field Histogram (VFH) by Borenstein *et al.* (1991). In VFH, a polar histogram is generated at every discrete point step to represent the polar density of the obstacles around the robot. The robot's steering direction is chosen based on the least polar density and closeness to the current steering direction. The VFH algorithm is fast, very robust, and insensitive to misreadings, allowing continuous and fast motion of the mobile robot without stopping for obstacles. But the VFH-controlled robot may get "trapped" in dead-end situations (as is the case with other local path planners). When trapped, mobile robots usually exhibit "cyclic behavior". Another limitation of this technique is that the polar histogram must be regularly generated for every time step. Hence in narrow hallways, the robot may move in an oscillatory fashion. Also, this method is suited for environments with sparse moving obstacles. Ulrich and Borenstein (1998) proposed a method known as **VFH+** that introduces some of the parameters tuning to accommodate the robot's width, also.

This paper is organized as follows: first it describes the materials and methods used for this study followed by the 2D simulation method of VFH with static obstacles and start and goal positions. Then it describes the proposed Free Configuration Technique(FCE) path planner proposed by us Zaheer *et al.* (2022). Detailed simulation results are included in section 4 and followed by result analysis and conclusion.

2. Materials and methodology

This section describes the materials and methods used for this study. The robot model used here is a differential drive robot and the sensor used is a laser sensor. In this 2D simulation, it is assumed that a vehicle is equipped with a scanning laser range sensor with a field of view of 240°. Also, vehicle location is known and only kinematic motion of the vehicle is considered. The simulation experiments are done with Start and Goal positions on simulated 2D maps with different obstacle courses. The performance analysis is done

for different scenarios: namely **Scenario-I** with 3 separate obstacles with and L-shaped wall obstacle and **Scenario-II and III** with differently shaped obstacles. The respective trajectories for different Start and Goal positions were plotted on the map and path lengths analyzed.

2.1. Differential drive kinematics

A differential drive robot consists of 2 drive wheels mounted on a common axis, and each wheel can independently be driven either forward or backward. For the robot to perform direction change in it's translational motion, the velocity of each wheel may be varied appropriately. The robot actually performs rotatory motion about a point along the common left and right wheel axis which is known as the ICC (Instantaneous Center of Curvature) as seen in Figure 1. Hence by varying the velocities of the two wheels, we can vary the trajectories that the robot take

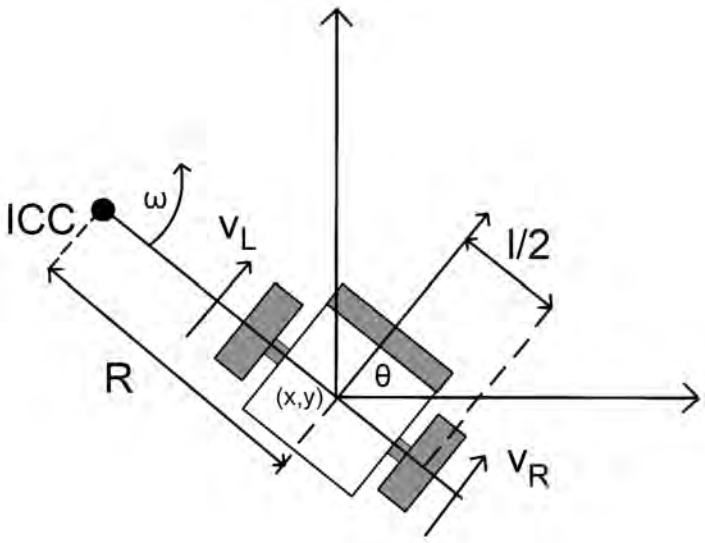


Figure 1: Differential drive kinematics

Since the rate of rotation ω about the ICC must be the same for both wheels, we can write the following equations:

$$\omega(R + \frac{l}{2}) = V_{\mathrm{R}} \tag{1}$$

$$\omega(R - \frac{l}{2}) = V_{\mathrm{L}} \tag{2}$$

where l is the distance between the wheels, V_R and V_L are the right and left wheel velocities along the ground respectively, and R is the signed distance from the ICC to the midpoint between the wheels. At any instance in time we can solve for R and ω:

$$R = \frac{l(V_{\mathrm{L}} + V_{\mathrm{R}})}{2(V_{\mathrm{R}} - V_{\mathrm{L}})} \tag{3}$$

$$\omega = \frac{V_{\mathrm{R}} - V_{\mathrm{L}}}{l} \tag{4}$$

In Figure 2, the velocity of the robot can be represented as a pair of vectors, $\vec{v}$ and $\vec{\omega}$, where $\vec{v}$ represents the linear velocity (forwards and backwards) of the robot and $\vec{\omega}$ represents the angular velocity of the robot. Given angular veocities of the right and left wheels ω_R and ω_L respectively, the linear and angular velocities of the differential drive robot are represented as shown in Equations 5 and 6.

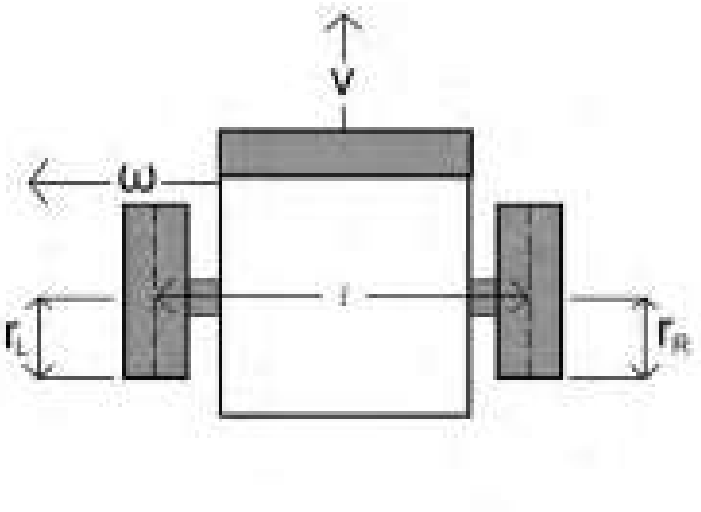


Figure 2: Physical configuration of the robot

$$v = \frac{r_R}{2}\omega_R + \frac{r_L}{2}\omega_L \tag{5}$$

$$\omega = \frac{r_R}{l}\omega_R - \frac{r_L}{l}\omega_L \tag{6}$$

where r_R and r_L are the wheel radii of the left and right wheels, respectively, and l is the width of the wheelbase as shown in Figure 2. The robot in the global coordinate frame is represented in Figure 3. The equations of motion are shown in Equation 7.

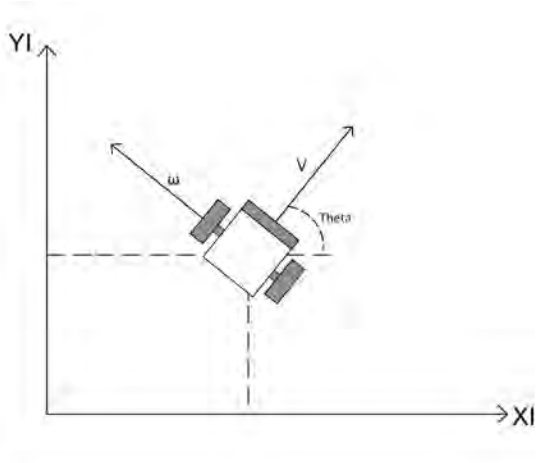


Figure 3: Robot's kinematics in global frame

$$\begin{bmatrix} \dot{x} \\ \dot{y} \\ \dot{\theta} \end{bmatrix} = \begin{bmatrix} v\ cos\theta \\ v\ sin\theta \\ \omega \end{bmatrix} \tag{7}$$

Suppose, at the initial time T_0 the pose is $[x_0, y_0, \theta_0]$; the pose at time t is $[x(t), y(t), \theta(t)]$; to find the pose at time T, we will have to integrate the variables at the pose $t + \Delta_t$ from within the limit T_0 to T which is added to the initial pose coordinates $[x_0, y_0, \theta_0]$, as follows:

$$x(T) = \int_{T_0}^{T} v(t)\cos(\theta(t))dt + x_0 \tag{8}$$

$$y(T) = \int_{T_0}^{T} v(t)\sin(\theta(t))dt + y_0 \tag{9}$$

$$\theta(T) = \int_{T_0}^{T} \omega(t)dt + \theta_0 \tag{10}$$

2.2. Principal component analysis (PCA)

PCA is a way of identifying patterns in data and expressing the data in such a way as to highlight their similarities and differences. Since patterns can be hard to find in data of high dimension, PCA helps us to identify patterns in data based on the correlation between features. It aims to find the directions of maximum variance in high-dimensional data and projects it onto a new subspace with equal or fewer dimensions than the original one.

In our case, with a laser range sensor, individual range measurements can be considered as an independent dimension. With such an approach, a range scanner with a 1^0 resolution and a 240^0 field view generates 240 observations which are called as point cloud

sensor data. Analysis of the point cloud data from such a scanner will show that adjacent range measurements are highly correlated. Thus, it is possible to use principal component analysis to determine a linear subspace with a minimum number of dimensions for representing an environment using a range sensor.

PCA finds Principal Components (PCs) that are linear combinations of the original variables ranked in terms of the variability in the data given by the variances. The corresponding orthogonal directions are given by the eigenvectors of the covariance matrix (C) of the data. The steps involved in PCA analysis are as follows :

- Standardize the dataset
- Compute the covariance matrix of the dataset.
- Perform eigen decomposition on the covariance matrix.
- Order the eigenvectors in decreasing order based on the magnitude of their corresponding eigenvalues.

Let S be the 2D point cloud data,where, $S_j = (x_j, y_j)$; $(S_j \in S \in R^2)$ and $\bar{S}$ is the mean of k no of sensor data points as given below.

To perform the PCA transformation, we have to compute the covariance matrix C of point cloud data set S using the below equation:

$$C_{2x2} = \frac{1}{K}\sum_{j=1}^{k}(S_j - \bar{S})(S_j - \bar{S})^T; \bar{S} = \frac{1}{K}\sum_{j=1}^{k} S_j \tag{11}$$

To perform transformation we have to solve the eigenvalue Equation 12

$$C\,V = \lambda\,V \tag{12}$$

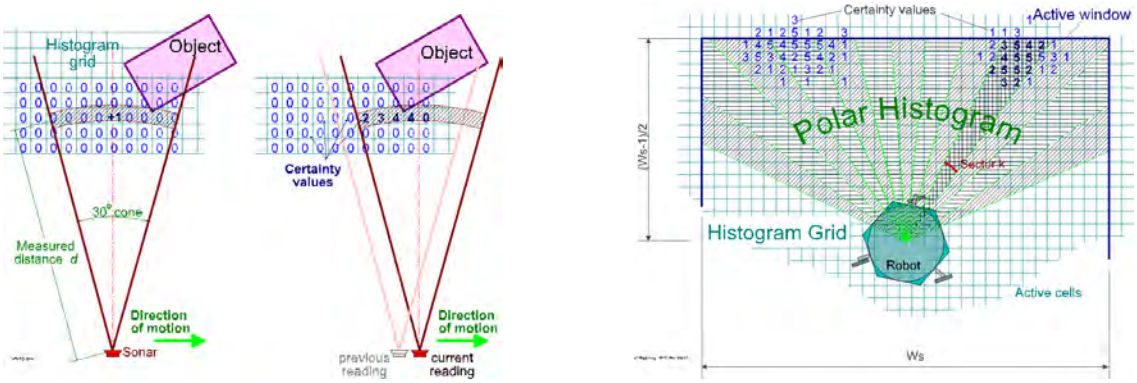
Solving the Equation 12 , we can get the eigenvalues λ, where $\lambda_1 \geq \lambda_2$ and eigenvectors V $[V_1, V_2]$.

These eigenvectors are called Principal Components (PCs). By applying a proper technique we can identify the pattern in 2D point cloud range data. In our case, the obstacle area or free area identification can be performed with this PCA technique .

3. Vector field histogram (VFH)

VFH method is executed in three main steps that are: Two dimensional cartesian histogram grid, Polar histogram sector and candidate valley selection. To begin with, on-board sensors such as ultrasonic sensor or laser rangefinder are used for mapping obstacles into histogram grid. In this step, the two-dimensional cartesian histogram grid is continuously updated in with range data sampled by the on-board range sensors as shown in Figure 4a.

At the second step, a one-dimensional polar histogram is constructed around the robot's momentary location by dividing the polar histogram into angular sectors of suitable width as shown in Figure 4b. At the third step the output of the VFH algorithm, which



(a) Cratesian histogram **(b) Polar histogram**

Figure 4: Algorithm steps - I and II

(a) Polar obstacle density of wide valley **(b) POD with narrow valley**

Figure 5: VFH Algorithm steps - III

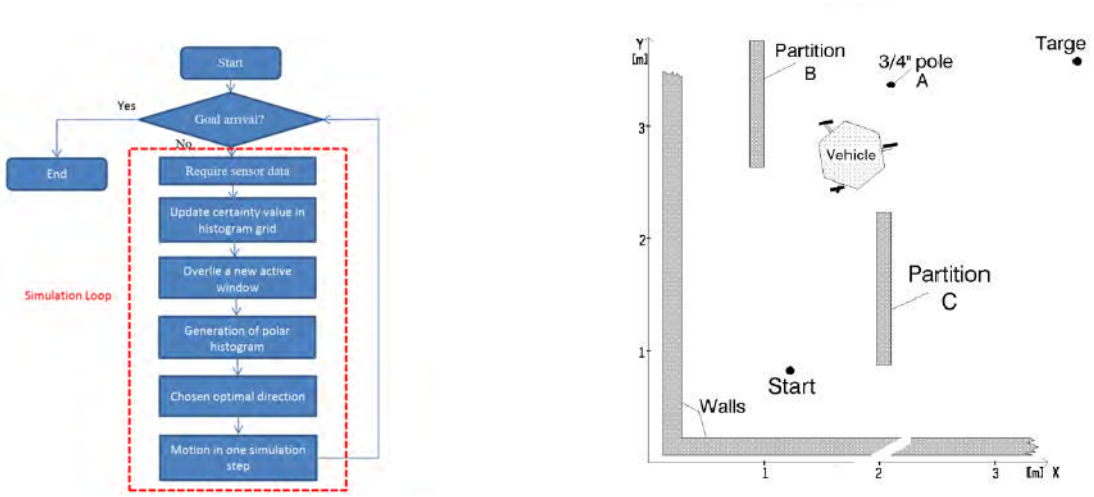
is the reference value for the new steering direction of the robot. The optimal direction is selected in each candidate valley such that every sector density is less than a suitable threshold value. The algorithm measures the size of the selected valley. Hence, two types of valleys are distinguished, namely, wide and narrow ones.A valley is considered wide if the no of consecutive polar sectors (S) are greater than 18 nos ($S_{max=18}$).

The sector nearest to target is denoted as **kn** and the far border sector is denoted as **kf** and is defined as $kf = kn + S_{max}$. The desired steering direction is then defined by $\theta = \frac{(kn+kf)}{2}$ and is closer to the goal or target directions as shown in Figure 5a.

In the second case, a narrow valley between closely spaced obstacles (shown in Pink), are shown in Figure 5b. Here the far border **kf** is less than S_{max} . However, the direction of travel is again chosen as $\theta = \frac{(kn+kf)}{2}$ and the robot maintains a course centered between obstacles as shown in Figure 5b.

3.1. VFH 2D simulation

In this section, we have done 2D Simulation of VFH algorithm in MATLAB. The **scenario-I** is an exact replica of the simulation done in the original paper by Borenstein *et al.* (1991). This scenario-I is shown in Figure 6b which contains three obstacles denoted as A,B,C with a L shaped wall.



(a) VFH simulation flow chart **(b) Original VFH simulation scenario**

Figure 6: VFH 2D Simulation

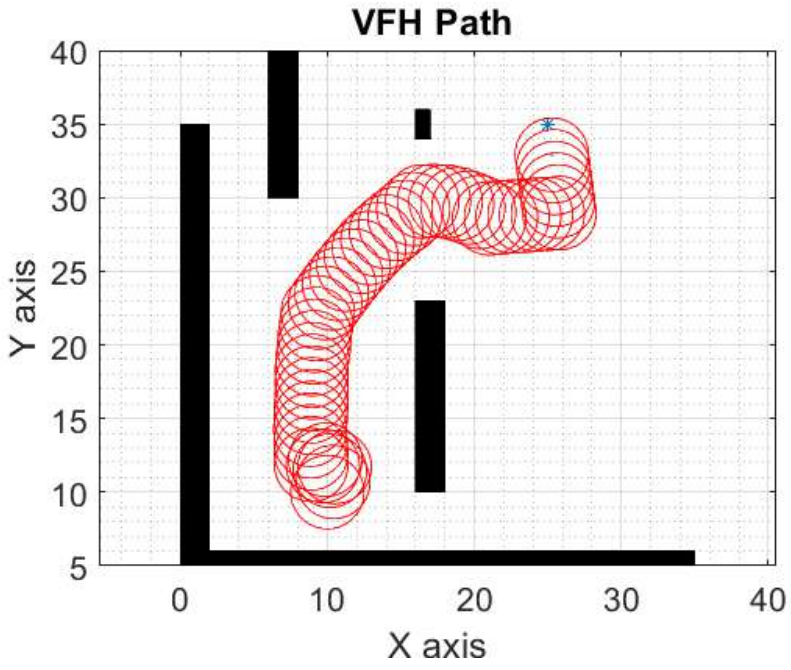


Figure 7: VFH Path: Start[10,10] , Goal[25,35]

The flow chart of VFH's 2D simulation is shown in below Figure 6a. 2D simulation is carried out for the VFH algorithm using MATLAB. The robot is assumed to be of circular shape of radius 0.025m and moving at a speed of 0.8 m/s. The Start position is at [10,10] and the Goal position is at [25,35] with a map scale 1 unit = 0.1m. The simulated robot trajectory is shown in Figure 7 and its path length is recorded in Table 1.

4. Free configuration eigenspace (FCE) formulation

The FCE method uses a two-stage sensor data reduction technique and three levels of sensor data representation :

- The first level represents the detailed description of the robot's environment. In this level, the two-dimensional cartesian map (world model) is created.
- At the second level, the high dimensional sensor space is reduced to a low-dimensional Eigenspace around the momentary location of the robot by computing Principle Components of sensor data. These PCs provides a spatial interpretation of the environment in terms of its variance of the sensor data.
- The third level of data representation is the output of the FCE algorithm, which selects the PC direction which is closest to the goal direction

4.1. Proposed FCE goal reaching algorithm

A 2D online pathplanning algorithm has been formulated with static obstacles using FCE philosophy by Zaheer *et al.*(2014). The below sections describe the algorithm formulation and implementation of FCE's 2D Trajectory generation with Start and Goal positions for a robot **A**. The flow chart of 2D simulation is shown in Figure 8.

The flowchart has the followings steps:

- The algorithm starts by computing the distance and the angle from the **Start** to **Goal** positions.
- If the goal is not reached, then the sensor cartesian data is acquired from the map's obstacle positions.
- The two PCs of sensor data are computed by applying PCA
- From the two PCs, the PC direction closer to the goal position is found.
- Then the velocity components of the new PC angle is calculated
- The new position is computed from the current position and the velocity components.
- The new position is added to robot path array .
- Finally, the robot traverses the generated path.

4.2. FCE's 2D simulation algorithm

Algorithm 1 FCE - Eigen vector trajectory

Input:
RobotA Start Position: **PosA**[ ]
RobotA Goal Position : **GoalA**[ ]
Scan Data: **S**[ ]
Initialize Simulation parameters :($v = 0.8m/s, r_A = 0.025m, t = 1s, P = 1, N = 100$)
Output: **Path**[ ] Path from **Start** to **Goal**

Compute distance (D) and angle between from Start to Goal(angA)
Path$[P, 1] = PosA(1)$
Path$[P, 2] = PosA(2)$
if (D$\geq$0) and(D $<$ 2.5) **then**
    GoalReached=1

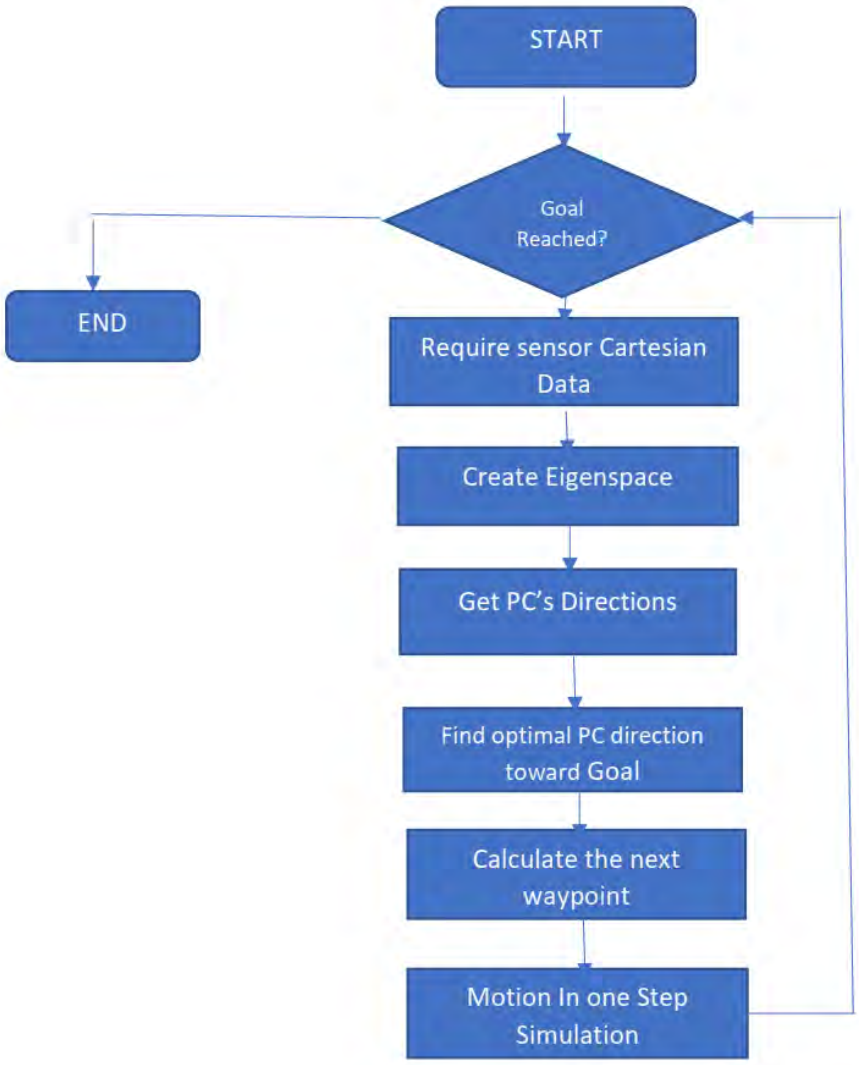


Figure 8: FCE 2D simulation flow chart

```
6: else
7:    GoalReached=0
8: end if
9: while (GoalReached==0) do
10:   repeat
11:      S=Get Cartetian Coordiante of Scan data
12:      [ PC ] =  eig (covariance (S));
13:      DirPC1 = atan2(PC(2, 1), PC(1, 1))
14:      DirPC2 = atan2(PC(2, 2), PC(1, 2))
```

newangleA=[DirPC1,DirPC2]
angleFuture=findFutureAngle(newangleA, angA);
angA=angleFuture
$V = [vcos(angA), vsin(angA)]$
$newX = PosA(1) + V(1) * t$
$newY = PosA(2) + V(2) * t$
$PosA = [newX, newY]$
$P = P + 1$
$Path[P, 1] = PosA(1)$
$Path[P, 2] = PosA(2)$
until (GoalReached==1) or (P==N)
end while
Plot the Path Points with robot **A** as circular shape

Algorithm 2 : Function: findFutureAngle()

Input: newangleA, angA
Output:angleFuture

$X1 = \cos(newangleA(1)) - \cos(angA)$
$Y1 = \sin(newangleA(1)) - \sin(angA)$
$Point1 = sqrt(X1^2 + Y1^2)$
$X2 = \cos(newangleA(2)) - \cos(angA)$
$Y2 = \sin(newangleA(2)) - \sin(angA)$
$Point2 = sqrt(X2^2 + Y2^2)$
if (Point1¡Point2) **then**
angleFuture=newangleA(1)
else
angleFuture=newangleA(2)
end if
Return (angleFuture)

In this simulation, the robot is assumed to be of circular shape with radius 0.025m and moving at a speed of 0.8 m/s. The simulation assumes that the robot is equipped with range sensor. The simulation starts with the input of initial Start position as (PosA) and the final Goal position as (GoalA). The initial step is to compute the distance (D) and the angle between the Start and the Goal Position (angA). If the distance value is greater than zero, then the obstacle free position which is more close to the Goal position has to be found from scan data. With FCE, this is achieved by computing the eigenvectors of the covariance matrix of the sensor cartesian data. This will give two PCs as shown in Figure 9. From these two PC's, the next suitable direction (angleFuture) to move towards the Goal with out hitting obstacles is identified by computing the distance between the new PCs direction point and the initial angle point (angA) and then selecting the minimum distance point among them as given in algorithm 2. Once the closest direction to the Goal is selected, the next step is to find the velocity components and the next position of the robot to move on. The next waypoint is computed by differential drive robot's kinematic equations and the values will be stored in the path array, which will give the obstacle free path from Start to

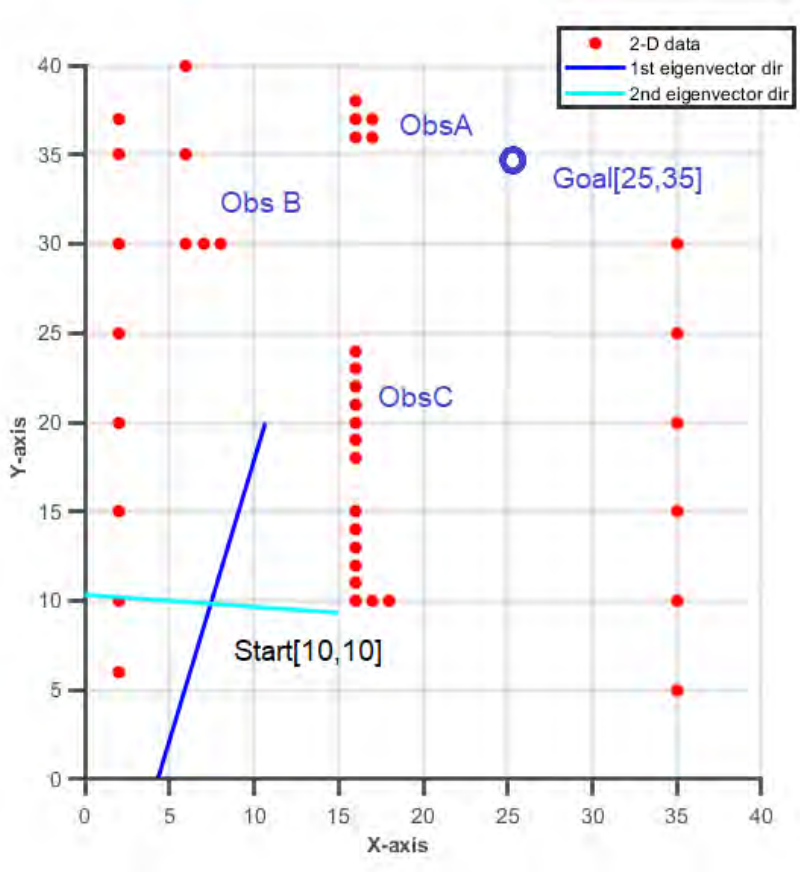


Figure 9: FCE cartesian map with eigenvectors

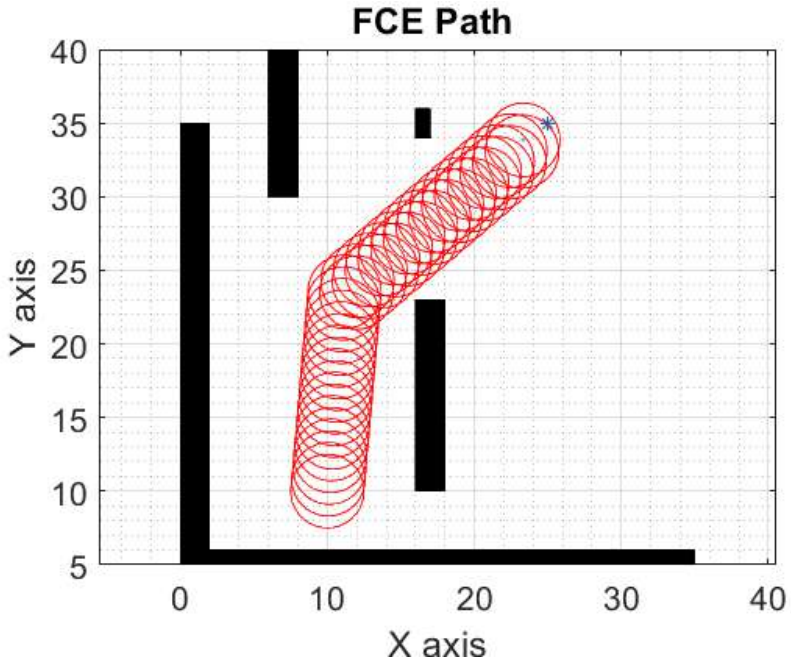


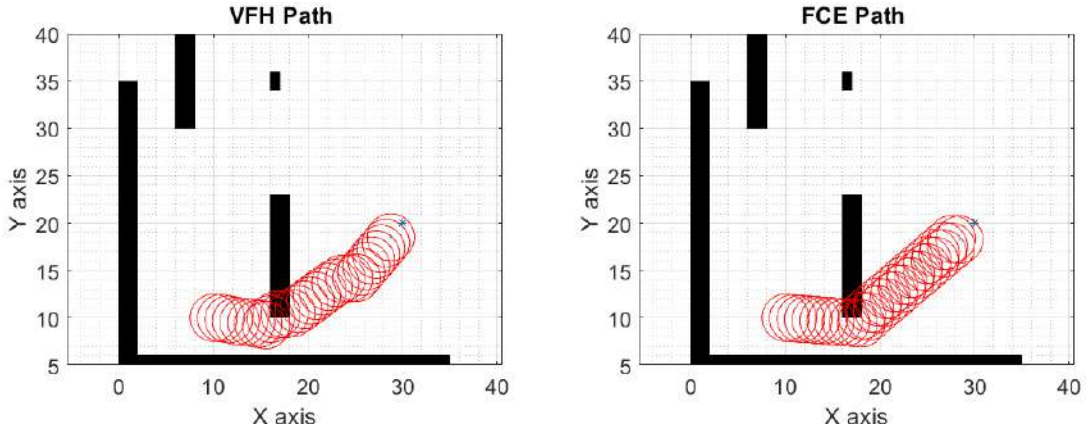
Figure 10: FCE result of scenario-I : Start[10,10], Goal[25,35]

Goal. Once the path is generated, the simulated robot trajectory can be plotted, assuming robot to be circular in shape.

The simulation scenario here is also the same as in the case of VFH described in the above section (Scenario-I) with **Start [10,10]** and **Goal [25,35]**. The FCE simulation result is shown in Figure 10 and the map scale is taken as 1 unit = 0.1m. Since the eigenspace generates only two PCs, the robot trajectory comprises of straight line segments as compared to the curved trajectory of VFH. This is shown in Figure 10 and the path length is recorded in Table 1.

5. Result analysis

Trajectory analysis of VFH and FCE technique was carried out by creating different obstacle configurations with different Start and Goal positions and the path lengths generated by each algorithm are computed. The analysis is done in two scenarios; Scenario-I, scenario-II and scenario-III having different Start and Goal locations. Then the VFH and FCE trajectory are plotted as shown in Figures 11, 12, 13 and 14 and the path lengths are recorded in Table 1. From the table it's can conclude that the VFH performance is better in terms of path length but have lots of abrupt change in directions. The FCE has straight line trajectory and shows some oscillations at some segments, as well.

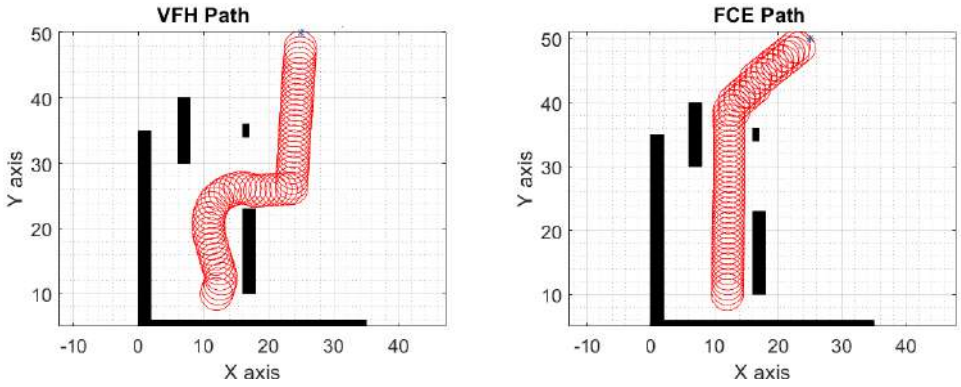


(a) Start[10,10],Goal[30,20] **(b) Start[10,10], Goal[30,20]**

Figure 11: Scenario -I

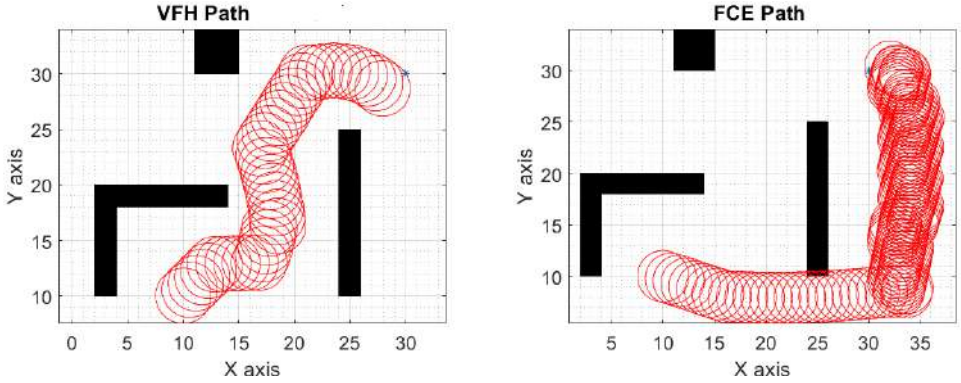
As shown in Figure 11 for Scenario-I with Start[10,10] and Goal[30,20], the path generated by both VFH and FCE are in the same direction but the VFH path length is shorter than FCE. Also, it is seen that the both trajectories are colliding with obstacle C.

As shown in Figure 12 for Scenario-I with Start[12,10] and Goal[25,50], the path generated by both VFH and FCE are in different coordinates but the VFH path length is shorter than FCE path length as seen in Table 1.



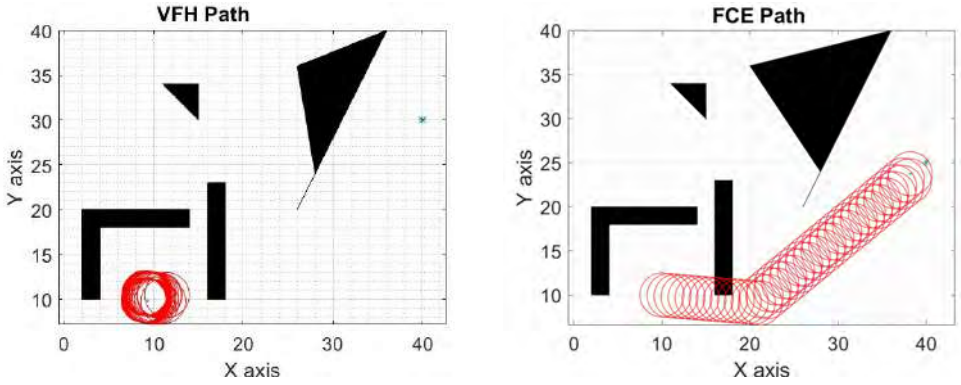
(a) Start[12,10],Goal[25,50] **(b) Start[12,10], Goal[25,50]**

Figure 12: Scenario -I



(a) Start[10,10],Goal[30,20] **(b) Start[10,10], Goal[30,20]**

Figure 13: Scenario -II



(a) Start[10,10],Goal[40,30] **(b) Start[10,10], Goal[40,30]**

Figure 14: Scenario -III

As shown in Figure 14 for Scenario-III which has more clutterd obastacles with Start[10,10] and Goal[40,30], VFH method robot is getting "trapped" in this Senario. When trapped, mobile robots exhibit "cyclic behavior", which is evident in Figure 14. But FCE method finds a path from start to goal as we can see in the Figure 14.

Table 1: 2D Simulation results

Technique	Scenario	Start	Goal	Path length(m)
VFH	Scenario -I	[10,10]	[25,35]	3.52000
FCE	Scenario -I	[10,10]	[25,35]	3.84000
VFH	Scenario -I	[10,10]	[30,20]	2.40000
FCE	Scenario -I	[10,10]	[30,20]	2.88000
VFH	Scenario -I	[12,10]	[25,50]	4.88000
FCE	Scenario -I	[12,10]	[25,50]	5.04000
VFH	Scenario -II	[10,10]	[30,30]	3.44000
FCE	Scenario -II	[10,10]	[30,30]	11.28000
VFH	Scenario -III	[10,10]	[30,30]	0.14000
FCE	Scenario -III	[10,10]	[30,30]	4.64000

6. Conclusion

Performance analysis with FCE and VFH for different scenrios shows that VFH gives the shortest path but have many changes in directions. But in the case of FCE, trajectory segments are almost straight lines but show some ocilllations in certain situations. The result analysis shows that the VFH perfomance is better in environments with unclutterd static obstacles. For exploring future scope, this research can be extended to path planning with dynamic obstacles.

References

Borenstein, J., Koren, Y., et al. (1991). The vector field histogram-fast obstacle avoidance for mobile robots. *IEEE Transactions on Robotics and Automation*, **7**, 278–288.

Chakravarthy, A. and Ghose, D. (1998). Obstacle avoidance in a dynamic environment: A collision cone approach. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, **28**, 562–574.

Fox, D., Burgard, W., and Thrun, S. (1997). The dynamic window approach to collision avoidance. *IEEE Robotics & Automation Magazine*, **4**, 23–33.

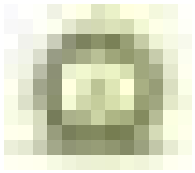
Khatib, O. (1986). Real-time obstacle avoidance for manipulators and mobile robots. *The International Journal of Robotics Research*, **5**, 90–98.

Lumelsky, V. J. and Stepanov, A. A. (1987). Path-planning strategies for a point mobile automaton moving amidst unknown obstacles of arbitrary shape. *Algorithmica*, **2**, 403–430.

Ulrich, I. and Borenstein, J. (1998). Vfh+: Reliable obstacle avoidance for fast mobile robots. In *Proceedings. 1998 IEEE International Conference on Robotics and Automation (Cat. No. 98CH36146)*, volume 2, pages 1572–1577. IEEE.

Zaheer, S., Gulrez, T., and Thythodath Paramabath, I. A. (2022). From sensor-space to eigenspace–a novel real-time obstacle avoidance method for mobile robots. *IETE Journal of Research*, **68**, 1512–1524.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 22, No. 1, 2024 (New Series), pp 73–82
http://www.ssca.org.in/journal



Development of Survey Weighted Composite Indices under Complex Surveys

Deepak Singh[1], Pradip Basak[2], Tauqueer Ahmad[1], Raju Kumar[1] and Anil Rai[1]
[1]*ICAR-Indian Agricultural Statistics Research Institute, New Delhi*
[2]*Department of Agricultural Statistics, Uttar Banga Krishi Viswavidyalaya, Cooch Behar, West Bengal*

Received: 01 July 2022; Revised: 01 June 2023; Accepted: 05 June 2023

Abstract

An index is constructed by a mathematical model representing multi-dimensional variables into a single value. Multi-dimensional variables are often correlated with each other which is referred as the problem of multicollinearity. Most of the present indices except principal component analysis (PCA) based method do not consider the effect of multicollinearity among the variables. For survey data, even though the PCA based indices are able to tackle the problem of multicollinearity but do not use survey weights and auxiliary information which leads to erroneous ranking of the survey units like households, districts, states, *etc.* Therefore, the present study proposes some new methods of index construction which are capable to incorporate the survey weights and auxiliary information available in the complex survey data as well as removes the effect of multicollinearity among variables.

Key words: Index; Multicollinearity; Survey weight; Auxiliary information; Complex surveys.

AMS Subject Classifications: 62K05, 05B05

1. Introduction

Indicators are helpful in recognizing patterns and attracting consideration regarding specific issues. A composite indicator is framed when singular indicators are assembled into a single index on the premise of a basic model. Composite indicators/index are much similar to mathematical or computational models which should ideally measure multi-dimensional concepts, which can't be caught by a single indicator alone, *e.g.*, intensity, industrialization, *etc.*

An index is a single measure to rank the units based on the multi-dimensional concepts which are presented in form of multivariable. Indices are quite useful to measure multi-dimensional concepts, which can't be caught by a single indicator alone like to summarize complex (elusive) or multi-dimensional processes into a single figure to benchmark the performance of countries, states, districts, *etc.* for policy formulation. The primary

Corresponding Author: Pradip Basak
Email: pradipbasak.99@gmail.com

role of index is to simplify otherwise complicated comparisons. Niti Aayog, Government of India develops composite indices like water management index, district hospital index, export preparedness index, India innovation index, multidimensional poverty index, school education quality index, SDG Index, state energy index and state health index. A comprehensive survey of different indicators of economic and social well-being has been provided by Sharpe (1999). The quantification of development efforts affected in various socio-economic fields was studied by constructing composite index of development based on information on fourteen important indicators by Narain *et al.,* (1991). The economic growth of China using social indicators has been estimated by Klein and Ozmucur (2003). Potential agro-forestry areas using Objective Analytic Hierarchy Process was identified by Ahmad *et al.,* (2003). Economic development in Karnataka, hilly states and Jammu and Kashmir was evaluated by Narain *et al.,* (2003, 2004, 2005). Livelihood index for different agro-climatic zones of India was developed by Rai *et al.,* (2008). The food insecurity in urban India was reported by developing the food insecurity index by Athreya *et al.,* (2010). The Human Development Index (HDI) developed by United Nations Development Programme is the geometric mean of the three-dimension indices *i.e.*, Health, Education and Income (Human Development Report, 2016).

Therefore, in most of the situations, composite index are based on simple or weighted average method which does not consider the effect of multicollinearity among the indicator variables that are used for index construction. PCA based index accounts for the effect of multicollinearity among the indicator variables through the eigen values and eigen vectors derived from the variance-covariance matrix using maximum likelihood or ordinary least squares methods of estimation. Dahal (2007) developed soil quality index by using PCA in which all those principal components (PCs) for which the eigen value is greater than one are retained. Agricultural development index was developed by Kumar (2008) using the principal component technique. Medical expenditure panel survey from 1996 to 2011 was used to develop principal component-based index by Chao and Wu (2017). Water poverty index was developed by Senna *et al.,* (2019) using PCA.

However, the above PCA based index methods are based on the assumption that sample elements, on which the indicator variables are measured, are independent and identically distributed. This assumption of independence holds good if the data are collected through simple random sampling with replacement. However, it does not hold good for other sampling designs where the inclusion probability and survey weight are attached with each sampling unit and hence the above PCA based index methods lacks representativeness of the population when the data is collected with complex survey design. Now a day, most of the survey designs are complex in nature involving stratification, unequal probabilities of selection, clustering, multi-stages, multi-phases and auxiliary information. In case of large-scale surveys, stratified multistage sampling design is widely used where the units in a stratum are relatively homogenous which violates the assumption of independence of sample elements. Any deviation from independence assumption leads to erroneous estimation of variance covariance matrix which in turn leads to erroneous estimation of eigenvalues and eigenvectors, and thereby resulting in poor PCA based index. Therefore, in case of complex survey data there is a need to develop PCA based index using survey weights to tackle the problem of representativeness of population and auxiliary information which leads to development of efficient indices.

2. Methodology

2.1. Estimators of variance-covariance matrix

Let us consider a finite population $U = (1, 2, \ldots, k, \ldots, N)$ of size N units having l subpopulations/blocks/states such that the h^{th} subpopulation has N_h units and $\sum_{h=1}^{l} N_h = N$, $(h = 1, 2, \ldots, l)$. Let s be a probabilistic sample of size n drawn from this population such that $\sum_{h=1}^{l} n_h = n$ where n_h is the number of units belongs to the h^{th} sub-population with assumption that $n_h \neq 0$ and d_{hi} denotes the survey weight associated with i^{th} unit of the sample in h^{th} subpopulation such that $\sum_{h=1}^{l} \sum_{i=1}^{n_h} d_{hi} = 1$. Let $\mathbf{y} = (y_{1,}y_{2,}..., y_q)'$ and $\mathbf{x} = (x_{1,}x_{2,}..., x_q)'$ be the p and q set of standardised indicators and auxiliary variables respectively. Let, $\mathbf{y}_{hi} = (y_{hi1,}y_{hi2,}..., y_{hip})'$ and $\mathbf{x}_{hi} = (x_{hi1,}x_{hi2,}..., x_{hip})'$ be values of the variables $\mathbf{y}$ and $\mathbf{x}$ corresponding to i^{th} sample unit of h^{th} sub-population where, $h = 1, 2, \ldots, l$ and $i = 1, 2, \ldots, n_h$. The ordinary least squares estimator of variance-covariance matrix, Σ_{yy} is given as

$$\hat{\sum}_{yy} = \mathbf{V}_{yys} = (n-1)^{-1} \sum_h \sum_i (\mathbf{y}_{hi} - \bar{\mathbf{y}}_s)(\mathbf{y}_{hi} - \bar{\mathbf{y}}_s)^T \tag{1}$$

where, $\bar{\mathbf{y}}_s = \sum_h \sum_i \mathbf{y}_{hi}/n$.

Following Skinner *et al.,*(1986) and Smith & Holmes (1989), the survey-weighted estimator of $\sum_{yy}$ is given by

$$\hat{\sum}_{yyw} = \mathbf{V}^*_{yys} = \sum_h \sum_i d_{hi} \mathbf{y}_{hi} \mathbf{y}^T_{hi} - \mathbf{y}^*_s \mathbf{y}^{*T}_s \tag{2}$$

where, $\bar{\mathbf{y}}^*_s = \sum_h \sum_i d_{hi} \mathbf{y}_{hi}$, and in the presence of auxiliary information $\mathbf{x}$, unweighted regression estimator is given by

$$\hat{\sum}_{yyr} = \mathbf{V}_{yys} + \mathbf{b}_{yx}(\sum_{xx} - V_{xxs})\mathbf{b}^T_{yx} \tag{3}$$

where,

$$\mathbf{b}_{yx} = \mathbf{V}_{xys} \mathbf{V}^{-1}_{xxs},$$
$$\mathbf{V}_{xxs} = (n-1)^{-1} \sum_h \sum_i (\mathbf{x}_{hi} - \bar{\mathbf{x}}_s)(\mathbf{x}_{hi} - \bar{\mathbf{x}}_s)^T,$$
$$\mathbf{V}_{xys} = (n-1)^{-1} \sum_h \sum_i (\mathbf{x}_{hi} - \bar{\mathbf{x}}_s)(\mathbf{y}_{hi} - \bar{\mathbf{y}}_s)^T.$$

In the case of survey data with auxiliary information, following Skinner *et al.,* (1986) and Smith & Holmes (1989), survey-weighted regression estimator is given by

$$\hat{\sum}_{yywr} = \mathbf{V}^*_{yys} + \mathbf{b}_{yxw}(\sum_{xx} - \mathbf{V}^*_{xxs})\mathbf{b}^T_{yxw} \tag{4}$$

where,

$$\mathbf{b}_{yxw} = \mathbf{V}^*_{xys} \mathbf{V}^{*-1}_{xxs},$$

$$\mathbf{V}^*_{xxs} = \sum_h \sum_i d_{hi}\mathbf{x}_{hi}\mathbf{x}^T_{hi} - \bar{\mathbf{x}}^*_s\bar{\mathbf{x}}^{*T}_s,$$

$$\mathbf{V}^*_{xys} = \sum_h \sum_i d_{hi}\mathbf{x}_{hi}\mathbf{y}^T_{hi} - \bar{\mathbf{x}}^*_s\bar{\mathbf{y}}^{*T}_s.$$

2.2. Methodology of proposed indices

Let us assume that $\hat{\Sigma}_{yy}$ is a real positive definite matrix. Let, the non-zero eigenvalues of $\hat{\Sigma}_{yy}$ are $\lambda_1 > \lambda_2 > \lambda_3 \ldots > \lambda_{\mathrm{p}}$ and the corresponding eigen vectors are $\gamma_1, \gamma_2, \gamma_3, \ldots, \gamma_{\mathrm{p}}$. For distinct λ_{j}'s $(j = 1, 2, 3, \ldots, p)$, an orthogonal matrix of order pxp can be formed as

$$\Gamma = [\gamma_1, \gamma_2, \gamma_3, \ldots, \gamma_{\mathrm{p}}], \tag{5}$$

such that, $\hat{\Sigma}_{yy} = \Gamma A \Gamma^T$, where $\mathrm{A} = \mathrm{diag}(\lambda_1, \lambda_2, \lambda_3, \ldots, \lambda_{\mathrm{p}}) = \Gamma^T \hat{\Sigma}_{yy} \Gamma$. Now let us consider an orthogonal transformation of $\mathbf{y}$ such that

$$\mathbf{P} = \Gamma^T \mathbf{y} \tag{6}$$

where $\mathrm{PC}_1, \mathrm{PC}_2, \ldots, \mathrm{PC}_{\mathrm{p}}$ are the p components of $\mathbf{P}$ and are called as PCs. The composite index corresponding to i^{th} sample unit of h^{th} sub-population is given as

$$C_{hi} = \frac{\sum_{j=1}^{p} \lambda_j PC_{hij}}{\sum_{j=1}^{p} \lambda_j} \tag{7}$$

where the PC_{hij}'s are principal component scores of j^{th} variable corresponding to the i^{th} sample unit of h^{th} sub-population $\forall\ h = 1, 2, \ldots, l; i = 1, 2, \ldots, n_h; j = 1, 2, \ldots, p$. The average of C_{hi}'s within h^{th} sub-population gives the composite index value for h^{th} sub-population as

$$C_h = \sum_{i=1}^{n_h} C_{hi} \big/ n_h. \tag{8}$$

The composite index values of sub-populations are re-scaled by using the following formula as

$$CI_h = \frac{C_h - \min(C_h)}{\max(C_h) - \min(C_h)}. \tag{9}$$

The ranking of l sub-populations is done based on the re-scaled composite index values (CI_h). All the composite index values (CI_h) lie between 0 and 1, where one denotes the highest rank and zero denotes the lowest rank.

The existing PCA based index uses the non-zero eigenvalues derived from the ordinary least squares estimator of variance-covariance matrix, $\hat{\Sigma}_{yy}$. Here, indices are proposed based

on survey weighted estimator, unweighted regression estimator and survey weighted regression estimator of variance-covariance matrix. The index that uses the non-zero eigenvalues derived from the survey weighted estimator of variance-covariance matrix, $\hat{\Sigma}_{yyw}$ is referred as the survey weighted PCA based index and it is used when the data are collected through complex survey designs in which the inclusion probability for all the units is not same. The index developed based on the non-zero eigenvalues derived from unweighted regression estimator of variance-covariance matrix, $\hat{\Sigma}_{yyr}$ is referred as the unweighted regression PCA based index and it is useful when auxiliary information is available in the data. One more index has been developed that uses the non-zero eigenvalues derived from survey weighted regression estimator of variance-covariance matrix, $\hat{\Sigma}_{yywr}$ and is referred as the survey-weighted regression PCA based index. This index is particularly useful when there is the presence of auxiliary information under complex survey designs.

3. Empirical evaluations

This Section summarizes the simulation studies conducted to evaluate the empirical performance of the developed indices. Two types of simulation studies, namely design based simulation and model-based simulation are considered. In case of design-based simulation, real survey dataset is used as a finite population. From this fixed population, repeated random samples are drawn. In the case of model-based simulation, at each simulation run a synthetic population data is first generated under the model and then a sample is drawn from this simulated population, and process is repeated several times. In the simulation studies, the following indices are considered

i) Unweighted PCA based index (denoted as PCA Index),

ii) Survey weighted PCA based index (denoted as SW-PCA Index),

iii) Unweighted regression PCA based index (denoted as REG-PCA Index), and

iv) Survey weighted regression PCA based index (denoted as SW-REG-PCA Index).

3.1. Design-based simulation

The household consumer expenditure survey data of NSS 68^{th} round is used for design based simulation study. The data of five states namely, Jammu and Kashmir, Orissa, Kerala, Sikkim and Jharkhand, and one union territory, *i.e.*, Andaman and Nicobar Islands have been considered for the study. The survey data of these five states and one union territory are considered as independent populations and then samples (10 % of the population) are drawn from each of these populations. The primary units of the survey are households. Therefore, within a state, sample size is allocated among the districts using proportional allocation and then from each of the districts, households are selected by simple random sampling without replacement (SRSWOR). Here, the variables considered are Cereals (Z1), Pulses and pulse products (Z2), Milk and milk products (Z3), Salt and sugar (Z4), Edible oil (Z5), Egg, Fish and meat (Z6), Vegetables (Z7), Fruits (fresh) (Z8), Spices (Z9), Beverages (Z10), Served processed food (Z11) and Packaged processed food (Z12). Following Smith and Holmes (1989), a new variable is created by summing up all the twelve variables which is considered as auxiliary variable.

The composite index values are computed for each households using different methods of index construction and the average of composite index values of all the households within a district is taken as the index value for that district. The index values computed for each of the districts within a state are compared with the composite index value of districts based on the population variance covariance matrix. The Monte Carlo simulation was run S=5000 times. Simulation studies are carried out in R software. The developed indices are evaluated by percentage relative root mean squared error (RRMSE), defined by

$$RRMSE(\widehat{\theta}) = \sqrt{\frac{1}{s}\sum_{s=1}^{S}\left[\sum_{i=1}^{k}\left(\frac{\widehat{\theta}_{si}-\theta_i}{\theta_i}\right)^2\right]} * 100 \tag{10}$$

where, $\widehat{\theta}_{si}$ is the sample index value for i^{th} district at s^{th} simulation run and θ_i is the population index value for i^{th} district. The values of the percentage relative root mean square error of different indices are reported in Table 1.

Table 1: Percentage relative root mean square error (RRMSE %) of different indices considered in the design-based simulation

State	PCA Index	SW-PCA Index	REG-PCA Index	SW-REG-PCA Index
Jammu & Kashmir	1242.00	879.14	1230.95	863.70
Orissa	363.18	362.21	357.90	352.35
Kerala	228.27	227.91	222.10	223.26
Andaman & Nicobar Islands	208.97	207.47	206.27	205.00
Sikkim	126.49	125.59	121.51	118.10
Jharkhand	466.21	452.80	432.25	424.95

From Table 1, it is clear that the proposed indices perform better than the unweighted PCA based index in terms of RRMSE. Among the proposed indices, SW-REG-PCA Index performs best followed by REG-PCA Index and SW-PCA Index. Since SW-REG-PCA Index utilises the auxiliary information as well as survey weights available through survey design, therefore, it performs best. However, for the state of Jammu & Kashmir, SW-PCA Index performs better than the REG-PCA Index. The Table 1 indicates that the proposed methodologies of indices development are efficient in comparison to the existing PCA based index method for complex survey designs.

3.2. Model based simulation

In model-based simulation, the methodology given by Smith & Holmes (1989) were followed where they have assumed the design variable Z as a sum of other variables. Thus, Z is a continuous variable to form population design groups such as strata to investigate the performance of different estimators of population variance covariance matrix for complex survey design. In this simulation study, an artificial population is generated using multivariate normal distribution $\mathbf{X} = \left(\mathbf{Y}^T, Z\right)^T$ having mean vector $\mathbf{u}_x$ and variance covariance matrix $\sum_{xx}$ satisfying the linearity and homoscedasticity assumptions. The vector $\mathbf{Y}$ comprises of

twelve variables and Z is the design variable which is the sum of all the twelve variables. The mean vector $\mathbf{u}_x$ and variance-covariance matrix $\sum_{xx}$ are estimated from the NSS 68th round household consumption expenditure survey data on the twelve set of variables, *i.e.*, Cereals (Z1), Pulses and pulse products (Z2), Milk and milk products (Z3), Salt and sugar (Z4), Edible oil (Z5), Egg, Fish and meat (Z6), Vegetables (Z7), Fruits (fresh) (Z8), Spices (Z9), Beverages (Z10), Served processed food (Z11) and Packaged processed food (Z12). A finite population of one lakh units is generated at each simulation run. Then the population is stratified into five strata, each having equal number of units based on the ordered z-values of the design variable Z.

Table 2: Mean of variables considered for population data generation

Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8	Y9	Y10	Y11	Y12
807.93	213.31	721.86	127.64	65.13	166.69	120.73	70.09	58.70	49.10	97.85	56.63

Table 3: Variance-covariance matrix of variables considered for population data generation

	Y1	Y2	Y3	Y4	Y5	Y6	Y7	Y8	Y9	Y10	Y11	Y12
Y1	263931	34252	76232	19551	11007	26131	19112	6828	6776	6280	3673	7081
Y2	34252	22964	41505	9415	4393	4090	4918	3003	2995	2189	1594	2829
Y3	76232	41505	620633	54028	14295	8103	18088	16563	8402	11975	10450	15495
Y4	19551	9415	54028	14862	3476	2290	3434	1884	1883	1560	896	1835
Y5	11007	4393	14295	3476	3891	3192	2433	973	1508	982	555	1601
Y6	26131	4090	8103	2290	3192	28266	4843	3285	2466	2187	3389	2588
Y7	19112	4918	18088	3434	2433	4843	6531	1726	1520	1361	924	1772
Y8	6828	3003	16563	1884	973	3285	1726	4607	887	1316	1763	1731
Y9	6776	2995	8402	1883	1508	2466	1520	887	1852	717	733	890
Y10	6280	2189	11975	1560	982	2187	1361	1316	717	2733	1582	1184
Y11	3673	1594	10450	896	555	3389	924	1763	733	1582	101389	2261
Y12	7081	2829	15495	1835	1601	2588	1772	1731	890	1184	2261	6203

Samples of size 2000 are selected from this population using SRSWOR within each stratum and allocated sample sizes in the strata are provided in Table 4.

Table 4: Allocation of sample size in the strata

Stratum	**1**	**2**	**3**	**4**	**5**
Sample size	600	300	200	300	600

Then the various indices are computed using this sample data. The Monte Carlo simulation was run S=5000 times. Simulation studies are carried out in R software. The developed indices are evaluated by the criterion of percentage relative root mean squared error (RRMSE, %), defined by

$$RRMSE(\widehat{\theta}) = \sqrt{\frac{1}{s}\sum_{s=1}^{S}\left[\sum_{i=1}^{k}\left(\frac{\widehat{\theta}_{si}-\theta_{si}}{\theta_{si}}\right)^2\right]} * 100 \tag{11}$$

where, $\widehat{\theta}_{si}$ and θ_{si} are the sample and population index values of i^{th} strata at s^{th} simulation run.

Table 5: Percentage relative root mean squared error (RRMSE, %) of different indices considered in model-based simulation

Indices	% RRMSE
Unweighted PCA based index	22.84
Survey weighted PCA based index	19.46
Unweighted regression PCA based index	19.20
Survey weighted regression PCA based index	19.03

Table 5 reports the performance of all the proposed indices obtained from the simulation study. From Table 5, it is clear that the SW-REG-PCA Index, which utilises the auxiliary information as well as survey weights, performs best followed by REG-PCA Index and SW-PCA Index. Therefore, all the developed indices performs better than the existing PCA based Index in terms of the criterion of RRMSE, %.

From the empirical evaluations in section 3, it is inferred that when sample is selected through complex survey design in which there is unequal selection probabilities of sample units, the indices that incorporate survey weights perform better in comparison to the traditional PCA based index method which is incapable to incorporate the survey weights. The proposed REG-PCA Index which is capable to incorporate the auxiliary information performs better than the traditional PCA based index method which does not utilize the auxiliary information even when it is available.

4. Conclusions

Most of the large-scale surveys conducted by different Government agencies, NGOs, research organisations and private firms use complex survey designs which involve unequal probabilities of selection, stratification, clustering, multistage, multiphase, nonresponse and other post stratification adjustments. Ignoring these aspects of complex survey data while constructing indices may lead to biased estimates and greater standard errors which leads to erroneous ranking of the survey units under consideration. Thus, it may result in erroneous inferences. Therefore, in the present study, different indices are developed which are capable to incorporate the survey weights and auxiliary information available in the complex survey data as well as removes the effect of multicollinearity among the index variables. The improved performance of the developed indices in comparison to the existing PCA based index have been demonstrated through simulation studies using both real and artificially generated data. Therefore, the developed indices will give better inferences in the case of complex survey data.

Acknowledgements

The authors are profoundly thankful to the editors and the anonymous referee for their valuable comments and suggestions which led to improvements in the article. The authors also gratefully acknowledge the valuable contributions made by Late Dr. Hukum Chandra, National Fellow and Principal Scientist, ICAR-Indian Agricultural Statistics Research Institute (ICAR-IASRI), New Delhi.

References

Ahmad, T., Singh, R., and Rai, A. (2003). Development of GIS based technique for identification of potential agro forestry areas. *Project Report*, ICAR-Indian Agricultural Statistics Research Institute, New Delhi, India.

Athreya, V. B., Rukmani, R., Bhavani, R. V., Anuradha, G., Gopinath, R., and Velan S. A. (2010). Report on the state of food insecurity in urban India. *Report*, M. S. Swaminathan Research Foundation.

Chao, Y. S. and Wu, C. J. (2017). Principal component-based weighted indices and a framework to evaluate indices: Results from the medical expenditure panel survey 1996 to 2011. *PLoS ONE*, **12**, e0183997.

Dahal, H. (2007). Factor analysis for soil test data: A methodological approach in environment friendly soil fertility management. *Journal of Agriculture and Environment*, **8**, 8–19.

Human Development Report (2016). *Human Development for Everyone*. United Nations Development Programme.

Klein, L. R. and Ozmucur, S. (2003). The estimation of China's economic growth. *Journal of Economic and Social Measurement*, **28**, 187–202.

Kumar, M. (2008). *A Study on Development Indices and their Sensitivity Analysis*. M.Sc. Thesis, ICAR-Indian Agricultural Research Institute, New Delhi, India.

Narain, P., Rai, S. C., and Sarup, S. (1991). Statistical evaluation on development on socio-economic front. *Journal of the Indian Society of Agricultural Statistics*, **43**, 329–345.

Narain, P., Sharma, S. D., Rai, S. C., and Bhatia, V. K. (2004). Estimation of socio-economic development in hilly States.*Journal of the Indian Society of Agricultural Statistics*, **58**, 126–135.

Narain, P., Sharma, S. D., Rai, S. C., and Bhatia, V. K. (2005). Dimension of socio-economic development in Jammu and Kashmir. *Journal of the Indian Society of Agricultural Statistics*, **59**, 243–250.

Narain, P., Sharma, S. D., Rai, S. C., and Bhatia, V. K. (2003). Evaluation of economic development at micro level in Karnataka. *Journal of the Indian Society of Agricultural Statistics*, **56**, 52–63.

Rai, A., Sharma, S. D., Sahoo, P. M., and Malhotra, P. K. (2008). Development of livelihood index for different agro-climatic zones of India. *Agricultural Economics Research Review*, **21**, 173–182.

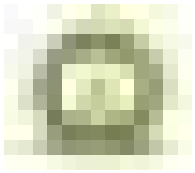
Senna, L. D., Maia, A. G., and Medeiros, J. D. F. (2019). The use of principal component analysis for the construction of the Water Poverty Index. *Brazilian Journal of Water Resources*, **24**, 1–14.

Sharpe, A. (1999). A Survey of Indicators of Economic and Social Well-being. *CSLS Research Reports*.

Skinner, C. J., Holmes, D. J., and Smith, T. M. F. (1986). The Effect of Sample Design on Principal Component Analysis. *Journal of the American Statistical Association*, **81**, 789–798.

Smith, T. M. F. and Holmes, D. J. (1989). *Multivariate Analysis*. In Analysis of Complex Surveys (Eds: Skinner, C. J., Holt, D. and Smith, T. M. F.), 165–190, New York: John Wiley and Sons.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 22, No. 1, 2024 (New Series), pp 83–100
http://www.ssca.org.in/journal



Distribution of Runs of Length Exactly k_1 Until a Stopping Time for Higher Order Markov Chain

Anuradha
Department of Statistics, Lady Shri Ram College for Women, University of Delhi.

Received: 02 November 2022; Revised: 05 June 2023; Accepted: 10 June 2023

Abstract

Consider an m^{th} order Markov chain $\{X_j : j \geq -m+1\}$ taking values in $\{\mathbf{0}, \mathbf{1}\}$. We set $R_i = 0$ for $i = 0, -1, \ldots, -l+1$. A l-look-back run of length k starting at i, R_i is defined inductively as a run of $\mathbf{1}$'s starting at i, provided that no l-look-back run of length k occurs, starting at time $i-1, i-2, \ldots, i-l$, *i.e.*, $R_i = \prod_{j=i-1}^{i-l}(1-R_j)\prod_{j=i}^{i+k-1} X_j$. We study the conditional distribution of the number of runs of length exactly k_1, till the r-th occurrence of the l-look-back run of length k where $k_1 \leq k-1$ and obtain the explicit expression of it's probability generating function. We establish that the number of runs can be written as sum of r independent random variables with the first term having a slightly different distribution. We further establish the strong law of large numbers for the number of runs of length exactly k_1.

Key words: Runs; Markov chain; Stopping time; Probability generating function; Strong Markov property: Strong law of large numbers.

AMS Subject Classifications: 60C05, 60E05, 60F05

1. Introduction

Theory of distributions of runs has been studied, since Feller (1968) introduced runs as an example of a renewal event. In recent years, this field has received a lot of interest among researchers. Many powerful techniques such as Markov embedding technique, method of conditional p.g.f.s *etc.* have been developed which enabled us to study new features of the distributions of various run statistics. For a more detailed discussion on the run statistics and its application, we refer the readers to Balakrishnan and Koutras [2002].

We consider an m-th order homogeneous $\{\mathbf{0}, \mathbf{1}\}$-valued Markov chain. Further, we assume that the initial condition $\{X_0 = x_0, X_{-1} = x_1, \ldots, X_{-m+1} = x_{m-1}\}$ is given to us. The state $\mathbf{1}$ can be thought as success in an experiment while $\mathbf{0}$ as failure. A run of length k is a consecutive occurrence of k successes. Anuradha (2022) introduced the l-look-back counting scheme for runs. In this scheme a run is counted starting at time i, if $X_i = X_{i+1} = \cdots = X_{i+k-1} = \mathbf{1}$, and no runs can be counted till the time point $i+l$.

Corresponding Author: Anuradha
Email: sarkar.anuradha@gmail.com

The next counting of run can start only from the time point $i + l + 1$. This mechanism is repeated every time a run is counted. In other words, if a run is counted starting at time i, there are k-consecutive successes from the time point i and no runs of length k has been counted which had the starting time as $i-1, i-2, \ldots, i-l$. Such a run will be referred as a l-look-back run of length k. Clearly, if $l = 0$, this counting of run matches exactly with the number of overlapping runs of length k, while if we set $l = k-1$, this counting results in the number of non-overlapping runs of length k. Aki and Hirano (2000) also defined a counting scheme which they referred as μ-overlapping counting. It should be noted that both these concepts match if we set $l = k - \mu - 1$.

The following example illustrates the practical usage of the l-look-back counting scheme for runs of length k. Consider an experiment of a drug administration where observations are taken every hour for the presence or absence (success or failure) of a particular symptom, say, fever exceeding a specified temperature. If we observe the presence of the symptom for k-successive time points, a drug has to be administered; however, as is the case with most drugs, once the drug is administered, we have to wait for l-hours for the next administration of the drug. But the process of the observation for the presence or absence of the symptom is continued as ever. In such a case, the number of administrations of the drug until time point n, is the number of l-look-back runs of length k up to time n.

Aki and Hirano (1994) studied the marginal distributions of failures, successes and success-runs of length less than k until the first occurrence of consecutive k successes where the underlying random variables are either *i.i.d.* or homogeneous Markov chain or binary sequence of order k. Aki and Hirano (1995) derived the joint distributions of failures, successes and success-runs for the same set-up. Hirano *et. al.* (1997) obtained the distributions of number of success-runs of a specific length for various counting schemes (*e.g.* runs of length k_1, overlapping runs of length k_1, non-overlapping runs of length k_1 *etc.*) until the first occurrence of the success-run of length k for a m-th order homogeneous Markov chain where $m \leq k_1 < k$. Uchida (1998) studied the joint distributions of the waiting time and the number of outcomes such as failures, successes and success-runs of length less than k for various counting schemes of runs for an m^{th} order homogeneous Markov chain. Chadjiconstantindis and Koutras (2001) also obtained the distribution of number of failures and successes in a waiting time problem.

In this paper, we study the distribution of runs of successes of exact length. A run of length exactly k can be described as an occurrence of a failure, followed by k consecutive successes, followed by another failure. The literature on runs of exact length is rather limited. This is indeed a difficult problem, specially when the underlying distribution of random variables has a dependent structure.

In recent years, the runs of exact length has found usage in very important areas. We site one such example here. The study of random sequences constitutes an important part of cryptography specially in the areas of challenge and response authentication systems, generation of digital signatures, and zero-knowledge protocols. Many protocols in cryptography depend on the assumption that the resulting ciphertext from a cipher (cryptographic algorithm) should appear to be as random as possible. Various tests are used for testing the randomness of such ciphertexts, which in turn help in deciding whether a given protocol leaks information or not. Doganaksoy *et. al.* (2015) developed three statistical randomness

tests based on runs of exact length and named them as runs of length one, runs of length two, and runs of length three tests respectively and showed that they work better than the traditional tests. However, the main challenge in the wider application of their work was that the distribution of the resulting statistic is not tractable when the (exact) length of run is large. In fact they could use only lengths 1, 2 and 3. Hence there is an imperative need to study the distribution, or at least find good approximation of the distribution, of runs of exact length for larger values of length, specially when the underlying random variables are not *i.i.d.* but have some dependent structure.

Since the study of the exact distribution of runs of exact length is complicated, it is prudent to find a simpler structure embedded in this set up. We study the conditional distribution of the number of runs of length exactly k_1, until a specified stopping time, namely the r^{th} occurrence of the l-look-back run of length k where $k_1 < k$. The study of distributions of runs until a stopping time brings out many salient features of various run statistic and establishes new connection between various discrete distributions. Indeed, our results exhibit an independence structure in the number of runs until the stopping time where we may explicitly write the distribution in terms of simpler random variables following Bernoulli and geometric distributions. (see Corollary 1 for details).

The novelty of our method lies in translating our problem into a first order homogeneous Markov chain. Indeed, we define a new first order Markov chain taking values in a finite set in such a way that the states of the new chain combines the last k_1 states of the previous chain (refer to the third section for exact definition). Further, the states of the original m-th order Markov chain may be recovered from the states of the newly defined Markov chain. This allows us to translate the problem in terms of the new Markov chain. For a simple Markov chain, the powerful results such as the strong Markov property can now be used to derive a recurrence relation between the probabilities. We now employ the method of conditional probability generating functions. We use this basic relation involving the probabilities to obtain a recurrence relation involving probability generating functions. This, in turn, provides a simple linear equation involving the generating function of the probability generating functions which can be solved to obtain its expression.

The explicit expression of the probability generating function implies that the distribution of the number of runs of length exactly k_1 until the stopping time has a renewal structure. Hence the number of runs until the stopping time splits into sum of independent random variables, which may be interpreted as arrival times in a renewal process. Further, we have shown that the arrival times are identical except the first arrival time. In other words, it admits a delayed renewal structure. We are also able to identify the arrival times through geometric and Bernoulli random variables. Thus we are able to approximate the number of runs of length exactly k_1 through simpler random variables.

We may apply our results to obtain an approximation of the number of runs of length exactly k_1 until time n in the following way: we choose some $k > k_1$ and find the number of non-overlapping runs of length k, *i.e.*, number of $(k-1)$-look-back runs of length k, until time n, say r. Clearly the total number of runs of length exactly k_1 until time n lies between the number of runs of length exactly k_1 until the r^{th} and $(r+1)^{th}$ occurrence of non-overlapping runs of length k respectively. Now we use our main result to compute the distribution of the number of runs of length exactly k_1 until this r^{th} as well as $(r+1)^{th}$ occurrences of

non-overlapping runs of length k. For large values of n, this works quite well. We use the Markov inequality and this method of approximation effectively to derive the strong law of large numbers (see Theorem 2) for the number of runs of length exactly k_1. We hope that the methods of approximation can, in future, be extended to obtain a central limit theorem as well as the law of iterated logarithm for the same.

In the next section, we give the important definitions and state the main Theorem and Corollary related to the distribution of the number of runs of length exactly k_1 until a stopping time, where $k_1 < k$. Section 3 is devoted towards formalizing the underlying set up for deriving the results. In Section 4, we prove the main theorem, while in Section 5, we prove the strong law for the number of runs.

2. Definitions and statement of results

Let $X_{-m+1}, X_{-m+2}, \ldots, X_0, X_1, \ldots$ be a sequence of stationary m-order $\{\mathbf{0}, \mathbf{1}\}$ valued Markov chain. It is assumed that the states of $X_{-m+1}, X_{-m+2}, \ldots, X_0$ are known, *i.e.*, we are given the initial condition $\{X_0 = x_0, X_{-1} = x_1, \ldots, X_{-m+1} = x_{m-1}\}$.

To make things formal, for any $i \geq 0$, define $C_i = \{0, 1, \ldots, 2^i - 1\}$. It is clear that C_i and $\{\mathbf{0}, \mathbf{1}\}^i$ can be identified easily by the mapping $x = (x_0, x_1, \ldots, x_{i-1}) \longrightarrow \sum_{j=0}^{i-1} 2^j x_j$. Since, $\{X_n : n \geq -m + 1\}$ is m^{th} order Markov chain, we have, for any $n \geq 0$,

$$p_x = \mathbb{P}(X_{n+1} = 1 | X_n = x_0, X_{n-1} = x_1, \ldots, X_{n-m+1} = x_{m-1}) \tag{1}$$

where $x = \sum_{j=0}^{m-1} 2^j x_j \in C_m$. Consequently, we have $q_x = \mathbb{P}(X_{n+1} = 0 | X_n = x_0, X_{n-1} = x_1, \ldots, X_{n-m+1} = x_{m-1}) = 1 - p_x$. We assume that $0 < p_x < 1$ for all $x \in C_i$. One particular case will be of importance in our study, when $\{X_n = 1, X_{n-1} = 1, \ldots, X_{n-m+1} = 1\}$. In our notation, this condition will become $x = \sum_{j=0}^{m-1} 2^j 1 = 2^m - 1$. Thus, using (1), for all $n \geq 0$, we have

$$p_{2^m-1} = \mathbb{P}(X_{n+1} = 1 | X_n = 1, X_{n-1} = 1, \ldots, X_{n-m+1} = 1) = 1 - q_{2^m-1}.$$

Definition 1: (**l-look-back run**) Fix two integers $k \geq 1$ and $1 \leq l \leq k - 1$. We set $R_i(k, l) = 0$ for $i = 0, -1, \ldots, -l + 1$ and for any $i \geq 1$, define inductively,

$$R_i(k, l) = \prod_{j=i-l}^{i-1} (1 - R_j(k, l)) \prod_{j=i}^{i+k-1} X_j. \tag{2}$$

If $R_i(k, l) = 1$, we say that an *l-look-back run of length k* has been recorded which started at time i.

It should be noted that for an l-look-back run to start at the time point i, we need to look back at the preceding l many time points, *i.e.*, $i - 1$ to $i - l$, none of which can be the starting point of an l-look-back run of length k.

Next we define the stopping times where the r-th occurrence of l-look-back run of length k is completed.

Definition 2: For $r \geq 1$, the stopping time $\tau_r(k,l)$ be the (random) time point at which the r-th occurrence of l-look-back run of length k is completed. In other words,

$$\tau_r(k,l) = k - 1 + \inf\{n : \sum_{i=1}^{n} R_i(k,l) = r\}. \tag{3}$$

Now we define the runs of length exactly k.

Definition 3: When $k(\geq 1)$ consecutive successes, either occur at the beginning of the sequence or end of the sequence or bordered on both sides by failures, contribute towards the counting of a run then we call it run of length exactly k.Note that when there are more than k consecutive successes then it is not counted as run of length exactly k.

We may represent this mathematically as follows:

$$\epsilon_i(k) = \begin{cases} \prod_{j=1}^{k} X_j(1 - X_{k+1}) & \text{if } i = 1 \\ (1 - X_{i-1}) \prod_{j=i}^{i+k-1} X_j(1 - X_{i+k}) & \text{if } 1 < i < n - k + 1 \\ (1 - X_{n-k}) \prod_{j=n-k+1}^{n} X_j & \text{if } i = n - k + 1. \end{cases}$$

Note here that $\epsilon_i(k) = 1$ if and only if a run of length exactly k starts at time point i. Now, we define the total number of runs of length exactly k by

$$N_n(k) = \sum_{i=1}^{n} \epsilon_i(k). \tag{4}$$

In this paper, we study the number of runs of length exactly k till the stopping time $\tau_r(k,l)$ (see Definition (2)). Fix any constant $k_1 \leq k$. For each $r \geq 1$, we define the random variable

$$N_r(= N_r(k_1)) := N_{\tau_r(k,l)}(k_1) = \sum_{i=1}^{\tau_r(k,l)} R_i(k_1) \tag{5}$$

as the number of runs of length exactly k_1 until the stopping time $\tau_r(k,l)$.

Before we proceed, we present an example to facilitate the understanding. Consider the following sequence of **0**'s and **1**'s of length 20

11010111011111011101.

For $k = 3$ and $l = 1$, it should be noted that, $R_6(3,1) = R_{10}(3,1) = R_{12}(3,1) = R_{16}(3,1) = 1$, while for other values of i, $R_i(3,1) = 0$. Thus, $\tau_1(3,1) = 8, \tau_2(3,1) = 12, \tau_3(3,1) = 14$ and $\tau_4(3,1) = 18$. For $k_1 = 2$, the number of runs of length exactly k_1 are given by $N_1 = N_2 = N_3 = N_4 = 1$ respectively.

Let us define the probability generating function of N_r, *i.e.*,

$$\zeta_r(s; k_1) := \sum_{n=0}^{\infty} \mathbb{P}(N_r = n)s^n. \tag{6}$$

Theorem 1: For any initial condition $x \in C_i$, $k_2 = k - k_1 > 0$ and $k_1 \geq m$, the probability generating function of N_r is given by

$$\zeta_r(s;k_1) = \left[\frac{\left(p_{2^m-1}\right)^{k_2}}{q_{2^m-1} + \left(p_{2^m-1}\right)^{k_2} - q_{2^m-1}s}\right]\left[\left(p_{2^m-1}\right)^{l+1} + \frac{\left(p_{2^m-1}\right)^{k_2}}{q_{2^m-1} + \left(p_{2^m-1}\right)^{k_2} - q_{2^m-1}s}\left(1 - \left(p_{2^m-1}\right)^{l+1}\right)\right]^{r-1}.$$

Theorem 1 provides a useful representation of N_r in terms of Bernoulli and geometric random variables when $k_2 > 0$. Let us set,

$$p_E = \frac{\left(p_{2^m-1}\right)^{k_2}}{1 - \sum_{t=1}^{k_2-1}\left(p_{2^m-1}\right)^t q_{2^m-1}} = \frac{\left(p_{2^m-1}\right)^{k_2}}{q_{2^m-1} + \left(p_{2^m-1}\right)^{k_2}}. \tag{7}$$

Corollary 1: Let $\{G_i : i = 1, \ldots, r\}$ and $\{B_i : i = 1, \ldots, r\}$ be two independent sets of random variables with each G_i having a geometric distribution (taking values in $\{0, 1, \ldots, \}$) with parameter p_E and each B_i having a Bernoulli distribution with parameter $\left(1 - \left(p_{2^m-1}\right)^{l+1}\right)$, then we have

$$N_r \stackrel{d}{=} G_1 + \sum_{i=2}^{r} G_i B_i. \tag{8}$$

Indeed, it is easy to see that the probability generating function of G_i, for $i \geq 1$, is given by

$$\frac{\left(p_{2^m-1}\right)^{k_2}}{q_{2^m-1} + \left(p_{2^m-1}\right)^{k_2} - q_{2^m-1}s}$$

and the probability generating function of B_i, for $i \geq 1$, is given by

$$\left(p_{2^m-1}\right)^{l+1} + s\left(1 - \left(p_{2^m-1}\right)^{l+1}\right).$$

Therefore, the probability generating function of $G_i B_i$ is given by

$$\left(p_{2^m-1}\right)^{l+1} + \frac{\left(p_{2^m-1}\right)^{k_2}}{q_{2^m-1} + \left(p_{2^m-1}\right)^{k_2} - q_{2^m-1}s}\left(1 - \left(p_{2^m-1}\right)^{l+1}\right). \tag{9}$$

From the independence of G_i and B_i for $i \geq 1$, the corollary easily follows.

When $k_2 = 0$, *i.e.*, $k = k_1$, we can obtain the the probability generating function, but it is difficult to identify the exact distribution (see Section 4).

The delayed renewal structure of the number of runs of exact length until the stopping time, observed in equation (8), can be used for approximating the original distribution when

the number of trials are large. Indeed we may obtain a strong law for the number of runs of exact length using this.

Let us set $k = k_1 + 1$ and $l = k - 1 = k_1$. Then, the expectation of G_1B_1 can be easily computed from the expression of the probability generating function in (9). Indeed, it is given by

$$\mu_1 = \frac{q_{2^m-1}}{p_{2^m-1}}\Big(1 - \Big(p_{2^m-1}\Big)^{k_1+1}\Big). \tag{10}$$

We will further define a constant μ. Let S be the first time when k successive heads have occurred given the initial condition of k successive heads. In section 5, We will show that S is finite with probability 1. Further, its expectation is also finite. We denote

$$\mu = \mathbb{E}(S). \tag{11}$$

Theorem 2: For any initial condition $x \in C_i$ and $k_1 \geq m$, we have

$$\frac{1}{n}N_n(k_1) \to \frac{\mu_1}{\mu}$$

as $n \to \infty$ with probability 1.

3. Formal set-up

In this section, we outline the basic set up which will be used in the subsequent section to establish the results. Let us define two functions $f_0, f_1 : C_{k_1} \to C_{k_1}$ by

$$f_1(x) = 2x + 1 \pmod{2^{k_1}} \text{ and } f_0(x) = 2x \pmod{2^{k_1}}.$$

Further define a projection $\theta_m : C_{k_1} \to C_m$ by $\theta_m(x) = x \pmod{2^m}$. Now, set $X_{-m} = X_{-m-1} = \cdots = X_{-k_1+1} = 0$. Define a sequence of random variables $\{Y_n : n \geq 0\}$ as follows:

$$Y_n = \sum_{j=0}^{k_1-1} 2^j X_{n-j}.$$

Since $X_i \in \{\mathbf{0}, \mathbf{1}\}$ for all i, Y_n assumes values in the set C_{k_1}. Further, the random variables X_n's are stationary and forms a m^{th} order Markov chain, hence we have that $\{Y_n : n \geq 0\}$ is a homogeneous Markov chain with transition matrix given by

$$\mathbb{P}(Y_{n+1} = y | Y_n = x) = \begin{cases} p_{\theta_m(x)} & \text{if } y = f_1(x) \\ 1 - p_{\theta_m(x)} & \text{if } y = f_0(x) \\ 0 & \text{otherwise.} \end{cases}$$

It should be noted that Y_n is even if and only if $X_n = 0$. This motivates us to define the function $\kappa : C_{k_1} \to \{0, 1\}$ by

$$\kappa(x) = \begin{cases} 1 & \text{if } x \text{ is odd} \\ 0 & \text{if } x \text{ is even.} \end{cases}$$

Therefore, $\kappa(Y_n) = 1$ if and only if $X_n = 1$. Hence, the definition of l-look-back run can be described in terms of Y_n's as

$$R_i(k,l) = \prod_{j=i-l}^{i-1} (1 - R_j(k,l)) \prod_{j=i}^{i+k-1} \kappa(Y_j).$$

Let us fix any initial condition $x \in C_m$. We denote the probability measure governing the distribution of $\{Y_n : n \geq 1\}$ with $Y_0 = x \in C_k$ by $\mathbb{P}_x$. Since we have set $X_{-m} = X_{-m-1} = \cdots = X_{-k+1} = 0$, we have $Y_0 = x$.

In order to obtain the recurrence relation for the probabilities, we will condition the process after the first occurrence of the run of length k_1. Therefore, we consider the stopping time T when the first occurrence of a run of length k_1 ends, *i.e.*, when we observe k_1 successes consecutively for the first time. More precisely, define

$$T := \inf\{i \geq k_1 : \prod_{j=i-k_1+1}^{i} X_j = 1\}. \tag{12}$$

We would like to translate the above definition in terms of Y_i's. It must be the case that when T occurs, last k_1 trials have resulted in success, which may be described by $\kappa(Y_j) = 1$ for $j = i - k_1 + 1$ to i. Therefore, Y_T must equal $2^{k_1} - 1$. Since this is the first occurrence and this has not happened earlier. So, T can be better described as

$$T = \inf\{i \geq k_1 : Y_i = 2^{k_1} - 1\}, \tag{13}$$

i.e., the first visit of the chain to the state $2^{k_1} - 1$ after time $k_1 - 1$. Now, we note that $\{Y_n : n \geq 0\}$ is a Markov chain with finite state space. Further, since $0 < p_u < 1$ for $u \in C_m$, this is an irreducible chain; hence, it is positive recurrent. So we must have $\mathbb{P}_x(T < \infty) = 1$. We observe that when the first occurrence of k consecutive successes happens, then k_1 consecutive successes must have occured previously since $k_1 \leq k$. Therefore, we have $P_x(T < \tau_1(k,l)) = 1$.

4. Number of runs of exact length until stopping time

First we establish the basic recurrence relation which is central to our result. Define the probability $g_r^{(x)}(n)$ by

$$g_r^{(x)}(n) := \mathbb{P}_x(N_r = n) \tag{14}$$

for $n \in \mathbb{Z}$. We note that since $N_r \geq 0$, $\mathbb{P}_x(N_r = n) = 0$ for $n < 0$. Our first task is to show that $g_r^{(x)}(n)$ is independent of x.

Theorem 3: Suppose that $k_2 = k - k_1 > 0$. For any $x \in C_{k_1}$ and any $n \geq 0$, we have

$$\begin{aligned} g_1^{(x)}(n) = q_{2^m-1} g_1^{(2^m-2)}(n-1) + \sum_{t=1}^{k_2-1} q_{2^m-1} \left(p_{2^m-1}\right)^t g_1^{(2^m-2)}(n) \\ + \left(p_{2^m-1}\right)^{k_2} \mathbb{I}_n(0) \end{aligned} \tag{15}$$

where $\mathbb{I}_{u_1}(u_2)$ is the indicator function defined by

$$\mathbb{I}_{u_1}(u_2) = \begin{cases} 1 & \text{if } u_1 = u_2 \\ 0 & \text{otherwise.} \end{cases}$$

Proof: When $k_2 = k - k_1 > 0$ and $r = 1$, we have

$$\begin{aligned} g_1^{(x)}(n) &= \mathbb{P}_x(N_1 = n) = \mathbb{P}_x(N_1 = n, Y_{T+1} = 2^{k_1} - 2) \\ &+ \sum_{t=1}^{k_2-1} \mathbb{P}_x(N_1 = n, Y_{T+1} = 2^{k_1} - 1, \ldots, Y_{T+t} = 2^{k_1} - 1, Y_{T+t+1} = 2^{k_1} - 2) \\ &+ \mathbb{P}_x(N_1 = n, Y_{T+1} = 2^{k_1} - 1, Y_{T+2} = 2^{k_1} - 1, \ldots, \\ &\qquad Y_{T+k_2-1} = 2^{k_1} - 1, Y_{T+k_2} = 2^{k_1} - 1). \end{aligned} \tag{16}$$

We simplify the terms in the summation first. For any $1 \le t \le k_2 - 1$, we have,

$$\begin{aligned} &\mathbb{P}_x(N_1 = n, Y_{T+1} = 2^{k_1} - 1, Y_{T+2} = 2^{k_1} - 1, \ldots, \\ &\qquad Y_{T+t} = 2^{k_1} - 1, Y_{T+t+1} = 2^{k_1} - 2) \\ &= \mathbb{P}_x(N_1 = n \mid Y_{T+1} = 2^{k_1} - 1, Y_{T+2} = 2^{k_1} - 1, \ldots, \\ &\qquad Y_{T+t} = 2^{k_1} - 1, Y_{T+t+1} = 2^{k_1} - 2) \\ &\times \mathbb{P}_x(Y_{T+1} = 2^{k_1} - 1, Y_{T+2} = 2^{k_1} - 1, \ldots, \\ &\qquad Y_{T+t} = 2^{k_1} - 1, Y_{T+t+1} = 2^{k_1} - 2). \end{aligned} \tag{17}$$

The second term in (17) can be written as

$$\begin{aligned} &\mathbb{P}_x(Y_{T+1} = 2^{k_1} - 1, \ldots, Y_{T+t} = 2^{k_1} - 1, Y_{T+t+1} = 2^{k_1} - 2) \\ &= \mathbb{P}_x(Y_{T+t+1} = 2^{k_1} - 2 \mid Y_{T+1} = 2^{k_1} - 1, \ldots, Y_{T+t} = 2^{k_1} - 1) \\ &\times \prod_{j=1}^{t} \mathbb{P}_x(Y_{T+j} = 2^{k_1} - 1 \mid Y_{T+1} = 2^{k_1} - 1, \ldots, Y_{T+j-1} = 2^{k_1} - 1). \end{aligned}$$

Now, for any $1 \le j \le t, T + j - 1$ is also a stopping time. We denote by $\mathcal{F}_{T+j-1}$, the σ-algebra generated by the process Y_n up to the stopping time $T + j - 1$, and by $\mathcal{F}_{(T+j-1)+}$, the σ-algebra generated by the process after the stopping time $T + j - 1$. Clearly, $\{Y_{T+1} = 2^{k_1} - 1, \ldots, Y_{T+j-1} = 2^{k_1} - 1\} \in \mathcal{F}_{T+j-1}$ and $\{Y_{T+j} = 2^{k_1} - 1\} \in \mathcal{F}_{(T+j-1)+}$. Thus, using the strong Markov property, we can write

$$\begin{aligned} &\mathbb{P}_x(Y_{T+j} = 2^{k_1} - 1 \mid Y_{T+1} = 2^{k_1} - 1, \ldots, Y_{T+j-1} = 2^{k_1} - 1) \\ &= \mathbb{P}_{Y_{T+j-1}}(Y_{T+j} = 2^{k_1} - 1) = \mathbb{P}_{2^{k_1}-1}(Y_1 = 2^{k_1} - 1) = p_{2^m-1}. \end{aligned} \tag{18}$$

A similar argument shows that

$$\mathbb{P}_x(Y_{T+t+1} = 2^{k_1} - 2 \mid Y_{T+1} = 2^{k_1} - 1, \ldots, Y_{T+t} = 2^{k_1} - 1) = q_{2^m-1}. \tag{19}$$

For the first term in (17), we note that $T + t + 1$ is also a stopping time and $\{Y_{T+1} = 2^{k_1} - 1, \ldots, Y_{T+t} = 2^{k_1} - 1, Y_{T+t+1} = 2^{k_1} - 2\} \in \mathcal{F}_{T+t+1}$. Since $Y_T = 2^{k_1} - 1$, we must have either $X_{T-k_1} = 0$ and $X_{T-j} = 1$ for $j = 0, 1, \ldots, k_1 - 1$ or $T = k_1$. Further, since $Y_{T+j} = 2^{k_1} - 1$ for $j = 1, \ldots, t$ and $Y_{T+t+1} = 2^{k_1} - 2$, we also have $X_{T+j} = 1$ for $j = 0, 1, \ldots, t$ and $X_{T+t+1} = 0$. Therefore, we have a sequence of $\mathbf{1}'$s of length $k_1 + t$ with $t > 0$ which

contributes to 0 runs of length exactly k_1 and since there are no runs of length k_1 before T, by the very definition of T, we have that the number of runs of length exactly k_1 up to time $T+t+1$ is 0. Since $t \leq k_2 - 1$, we have that $T + t + 1 < \tau_1(k,l)$. Let us define $Y_i' = Y_{i+T+t+1}$ for $i \geq 0$. Now, using the strong Markov property, we have that $\{Y_i' : i \geq 0\}$ is a homogeneous Markov chain with same transition matrix as that of $\{Y_i : i \geq 0\}$ with $Y_0' = 2^{k_1} - 2$. Now, define $\tau_1'(k,l)$ as the stopping time for the process $\{Y_i' : i \geq 0\}$. From the above discussion, we have that $\tau_1(k,l) = T + t + 1 + \tau_1'(k,l)$. Further, if we define, N_1' as the number runs of length exactly k_1 up to time $\tau_1'(k,l)$ for the process $\{Y_i' : i \geq 0\}$, we must have that $N_1' = n$. Therefore, we have,

$$\begin{aligned}&\mathbb{P}_x(N_1 = n \mid Y_{T+1} = 2^{k_1} - 1, \ldots, Y_{T+t} = 2^{k_1} - 1, Y_{T+t+1} = 2^{k_1} - 2)\\ &= \mathbb{P}_{(2^m-2)}(N_1' = n) = g_1^{(2^m-2)}(n). \end{aligned} \tag{20}$$

Now, the first term in (16) can be written as

$$\begin{aligned}&\mathbb{P}_x(N_1 = n, Y_{T+1} = 2^{k_1} - 2)\\ &= \mathbb{P}_x(N_1 = n \mid Y_{T+1} = 2^{k_1} - 2, Y_T = 2^{k_1} - 1)\mathbb{P}_x(Y_{T+1} = 2^{k_1} - 2 \mid Y_T = 2^{k_1} - 1)\\ &= q_{2^m-1}\mathbb{P}_x(N_1 = n \mid Y_{T+1} = 2^{k_1} - 2, Y_T = 2^{k_1} - 1). \end{aligned} \tag{21}$$

The arguments leading to equation (20) can now be repeated to conclude that

$$\mathbb{P}_x(N_1 = n \mid Y_{T+1} = 2^{k_1} - 2, Y_T = 2^{k_1} - 1) = \mathbb{P}_{(2^m-2)}(N_1 = n-1) = g_1^{(2^m-2)}(n-1). \tag{22}$$

Using the equivalent characterisation of T (see equation (13)) we note that $Y_T = 2^{k_1} - 1$ with probability 1. Hence, for the last term in (16) becomes

$$\begin{aligned}&\mathbb{P}_x(N_1 = n, Y_{T+1} = 2^{k_1} - 1, \ldots, Y_{T+k_2-1} = 2^{k_1} - 1, Y_{T+k_2} = 2^{k_1} - 1)\\ &= \mathbb{P}_x(N_1 = n, Y_T = 2^{k_1} - 1, Y_{T+1} = 2^{k_1} - 1, \ldots, Y_{T+k_2-1} = 2^{k_1} - 1, Y_{T+k_2} = 2^{k_1} - 1)\\ &= \prod_{j=1}^{k_2} \mathbb{P}_x(Y_{T+j} = 2^{k_1} - 1 \mid Y_T = 2^{k_1} - 1, Y_{T+1} = 2^{k_1} - 1, \ldots, Y_{T+j-1} = 2^{k_1} - 1)\\ &\quad \times \mathbb{P}_x(N_1 = n \mid Y_T = 2^{k_1} - 1, Y_{T+1} = 2^{k_1} - 1, \ldots, Y_{T+k_2-1} = 2^{k_1} - 1, Y_{T+k_2} = 2^{k_1} - 1)\\ &= \left(p_{2^m-1}\right)^{k_2}\mathbb{P}_x(N_1(k_1) = n \mid Y_{T+1} = 2^{k_1} - 1, \ldots, Y_{T+k_2-1} = 2^{k_1} - 1, Y_{T+k_2} = 2^{k_1} - 1). \end{aligned}$$

Note that given $\{Y_{T+1} = 2^{k_1} - 1, \ldots, Y_{T+k_2-1} = 2^{k_1} - 1, Y_{T+k_2} = 2^{k_1} - 1\}$, we have $\tau_1(k,l) = T + k_2$. Therefore, $N_1 = n$ if and only if $n = 0$. In other words, $\mathbb{P}_x(N_1 = n \mid Y_{T+1} = 2^{k_1} - 1, \ldots, Y_{T+k_2-1} = 2^{k_1} - 1, Y_{T+k_2} = 2^{k_1} - 1) = \mathbb{I}_n(0)$ where $\mathbb{I}$ is the indicator function as defined in the statement of the Theorem.

Thus combining the above expression with the equations (16) - (22), we have

$$g_1^{(x)}(n) = q_{2^m-1}g_1^{(2^m-2)}(n-1) + \sum_{t=1}^{k_2-1} q_{2^m-1}\left(p_{2^m-1}\right)^t g_1^{(2^m-2)}(n) + \left(p_{2^m-1}\right)^{k_2}\mathbb{I}_n(0).$$

This completes the proof. □

We note that the right hand side of (15) does not involve the initial condition $x \in C_m$. Therefore $g_1^{(x)}(n)$ must be independent of x. So, we will drop x and denote the above probability by $g_1(n)$, *i.e.*,

$$g_1(n) = \mathbb{P}_x(N_1 = n).$$

Hence, we may rewrite the equation (15) as follows: for any $k_2 = k - k_1 > 0$, $x \in C_{k_1}$ and any $n \geq 0$,

$$g_1(n) = q_{2^m-1} g_1(n-1) + \sum_{t=1}^{k_2-1} q_{2^m-1}\left(p_{2^m-1}\right)^t g_1(n) + \left(p_{2^m-1}\right)^{k_2} \mathbb{I}_n(0). \tag{23}$$

Now, the equation (23) can be easily solved.

Corollary 2: Suppose that $k_2 = k - k_1 > 0$. For any $x \in C_{k_1}$ and any $n \geq 0$, we have

$$g_1(n) = \left[\frac{q_{2^m-1}}{1 - \sum_{t=1}^{k_2-1}\left(p_{2^m-1}\right)^t q_{2^m-1}}\right]^n \frac{\left(p_{2^m-1}\right)^{k_2}}{1 - \sum_{t=1}^{k_2-1}\left(p_{2^m-1}\right)^t q_{2^m-1}}. \tag{24}$$

Indeed, for $n = 0$, we have

$$g_1(0) = \frac{\left(p_{2^m-1}\right)^{k_2}}{1 - \sum_{t=1}^{k_2-1}\left(p_{2^m-1}\right)^t q_{2^m-1}} = \frac{\left(p_{2^m-1}\right)^{k_2}}{q_{2^m-1} + \left(p_{2^m-1}\right)^{k_2}}. \tag{25}$$

For $n \geq 1$, inductively we have

$$g_1(n) = \frac{g_1(n-1)\ q_{2^m-1}}{1 - \sum_{t=1}^{k_2-1}\left(p_{2^m-1}\right)^t q_{2^m-1}} = \left[\frac{q_{2^m-1}}{1 - \sum_{t=1}^{k_2-1}\left(p_{2^m-1}\right)^t q_{2^m-1}}\right]^n g_1(0)$$

which proves the corollary.

We observe that N_1 follows a geometric distribution with parameter p_E where p_E is given in (7). The generating function of N_1 is given by

$$\zeta_1(s; k_1) = \frac{p_E}{1 - (1 - p_E)s} = \frac{\left(p_{2^m-1}\right)^{k_2}}{q_{2^m-1} + \left(p_{2^m-1}\right)^{k_2} - q_{2^m-1}s}. \tag{26}$$

For $r \geq 2$, we can also derive a similar recurrence relation.

Theorem 4: Suppose that $k_2 = k - k_1 > 0$. For any $x \in C_{k_1}$ and any $n \geq 0, r \geq 2$, we have

$$\begin{aligned} g_r^{(x)}(n) &= q_{2^m-1} g_r^{(2^m-2)}(n-1) + \sum_{t=1}^{k_2-1} q_{2^m-1}\left(p_{2^m-1}\right)^t g_r^{(2^m-2)}(n) \\ &\quad + \sum_{j_1=0}^{r-2}\sum_{j_2=0}^{l} q_{2^m-1}\left(p_{2^m-1}\right)^{k_2+j_1(l+1)+j_2} g_{r-1-j_1}^{(2^m-2)}(n) + \left(p_{2^m-1}\right)^{k_2+(r-1)(l+1)} \mathbb{I}_n(0). \end{aligned} \tag{27}$$

where $\mathbb{I}$ is the indicator function as defined earlier.

The proof is very similar to the proof of Theorem 3. Again, conditioning on the process when T occurs, we obtain for $k_2 > 0$, as in Theorem 3,

$$\begin{aligned} g_r^{(x)}(n) &= \mathbb{P}_x(N_r = n, Y_{T+1} = 2^{k_1} - 2) \\ &\quad + \sum_{t=1}^{k_2-1} \mathbb{P}_x(N_r = n, Y_{T+1} = 2^{k_1} - 1, \ldots, Y_{T+t} = 2^{k_1} - 1, Y_{T+t+1} = 2^{k_1} - 2) \\ &\quad + \sum_{t=k_2}^{k_2+(r-1)(l+1)-1} \mathbb{P}_x(N_r = n, Y_{T+1} = 2^{k_1} - 1, \ldots, \\ &\qquad Y_{T+t} = 2^{k_1} - 1, Y_{T+t+1} = 2^{k_1} - 2) \\ &\quad + \mathbb{P}_x(N_r = n, Y_{T+1} = 2^{k_1} - 1, Y_{T+2} = 2^{k_1} - 1, \ldots, \\ &\qquad Y_{T+k_2+(r-1)(l+1)-1} = 2^{k_1} - 1, Y_{T+k_2+(r-1)(l+1)} = 2^{k_1} - 1). \end{aligned}$$

The above expression is similar to the expression given in (16) obtained in Theorem 3. Hence following the similar calculations, we get the required result in Theorem 4.

The recurrence relation in (27) cannot be solved directly. However, we may easily check that $g_r^{(x)}(\cdot)$ is independent of x. We have shown that $g_r^{(x)}(\cdot)$ is independent of x for $r = 1$. By induction, assume that $g_r^{(x)}(\cdot)$ is independent of $x \in C_m$. Clearly, from the relation (27), we have that $g_{r+1}^{(x)}(\cdot)$ can be expressed as weighted sums of $g_i^{(x)}(\cdot)$ for $i = 1, 2, \ldots, r$ and other terms which do not involve x. Since the right hand side of the above relation does not involve any $x \in C_m$, the left hand side, *i.e.*, $g_{r+1}^{(x)}(\cdot)$ must be independent of x. Therefore, from now on, we will drop the superscript x from the notation and denote it by $g_r(\cdot)$.

The equation in (27) may now be simplified. Transferring terms containing $g_r(n)$ in the right hand side to the left hand side, we have the following result.

Lemma 1: Suppose that $k_2 = k - k_1 > 0$. For any $x \in C_{k_1}$ and any $n \geq 0, r \geq 1$, $g_r^{(x)}(n) = \mathbb{P}_x(N_r = n)$ is independent of x. For $r \geq 2$, it satisfies the recurrence relation

$$\begin{aligned} &\Big(1 - \sum_{j=1}^{k_2-1} q_{2^m-1}\big(p_{2^m-1}\big)^j\Big) g_r(n) \\ &= q_{2^m-1} g_r(n-1) + \big(p_{2^m-1}\big)^{k_2}\Big(1 - \big(p_{2^m-1}\big)^{l+1}\Big)\sum_{j=0}^{r-2}\big(p_{2^m-1}\big)^{j(l+1)} g_{r-1-j}(n) \\ &\qquad + \big(p_{2^m-1}\big)^{k_2+(r-1)(l+1)} \mathbb{I}_n(0). \end{aligned} \tag{28}$$

Now, using relation (28), we develop the recurrence relation between the probability generating functions of N_r. The probability generating function $\zeta_r(s; k_1)$, for $r \geq 2$ and $k_2 > 0$, is given by

$$\begin{aligned} &\Big(1 - \sum_{j=1}^{k_2-1} q_{2^m-1}\big(p_{2^m-1}\big)^j\Big)\zeta_r(s; k_1) = \sum_{n=0}^{\infty}\Big(1 - \sum_{j=1}^{k_2-1} q_{2^m-1}\big(p_{2^m-1}\big)^j\Big) g_r(n) s^n \\ &= \big(p_{2^m-1}\big)^{k_2+(r-1)(l+1)} + \sum_{n=0}^{\infty} q_{2^m-1} g_r(n-1) s^n \end{aligned}$$

$$
\begin{aligned}
&+\left(p_{2^m-1}\right)^{k_2}\left(1-\left(p_{2^m-1}\right)^{l+1}\right)\sum_{n=0}^{\infty}\sum_{j=0}^{r-2}\left(p_{2^m-1}\right)^{j(l+1)}g_{r-1-j}(n)s^n\\
&=\left(p_{2^m-1}\right)^{k_2+(r-1)(l+1)}+q_{2^m-1}s\zeta_r(s;k_1)\\
&\quad+\left(p_{2^m-1}\right)^{k_2}\left(1-\left(p_{2^m-1}\right)^{l+1}\right)\sum_{j=0}^{r-2}\left(p_{2^m-1}\right)^{j(l+1)}\sum_{n=0}^{\infty}g_{r-1-j}(n;k_1)s^n\\
&=\left(p_{2^m-1}\right)^{k_2+(r-1)(l+1)}+q_{2^m-1}s\zeta_r(s;k_1)\\
&\quad+\left(p_{2^m-1}\right)^{k_2}\left(1-\left(p_{2^m-1}\right)^{l+1}\right)\sum_{j=0}^{r-2}\left(p_{2^m-1}\right)^{j(l+1)}\zeta_{r-1-j}(s;k_1).
\end{aligned}
$$

Thus, we have proved the following lemma.

Lemma 2: For $r \geq 2$ and $k_2 > 0$, the sequence of probability generating functions satisfy the recurrence relation

$$
\begin{aligned}
&\left(1-q_{2^m-1}s-\sum_{j=1}^{k_2-1}q_{2^m-1}\left(p_{2^m-1}\right)^j\right)\zeta_r(s;k_1)\\
&=\left(p_{2^m-1}\right)^{k_2}\left(1-\left(p_{2^m-1}\right)^{l+1}\right)\sum_{j=0}^{r-2}\left(p_{2^m-1}\right)^{j(l+1)}\zeta_{r-1-j}(s;k_1)\\
&\qquad+\left(p_{2^m-1}\right)^{k_2+(r-1)(l+1)}.
\end{aligned}
\tag{29}
$$

Now we can use the above results to prove the main Theorem 1 as follows:

Proof: (**Theorem 1**) Let the generating function of the sequence $\{\zeta_r(s;k_1) : r \geq 1\}$ be denoted by $\Xi(z;k_1)$, *i.e.*, $\Xi(z;k_1) = \sum_{r=1}^{\infty}\zeta_r(s;k_1)z^r$. For $k_2 > 0$, we have

$$
\begin{aligned}
&\left(1-q_{2^m-1}s-\sum_{j=1}^{k_2-1}q_{2^m-1}\left(p_{2^m-1}\right)^j\right)\Xi(z;k_1)\\
&=\sum_{r=1}^{\infty}\left(1-q_{2^m-1}s-\sum_{j=1}^{k_2-1}q_{2^m-1}\left(p_{2^m-1}\right)^j\right)\zeta_r(s;k_1)z^r\\
&=\left(1-q_{2^m-1}s-\sum_{j=1}^{k_2-1}q_{2^m-1}\left(p_{2^m-1}\right)^j\right)\zeta_1(s;k_1)z+\sum_{r=2}^{\infty}\left(p_{2^m-1}\right)^{k_2+(r-1)(l+1)}z^r\\
&\quad+\left(p_{2^m-1}\right)^{k_2}\left(1-\left(p_{2^m-1}\right)^{l+1}\right)\sum_{r=2}^{\infty}\sum_{j=0}^{r-2}\left(p_{2^m-1}\right)^{j(l+1)}\zeta_{r-1-j}(s;k_1)z^r\\
&=\left(p_{2^m-1}\right)^{k_2}z+\left(p_{2^m-1}\right)^{k_2}z\sum_{r=1}^{\infty}\left(p_{2^m-1}\right)^{r(l+1)}z^r\\
&\quad+\left(p_{2^m-1}\right)^{k_2}\left(1-\left(p_{2^m-1}\right)^{l+1}\right)\sum_{j=0}^{\infty}\sum_{r=j}^{\infty}\left(p_{2^m-1}\right)^{j(l+1)}\zeta_{r-j+1}(s;k_1)z^{r+2}
\end{aligned}
$$

$$= \frac{\left(p_{2^m-1}\right)^{k_2} z}{1-\left(p_{2^m-1}\right)^{(l+1)} z} + z\Xi(z;k_1)\left(p_{2^m-1}\right)^{k_2}\left(1-\left(p_{2^m-1}\right)^{l+1}\right)\sum_{j=0}^{\infty}\left(p_{2^m-1}\right)^{j(l+1)} z^j$$

$$= \frac{\left(p_{2^m-1}\right)^{k_2} z}{1-\left(p_{2^m-1}\right)^{(l+1)} z} + \frac{z\Xi(z;k_1)\left(p_{2^m-1}\right)^{k_2}\left(1-\left(p_{2^m-1}\right)^{l+1}\right)}{1-\left(p_{2^m-1}\right)^{(l+1)} z}. \tag{30}$$

Now, from the above equation (30), we can easily solve for $\Xi(z;k_1)$ to obtain

$$\begin{aligned}
&\Xi(z;k_1)\\
&= \left[\left(p_{2^m-1}\right)^{k_2} z\right]\left[\left(1-\left(p_{2^m-1}\right)^{l+1} z\right)\left(1-q_{2^m-1}s-\sum_{j=1}^{k_2-1} q_{2^m-1}\left(p_{2^m-1}\right)^j\right)\right.\\
&\qquad \left.-z\left(p_{2^m-1}\right)^{k_2}\left(1-\left(p_{2^m-1}\right)^{l+1}\right)\right]^{-1}\\
&= \frac{z\left(p_{2^m-1}\right)^{k_2}}{1-q_{2^m-1}s-\sum_{j=1}^{k_2-1} q_{2^m-1}\left(p_{2^m-1}\right)^j}\\
&\qquad \times\left[1-\left(p_{2^m-1}\right)^{l+1} z-\frac{z\left(p_{2^m-1}\right)^{k_2}\left(1-\left(p_{2^m-1}\right)^{l+1}\right.}{1-q_{2^m-1}s-\sum_{j=1}^{k_2-1} q_{2^m-1}\left(p_{2^m-1}\right)^j}\right]^{-1}\\
&= \frac{z\left(p_{2^m-1}\right)^{k_2}}{1-q_{2^m-1}s-\sum_{j=1}^{k_2-1} q_{2^m-1}\left(p_{2^m-1}\right)^j}\\
&\qquad \times\left[1-z\left[\left(p_{2^m-1}\right)^{l+1}+\frac{\left(p_{2^m-1}\right)^{k_2}\left(1-\left(p_{2^m-1}\right)^{l+1}\right.}{1-q_{2^m-1}s-\sum_{j=1}^{k_2-1} q_{2^m-1}\left(p_{2^m-1}\right)^j}\right]\right]^{-1}.
\end{aligned} \tag{31}$$

From the expression of generating function $\Xi(z;k_1)$, $\zeta_r(s;k_1)$ is obtained by computing the coefficient of z^r. Observe that first term in the right hand side of (31) has a power of z. Therefore, we need a power of z^{r-1} from the second term. Using the expansion $(1-az)^{-1} = \sum_{n=0}^{\infty} a^n z^n$, we have

$$\begin{aligned}
\zeta_r(s;k_1) = &\left[\frac{\left(p_{2^m-1}\right)^{k_2}}{q_{2^m-1}+\left(p_{2^m-1}\right)^{k_2}-q_{2^m-1}s}\right]\left[\left(p_{2^m-1}\right)^{l+1}\right.\\
&\left.+\frac{\left(p_{2^m-1}\right)^{k_2}}{q_{2^m-1}+\left(p_{2^m-1}\right)^{k_2}-q_{2^m-1}s}\left(1-\left(p_{2^m-1}\right)^{l+1}\right)\right]^{r-1}.
\end{aligned} \tag{32}$$

This completes the proof. □

If we consider the case, when $k_2 = 0$, *i.e.*, $k = k_1$, then for $r = 1$, we must have

$\mathbb{P}_x(N_1(k) = 1) = 1$. Thus the probability generating function is given by

$$\zeta_1(s; k) = s$$

However for $r \geq 2$, using the similar arguments, we obtain,

$$g_r(n) = q_{2^m-1} g_{r-1}(n-1) + \Big(1 - q_{2^m-1} - \Big(p_{2^m-1}\Big)^{l+1}\Big) g_{r-1}(n)$$
$$+ \sum_{j=1}^{r-2} \Big(p_{2^m-1}\Big)^{j(l+1)} \Big(1 - \Big(p_{2^m-1}\Big)^{l+1}\Big) g_{r-1-j}(n) + \Big(p_{2^m-1}\Big)^{(r-1)(l+1)} \mathbb{I}_n(0).$$

This again can be used to obtain the recurrence relation between the probability generating functions $\zeta_r(s; k)$ for $r \geq 2$. Indeed, we would obtain

$$\zeta_r(s; k) = \Big(p_{2^m-1}\Big)^{(r-1)(l+1)} + q_{2^m-1} s \zeta_{r-1}(s; k) + \Big(1 - q_{2^m-1} - \Big(p_{2^m-1}\Big)^{l+1}\Big) \zeta_{r-1}(s; k)$$
$$+ \Big(1 - \Big(p_{2^m-1}\Big)^{l+1}\Big) \sum_{j=1}^{r-2} \Big(p_{2^m-1}\Big)^{j(l+1)} \zeta_{r-1-j}(s; k).$$

Using the above expression, we obtain the generating function $\Xi(z; k)$ as follows:

$$\Xi(z; k) = \frac{sz + \Big(p_{2^m-1}\Big)^{l+1}(1-s)}{1 - z(p_{2^m-1} + q_{2^m-1}s) - z^2\Big(p_{2^m-1}\Big)^{l+1}(1 - p_{2^m-1} - q_{2^m-1}s)}. \tag{33}$$

However, the explicit expression for $\zeta_r(s; k)$, *i.e.*, the coefficient of z^r in (33), will be complicated and it would be difficult to identify the distribution of the underlying random variables in terms of the known probability distributions..

5. Strong law of large numbers

In this section, we show how we may use our main result to establish the strong law of large numbers for the number of runs of exact length. Given k_1, we may fix $k = k_1 + 1$. For simplicity of the calculations, we will consider here the non-overlapping runs, *i.e.*, $l = k - 1 = k_1$. Let us define

$$\theta(n) = \sup\{r \geq 0 : \tau_r(k_1 + 1, k_1) \leq n\}. \tag{34}$$

Clearly, $\theta(n)$ represents the number of non-overlapping runs of length k that have been observed until time n. Also, we must have

$$\tau_{\theta(n)}(k_1 + 1, k_1) \leq n < \tau_{\theta(n)+1}(k_1 + 1, k_1).$$

First we observe that the occurrence of a non-overlapping run is a renewal event in our set up. Let E_t denote the event that a non-overlapping run has finished at time t. Then, for $t, s \geq 1$, we have

$$\mathbb{P}_x(E_t \cap E_{t+s}) = \mathbb{P}_x(E_t)\mathbb{P}_x\Big(E_{t+s} \mid (Y_u; u \leq t)\Big)$$

$$= \mathbb{P}_x(E_t)\mathbb{P}_x\Big(E_{t+s} \mid Y_t = 2^{k_1-1}\Big)$$
$$= \mathbb{P}_x(E_t)\mathbb{P}_{(2^m-1)}\Big(E_s\Big)$$

where we have used the strong Markov property on the expression in second step and the fact that at time t, a non-overlapping run is finished and hence we must have $Y_t = 2^{k_1-1}$. Further, this shows that the events again have the structure of a delayed renewal event.

Since we have assumed that $0 < p_x < 1$, for all $x \in C_i$, it is the case that the Markov chain $\{Y_t : t \geq 0\}$ is an irreducible chain and hence positive recurrent. This implies that the renewal event is also positive recurrent. Therefore, the expected time for getting $k_1 + 1$ consecutive successes from any state is finite and have finite expectation. In other words, we must have

$$\mathbb{E}_{(2^m-1)}(\tau_1(k_1+1, k_1)) = \mu < \infty. \tag{35}$$

The value of μ will depend upon the values of $\{p_x : x \in C_i\}$. For the *i.i.d.* case, it is known that (see Feller (1968), page 324),

$$\mu = \frac{1 - p^{k_1+1}}{qp^{k_1+1}}.$$

Using the results of renewal theory (see Feller (1968)), we further have that

$$\frac{1}{n}\theta(n) \to \frac{1}{\mu} \tag{36}$$

with probability 1. Now, we prove Theorem 2 which establishes the strong law of large numbers.

Proof: (**Theorem 2**) For any $r \geq 1$, we can represent, using Corollary 1,

$$N_{\tau_r(k_1+1,k_1)}(k_1)(= N_r(k_1)) \stackrel{d}{=} G_1 + \sum_{i=2}^{r} G_i B_i.$$

Since the equality is in distribution, we cannot directly apply the strong law on this family to conclude our result.

Now, expectation of the random variable G_1 as well as G_1B_1 may be computed from the probability generating function given in equation (9). Indeed, we have $\mathbb{E}(G_1B_1) = \mu_1$ (see equation (10)). Further observe that all moments of G_1B_1 are finite.

Let us set $\mu_1(r) = \Big[\mathbb{E}(G_1) + (r-1)\mathbb{E}(G_1B_1)\Big]$. Then, we have

$$\frac{1}{r}\mu_1(r) = \frac{1}{r}\Big[\mathbb{E}(G_1) + (r-1)\mu_1\Big] \to \mu_1 \tag{37}$$

as $r \to \infty$. Note that, from the representation, we have $\mathbb{E}(N_{\tau_r(k_1+1,k_1)}(k_1)) = \mu_1(r)$. Furthermore, for any $\epsilon > 0$, we have

$$\mathbb{P}\Big(\frac{1}{r}|N_{\tau_r(k_1+1,k_1)}(k_1) - \mu_1(r)| \geq \epsilon\Big) = \mathbb{P}\Big(\frac{1}{r}|G_1 + \sum_{i=2}^{r} G_i B_i - \mu_1(r)| \geq \epsilon\Big)$$

$$= \mathbb{P}\Big(|G_1 - \mathbb{E}(G_1) + \sum_{i=2}^{r}(G_iB_i - \mathbb{E}(G_1B_1))| \geq r\epsilon\Big).$$

Now, we may estimate the probability using the Markov inequality. Indeed, we have

$$\begin{aligned}
&\mathbb{P}\Big[|G_1 - \mathbb{E}(G_1) + \sum_{i=2}^{r} G_iB_i - \mathbb{E}(G_1B_1)| \geq r\epsilon\Big] \\
&\leq \frac{1}{r^4\epsilon^4}\mathbb{E}\Big[\Big(G_1 - \mathbb{E}(G_1) + \sum_{i=2}^{r} G_iB_i - \mathbb{E}(G_1B_1)\Big)^4\Big] \\
&\leq \frac{1}{r^4\epsilon^4}\Big[\mathbb{E}\Big(G_1 - \mathbb{E}(G_1)\Big)^4 + 3(r-1)\mathbb{E}\Big(G_1 - \mathbb{E}(G_1)\Big)^2 E\Big(G_1B_1 - \mathbb{E}(G_1B_1)\Big)^2 \\
&\qquad + 6(r-1)^2\Big(E\Big(G_1B_1 - \mathbb{E}(G_1B_1)\Big)^2\Big)^2 + (r-1)\mathbb{E}\Big(G_1B_1 - \mathbb{E}(G_1B_1)\Big)^4\Big] \\
&\leq \frac{C}{r^2\epsilon^4}
\end{aligned}$$

for a suitably chosen constant $C > 0$.

Thus, by Borel-Cantelli lemma, we conclude that $\frac{1}{r}\Big(N_{\tau_r(k_1+1,k_1)}(k_1) - \mu_1(r)\Big) \to 0$ with probability 1. This along with equation (37) implies that

$$\frac{1}{r}N_{\tau_r(k_1+1,k_1)}(k_1) \to \mu_1 \tag{38}$$

as $r \to \infty$ with probability 1.

Since $\tau_{\theta(n)}(k_1+1,k_1) \leq n < \tau_{\theta(n)+1}(k_1+1,k_1)$, we must have $N_{\theta(n)}(k_1) \leq N_n(k_1) \leq N_{\theta(n)+1}(k_1)$. Therefore, we obtain that

$$\begin{aligned}
&\frac{1}{n}N_{\theta(n)}(k_1) \leq \frac{1}{n}N_n(k_1) \leq \frac{1}{n}N_{\theta(n)+1}(k_1) \\
&\implies \frac{\theta(n)}{n} \times \frac{1}{\theta(n)}N_{\theta(n)}(k_1) \leq \frac{1}{n}N_n(k_1) \leq \frac{\theta(n)+1}{n} \times \frac{1}{\theta(n)+1}N_{\theta(n)+1}(k_1).
\end{aligned}$$

Since $\theta(n)/n \to \frac{1}{\mu}$, we have that $\theta(n) \to \infty$ as $n \to \infty$. Hence, we may apply the equation (38) along the sub-sequence $\theta(n)$ and equation (36) to conclude that both the upper bound as well as the lower bound will converge to μ_1/μ. This proves the result. □

Acknowledgement

The author wishes to thank the referees for their helpful comments which have improved the presentation of the paper.

References

Aki, S. and Hirano, K. (1994). Distributions of numbers of failures and successes until the first consecutive k successes. *Annals of the Institute of Statistical Mathematics*, **46**, 193-202.

Aki, S. and Hirano, K. (1995). Joint distributions of numbers of success-runs and failures until the first k consecutive successes. *Annals of the Institute of Statistical Mathematics*, **47**, 225-235.

Aki, S. and Hirano, K. (2000). Numbers of success-runs of specified length until certain stopping time rules and generalized binomial distributions of order k. *Annals of the Institute of Statistical Mathematics*, **52**, 767-777.

Anuradha. (2022). Asymptotic results for generalized runs in higher order Markov Chains. *Statistics and Applications*, **21**, 189-207.

Balakrishnan, N. and Koutras, M. V. (2002). *Runs and Scans with Applications.* John Wiley and Sons, New York.

Chadjiconstantinidis, S. and Koutras, M. V. (2001). Distributions of the numbers of failures and successes in a waiting time problem. *Annals of the Institute of Statistical Mathematics*, **53**, 576-598.

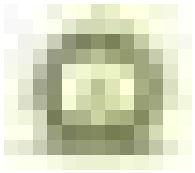
Doganaksoy, A., Sulak, F., Uguz, M., Seker, O., and Akcengiz, Z. (2015). New statistical randomness tests based on length of runs. *Mathematical Problems in Engineering*, **2015**, 1-14.

Feller, W. (1968) *An Introduction to Probability Theory and its Applications. Vol - I.* John Wiley, New York, 3rd ed.

Hirano, K., Aki, S., and Uchida, M. (1997). Distributions of numbers of success-runs until the first consecutive k successes in higher order Markov dependent trials. *Advances in Combinatorial Methods and Applications to Probability and Statistics,* 401–410, Statistics for Industry and Technology, Birkhäuser Boston, Boston, MA.

Uchida, M. (1998). On number of occurrences of success runs of specified length in a higher order two-state Markov chain. *Annals of the Institute of Statistical Mathematics*, **50**, 587-601.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 22, No. 1, 2024 (New Series), pp 101–122
http://www.ssca.org.in/journal



A Discrete Analogue of Intervened Poisson Compounded Family of Distributions: Properties with Applications to Count Data

K. Jayakumar and Jiji Jose
Department of Statistics, University of Calicut, Kerala-673635, India

Received: 26 September 2022; Revised: 20 May 2023; Accepted: 15 June 2023

Abstract

There are several discrete distributions have been developed in statistical literature. Even though, it is inadequate to analyse the real data produced from different fields through the various discrete distributions available in the existing literature. According to this motives, we have proposed a new family of discrete models called discrete intervened Poisson compounded (DIPc) family. A key feature of the proposed family is its hazrad rate function can take variety of shapes for distinct values of the parameters like decreasing, constant, bathtub shaped. Furthermore, several distributional characteristics are extensively studied for the particular distribuiton of DIPc family. Certain characterizations of the new distribution are obtained. An integer valued autoregressive process with the distribution as marginal is introduced. The unknown parameters of the distribution are estimated using different methods of estimation. Finally, we have explained the usefullness of the proposed family by using a real data set.

Key words: Characterizations; Exponential Intervened Poisson (EIP) distribution; Discrete Intervened Poisson (DEIP) distribution; INAR(1) process; Stress- strength parameter.

AMS Subject Classifications: 60E05, 62E10

1. Introduction

The intervened Poisson distribution (IPD) is introduced by Shanmugam (1985) which provides stochastic models to study the effect of such actions as they are closer to real life situations. The IPD is a modified version of zero truncated Poisson (ZTP) distribution, which is applicable in reliability analysis, queueing problems, epidemiological problems where ZTP fails. Jayakumar and Sankaran (2019) introduce a new family of distributions generated using IPD and this distributions helps to develop a rich class of families which contain Marshall and Olkin (1997) extended families of distribution. The intervened Poisson compounded (IP) family of continuous distributions is one among them. The cumulative distribution function

Corresponding Author: Jiji Jose
Email: jijijose2793@gmail.com

(CDF) of IP family of distributions is given by

$$G(x;\lambda,\rho;\phi) = 1 - [\frac{e^{\lambda(1+\rho)\bar{F}(x;\phi)} - e^{\lambda\rho\bar{F}(x;\phi)}}{e^{\lambda\rho}(e^{\lambda}-1)}]\ ; x\ \in\ \mathbb{R} \tag{1}$$

where $\bar{F}(x;\phi)$ is the CDF of base line continuous distribution and ϕ is the vector of the given model parameters.

Here we establish the discretization of continuous distribution. Discretization of a continuous lifetime model is an interesting and intuitively appealing approach to derive a lifetime model corresponding to the continuous one. Meanwhile, it is difficult or inconvenient to get samples from a continuous distribution in real life situations. In modelling, the observed values are actually discrete because they are measured to only a finite number of decimal places and cannot really constitute all points in a continuum. For example, in case of survival analysis, the number of days of survival for lung cancer patients since therapy are usually recorded in discrete values. In the recent, special role of discrete distributions are getting recognition in the field of reliability. In this way, one of the active areas of research is to model discrete data by developing discretized distributions.
Chakraborty (2015) surveyed different methods for generating discrete analogues of continuous probability distributions. One of the methods is described as follows:
Let X be a continuous random variable, then the discerte analogue Y of X can be derived by using the survival function as follows, $S(.)$ is the survival function of the random variable X, then

$$P(Y = y) = P(X \geq y) - P(X \geq y+1) = S(y) - S(y+1); y = 0,1,2,3,.... \tag{2}$$

where $Y = \lfloor X \rfloor$ largest integer less than or equal to X. The first and easiest in this approach is the geometric distribution with pmf

$$p(x) = \theta^x - \theta^{x+1}; x = 0,1,2,...$$

which is derived by discretizing exponential distribution with survival function $S(x) = e^{-\lambda x}; \lambda, x > 0$ and $\theta = e^{-\lambda}, (0 < \theta < 1)$.
Following this approach, discretization of some known continuous distributions for use as lifetime distribution was studied by different researchers. Nakagawa and Osaki (1975) proposed discrete Weibull distribution with pmf

$$P(Y = v) = q^{v^{\beta}} - q^{(v+1)^{\beta}}, v = 0,1,2,...; \beta > 0, 0 < q < 1. \tag{3}$$

Stein and Dattero (1984) presented another discretization of Weibull distribution. Roy (2003) proposed discrete normal distribution and also studied discrete Rayleigh distribution (Roy (2004)). Krishna and Pundir (2009) studied discrete Burr distribution, and obtained the discrete Pareto distribution as its particular case.

The discretization of a continuous distribution using this method retains the same functional form of the survival function. As a result, many reliability characteristics remain unchanged. As such there is enough motivation to use this technique of generating discretized version of continuous distribution with this approach to develop new discrete lifetime models corresponding to the existing continuous one. In this article, we propose a family of

discrete univariate distributions using survival discretization method. Thus the objective of proposing Discrete Intervned Poisoon compounded (DIPc) family are to generate models for modelling probability distribution of count data and produce consistently superior fits than other developed discrete distributions in the existing literature.

The remaining parts of the article are as follows: Section 2 introduces the DIPc family and some statistical properties are derived. In Section 3, the special model of the proposed family is extensively studied. The expression for moments, stress - strength reliability are derived. Also, using the proposed distribution, an integer valued autoregressive process with the distribution as marginal is introduced. In Section 4, three characterizations of the new distribution are obtained and in Section 5, an extensive estimation and simulation study is conducted to investigate the behaviour of different estimation methods. The flexibility of the proposed model is illustrated by using a real data set in Section 6. Finally, some important remarks about the presented study are discussed in Section 7.

2. Genesis of the family

The random variable Y is said to follow Discrete Intervened Poisson compounded (DIPc) family, its probability mass function (pmf) is given by

$$P_Y(y;\lambda,\rho,\phi)=\left[\frac{e^{\lambda(1+\rho)\bar{F}(y;\phi)}-e^{\lambda\rho\bar{F}(y;\phi)}}{e^{\lambda\rho}(e^{\lambda}-1)}\right]-\left[\frac{e^{\lambda(1+\rho)\bar{F}(y+1;\phi)}-e^{\lambda\rho\bar{F}(y+1;\phi)}}{e^{\lambda\rho}(e^{\lambda}-1)}\right]. \tag{4}$$

The corresponding CDF of DIPc is obtained as

$$\begin{aligned}G_Y(y;\lambda,\rho,\phi)&=1-G_X(y;\lambda,\rho,\phi)+P_Y(y;\lambda,\rho,\phi)\\&=1-\left[\frac{e^{\lambda(1+\rho)\bar{F}(y;\phi)}-e^{\lambda\rho\bar{F}(y;\phi)}}{e^{\lambda\rho}(e^{\lambda}-1)}\right]+\left[\frac{e^{\lambda(1+\rho)\bar{F}(y;\phi)}-e^{\lambda\rho\bar{F}(y;\phi)}}{e^{\lambda\rho}(e^{\lambda}-1)}\right]-\left[\frac{e^{\lambda(1+\rho)\bar{F}(y+1;\phi)}-e^{\lambda\rho\bar{F}(y+1;\phi)}}{e^{\lambda\rho}(e^{\lambda}-1)}\right]\\&=1-\left[\frac{e^{\lambda(1+\rho)\bar{F}(y+1;\phi)}-e^{\lambda\rho\bar{F}(y+1;\phi)}}{e^{\lambda\rho}(e^{\lambda}-1)}\right];y\in\mathbb{N}\end{aligned} \tag{5}$$

where $\mathbb{N}=\{0,1,2,...\}$, $(\lambda,\rho)\in(0,\infty)$ and $G_X(y;\lambda,\rho,\phi)=\left[\frac{e^{\lambda(1+\rho)\bar{F}(y;\phi)}-e^{\lambda\rho\bar{F}(y;\phi)}}{e^{\lambda\rho}(e^{\lambda}-1)}\right]$ is the CDF of X.

The survival function of DIPc family is given by

$$S_Y(y;\lambda,\rho,\phi)=\frac{e^{\lambda(1+\rho)\bar{F}(y+1;\phi)}-e^{\lambda\rho\bar{F}(y+1;\phi)}}{e^{\lambda\rho}(e^{\lambda}-1)};y\in\mathbb{N}. \tag{6}$$

The hazard rate and reverse hazard rate are

$$h_Y(y;\lambda,\rho,\phi)=1-\left[\frac{e^{\lambda\rho[\bar{F}(y+1;\phi)-\bar{F}(y,\phi)]}(e^{\lambda\bar{F}(y;\phi)}-1)}{e^{\lambda\bar{F}(y+1;\phi)}-1}\right] \tag{7}$$

and

$$r_Y(y;\lambda,\rho,\phi)=\frac{\left[e^{\lambda(1+\rho)\bar{F}(y;\phi)}-e^{\lambda\rho\bar{F}(y;\phi)}\right]-\left[e^{\lambda(1+\rho)\bar{F}(y+1;\phi)}-e^{\lambda\rho\bar{F}(y+1;\phi)}\right]}{[e^{\lambda\rho}(e^{\lambda}-1)]-\left[e^{\lambda(1+\rho)\bar{F}(y+1;\phi)}-e^{\lambda\rho\bar{F}(y+1;\phi)}\right]} \tag{8}$$

respectively.

2.1. Moments

Let the random variable $Y \sim DIPc(\lambda, \rho, \phi)$, then the r^{th} moment is given by

$$\mu'_r = \sum_{y=0}^{\infty}((y+1)^r - y^r)\left[\frac{e^{\lambda(1+\rho)\bar{F}(y+1;\phi)} - e^{\lambda\rho\bar{F}(y+1;\phi)}}{e^{\lambda\rho}(e^{\lambda}-1)}\right] ; y \in \mathbb{N}. \tag{9}$$

Using the Equation 9, the mean and variance of DIPc can be obtained as follows, respectively,

$$\mu'_1 = \sum_{y=0}^{\infty}\left[\frac{e^{\lambda(1+\rho)\bar{F}(y+1;\phi)} - e^{\lambda\rho\bar{F}(y+1;\phi)}}{e^{\lambda\rho}(e^{\lambda}-1)}\right] \tag{10}$$

and

$$variance = \sum_{y=0}^{\infty}(2y+1)\left[\frac{e^{\lambda(1+\rho)\bar{F}(y+1;\phi)} - e^{\lambda\rho\bar{F}(y+1;\phi)}}{e^{\lambda\rho}(e^{\lambda}-1)}\right] - (\mu'_1)^2. \tag{11}$$

The index of dispersion (DI), (variance/ mean), determines whether the given distribution is suited for under, over or equi-dispersed data sets. If $DI > 1$, then distribution is overdispersed whereas $DI < 1$, then distribution is underdispersed. If $DI = 1$, then distribution is equidispersed.
The moment generating function of the distribution is given by

$$M_Y(t) = \sum_{y=0}^{t}\sum_{r=0}^{\infty}\frac{(yt)^r}{r!}\left[\frac{e^{\lambda(1+\rho)\bar{F}(y;\phi)} - e^{\lambda\rho\bar{F}(y;\phi)}}{e^{\lambda\rho}(e^{\lambda}-1)}\right] - \left[\frac{e^{\lambda(1+\rho)\bar{F}(y+1;\phi)} - e^{\lambda\rho\bar{F}(y+1;\phi)}}{e^{\lambda\rho}(e^{\lambda}-1)}\right]. \tag{12}$$

From the Equation (12), it can be obtained first four raw moments about the origin when $t = 0$. Also skewness and kurtosis based on moments can be computed by using the moment generating function.

3. Special model

In this section, we study a particular distribution of DIPc family to establish its viability. The main objective of establishing new model is to study the properties of the particular model of the presented family, to illustrate the flexibility of the developed family through real data sets.

3.1. Discrete Exponential Intervened Poisson (DEIP) distribution

Using the CDF of the exponential distribution, the pmf of DEIP can be formulated as

$$P(Y = y) = \frac{\left[e^{\lambda(1+\rho)e^{-\theta y}} - e^{\lambda\rho e^{-\theta y}}\right] - \left[e^{\lambda(1+\rho)e^{-\theta(y+1)}} - e^{\lambda\rho e^{-\theta(y+1)}}\right]}{e^{\lambda\rho}(e^{\lambda}-1)} \tag{13}$$

where $y = 0, 1, 2, \ldots$, $\lambda > 0$, $\rho \geq 0$, $\theta > 0$.

Theorem 1: The pmf of DEIP distribution is unimodal.

Proof: The pmf of DEIP is log concave, where $P(y+1;\lambda,\rho,\theta)/P(y;\lambda,\rho,\theta)$ is a decreasing function in y for all model parameters. As a direct consequence of log concavity, the DEIP is unimodal.

Figures 1 and 2 show the pmf and hazard rate plots of the DEIP model respectively. The pmf is unimodal and can be used to analyze positively skewed data set. Furthermore, the hazard rate can be either decreasing, constant, decreasing- constant and bathtubshaped. Therefore, the parameters of the DEIP model can be fixed to fit most data sets.

3.2. Structural Properties

The CDF of DEIP is given by

$$\begin{aligned} F(y;\lambda,\rho,\theta) =& P(Y \leq y) \\ =& 1 - \left[\frac{e^{\lambda(1+\rho)e^{-\theta(y+1)}} - e^{\lambda\rho e^{-\theta(y+1)}}}{e^{\lambda\rho}(e^{\lambda}-1)}\right]. \end{aligned} \tag{14}$$

The survival function of DEIP is given by

$$S(y;\lambda,\rho,\theta) = \left[\frac{e^{\lambda(1+\rho)e^{-\theta(y+1)}} - e^{\lambda\rho e^{-\theta(y+1)}}}{e^{\lambda\rho}(e^{\lambda}-1)}\right]. \tag{15}$$

The hazard rate of DEIP distribution is

$$\begin{aligned} h(y) = P(Y=y|Y\geq y) =& \frac{P(Y=y)}{P(Y\geq y)} \\ =& \frac{\left[e^{\lambda(1+\rho)e^{-\theta y}} - e^{\lambda\rho e^{-\theta y}}\right] - \left[e^{\lambda(1+\rho)e^{-\theta(y+1)}} - e^{\lambda\rho e^{-\theta(y+1)}}\right]}{e^{\lambda\rho}(e^{\lambda}-1)} \\ & \overline{\left[\frac{e^{\lambda(1+\rho)e^{-\theta(y+1)}} - e^{\lambda\rho e^{-\theta(y+1)}}}{e^{\lambda\rho}(e^{\lambda}-1)}\right]} \\ =& \frac{\left[e^{\lambda(1+\rho)e^{-\theta y}} - e^{\lambda\rho e^{-\theta y}}\right] - \left[e^{\lambda(1+\rho)e^{-\theta(y+1)}} - e^{\lambda\rho e^{-\theta(y+1)}}\right]}{e^{\lambda(1+\rho)e^{-\theta(y+1)}} - e^{\lambda\rho e^{-\theta(y+1)}}} \\ =& \left[\frac{e^{\lambda(1+\rho)e^{-\theta y}} - e^{\lambda\rho e^{-\theta y}}}{e^{\lambda(1+\rho)e^{-\theta(y+1)}} - e^{\lambda\rho e^{-\theta(y+1)}}}\right] - 1. \end{aligned} \tag{16}$$

The reverse hazard rate is

$$\begin{aligned} r(y) =& P(Y=y)/P(Y\leq y) \\ =& \frac{\left[e^{\lambda(1+\rho)e^{-\theta y}} - e^{\lambda\rho e^{-\theta y}}\right] - \left[e^{\lambda(1+\rho)e^{-\theta(y+1)}} - e^{\lambda\rho e^{-\theta(y+1)}}\right]}{[e^{\lambda\rho}(e^{\lambda}-1)] - \left[e^{\lambda(1+\rho)e^{-\theta(y+1)}} - e^{\lambda\rho e^{-\theta(y+1)}}\right]} \end{aligned} \tag{17}$$

and the second rate of failure of DEIP distribution is given by,

$$\begin{aligned} h^{**}(y) =& log\left[\frac{S(y)}{S(y+1)}\right] \\ =& log\left[\frac{e^{\lambda(1+\rho)e^{-\theta(y+1)}} - e^{\lambda\rho e^{-\theta(y+1)}}}{e^{\lambda(1+\rho)e^{-\theta(y+2)}} - e^{\lambda\rho e^{-\theta(y+2)}}}\right]. \end{aligned} \tag{18}$$

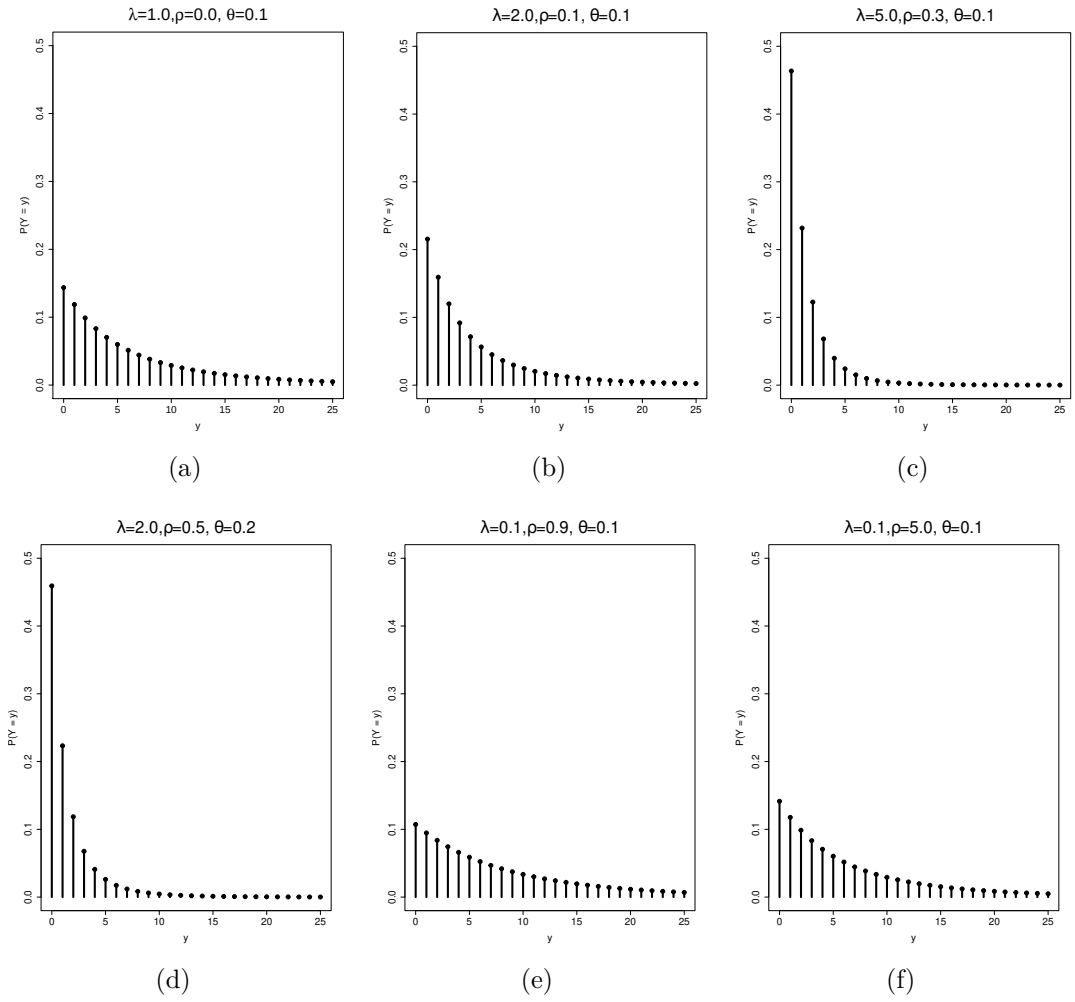


Figure 1: The pmf plots of DEIP(λ, ρ, θ) for different values of λ, ρ and θ

3.3. Recurrence relation for probabilities

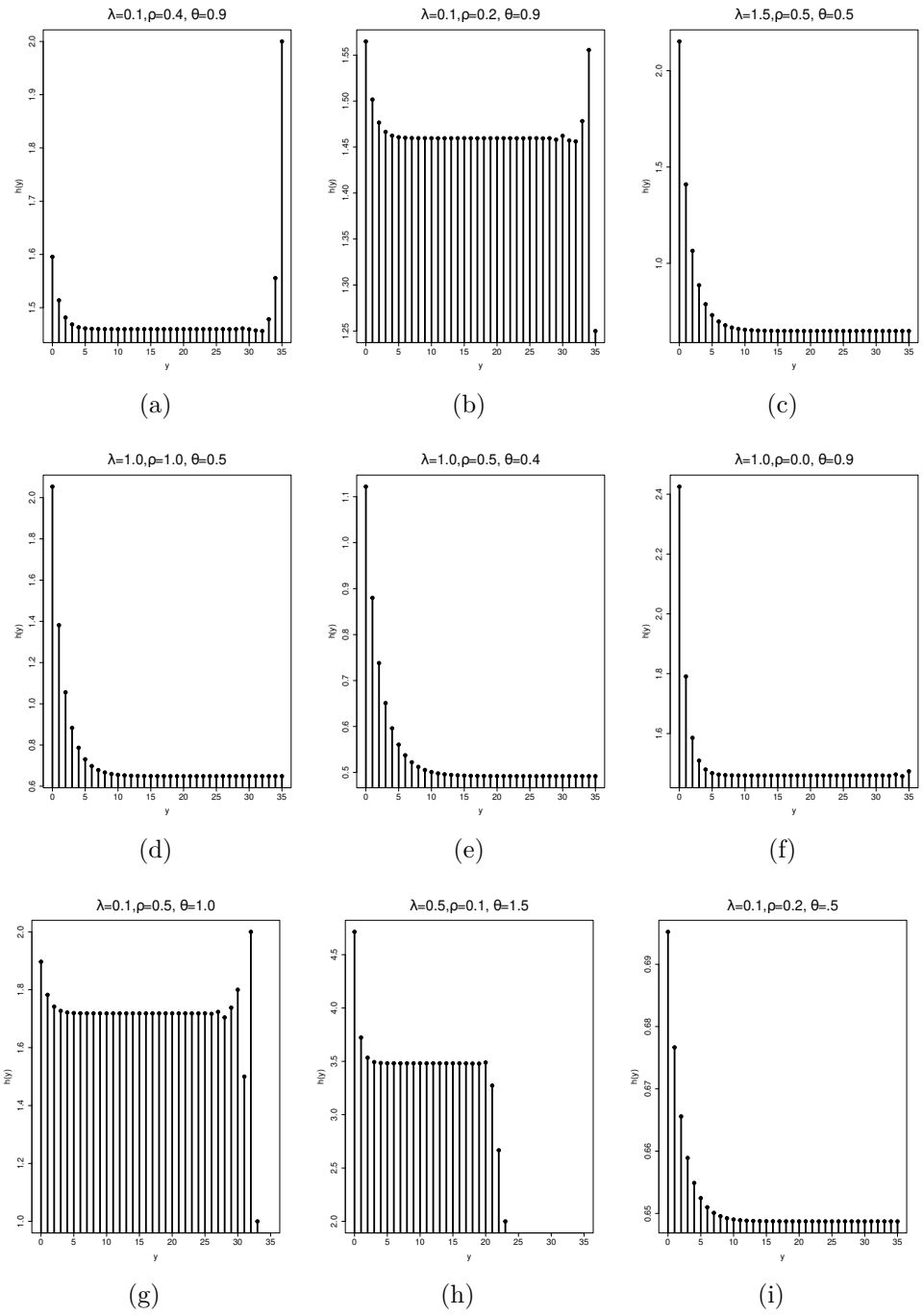
The recurrence relation for generating probabilities of DEIP $(\lambda,\ \rho,\ \theta)$ is given by

$$\frac{p(y+1;\lambda,\rho,\theta)}{p(y;\lambda,\rho,\theta)} = \frac{\left[e^{\lambda(1+\rho)e^{-\theta(y+1)}} - e^{\lambda\rho e^{-\theta(y+1)}}\right] - \left[e^{\lambda(1+\rho)e^{-\theta(y+2)}} - e^{\lambda\rho e^{-\theta(y+2)}}\right]}{\left[e^{\lambda(1+\rho)e^{-\theta y}} - e^{\lambda\rho e^{-\theta y}}\right] - \left[e^{\lambda(1+\rho)e^{-\theta(y+1)}} - e^{\lambda\rho e^{-\theta(y+1)}}\right]} \tag{19}$$

3.4. Moments

The r^{th} moment of DEIP distribution is given by

$$\begin{aligned} E(Y^r) &= \sum_{y=0}^{\infty} y^r P(Y=y) \\ &= \sum_{y=0}^{\infty} \left[(y+1)^r - y^r\right] S(y). \end{aligned} \tag{20}$$



(a) (b) (c)

(d) (e) (f)

(g) (h) (i)

Figure 2: Hazard rate plots of DEIP(λ, ρ, θ) for different values of λ, ρ and θ

$$\begin{aligned} E(Y) &= \sum_{y=0}^{\infty} S(y) \\ &= \sum_{y=0}^{\infty} \left[\frac{e^{\lambda(1+\rho)e^{-\theta(y+1)}} - e^{\lambda\rho e^{-\theta(y+1)}}}{e^{\lambda\rho}(e^{\lambda}-1)} \right]. \end{aligned} \tag{21}$$

Now

$$\begin{aligned} E(Y^2) &= \sum_{y=0}^{\infty}(2y+1)S(y) \\ &= \sum_{y=0}^{\infty}(2y+1)\left[\frac{e^{\lambda(1+\rho)e^{-\theta(y+1)}} - e^{\lambda\rho e^{-\theta(y+1)}}}{e^{\lambda\rho}(e^{\lambda}-1)}\right]. \end{aligned} \tag{22}$$

$$\begin{aligned} V(Y) =& E(Y^2) - E(Y) \\ =& \sum_{y=0}^{\infty}(2y+1)S(y) - \sum_{y=0}^{\infty}S(y) \\ =& \sum_{y=0}^{\infty}(2y+1)\left[\frac{e^{\lambda(1+\rho)e^{-\theta(y+1)}} - e^{\lambda\rho e^{-\theta(y+1)}}}{e^{\lambda\rho}(e^{\lambda}-1)}\right] - \left[\sum_{y=0}^{\infty}\left[\frac{e^{\lambda(1+\rho)e^{-\theta(y+1)}} - e^{\lambda\rho e^{-\theta(y+1)}}}{e^{\lambda\rho}(e^{\lambda}-1)}\right]\right]^2. \end{aligned} \tag{23}$$

Table 1 shows mean and variance (given in parenthesis) of DEIP distribution using given values of λ, ρ and θ. For fixed θ, as ρ increases mean and variance decreases. Also, as λ increases, mean and variance decreases. From the Table 1, it can also seen that mean is always less than variance for different set of the parameters λ, ρ and θ. Therefore, DEIP is suited better for modelling over-dispersed data.

3.5. Stress-Strength analysis

The stress - strength analysis is used in mechanical component analysis and the stress - strength parameter R measures component reliabilty. Let the random variable Y be strength of a component which is subjected to a random stress Z. For a detailed review of stress- strength models, one may refer Choudhary *et al.* (2021). The stress-strength model defined in discrete case as,

$$P(Y > Z) = \sum_{y=0}^{\infty} p_Y(y)F_Z(y). \tag{24}$$

Let Y and Z be independent stress and strength random variables from $Y \sim \text{DEIP}(\lambda_1, \rho_1, \theta_1)$ and $Z \sim \text{DEIP}(\lambda_2, \rho_2, \theta_2)$ respectively. Also p_Y and F_Y denote the pmf and CDF of the distribution respectively.

Then the stress - strength parameter for the model DEIP is given by,

$$\begin{aligned} R = P(Y > Z) =& \sum_{y=0}^{\infty}\left[\frac{e^{\lambda_1\rho_1 e^{-\theta_1 y}}\left[e^{\lambda_1 e^{-\theta y}} - 1\right] - e^{\lambda_1\rho_1 e^{-\theta_1(y+1)}}\left[e^{\lambda_1 e^{-\theta_1(y+1)}} - 1\right]}{e^{\lambda_1\rho_1}(e^{\lambda_1}-1)}\right] \\ &\times \left[1 - \left(\frac{e^{\lambda_2(1+\rho_2)e^{-\theta_2(y+1)}} - e^{\lambda_2\rho_2 e^{-\theta_2(y+1)}}}{e^{\lambda_2\rho_2}(e^{\lambda_2}-1)}\right)\right] \\ &= \delta(\lambda_1, \rho_1, \theta_1, \lambda_2, \rho_2, \theta_2). \end{aligned} \tag{25}$$

Obviously, the solution of the summation in Equation (25) cannot be obtained explicitly. That is, there is no closed form expression of $\delta(.)$, therefore, we resort to numerical method

Table 1: Mean and Variance of DEIP for different values of λ, ρ and θ

			$\theta = 0.5$		
$\lambda \mid \rho$	0.25	0.5	1.0	2.0	3.0
0.50	1.2141	1.1259	0.9703	0.7261	0.5487
	(3.1536)	(2.9273)	(2.5116)	(1.8282)	(1.3212)
0.75	1.0734	0.9604	0.7721	0.5074	0.3403
	(2.7956)	(2.4901)	(1.9633)	(1.2059)	(0.7405)
1.00	0.9468	0.8182	0.6157	0.3585	0.2156
	(2.4604)	(2.1002)	(1.5200)	(0.7918)	(0.4192)
2.00	0.5620	0.4276	0.2540	0.0972	0.04007
	(1.3905)	(1.0006)	(0.5234)	(0.1564)	(0.0535)
			$\theta = 1.0$		
$\lambda \mid \rho$	0.25	0.5	1.0	2.0	3.0
0.50	0.4388	0.4008	0.3348	0.2342	0.1645
	(0.7198)	(0.6619)	(0.5573)	(0.3901)	(0.2702)
0.75	0.3787	0.3309	0.25309	0.1489	0.0884
	(0.6282)	(0.5517)	(0.4225)	(0.2436)	(0.1391)
1.00	0.3255	0.2722	0.1908	0.0948	0.0477
	(0.5440)	(0.4556)	(0.3166)	(0.1503)	(0.0712)
2.00	0.1714	0.1206	0.0604	0.0156	0.0041
	(0.2860)	(0.1969)	(0.0929)	(0.0211)	(0.0051)
			$\theta = 3.0$		
$\lambda \mid \rho$	0.25	0.5	1.0	2.0	3.0
0.50	0.0363	0.0322	0.0253	0.0157	0.0097
	(0.0387)	(0.0344)	(0.0273)	(0.0170)	(0.0106)
0.75	0.0299	0.0250	0.0175	0.0085	0.0042
	(0.0321)	(0.0269)	(0.0189)	(0.0093)	(0.0045)
1.00	0.0246	0.0193	0.0120	0.0046	0.0017
	(0.0264)	(0.0209)	(0.0131)	(0.0050)	(0.0019)
2.00	0.0106	0.0066	0.0025	0.0003	0.0005
	(0.0116)	(0.0072)	(0.0027)	(0.0004)	(0.0006)

to calculate the system reliability.

To find the maximum likelihood (ML) estimator of the system reliability, we consider $Y_i, i = (1, 2, ...n)$ and $Z_j, j = (1, 2, ..., m)$ two independent samples from DEIP$(\lambda_1, \rho_1, \theta_1)$ and DEIP$(\lambda_2, \rho_2, \theta_2)$ respectively. Then the likelihood function is given by,

$$\begin{aligned} L =& \prod_{i=1}^{n} P(Y = y_i) \prod_{j=1}^{m} P(Y = z_j) \\ =& e^{-n\lambda_1\rho_1}(e^{\lambda} - 1)^{-n} \prod_{i=1}^{n} [L_1 - L_2] \times e^{-m\lambda_2\rho_2}(e^{\lambda_2} - 1)^{-m} \prod_{j=1}^{m} [L_3 - L_4] \end{aligned} \tag{26}$$

where,
$L_1 = e^{\lambda_1(1+\rho_1)e^{-\theta_1 y_i}} - e^{\lambda_1\rho_1 e^{-\theta y_i}}$, $L_2 = e^{\lambda_1(1+\rho_1)e^{-\theta(y_i+1)}} - e^{\lambda_1\rho_1 e^{-\theta_1(y_i+1)}}$
$L_3 = e^{\lambda_2(1+\rho_2)e^{-\theta_2 z_j}} - e^{\lambda_2\rho_2 e^{-\theta z_j}}$ and $L_4 = e^{\lambda_2(1+\rho_2)e^{-\theta_2(z_j+1)}} - e^{\lambda_2\rho_2 e^{-\theta_2(z_j+1)}}$. In order to obtain the ML estimators of $\lambda_1, \rho_1, \theta_1, \lambda_2, \rho_2$ and θ_2, we first derive the log-likelihood (LogL) function by taking the logarithm of Equation (26). Then, we take the derivatives of the logL function with respect to the parameters of interest and obtain the likelihood equations. The solutions of these equations cannot be obtained in closed form, and the estimates of the unknown parameters are found by using numerical methods with the help of **R** programming. Then by using the invariance property of ML estimators, the ML estimate of system reliability is obtained as

$$\hat{R} = \delta(\hat{\lambda}_1, \hat{\rho}_1, \hat{\theta}_1, \hat{\lambda}_2, \hat{\rho}_2, \hat{\theta}_2).$$

Some numerical results of R are reported in Table 2 using DEIP distribution for the parameters $\lambda_1 = \lambda_2 = \rho_1 = \rho_2 = 0.5$. It is clear that R decreases(increases) when θ_1 increases (θ_2 increases).

Table 2: Some numerical results of R for different values of θ_1 and θ_2

θ_1—θ_2	0.1	0.5	0.9	1.0
0.1	0.5475	0.6027	0.6093	0.6100
0.5	0.3105	0.5995	0.7066	0.7229
0.9	0.2705	0.5482	0.6707	0.6906
1.0	0.2653	0.5404	0.6651	0.6854

3.6. Infinite divisibility

The famous structural property of infinite divisibility of the distribution is an interesting area to the researchers. Such a characteristic has a close relation to the Central Limit Theorem and waiting time distributions. According to Steutel and van Harn (2003), if $p_x, x \in \mathbb{N}_0$ is infinitely divisible, then $p_x \leq e^{-1}$ for all $x \in \mathbb{N}$. Also from Theoreom 3.2 of Steutel and van Harn (2003), if for atleast one case for which p_x is greater than $1/e$, then pmf cannot be compound Poisson and hence it cannot be infinitely divisible. In DEIP distribution, $\lambda = 3$, $\rho = 0.6$ and $\theta = 0.1$, then $p_0 = 0.3776 > e^{-1} = 0.367$. Therefore we can conclude that DEIP distribution is not infinitely divisible. The classes of self-decomposable and stable distributions are subclasses of infinitely divisible distributions, in their discrete concepts. So in this case, DEIP distribution can be neither self-decomposable nor stable in general.

3.7. Application in first order integer valued autoregressive (INAR(1)) process

There has been a growing interest in discrete-valued time series models and several models for stationary processes with discrete marginal distributions have been proposed in the literature. A simple model for a stationary sequence of integer-valued random variables with lag-one dependence is given and is referred to as the integer-valued autoregressive of order one (INAR(1)) process. It is widely used to model the time series of counts in different applied sciences such as actuarial, finance and medical sciences. The INAR(1)

process differs from the first-order autoregressive, shortly AR(1), process by applying the binomial thinning operator. The first INAR(1) process was introduced by McKenzie (1985) based on the Poisson innovations and is called as INAR(1)P.

$$Y_t = \alpha \ o \ Y_{t-1} + \epsilon_t, t \in \mathbb{Z} \tag{27}$$

where $\alpha \in (0,1)$ and ϵ_t is an innovation process with mean $E(\epsilon_t) = \mu_\epsilon$ and variance $Var(\epsilon_t) = \sigma^2_{\epsilon_t}$.
According to Steutel and van Harn (1979), the binomial thinning operator "o" is defined as

$$\alpha \ o \ Y_t = \sum_{j=i}^{Y_t} Z_j \tag{28}$$

where Z_j is the Bernoulli random variable with $P(Z_j = 1) = p = 1 - P(Z_j = 0)$. The one-step transition probability of INAR(1) process is

$$P(Y_t = k|Y_{t-1} = l) = \sum_{\substack{i=1 \\ k,l \geq 0}}^{min(k,l)} P(B_l^p = i) P(\epsilon_t = k - i) \tag{29}$$

where $B_n^p \sim$ Binomial(n,p) and $p \in (0,1)$.
Following the results of McKenzie (1985) and Al-Osh and Alzaid (1987), we propose an INAR(1) process with DEIP innovations by assuming that the $\{\epsilon_t\}_{t\in\mathbb{Z}}$ innovations follow DEIP distribution, given in Equation(13). Thus, one-step transition probability of INAR(1) DEIP process is given by

$$\begin{aligned} P(Y_t = k|Y_{t-1} = l) = & \sum_{i=1}^{min(k,l)} \binom{l}{i} \alpha^i (1-\alpha)^{l-i} \times \\ & \frac{\left[e^{\lambda(1+\rho)e^{-\theta(k-i)}} - e^{\lambda\rho e^{-\theta(k-i)}}\right] - \left[e^{\lambda(1+\rho)e^{-\theta(k-i+1)}} - e^{\lambda\rho e^{-\theta(k-i+1)}}\right]}{e^{\lambda\rho}(e^\lambda - 1)}. \end{aligned} \tag{30}$$

The mean and variance of the Y_t process are respectively given by,

$$E(Y_t) = \frac{\mu_\epsilon}{1-\alpha} \tag{31}$$

$$V(Y_t) = \frac{\alpha\mu_\epsilon + \sigma^2_\epsilon}{1-\alpha^2}. \tag{32}$$

The mean and variance of the INAR(1)DEIP process can be computed by replacing μ_ϵ and σ_ϵ in Equation(31) and Equation(32) with Equation (21) and Equation(23) respectively. The conditional expectation and variance of INAR(1) DEIP process are given, respectively, as (see Weiß (2018) and Al-Osh and Alzaid (1988))

$$E(Y_t|Y_{t-1}) = pY_{t-1} + \mu_\epsilon \tag{33}$$

and

$$V(Y_t|Y_{t-1}) = p(1-p)Y_{t-1} + \sigma^2_\epsilon \tag{34}$$

where μ_ϵ and σ^2_ϵ are given in Equation(21) and Equation(23).

4. Characterizations

Characterizations of distributions is an important research area which has attracted the attention of many researchers. The problem of characterizing a distribution is an important problem, where an investigator is vitally interested to know if their model follows the right distribution. Thus, various characterization results have been reported in the literature. These characterizations have been established in different directions. In this section, we obtain three characterizations of DEIP distribution based on: (i) the hazard rate function and (ii) the reverse hazard rate function and (iii) conditional expectation of certain function of the random variable.

4.1. Characterization based on hazard rate function

Proposition 2: Let $Y : \Omega \to \mathbb{N}$ be a random variable. The pmf of Y is in Equation(13) if and only if its hazard rate function satisfies the difference equation

$$h(k+1) - h(k) = \left[\frac{e^{\lambda(1+\rho)e^{-\theta(k+1)}} - e^{\lambda\rho e^{-\theta(k+1)}}}{e^{\lambda(1+\rho)e^{-\theta(k+2)}} - e^{\lambda\rho e^{-\theta(k+2)}}}\right] - \left[\frac{e^{\lambda(1+\rho)e^{-\theta k}} - e^{\lambda\rho e^{-\theta k}}}{e^{\lambda(1+\rho)e^{-\theta(k+1)}} - e^{\lambda\rho e^{-\theta(k+1)}}}\right], \tag{35}$$

$k \in \mathbb{N}$, with the boundary condition $h(0) = \frac{e^{\lambda\rho(e^{-\theta}-1)}(e^{\lambda}-1)}{e^{\lambda e^{-\theta}}-1} - 1$.

Proof: If Y has pmf in Equation(13), then clearly Equation(35) holds. Now, if Equation(35) holds, then for every $y \in \mathbb{N}$ we have

$$\sum_{k=1}^{y-1} h(k+1) - h(k) = \sum_{k=1}^{y-1} \left[\frac{e^{\lambda(1+\rho)e^{-\theta(k+1)}} - e^{\lambda\rho e^{-\theta(k+1)}}}{e^{\lambda(1+\rho)e^{-\theta(k+2)}} - e^{\lambda\rho e^{-\theta(k+2)}}}\right] - \left[\frac{e^{\lambda(1+\rho)e^{-\theta k}} - e^{\lambda\rho e^{-\theta k}}}{e^{\lambda(1+\rho)e^{-\theta(k+1)}} - e^{\lambda\rho e^{-\theta(k+1)}}}\right].$$

$$h(y) - h(0) = \left[\frac{e^{\lambda(1+\rho)e^{-\theta y}} - e^{\lambda\rho e^{-\theta y}}}{e^{\lambda(1+\rho)e^{-\theta(y+1)}} - e^{\lambda\rho e^{-\theta(y+1)}}}\right] - \left[\frac{e^{\lambda\rho(e^{-\theta}-1)}(e^{\lambda}-1)}{e^{\lambda e^{-\theta}}-1}\right].$$

In view of the fact that $h(0) = \frac{e^{\lambda\rho(e^{-\theta}-1)}(e^{\lambda}-1)}{e^{\lambda e^{-\theta}}-1} - 1$, from the last equation we have

$$h(y) = \left[\frac{e^{\lambda(1+\rho)e^{-\theta y}} - e^{\lambda\rho e^{-\theta y}}}{e^{\lambda(1+\rho)e^{-\theta(y+1)}} - e^{\lambda\rho e^{-\theta(y+1)}}} - 1\right]$$

which in view of Equation(16), implies Y has pmf in Equation(13).

4.2. Characterization based on reverse hazard rate function

Proposition 3: Let $Y : \Omega \to \mathbb{N}$ be a random variable. The pmf of Y is in Equation(13) if and only if its reverse hazard rate function satisfies the difference equation

$$\begin{aligned} r(k+1) - r(k) = &\frac{\left[e^{\lambda(1+\rho)e^{-\theta(k+1)}} - e^{\lambda\rho e^{-\theta(k+1)}}\right] - \left[e^{\lambda(1+\rho)e^{-\theta(k+2)}} - e^{\lambda\rho e^{-\theta(k+2)}}\right]}{\left[e^{\lambda\rho}(e^{\lambda}-1)\right] - \left[e^{\lambda(1+\rho)e^{-\theta(k+2)}} - e^{\lambda\rho e^{-\theta(k+2)}}\right]} \\ &- \frac{\left[e^{\lambda(1+\rho)e^{-\theta k}} - e^{\lambda\rho e^{-\theta k}}\right] - \left[e^{\lambda(1+\rho)e^{-\theta(k+1)}} - e^{\lambda\rho e^{-\theta(k+1)}}\right]}{\left[e^{\lambda\rho}(e^{\lambda}-1)\right] - \left[e^{\lambda(1+\rho)e^{-\theta(k+1)}} - e^{\lambda\rho e^{-\theta(k+1)}}\right]} \end{aligned} \tag{36}$$

with the boundary condition $r(0) = 1$.

Proof: If Y has pmf in Equation(13), then clearly Equation(36) holds. Now, if Equation(36) holds, then for every $y \in \mathbb{N}$ we have

$$\sum_{k=1}^{y-1} r(k+1) - r(k) = \sum_{k=1}^{y-1} \frac{\left[e^{\lambda(1+\rho)e^{-\theta(k+1)}} - e^{\lambda\rho e^{-\theta(k+1)}}\right] - \left[e^{\lambda(1+\rho)e^{-\theta(k+2)}} - e^{\lambda\rho e^{-\theta(k+2)}}\right]}{\left[e^{\lambda\rho}(e^{\lambda}-1)\right] - \left[e^{\lambda(1+\rho)e^{-\theta(k+2)}} - e^{\lambda\rho e^{-\theta(k+2)}}\right]} - \frac{\left[e^{\lambda(1+\rho)e^{-\theta k}} - e^{\lambda\rho e^{-\theta k}}\right] - \left[e^{\lambda(1+\rho)e^{-\theta(k+1)}} - e^{\lambda\rho e^{-\theta(k+1)}}\right]}{\left[e^{\lambda\rho}(e^{\lambda}-1)\right] - \left[e^{\lambda(1+\rho)e^{-\theta(k+1)}} - e^{\lambda\rho e^{-\theta(k+1)}}\right]}. \tag{37}$$

Or,

$$r(y) - r(0) = \frac{\left[e^{\lambda(1+\rho)e^{-\theta y}} - e^{\lambda\rho e^{-\theta y}}\right] - \left[e^{\lambda(1+\rho)e^{-\theta(y+1)}} - e^{\lambda\rho e^{-\theta(y+1)}}\right]}{\left[e^{\lambda\rho}(e^{\lambda}-1)\right] - \left[e^{\lambda(1+\rho)e^{-\theta(y+1)}} - e^{\lambda\rho e^{-\theta(y+1)}}\right]} - 1.$$

In view of the fact that $r(0) = 1$, from the last equation we have

$$r(y) = \frac{\left[e^{\lambda(1+\rho)e^{-\theta y}} - e^{\lambda\rho e^{-\theta y}}\right] - \left[e^{\lambda(1+\rho)e^{-\theta(y+1)}} - e^{\lambda\rho e^{-\theta(y+1)}}\right]}{\left[e^{\lambda\rho}(e^{\lambda}-1)\right] - \left[e^{\lambda(1+\rho)e^{-\theta(y+1)}} - e^{\lambda\rho e^{-\theta(y+1)}}\right]}$$

which in view of Equation(17), implies Y has pmf in Equation(13).

4.3. Characterization in terms of the conditional expectation of certain function of the random variable

Proposition 4: Let $Y : \Omega \to \mathbb{N}$ be a random variable. The pmf of Y is in Equation(13) if and only if

$$E\left\{\frac{\left[e^{\lambda(1+\rho)e^{-\theta Y}} - e^{\lambda\rho e^{-\theta Y}}\right] + \left[e^{\lambda(1+\rho)e^{-\theta(Y+1)}} - e^{\lambda\rho e^{-\theta(Y+1)}}\right]}{\left[e^{\lambda\rho}(e^{\lambda}-1)\right]} | Y > k\right\} = \frac{e^{\lambda(1+\rho)e^{-\theta(k+1)}} - e^{\lambda\rho e^{-\theta(k+1)}}}{e^{\lambda\rho}(e^{\lambda}-1)}. \tag{38}$$

Proof: If Y has pmf Equation(13), then LHS of Eqution(38) will be

$$
\begin{aligned}
&(1-F(k))^{-1}\sum_{y=k+1}^{\infty}\frac{\left[e^{\lambda(1+\rho)e^{-\theta Y}}-e^{\lambda\rho e^{-\theta Y}}\right]+\left[e^{\lambda(1+\rho)e^{-\theta(Y+1)}}-e^{\lambda\rho e^{-\theta(Y+1)}}\right]}{[e^{\lambda\rho}(e^{\lambda}-1)]}\times\\
&\qquad\frac{\left[e^{\lambda(1+\rho)e^{-\theta Y}}-e^{\lambda\rho e^{-\theta Y}}\right]-\left[e^{\lambda(1+\rho)e^{-\theta(Y+1)}}-e^{\lambda\rho e^{-\theta(Y+1)}}\right]}{[e^{\lambda\rho}(e^{\lambda}-1)]}\\
&=(1-F(k))^{-1}\sum_{y=k+1}^{\infty}\left(\frac{e^{\lambda(1+\rho)e^{-\theta y}}-e^{\lambda\rho e^{-\theta y}}}{e^{\lambda\rho}(e^{\lambda}-1)}\right)^2-\left(\frac{e^{\lambda(1+\rho)e^{-\theta(y+1)}}-e^{\lambda\rho e^{-\theta(y+1)}}}{e^{\lambda\rho}(e^{\lambda}-1)}\right)^2 \qquad (39)\\
&\qquad=\left(\frac{e^{\lambda\rho}(e^{\lambda}-1)}{e^{\lambda(1+\rho)e^{-\theta(k+1)}}-e^{\lambda\rho e^{-\theta(k+1)}}}\right)\left(\frac{e^{\lambda(1+\rho)e^{-\theta(k+1)}}-e^{\lambda\rho e^{-\theta(k+1)}}}{e^{\lambda\rho}(e^{\lambda}-1)}\right)^2\\
&\qquad=\frac{e^{\lambda(1+\rho)e^{-\theta(k+1)}}-e^{\lambda\rho e^{-\theta(k+1)}}}{e^{\lambda\rho}(e^{\lambda}-1)}.
\end{aligned}
$$

Conversely, if Equation (38) holds, then

$$
\begin{aligned}
&\sum_{y=k+1}^{\infty}\frac{\left[e^{\lambda(1+\rho)e^{-\theta y}}-e^{\lambda\rho e^{-\theta y}}\right]+\left[e^{\lambda(1+\rho)e^{-\theta(y+1)}}-e^{\lambda\rho e^{-\theta(y+1)}}\right]}{[e^{\lambda\rho}(e^{\lambda}-1)]}f(y)\\
&\quad=\sum_{y=k+1}^{\infty}\frac{\left[e^{\lambda(1+\rho)e^{-\theta y}}-e^{\lambda\rho e^{-\theta y}}\right]+\left[e^{\lambda(1+\rho)e^{-\theta(y+1)}}-e^{\lambda\rho e^{-\theta(y+1)}}\right]}{[e^{\lambda\rho}(e^{\lambda}-1)]}\times\\
&\frac{\left[e^{\lambda(1+\rho)e^{-\theta Y}}-e^{\lambda\rho e^{-\theta Y}}\right]-\left[e^{\lambda(1+\rho)e^{-\theta(Y+1)}}-e^{\lambda\rho e^{-\theta(Y+1)}}\right]}{[e^{\lambda\rho}(e^{\lambda}-1)]}\\
&\quad=\sum_{y=k+1}^{\infty}\left(\frac{e^{\lambda(1+\rho)e^{-\theta y}}-e^{\lambda\rho e^{-\theta y}}}{e^{\lambda\rho}(e^{\lambda}-1)}\right)^2-\left(\frac{e^{\lambda(1+\rho)e^{-\theta(y+1)}}-e^{\lambda\rho e^{-\theta(y+1)}}}{e^{\lambda\rho}(e^{\lambda}-1)}\right)^2 \qquad (40)\\
&\quad=\left(\frac{e^{\lambda(1+\rho)e^{-\theta(k+1)}}-e^{\lambda\rho e^{-\theta(k+1)}}}{e^{\lambda\rho}(e^{\lambda}-1)}\right)^2\\
&\quad=(1-F(k))\left(\frac{e^{\lambda(1+\rho)e^{-\theta(k+1)}}-e^{\lambda\rho e^{-\theta(k+1)}}}{e^{\lambda\rho}(e^{\lambda}-1)}\right)\\
&\quad=(1-F(k+1)+f(k+1))\left(\frac{e^{\lambda(1+\rho)e^{-\theta(k+1)}}-e^{\lambda\rho e^{-\theta(k+1)}}}{e^{\lambda\rho}(e^{\lambda}-1)}\right).
\end{aligned}
$$

From Equation (39), we also have,

$$
\begin{aligned}
&\sum_{y=k+2}^{\infty}\frac{\left[e^{\lambda(1+\rho)e^{-\theta y}}-e^{\lambda\rho e^{-\theta y}}\right]+\left[e^{\lambda(1+\rho)e^{-\theta(y+1)}}-e^{\lambda\rho e^{-\theta(y+1)}}\right]}{[e^{\lambda\rho}(e^{\lambda}-1)]}f(y)\\
&\qquad=(1-F(k+1))\left(\frac{e^{\lambda(1+\rho)e^{-\theta(k+2)}}-e^{\lambda\rho e^{-\theta(k+2)}}}{e^{\lambda\rho}(e^{\lambda}-1)}\right).
\end{aligned}
\qquad (41)
$$

Now, subtracting Equation (41) from Equation (40), we arrive at

$$\left(\frac{e^{\lambda(1+\rho)e^{-\theta(k+2)}}-e^{\lambda\rho e^{-\theta(k+2)}}}{e^{\lambda\rho}(e^{\lambda}-1)}\right) f(k+1) =$$
$$(1-F(k+1))\left(\frac{\left(e^{\lambda(1+\rho)e^{-\theta(k+1)}}-e^{\lambda\rho e^{-\theta(k+1)}}\right)-\left(e^{\lambda(1+\rho)e^{-\theta(k+2)}}-e^{\lambda\rho e^{-\theta(k+2)}}\right)}{e^{\lambda\rho}(e^{\lambda}-1)}\right).$$
$$h(y)=\frac{f(k+1)}{1-F(k+1)}=\left(\frac{\left(e^{\lambda(1+\rho)e^{-\theta(k+1)}}-e^{\lambda\rho e^{-\theta(k+1)}}\right)-\left(e^{\lambda(1+\rho)e^{-\theta(k+2)}}-e^{\lambda\rho e^{-\theta(k+2)}}\right)}{e^{\lambda(1+\rho)e^{-\theta(k+2)}}-e^{\lambda\rho e^{-\theta(k+2)}}}\right)$$

which, in view of Equation (16), implies that Y has pmf in Equation (13).

5. Estimation and simulation

In this section, some estimation methods are discussed. In particular, we considered the following estimation methods: Maximum likelihood (ML) estimation, ordinary least square (OLS) estimation, weighted least square (WLS) estimation and Cramer-von Mises (CVM) estimation.

5.1. Maximum likelihood estimation

We apply method of ML estimation for estimating the parameter vector $\beta=(\lambda,\rho,\theta)^T$ of DEIP distribution. Let $(y_1,y_2,...,y_n)$ be a random sample of size n, drawn from DEIP (λ,ρ,θ) distribution.
The log likelihood function is given below

$$\begin{aligned} logL =& -n(\lambda\rho+log(e^{\lambda}-1))+ \\ & \sum_{j=1}^{n} log\left\{e^{\lambda\rho e^{-\theta y_j}}\left[e^{\lambda e^{-\theta y_j}}-1\right]-e^{\lambda\rho e^{-\theta(y_j+1)}}\left[e^{\lambda e^{-\theta(y_j+1)}}-1\right]\right\}. \end{aligned} \tag{42}$$

By differentiating Equation 42 with respect to the parameters λ, ρ and θ, we get non linear likelihood equations as follows.

$$\begin{aligned} \frac{\partial logL}{\partial\lambda} =& -n\rho-\frac{ne^{\lambda}}{e^{\lambda}-1}+ \\ & \sum_{j=1}^{n}\frac{(\rho+1)e^{-\theta y_j}A_1-\rho e^{-\theta y_j}A_2-(\rho+1)e^{-\theta(y_j+1)}A_3+\rho e^{-\theta(y_j+1)}A_4}{A_1-A_2-A_3+A_4}. \end{aligned} \tag{43}$$

$$\frac{\partial logL}{\partial\rho}=-n\lambda+\sum_{j=1}^{n}\frac{\lambda e^{-\theta y_j}(A_1+A_2)-\lambda e^{-\theta(y_j+1)}(A_3-A_3)}{A_1-A_2-A_3+A_4}. \tag{44}$$

$$\frac{\partial logL}{\partial\theta}=\sum_{j=1}^{n}\frac{\lambda y_j e^{-\theta y_j}(A_2\rho-A_1(\rho+1))+\lambda(y_j+1)e^{-\theta(y_j+1)}(A_3(\rho+1)-A_4\rho)}{A_1-A_2-A_3+A_4} \tag{45}$$

where $A_1=e^{\lambda(\rho+1)e^{-\theta y_j}}$, $A_2=e^{\lambda\rho e^{-\theta y_j}}$, $A_3=e^{\lambda(\rho+1)e^{-\theta(y_j+1)}}$ and $A_4=e^{\lambda\rho e^{-\theta(y_j+1)}}$.
These Equations(43–45) cannot be solved analytically, therefore an iterative procedure like Newton Raphson is required to solve them numerically. The solutions of likelihood Equations (43–45) provide ML estimators of $\beta=(\lambda,\rho,\theta)^T$, say $\hat{\beta}=(\hat{\lambda},\hat{\rho},\hat{\theta})^T$.

The conditions for maximum are obtained as:

Let $g_1(\lambda;\rho,\theta,y)$ denote the function on the right hand side (RHS) of Equation (43) where ρ and θ are the true values of the parameters. Then there exist atleast one root for $g_1(\lambda;\rho,\theta,y)=0$ for $\lambda\in(0,\infty)$ and the solution is unique when
$\sum_{j=1}^{n}\frac{e^{-2\theta y_j}[(1+\rho)^2A_1-\rho^2A_2]-e^{-2\theta(y_j+1)}[(1+\rho)^2A_3-\rho^2A_4]}{A_1-A_2-A_3+A_4}<$
$\frac{ne^{\lambda}}{(e^{\lambda}-1)^2}+$

$\sum_{j=1}^{n}\frac{(\rho[e^{-\theta(y_j+1)}A_4-e^{-\theta y_j}A_2]-(1+\rho)[e^{-\theta(jy+1)}A_3-e^{-\theta y_j}A_1])(e^{-\theta y_j}((1+\rho)A_1-\rho A_2)-e^{-\theta(y_j+1)}((1+\rho)A_3-\rho A_4))}{(A_1-A_2-A_3+A_4)^2}$.

Let $g_2(\rho;\lambda,\theta,y)$ denote the function on the right hand side (RHS) of Equation (44) where λ and θ are the true values of the parameters. Then there exist atleast one root for $g_2(\rho;\lambda,\theta,y)=0$ for $\rho\in(0,\infty)$ when

$$-n+\sum_{j=1}^{n}\frac{e^{-\theta y}(1+e^{\lambda e^{-\theta y}})-e^{-\theta(y+1)}(e^{\lambda e^{-\theta(y+1)}}-1)}{e^{\lambda e^{-\theta y}}-e^{\lambda e^{-\theta(y+1)}}}>0$$

and the solution is unique when

$$\sum_{j=1}^{n}\frac{\lambda^2e^{-2\theta y_j}(A_2+A_4)-\lambda^2e^{-2\theta(y_j+1)}(A_3-A_1)}{A_1-A_2-A_3+A_4}<\sum_{j=0}^{n}\frac{(\lambda e^{-\theta y_j}(A_2+A_1)-\lambda e^{-\theta(y_j+1)}(A_3-A_4))^2}{(A_1-A_2-A_3+A_4)^2}.$$

Let $g_3(\theta;\lambda,\rho,y)$ denote the function on the right hand side (RHS) of Equation (45) where ρ and θ are the true values of the parameters. Then there exist atleast one root for $g_3(\theta;\lambda,\rho,y)=0$ for $\theta\in(0,\infty)$ and the solution is unique when

$\sum_{j=1}^{n}\frac{y_j^2\lambda^2(1+\rho)\rho e^{-2\theta y_j}(A_1-A_2)-(1+y_j)^2\lambda^2e^{-2\theta(y_j+1)}((1+\rho)^2A_3-\rho^2A_4)-\lambda y_j^2e^{-\theta y_j}((1+\rho)A_2-\rho A_1)-\lambda(1+y_j)^2e^{-\theta(y_j+1)((1+\rho)A_3-\rho A_4)}}{A_1-A_2-A_3+A_4}<$

$\sum_{j=1}^{n}\frac{((y_j+1)\lambda e^{-\theta(y_j+1)}[(1+\rho)A_3-\rho A_4]-y_j\lambda e^{-\theta y_j}[(1+\rho)A_1-\rho A_2])(\lambda y_je^{-\theta y_j}[(1+\rho)A_2-\rho A_1]+\lambda(y_j+1)e^{-\theta(y_j+1)}[(1+\rho)A_3-\rho A_4])}{(A_1-A_2-A_3+A_4)^2}$

where $A_1=e^{\lambda(\rho+1)e^{-\theta y_j}}$, $A_2=e^{\lambda\rho e^{-\theta y_j}}$, $A_3=e^{\lambda(\rho+1)e^{-\theta(y_j+1)}}$ and $A_4=e^{\lambda\rho e^{-\theta(y_j+1)}}$.

5.2. Ordinary least square estimation

This method is based on the observed sample $y_1,y_2,...,y_n$ from n ordered random sample of any distribution with CDF, where $F(.)$ denotes the CDF, we get

$$E(F(y_j))=\frac{j}{(n+1)}.$$

The OLS estimators are obtained by minimizing

$$OLS(\lambda,\rho,\theta)=\sum_{j=1}^{n}(F(y_j)-\frac{j}{n+1})^2. \tag{46}$$

Putting the CDF of DEIP in Equation (46) we get

$$OLS(\lambda,\rho,\theta)=\sum_{j=1}^{n}\left[1-\left[\frac{e^{\lambda(1+\rho)e^{-\theta(y_j+1)}}-e^{\lambda\rho e^{-\theta(y_j+1)}}}{e^{\lambda\rho}(e^{\lambda}-1)}\right]-\frac{j}{n+1}\right]^2. \tag{47}$$

After differentiating Equation (47) with respect to the parameters λ,ρ and θ and equating to zero, the normal equations are as follows:

$$\begin{aligned}\frac{\partial OLS}{\partial\lambda}=&2\sum_{j=1}^{n}\left(1-\left[\frac{e^{\lambda(1+\rho)e^{-\theta(y_j+1)}}-e^{\lambda\rho e^{-\theta(y_j+1)}}}{e^{\lambda\rho}(e^{\lambda}-1)}\right]-\frac{j}{n+1}\right)\frac{1}{(e^{\lambda}-1)^2}\\&A_4e^{-\theta(y_j+1)-\lambda\rho}\left(\rho e^{\theta(y_j+1)}+(1-e^{\lambda})((1+\rho)e^{\lambda e^{-\theta(y_j+1)}}-\rho)\right)+\\&e^{\theta(y_j+1)}(e^{\lambda}(1+\rho)(e^{\lambda e^{-\theta(y_j+1)}}-1)-\rho e^{\lambda e^{-\theta(y_j+1)}}).\end{aligned} \tag{48}$$

$$\begin{aligned}\frac{\partial OLS}{\partial\rho}=&2\sum_{j=1}^{n}\left(1-\left[\frac{e^{\lambda(1+\rho)e^{-\theta(y_j+1)}}-e^{\lambda\rho e^{-\theta(y_j+1)}}}{e^{\lambda\rho}(e^{\lambda}-1)}\right]-\frac{j}{n+1}\right)\\&\left(\frac{(A_4-A_3)(\lambda e^{\lambda\rho}(e^{-\theta(y_j+1)}-1))}{e^{\lambda}-1}\right).\end{aligned} \tag{49}$$

$$\begin{aligned}\frac{\partial OLS}{\partial\theta}=&2\sum_{j=1}^{n}\left(1-\left[\frac{e^{\lambda(1+\rho)e^{-\theta(y_j+1)}}-e^{\lambda\rho e^{-\theta(y_j+1)}}}{e^{\lambda\rho}(e^{\lambda}-1)}\right]-\frac{j}{n+1}\right)\\&\frac{e^{-\lambda\rho}\lambda(1+y_j)e^{-\theta(y_j+1)}((1+\rho)A_3-\rho A_4)}{e^{\lambda}-1}\end{aligned} \tag{50}$$

where $A_3=e^{\lambda(\rho+1)e^{-\theta(y_j+1)}}$ and $A_4=e^{\lambda\rho e^{-\theta(y_j+1)}}$. The above non-linear equations cannot be solved analytically. So the OLS estimators of λ, ρ and θ can be obtained by using some iterative techniques likes Newton-Raphson method.

5.3. Weighted least square estimation

The WLS estimators can be obtained by minimizing

$$WLS(\lambda,\rho,\theta)=\sum_{j=1}^{n}w_j(F(y_i)-\frac{j}{n+1})^2 \tag{51}$$

with respect to the unknown parameters, where $w_j=\frac{1}{Var(F(Y_j))}=\frac{(n+1)^2(n+2)}{j(n-j+1)}$.

Putting the CDF of DEIP distribution in Equation (51), we get

$$WLS(\lambda,\rho,\theta)=\sum_{j=1}^{n}w_j\left[1-\left[\frac{e^{\lambda(1+\rho)e^{-\theta(y_j+1)}}-e^{\lambda\rho e^{-\theta(y_j+1)}}}{e^{\lambda\rho}(e^{\lambda}-1)}\right]-\frac{j}{n+1}\right]^2. \tag{52}$$

The Equation(52) is differentiated with respect to the parameters λ, ρ and θ and then equating to zero, the normal equations are as follows:

$$\begin{aligned}\frac{\partial WLS}{\partial \lambda} =& 2\sum_{j=1}^{n}\left(1-\left[\frac{e^{\lambda(1+\rho)e^{-\theta(y_j+1)}}-e^{\lambda\rho e^{-\theta(y_j+1)}}}{e^{\lambda\rho}(e^{\lambda}-1)}\right]-\frac{j}{n+1}\right)\frac{1}{(e^{\lambda}-1)^2}\\ & A_4 e^{-\theta(y_j+1)-\lambda\rho} w_j(\rho e^{\theta(y_j+1)}+(1-e^{\lambda})((1+\rho)e^{\lambda e^{-\theta(y_j+1)}}-\rho)+\\ & e^{\theta(y_j+1)}(e^{\lambda}(1+\rho)(e^{\lambda e^{-\theta(y_j+1)}}-1)-\rho e^{\lambda e^{-\theta(y_j+1)}})).\end{aligned} \tag{53}$$

$$\begin{aligned}\frac{\partial WLS}{\partial \rho} =& 2\sum_{j=1}^{n}\left(1-\left[\frac{e^{\lambda(1+\rho)e^{-\theta(y_j+1)}}-e^{\lambda\rho e^{-\theta(y_j+1)}}}{e^{\lambda\rho}(e^{\lambda}-1)}\right]-\frac{j}{n+1}\right)\\ & w_j\left(\frac{(A_4-A_3)(\lambda e^{\lambda\rho}(e^{-\theta(y_j+1)}-1))}{e^{\lambda}-1}\right).\end{aligned} \tag{54}$$

$$\begin{aligned}\frac{\partial WLS}{\partial \theta} =& 2\sum_{j=1}^{n}\left(1-\left[\frac{e^{\lambda(1+\rho)e^{-\theta(y_j+1)}}-e^{\lambda\rho e^{-\theta(y_j+1)}}}{e^{\lambda\rho}(e^{\lambda}-1)}\right]-\frac{j}{n+1}\right)\\ & w_j\frac{e^{-\lambda\rho}\lambda(1+y_j)e^{-\theta(y_j+1)}((1+\rho)A_3-\rho A_4)}{e^{\lambda}-1}.\end{aligned} \tag{55}$$

where $A_3 = e^{\lambda(\rho+1)e^{-\theta(y_j+1)}}$ and $A_4 = e^{\lambda\rho e^{-\theta(y_j+1)}}$. These above nonlinear equations cannot be solved analytically. Therefore the WLS estimates can be obtained by using any iterative procedure techniques such as Newton-Raphson type algorithms.

5.4. Cramer-von Mises estimation

The CVM estimates of the parameter λ, ρ and θ are obtained by minimizing the following expression with respect to the parameters λ, ρ and θ respectively.

$$CVM_{\lambda,\rho,\theta} = \frac{1}{12n} + \sum_{j=1}^{n}(F(y_j) - \frac{-1+2j}{2n})^2. \tag{56}$$

For in the case of DEIP distribution, put CDF of DEIP in Equation(56).

$$CVM_{\lambda,\rho,\theta} = \frac{1}{12n} + \sum_{j=1}^{n}\left[1-\left[\frac{e^{\lambda(1+\rho)e^{-\theta(y_j+1)}}-e^{\lambda\rho e^{-\theta(y_j+1)}}}{e^{\lambda\rho}(e^{\lambda}-1)}\right]-\frac{-1+2j}{2n}\right)^2. \tag{57}$$

By differentiating Equation(57) with respect to the parameters λ, ρ and θ and equating to zero, we get the normal equations as follows:

$$\begin{aligned}\frac{\partial CVM}{\partial \lambda} =& 2\sum_{j=1}^{n}(\left(1-\left[\frac{e^{\lambda(1+\rho)e^{-\theta(y_j+1)}}-e^{\lambda\rho e^{-\theta(y_j+1)}}}{e^{\lambda\rho}(e^{\lambda}-1)}\right]-\frac{-1+2j}{2n}\right)\\ & \frac{1}{(e^{\lambda}-1)^2}A_4 e^{-\theta(y_j+1)-\lambda\rho})(\rho e^{\theta(y_j+1)}+(1-e^{\lambda})((1+\rho)e^{\lambda e^{-\theta(y_j+1)}}-\rho)+\\ & e^{\theta(y_j+1)}(e^{\lambda}(1+\rho)(e^{\lambda e^{-\theta(y_j+1)}}-1)-\rho e^{\lambda e^{-\theta(y_j+1)}})).\end{aligned} \tag{58}$$

$$\frac{\partial CVM}{\partial \rho} = 2\sum_{j=1}^{n}\left(1-\left[\frac{e^{\lambda(1+\rho)e^{-\theta(y_j+1)}}-e^{\lambda\rho e^{-\theta(y_j+1)}}}{e^{\lambda\rho}(e^{\lambda}-1)}\right]-\frac{-1+2j}{2n}\right)\left(\frac{(A_4-A_3)(\lambda e^{\lambda\rho}(e^{-\theta(y_j+1)}-1))}{e^{\lambda}-1}\right). \tag{59}$$

$$\frac{\partial CVM}{\partial \theta} = 2\sum_{j=1}^{n}\left(1-\left[\frac{e^{\lambda(1+\rho)e^{-\theta(y_j+1)}}-e^{\lambda\rho e^{-\theta(y_j+1)}}}{e^{\lambda\rho}(e^{\lambda}-1)}\right]-\frac{-1+2j}{2n}\right)\frac{e^{-\lambda\rho}\lambda(1+y_j)e^{-\theta(y_j+1)}((1+\rho)A_3-\rho A_4)}{e^{\lambda}-1} \tag{60}$$

where $A_3 = e^{\lambda(\rho+1)e^{-\theta(y_j+1)}}$ and $A_4 = e^{\lambda\rho e^{-\theta(y_j+1)}}$.
These Equations (58–60) cannot be solved analytically. The estimates of λ, ρ and θ can be obtained by setting the normal equations equations equal to zero and solving simultaneously with the help of statistical packages like *optim* or *nlm* in **R** programming.

5.5. Simulation

Here we examine the performance of the estimates of DEIP parameters using simulation study with 1000 replications. We calculate the estimates and mean square errors(MSE) using the **R** package. We used "*nlm*" function in **R** program for ML estimation and "*optim*" function is used for the estimation of OLS, WLS and CVM. The simulation procedure is given below.

1. Generate $N = 1000$ samples of sizes $n = 50, 100, 300$ from DEIP(0.1, 0.1, 0.1) and DEIP(0.5, 0.9, 0.1).

   Here, the random variable X possesses a continuous Exponential Intervened Poisson (EIP) distribution with parameters λ, ρ and θ. Then $Y = \lfloor X \rfloor$ follows the DEIP distribution with parameters λ, ρ and θ. To generate data from the DEIP distribution, first we have to generate data from EIP. Then take the integer values of each generated observation to get the simulated data set. The procedure of generating random samples from EIP distribution is explained in Jayakumar and Sankaran (2019). Initial values are chosen to compute the estimates in such way that the optimization function having minimum bias.

2. Compute the estimates for the 1000 samples, say $\hat{\beta}$ for j = 1, 2, ..., 1000.

3. Compute MSE by using the below quantity.

$$MSE(\hat{\beta}) = \frac{1}{1000}\sum_{j=1}^{1000}(\hat{\beta}-\beta)^2. \tag{61}$$

4. Compute the coverage probabilities [CP] of the estimates.

The empirical result from the Table 3 is when the sample size increases the MSEs of the parameter decreases. This shows the consistency of the estimators. Also, CVM estimates perform better when compared to other estimates.

Table 3: Estimates of of λ, ρ and θ

	True values→	$\lambda = 0.1$, $\rho = 0.1$, $\theta = 0.1$				$\lambda = 0.5$, $\rho = 0.9$, $\theta = 0.1$			
Sample size(n)	Parameter↓	ML	OLS	WLS	CVM	ML	OLS	WLS	CVM
50	$\hat{\lambda}$	0.1383	0.1761	0.1596	0.0961	0.5670	0.6458	0.6351	0.5293
	(MSE)	(0.2121)	(0.3562)	(0.4970)	(0.1831)	(0.4031)	(0.6234)	(0.6037)	(0.3960)
	[CP]	[0.841]	[0.617]	[0.700]	[0.863]	[0.796]	[0.635]	[0.641]	[0.801]
	$\hat{\rho}$	0.1312	0.1796	0.1623	0.1071	0.9606	0.8764	0.8823	0.9724
	(MSE)	(0.3210)	(0.4013)	(0.4433)	(0.0961)	(0.2683)	(0.4801)	(0.4573)	(0.2541)
	[CP]	[0.800]	[0.672]	[0.727]	[0.854]	[0.801]	[0.765]	[0.768]	[0.834]
	$\hat{\theta}$	0.1402	0.1634	0.1706	0.1324	0.1451	0.1937	0.1969	0.1433
	(MSE)	(0.2150)	(0.4146)	(0.5231)	(0.1612)	(0.1732)	(0.4176)	(0.5154)	(0.1365)
	[CP]	[0.829]	[0.631]	[0.786]	[0.847]	[0.896]	[0.719]	[0.703]	[0.902]
100	$\hat{\lambda}$	0.1238	0.1571	0.1431	0.0989	0.5312	0.6032	0.5993	0.5240
	(MSE)	(0.1972)	(0.3237)	(0.4130)	(0.1645)	(0.3632)	(0.5710)	(0.5651)	(0.3649)
	[CP]	[0.850]	[0.651]	[0.711]	[0.867]	[0.804]	[0.691]	[0.699]	[0.820]
	$\hat{\rho}$	0.1300	0.1586	0.1604	0.1009	0.9510	0.8923	0.8967	0.9813
	(MSE)	(0.2291)	(0.3913)	(0.4312)	(0.0801)	(0.2402)	(0.4154)	(0.4403)	(0.2130)
	[CP]	[0.838]	[0.699]	[0.732]	[0.861]	[0.864]	[0.791]	[0.793]	[0.856]
	$\hat{\theta}$	0.1352	0.1546	0.1695	0.1308	0.1363	0.1891	0.1893	0.1382
	(MSE)	(0.1676)	(0.3001)	(0.4961)	(0.1532)	(0.1565)	(0.4073)	(0.5035)	(0.1325)
	[CP]	[0.851]	[0.657]	[0.789]	[0.857]	[0.916]	[0.763]	[0.746]	[0.917]
300	$\hat{\lambda}$	0.1211	0.1503	0.1364	0.1061	0.5021	0.5638	0.5712	0.5009
	(MSE)	(0.1681)	(0.3146)	(0.3291)	(0.0261)	(0.1641)	(0.3173)	(0.3630)	(0.1512)
	[CP]	[0.891]	[0.672]	[0.780]	[0.893]	[0.899]	[0.747]	[0.731]	[0.902]
	$\hat{\rho}$	0.1281	0.1492	0.1470	0.1006	0.9371	0.8974	0.8982	0.9503
	(MSE)	(0.1441)	(0.3530)	(0.3021)	(0.0541)	(0.2121)	(0.3903)	(0.4102)	(0.2053)
	[CP]	[0.886]	[0.724]	[0.786]	[0.899]	[0.881]	[0.803]	[0.800]	[0.874]
	$\hat{\theta}$	0.1206	0.1488	0.1501	0.1201	0.1235	0.1709	0.1742	0.1292
	(MSE)	(0.1121)	(0.2943)	(0.3213)	(0.1036)	(0.2751)	(0.4156)	(0.5019)	(0.2601)
	[CP]	[0.894]	[0.786]	[0.779]	[0.864]	[0.930]	[0.796]	[0.758]	[0.929]

6. Application

In this section, we illustrate the flexibility of the proposed distribution using a real data set.

The fit of the proposed distribution is compared with the following distributions:

- Poisson (P) distribution having pmf

$$\mathrm{P(Y = y)} = \frac{e^{-\lambda}\lambda^{y}}{y!};\ \lambda \geq 0,\ \ y = 0, 1, 2, ...$$

- Discrete Burr (Krishna and Pundir (2009)) (DB) distribution having pmf

$$P(Y = y) = \theta^{log(1+y^{\alpha})} - \theta^{log(1+(1+y^{\alpha}))};\ 0 < \theta < 1,\ \alpha > 0\ ,\ y = 0, 1, 2, ...$$

- Discrete Gumbel (Chakraborty and Chakravarty (2014)) (DG) distribution having pmf

$$P(Y = y) = exp(-\alpha p^{y+1}) - exp(-\alpha p^{y}); \alpha > 0, 0 < p < 1,\ \ \ y = 0, 1, 2, ...$$

- A new three-parameter Poisson-Lindley (NTPPL) distribution (Das *et al.* (2018)) having pmf

$$P(Y=y)=\frac{\theta^2}{(\theta+1)^{x+2}}\left(1+\frac{\alpha+\beta x}{\theta\alpha+\beta}\right);\theta>0,\beta>0,\theta\alpha+\beta>0,y=0,1,2,...$$

The data set is taken from Efron (1988), a study of 51 patients with head and neck cancer conducted by the Northern California Oncology Group. We compare the fits of the DEIP distribution with the competitive models DG, DB, NTPPL and Poisson. The data is given below:

{0, 1, 1, 2, 2, 2, 2, 2, 2, 3, 3, 4, 4, 4, 4, 4, 4, 4, 4, 5, 5, 5, 5, 5, 5, 5, 6, 7, 7, 7, 8,8, 9, 9, 9, 10, 13, 13, 13, 14, 17, 17, 19, 19, 36, 36, 37, 40, 44, 46, 46}

Table 4: Goodness of fit for various models fitted for the dataset.

Model	P	DG	DB	NTPPL	DEIP
Estimates	$\hat{\lambda}=11.314$	$\hat{\alpha}=3.207$ $\hat{p}=0.668$	$\hat{\alpha}=2.551$ $\hat{\theta}=0.730$	$\hat{\alpha}=1.506$ $\hat{\beta}=1.282\times10^{6}$ $\hat{\theta}=0.088$	$\hat{\lambda}=2.212$ $\hat{\rho}=4.404$ $\hat{\theta}=0.007$
K-S	0.4787	0.9821	0.2955	6.665	0.1656
p-value	<0.000	<0.000	<0.000	0.000	0.1217

The K-S statistic given in Table 4 is smallest for the DEIP distribution with the value of 0.1656 and p- value is 0.1217, which is higher when compared to other models. That is, Table 4 gives that the DEIP distribution leads to a better fit for the data set compared to the other four models.

7. Conclusion

In the present article, we have introduced a new family of discrete distributions called DIPc family. One special model of the proposed family are studied in detail. Further we have noticed DIPc family can be used for modelling variety of failure data because its hazard rate can take different shapes. The methods of ML, OLS, WLS and CVM estimations have been utilized to estimate the unknown parametres of the models. Some characterizations of the proposed distribution have been also studied. An extensive simulation is carried out to evaluate the behaviour of the above stated estimation methods. The flexibility of the family has also been elucidated using a real data set. The new distribution can serve as a better alternative for modelling count data in various fields including reliability, insurance, medicine, engineering *etc.*

Acknowledgement

The authors wish to thank the Editor in Chief and the reviewer for the comments on an earlier version of the manuscript.

References

Al-Osh, M. and Alzaid, A. A. (1987). First-order integer-valued autoregressive (INAR (1)) process. *Journal of Time Series Analysis*, **8**, 261–275.

Al-Osh, M. and Alzaid, A. A. (1988). Integer-valued moving average (INMA) process. *Statistical Papers*, **29**, 281–300.

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

Chakraborty, S. and Chakravarty, D. (2014). A discrete gumbel distribution. *arXiv:1410.7568*, .

Choudhary, N., Tyagi, A., and Singh, B. (2021). Estimation of $R = P[Y < X < Z]$ under progressive type-II censored data from Weibull distribution. *Lobachevskii Journal of Mathematics*, **42**, 318–335.

Das, K. K., Ahmed, I., and Bhattacharjee, S. (2018). A new three-parameter Poisson-Lindley distribution for modelling over-dispersed count data. *International Journal of Applied Engineering Research*, **13**, 16468–16477.

Efron, B. (1988). Logistic regression, survival analysis, and the Kaplan-Meier curve. *Journal of the American statistical Association*, **83**, 414–425.

Jayakumar, K. and Sankaran, K. (2019). Exponential intervened Poisson distribution. *Communications in Statistics-Theory and Methods*, **47**, 1–31.

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

Marshall, A. W. and Olkin, I. (1997). A new method for adding a parameter to a family of distributions with application to the exponential and Weibull families. *Biometrika*, **84**, 641–652.

McKenzie, E. (1985). Some simple models for discrete variate time series 1. *JAWRA Journal of the American Water Resources Association*, **21**, 645–650.

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

Roy, D. (2003). The discrete normal distribution. *Communications in Statistics-Theory and Methods*, **32**, 1871–1883.

Roy, D. (2004). Discrete Rayleigh distribution. *IEEE Transactions on Reliability*, **53**, 255–260.

Shanmugam, R. (1985). An intervened Poisson distribution and its medical application. *Biometrics*, **41**, 1025–1029.

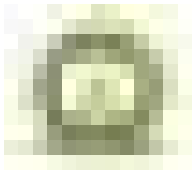
Stein, W. E. and Dattero, R. (1984). A new discrete Weibull distribution. *IEEE Transactions on Reliability*, **33**, 196–197.

Steutel, F. W. and van Harn, K. (1979). Discrete analogues of self-decomposability and stability. *The Annals of Probability*, **7**, 893–899.

Steutel, F. W. and van Harn, K. (2003). *Infinite Divisibility of Probability Distributions on the Real Line*. CRC Press.

Weiß, C. H. (2018). *An Introduction to Discrete-valued Time Series*. John Wiley & Sons.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 22, No. 1, 2024 (New Series), pp 123–136
http://www.ssca.org.in/journal



Estimation of Premium Cost for HIV/AIDS Patients Under ART in Presence of Prognostic Factors

Gurprit Grover and Parmeet Kumar Vinit
Department of Statistics, University of Delhi, Delhi, India

Received: 11 June 2022; Revised: 17 August 2022; Accepted: 22 June 2023

Abstract

This paper focuses to introduce an insurance plan for HIV/AIDS patients by incorporating the prognostic factors. High death rate among the HIV infected people always made a poor attention towards the insurance companies. The use of ART has declined the death rates drastically over the years which show a clear path for sustainable insurance plan. The long term survivors among HIV/AIDS patients can be treated as any other patient. The diagnosis involves a lot of time and financial investment. An affordable insurance plan will support the next of kin in case of patient's death. The survival probabilities in the presence of prognostic factors are obtained using COX-PH model and hence the cost of the premium is obtained using actuarial model.

Key words: HIV/AIDS; CD4; COX-PH; Regression; Premium; ART.

1. Introduction

Globally, there are 37.7 million people living with HIV/AIDS and 2.3 million in India in 2020. The incidence rate per 1000 uninfected population is 0.19 and 0.04, globally and India respectively. 6.8 million people died from HIV/AIDS-related illnesses in 2020 UNAIDS (2021),NACO (2019). The human immunodeficiency virus (HIV) targets the immune system and weakens people's defense against many infections and some types of cancer that people with healthy immune systems can fight off. As the virus destroys and impairs the function of immune cells, infected individuals gradually become immunodeficient. Immune function is typically measured by CD4 cell count and viral load. The most advanced stage of HIV infection is acquired immunodeficiency syndrome (AIDS), which can take many years to develop if not treated, depending upon the initial health of the individual. As per WHO, the stages of HIV/AIDS on the basis of CD4 count, WHO (2005) are presented in Table 1.

HIV can be managed by treatment regimens composed of a combination of three or more antiretroviral (ARV) drugs. Current antiretroviral therapy (ART) does not cure HIV infection but highly suppresses viral replication within a person's body and allows

Corresponding Author: Parmeet Kumar Vinit
Email: parmeet.vinit@gmail.com

Table 1: Clinical staging of HIV/AIDS infection established by WHO

Symptoms	Clinical stage	CD4 per microlitre
Asymptomatic	1	≥ 500
Mild symptoms	2	350-499
Advanced symptoms	3	200-349
Severe symptoms	4	< 200

an individual's immune system recovery to strengthen and regain the capacity to fight off opportunistic infections and some cancers.

Since 2016, WHO has recommended that all people living with HIV be provided with lifelong ART, including children, adolescents, adults and pregnant and breastfeeding women, regardless of clinical status or CD4 cell count WHO (2016). Around the world, 27.5 million people were able to access antiretroviral therapy (ART) in 2020 and 1.5 million in India UNAIDS (2021).

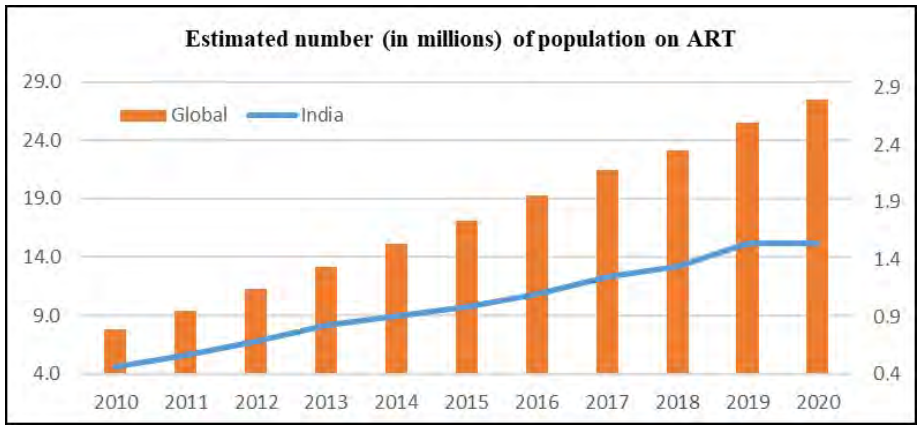


Figure 1: Estimated number (in millions) of population on ART, UNAIDS 2021

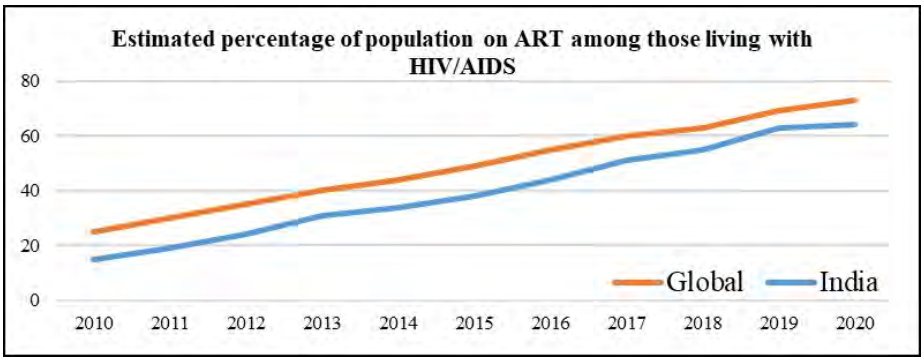


Figure 2: Estimated percentage of population on ART among those living with HIV/AIDS, UNAIDS 2021

Figure 1 shows the estimated population receiving ART globally and in India. The overall trend is upwards. Also, out of the infected population there is steady increase in

access to ART, Figure 2. The worldwide increased use of ART has contributed crucially for decline in rate of death, Collaborators *et al.* (2016), NACO and ICMR-NIMS (2017). This decrease in deaths shows long term survival in HIV/AIDS patients King Jr *et al.* (2003).

HIV/AIDS patients face a lot of problems in their day to day life. One such big problem is financial instability. The detection and treatment process exhausts a lot of time and money. The cost per new HIV diagnosis involves Rs. 866 to Rs. 1367 in England, Ong *et al.* (2016) and Rs. 1200 to Rs. 1610 in USA, Burns *et al.* (2013). According to NSSO 71st round data the average cost of diagnosis for the same is Rs. 1336 in India, Jain *et al.* (2015). Generally, the cost of diagnosis and other medical expenses are borne by the insurance company but there is no such provision for HIV/AIDS patients. This is due to high death rate in the infected population and expensive diagnosis. The current scenario of HIV/AIDS patients promises long term survival which supports the idea of providing an insurance plan for them.

Insurance is a means of protection from financial loss. It is a form of risk management, primarily used to hedge against the risk of a contingent or uncertain loss, such as death, severe illness, *etc.* The insured receives a contract, called the insurance policy, which details the conditions and circumstances under which the insurer will compensate the insured. The amount of money charged by the insurer to the Policy holder for the insurance policy is called the premium and the insured amount is paid once the event occurs. This premium cost is calculated by actuarial models. Actuarial models provide frameworks for analysis, allowing to project probable outcomes based on past experience adjusted for known material changes in circumstances. They are usually expressed in mathematical terms, and are typically designed to be consistent with fundamental principles of actuarial science. These models used are classified as either deterministic or stochastic. They are simplified representations of possible outcomes relative to future contingent events. A "contingent event" is an event whose occurrence, timing, or severity is uncertain. This contingent event can be death due to disease, sickness or accident, *etc.* Actuarial models may contain many elements and are usually based upon multiple interrelated assumptions about various aspects of risks associated with the event of interest. These models use probability of events and decrement models, Bowers *et al.* (1997).

Actuarial modeling is widely accepted to bring reliable methods for pricing the insurance contracts. As of the present there is no insurance company that provides any insurance policy to the HIV/AIDS patients. The HIV/AIDS patients and their dear ones involved in diagnosis process face lot of problems in terms of finance, so introducing such a plan will be beneficial in public interest. An insurance plan is introduced for patients suffering from Acute lymphoblastic leukemia (ALL), Grover *et al.* (2018) and HIV/AIDS Grover *et al.* (2021) based on the survival probabilities. This study considers the estimation the survival probability in presence of the prognostic factors and hence estimating the premium cost. The study suggests early enrolment in the insurance plan. The premium cost is lower in the earlier stage.

2. Methodology

A retrospective study is conducted for HIV/AIDS patients from a hospital located in Delhi. A total of 767 patients are selected for this study out of 5894 patients after a

complete-case analysis. The prognostic factors taken into study are age of the patients, smoking and drinking habit, drug addiction and modes of transmission. There are four modes of transmission 1- IDU (Injection Drug Users), 2- HOMO-MSM (sex with men to men), 3-HETERO (infected through sex) and 4-Blood (infected through blood transfusion). There are 556 males and 211 females. 84 of patients had smoking habit and 320 had drinking habit. Apart from this 685 were drug edict. Out of these four modes of transmission 10 got infected by mode 1, 6 by mode 2, 688 by mode 3 and remaining by mode 4.

The existing clinical staging of HIV/AIDS infection established by WHO in Table 1 is redefined into two categories for the study. The CD4 count is the primary factor considered for categorization and follow-up time is taken up for survival estimation. Based on the CD4 count the four categories mentioned in Table 1 are categorized further as:

Table 2: Categorising the existing stages of HIV/AIDS in symptomatic patients

HIV-associated symptoms	CD4 per microlitre	Category
Mild symptoms or Advanced symptoms	200-499	Category 1
Severe symptoms	<200	Category 2

The patients with mild or advanced symptoms having CD4 count from 200 to 499 are placed in Category 1 and patients with severe symptoms having CD4 count less than 200 are placed in Category 2 (Table 2). Throughout the article, three groups are considered for analysis. These are Category 1, Category 2 and the third is combined, when all the symptomatic patients are taken together including Category 1 and Category 2.

Let $\mu(t)$ be the hazard rate of a HIV/AIDS patient at time i. Cox-PH model is used to estimate the hazard rate in all three scenarios (viz, all the 767 patients and patients in the two categories described in Table 2), in the presence of prognostic factors. The purpose of this model is to evaluate simultaneously the effect of several prognostic factors on the survival of the patients. The Cox-PH model is defined as,

$$\mu(t) = \mu_0(t) \cdot \exp\left(b_1x_1 + b_2x_2 + ... + b_px_p\right)$$

Where, x_i is the $i^{th}(i = 1, 2, ..., p)$ prognostic factor, with coefficient b_i and $\mu_0(t)$ is baseline hazard rate.

Further, the probability of surviving one year (t) with hazard rate $\mu(t)$ is given by:

$$p_x = \exp\left\{-\int_x^{x+1} \mu(t)dt\right\}$$

These probabilities are used to obtain death probabilities in the presence of prognostic factors.

2.1. Premium model

Models for insurance are designed to reduce the financial impact of untimely death. Insurance systems are established to reduce the adverse financial impact of some type of ran-

dom event; here death is the primary event. To calculate the premium cost for HIV/AIDS patients a discrete model is considered Luptáková and Bilíková (2014), Haberman and Pitacco (1998). Deterministic model is derived from the principle of an unreal set and the equivalence principle which are basic principles of the classical methods Norberg (2000), Slud (2001). The estimated death probabilities of the HIV/AIDS patients (q_x) are used to calculate insurance premium for the given time period (t). We have considered the model for one year of insurance. The payable insurance premium cost ($\overline{A}^1_{(x,i)}$) for the patient, with a rate of interest i is calculated as

$$\overline{A}^1_{(x,i)} = q_x.v + p_x.q_(x+1).v^2$$

where, $v = (1+i)^{(-1)}$ is the discount factor used to calculate the value of unit currency after one year based on compound interest with the rate of interest i.

3. Results

3.1. Kaplan-Meier estimates

Kaplan-Meier survival estimates for HIV/AIDS patients are shown in Table 3. The survival estimates for the HIV/AIDS patients for the first year are 0.969 (SE=0.015), 0.909 (SE=0.12), and 0.923 (SE=0.010) for Category 1, Category 2, and combined (including patients of Category 1 and Category 2) respectively. The highest and lowest survival rates in the first year are for Category 1 and Category 2, respectively. For the patients taken combined the survival decreases by 4.98% in the next year. Further, from 7th to 8th year then it declines by 7.28%, and shows a sharp decline after this (Figure 3). For the patients in Category 1 the survival decreases by 6.91% in the very next year. In the year from 2nd to 3rd and 7th to 8th the survival declines by 5.65% and 5.26% respectively. In between it looks like a plateau (Figure 4). For the patients in Category 2, the survival drops by 4.51% in the next year, and 7th to 8th the survival declines by 9.62%. Figure 5 also show sharp decline in survival in the initial years and then after the 8th year. Eventually, after 5 years the survival estimates decline by 14.08%, 15.58% and 13.42% respectively.

Table 3: Kaplan-Meier survival estimates for the HIV/AIDS patients

Time (Years)	Combined	% Change	Category 1	% Change	Category 2	% Change
1	0.923		0.969		0.909	
2	0.877	4.98	0.902	6.91	0.868	4.51
3	0.839	4.33	0.851	5.65	0.834	3.92
4	0.817	2.62	0.818	3.88	0.814	2.40
5	0.793	2.94	0.818	0.00	0.787	3.32
6	0.776	2.14	0.8	2.20	0.767	2.54
7	0.742	4.38	0.761	4.88	0.738	3.78
8	0.688	7.28	0.721	5.26	0.667	9.62
9	0.508		0.451		0.523	

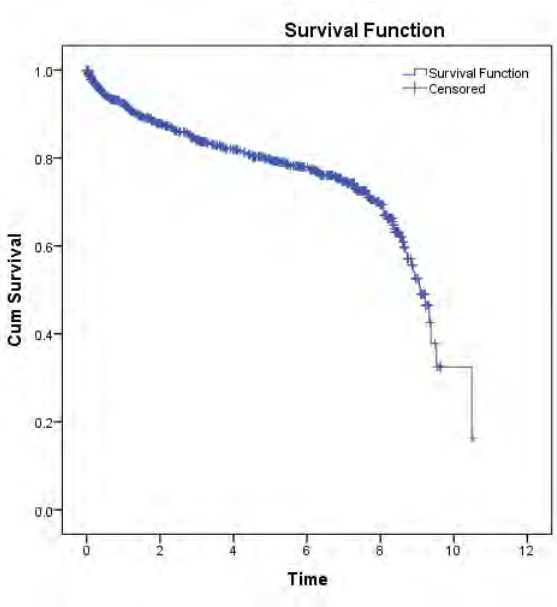


Figure 3: Kaplan Meier survival plot for the combined HIV/AIDS patients

3.2. Cox-PH regression survival estimates

Table 4 provides the Cox-PH model description for the combined HIV/AIDS patients. The model is highly significant with -2 log-likelihood value (1885.682) and Chi-square value as 38.362.

Table 4: Model description of Cox-PH for the combined HIV/AIDS patients

Omnibus Tests of Model Coefficients			
-2 Log Likelihood	**Overall (score)**		
	Chi-square	df	Sig.
1885.682	38.362	8	.000

Table 5: Model estimates of Cox-PH for the combined HIV/AIDS patients

Variables in the Equation								
	B	**SE**	**Wald**	**df**	**Sig.**	**Exp(B)**	**95.0% CI for Exp(B)**	
							Lower	**Upper**
Age	.018	.009	3.851	1	.050	1.018	1.000	1.036
Sex	-.285	.214	1.774	1	.183	.752	.495	1.144
Smoking	-.970	.204	22.672	1	.000	.379	.254	.565
DRUGS	.711	.323	4.836	1	.028	2.035	1.080	3.834
Alcohol	-.136	.194	.490	1	.484	.873	.597	1.277
MOT			1.497	3	.683			
MOT(1)	.042	.627	.005	1	.946	1.043	.305	3.563
MOT(2)	-.991	1.033	.921	1	.337	.371	.049	2.810
MOT(3)	-.211	.246	.736	1	.391	.810	.500	1.311

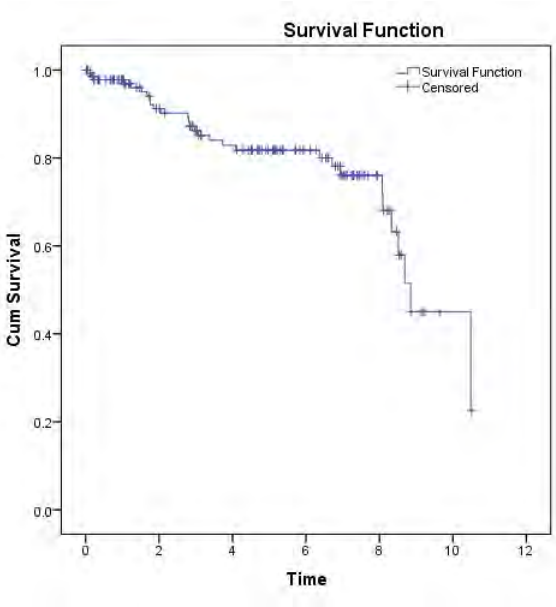


Figure 4: Kaplan Meier survival plot for the Category 1 HIV/AIDS patients

Table 5 explains the variables considered for the model of the combined HIV/AIDS patients. The variables include demography such as age and sex of the patient, smoking and drinking habit along with exposure to drugs. Apart from this method of transmission is one of the explanatory variables. Out of these variables we find that age (p-value= 0.050), habit of smoking (p-value= 0.000) and exposure to drugs (p-value= 0.028) are significant at 5% level of significance. Also, the hazard ratio suggests that for every one-year increase in age the risk will increase with a rate of 1.018. Similarly, for patients exposed to drugs will increase the risk with the rate 2.035.

Table 6: Cox-PH survival estimates for the combined HIV/AIDS patients

Time (Years)	Baseline Cum Hazard	Survival at mean of covariates	Std. Error	Cum Hazard
1	0.11	0.95	0.01	0.06
2	0.20	0.90	0.01	0.11
3	0.28	0.86	0.01	0.15
4	0.34	0.83	0.02	0.18
5	0.40	0.81	0.02	0.21
6	0.45	0.79	0.02	0.24
7	0.51	0.76	0.02	0.27
8	0.61	0.72	0.02	0.32
9	0.89	0.62	0.03	0.47

Table 6 provides the survival estimates for the combined HIV/AIDS patients obtained using Cox-PH regression. The estimated survival in the first year of follow-up is 0.95 and declines to 0.90. Also, after the 3rd year drops to 0.86 and further dips to 0.62 in the 9th year. Figure 6 shows the survival plot for the same and depicts a declining slope till 8th year and then drops.

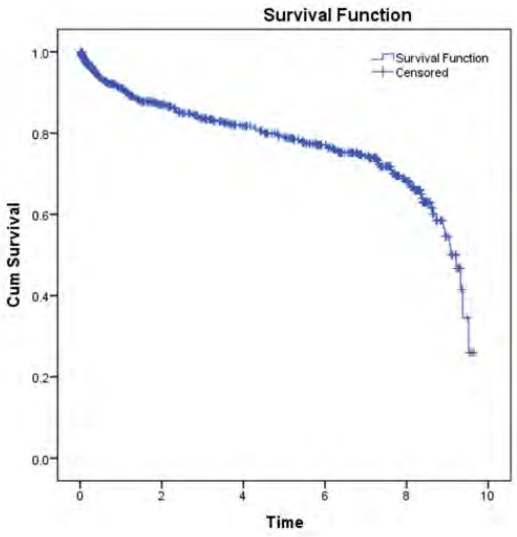


Figure 5: Kaplan Meier survival plot for the Category 2 HIV/AIDS patients

Table 7: Model description of Cox-PH for the Category 1 HIV/AIDS patients

Omnibus Tests of Model Coefficients			
-2 Log Likelihood	**Overall (score)**		
	Chi-square	df	Sig.
204.944	28.099	8	.000

Table 7 provides the Cox-PH model description for the HIV/AIDS patients in Category 1. The model is highly significant with lowest -2 log-likelihood value (204.944) and Chi-square value as 28.099.

Table 8: Model estimates of Cox-PH for the Category 1 HIV/AIDS patients

Variables in the Equation								
	B	**SE**	**Wald**	**df**	**Sig.**	**Exp(B)**	**95.0% CI for Exp(B)**	
							Lower	**Upper**
Age	.029	.025	1.392	1	.238	1.029	0.981	1.080
Sex	-.735	.847	0.753	1	.386	.480	.091	2.522
Smoking	-.351	.510	0.474	1	.491	.704	.259	1.912
DRUGS	1.892	.682	7.694	1	.006	6.634	1.742	25.257
Alcohol	-1.642	.766	4.599	1	.032	.194	.043	0.868
MOT			4.351	3	.226			
MOT(1)	.773	1.151	.451	1	.502	2.167	.227	20.689
MOT(2)	-13.104	464.703	.001	1	.978	.000	.000	
MOT(3)	-.991	.557	3.173	1	.075	.371	.125	1.105

Table 8 explains the variables considered for the model for the Category 1 HIV/AIDS patients. The variables include demography such as age and sex of the patient. Smoking and

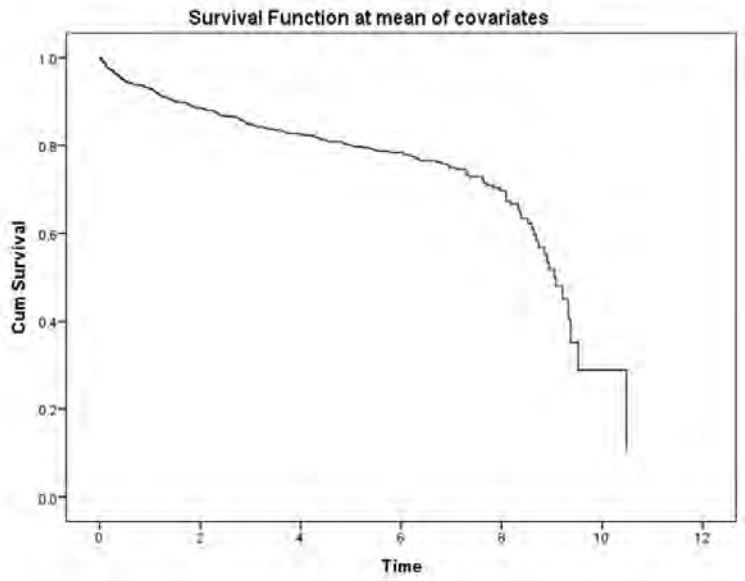


Figure 6: Cox-PH cumulative survival plot for the combined HIV/AIDS patients

drinking habit along with exposed to drugs. Apart from this method of transmission is also taken into care. Out of these variables we find that exposure of drugs (p-value= 0.006) and habit of alcohol intake (p-value= 0.032) are significant at 5% level of significance. Here, for the patient exposure to drugs, the risk will increase at a rate of 6.634 for each unit increase in drug exposer.

Table 9: Cox-PH survival estimates for the Category 1 HIV/AIDS patients

Time (Years)	Baseline Cum Hazard	Survival at mean of covariates	Std. Error	Cum Hazard
1	0.08	0.98	0.11	0.02
2	0.14	0.97	0.18	0.03
3	0.34	0.93	0.42	0.07
4	0.63	0.88	0.74	0.13
5	0.69	0.87	0.80	0.14
6	0.77	0.86	0.88	0.15
7	0.87	0.84	0.97	0.17
8	1.19	0.79	1.25	0.24
9	2.19	0.65	1.83	0.44
10	14.41	0.06	0.88	2.86

Table 9 provides the survival estimates for patients in Category 1 obtained by using Cox-PH regression. The estimated survival in the first year of follow-up is quite high (0.98) and reduces slightly in the next year. After 3rd year the survival estimate drops from 0.93 to 0.88 and further dips to 0.65.

Figure 7 shows the Cox-PH cumulative survival plot for the Category 1 HIV/AIDS

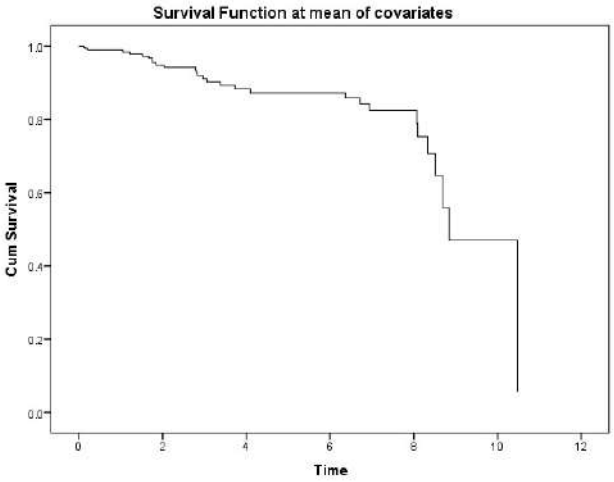


Figure 7: Cox-PH cumulative survival plot for the the Category 1 HIV/AIDS patients

patients indicting a drop in survival at 3rd and 8th year.

Table 10: Model description of Cox-PH for the Category 2 HIV/AIDS patients

Omnibus Tests of Model Coefficients			
-2 Log Likelihood	**Overall (score)**		
	Chi-square	df	Sig.
1512.359	32.131	8	.000

Table 10 provides the Cox-PH model description for the HIV/AIDS patients in Category 2. The model is highly significant with lowest -2 log-likelihood value (1512.359) and Chi-square value as 32.131.

Table 11: Model estimates of Cox-PH for the Category 1 HIV/AIDS patients

Variables in the Equation								
	B	**SE**	**Wald**	**df**	**Sig.**	**Exp(B)**	**95.0% CI for Exp(B)**	
							Lower	**Upper**
Age	.015	.010	2.383	1	.123	1.015	0.996	1.035
Sex	-.236	.230	1.057	1	.304	.790	.503	1.239
Smoking	-1.110	.223	24.761	1	.000	.330	.213	.510
DRUGS	.457	.388	1.388	1	.239	1.580	0.738	3.381
Alcohol	.082	.208	.155	1	.693	1.086	.722	1.633
MOT			0.119	3	.989			
MOT(1)	-.052	.762	.005	1	.946	0.950	.213	4.229
MOT(2)	-.357	1.040	.117	1	.732	.700	.091	5.379
MOT(3)	-.025	.282	.008	1	.930	.976	.562	1.695

Table 11 explains the variables considered for the model for the Category 2 HIV/AIDS

patients. The variables include demography such as age and sex of the patient. Smoking and drinking habit along with exposed to drugs. Apart from this method of transmission is also taken into care. Out of these variables we find that only habit of smoking (p-value = 0.000) is significant at 5% level of significance.

Table 12: Cox-PH survival estimates for the Category 2 HIV/AIDS patients

Time (Years)	Baseline Cum Hazard	Survival at mean of covariates	Std. Error	Cum Hazard
1	0.11	0.94	0.01	0.07
2	0.22	0.88	0.01	0.12
3	0.27	0.85	0.02	0.16
4	0.32	0.83	0.02	0.18
5	0.37	0.81	0.02	0.21
6	0.43	0.78	0.02	0.25
7	0.49	0.75	0.02	0.28
8	0.57	0.72	0.03	0.33
9	0.83	0.62	0.04	0.48

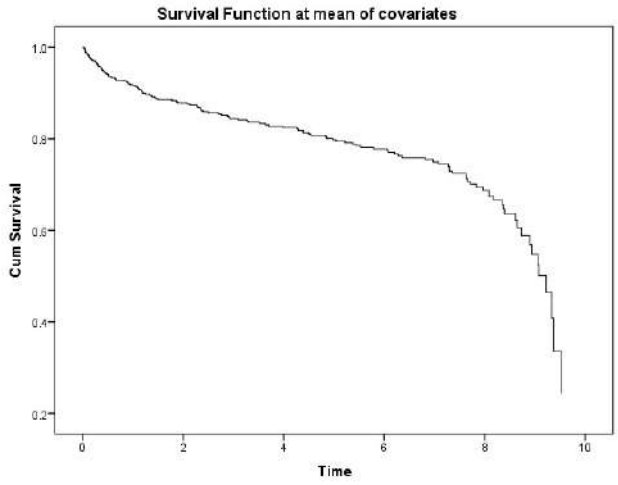


Figure 8: Cox-PH cumulative survival plot for the the Category 2 HIV/AIDS patients

Table 12 provides the survival estimates for patients in Category 2 obtained by using Cox-PH regression. The estimated survival in the first year of follow-up is 0.94 and declines to 0.88. Also, after the 5th year it goes from 0.81 to 0.78 in the 6th year and further dips to 0.62 in the 9th year.

Figure 8 shows the survival plot for the Category 2 depicts a drop in the survival first two years and then after the 8th year. In-between the survival declines slowly.

3.3. Calculation of premium cost

The survival estimates of Cox-PH regression are further utilised for the estimation of the premium cost. Table 13 shows the cost of the premium for all the HIV/AIDS patients

combined and also for the patients in Category 1 and 2. Here, rupees one hundred is taken as sum insured and the cost of the premium to be paid is estimated. The estimated premium cost for HIV/AIDS patients cumulatively, Category 1 and 2 is Rs. 14.59, Rs. 4.16 and Rs. 16.77 respectively. Cumulatively, there is 5% decline in survival from 1st to 2nd year and from 7th to 8th year. In the same manner the premium cost increases by 48.94% and 22.27% respectively. Here we see that the patients under Category 1 have lowest premium cost. But in the very next 2 subsequent years the premium cost for Category 1 is almost doubled to Rs. 8.82 (112.09% rise from the previous year) and then shoots to Rs. 16.99 (92.65% rise from the previous year. In Category 2 the premium cost increases from Rs. 16.77 to Rs. 23.92 (42.62% jump) in the first year. Overall, for the five years the rise in premium cost is 142.66%, 487.79% and 114.11% respectively for the three categories.

Figure 9 shows the comparison of survival estimates (primary-axis) and premium cost (secondary-axis) for all the patients combined, in Category 1 and 2. The increasing trend of premium cost is followed by the decreasing survival estimates over the follow-up years. It clearly reflects that throughout the year, premium cost for Category 1 is lowest of all.

Table 13: Cost of premium for the HIV/AIDS patients

	Survival Estimates			**Premium Cost**		
Time (Years)	**Combined**	**Category 1**	**Category 2**	**Combined**	**Category 1**	**Category 2**
1	0.95	0.98	0.94	14.59	4.16	16.77
2	0.90	0.97	0.88	21.73	8.82	23.92
3	0.86	0.93	0.85	27.12	16.99	28.06
4	0.83	0.88	0.83	31.67	22.36	31.86
5	0.81	0.87	0.81	35.40	24.44	35.91
6	0.79	0.86	0.78	38.97	27.00	39.92
7	0.76	0.84	0.75	43.61	32.59	44.53
8	0.72	0.79	0.72	53.33	47.47	53.75
9	0.62	0.65	0.62			

4. Conclusion

The survival estimates of the HIV/AIDS patients are obtained using Kaplan-Meier and Cox-PH regression methods. These estimates are obtained for the patients in three scenarios, Category 1, Category 2 and combined. The estimates obtained using Cox-PH regression gave better estimates than that of Kaplan-Meier. For the patients taken combined, age, smoking habit and exposure to drugs are significant predictors. In Category 1 the habit of smoking and alcohol consumption are the significant predictors whereas in Category 2 only smoking habit is the significant predictor. The fitted model in all three scenarios is highly significant. Till 7 years the overall survival estimates are more than 75%. This high survival estimates provide great evidence to introduce a yearly insurance plan for the HIV/AIDS patients.

The suggested yearly insurance plan assures a sum of Rs. 100 against the premium of Rs. 14.59, Rs. 4.16 and Rs. 16.77 in the three scenarios respectively, in case of death. The best premium is for the patient in Category 1 since they have to pay very less as compared

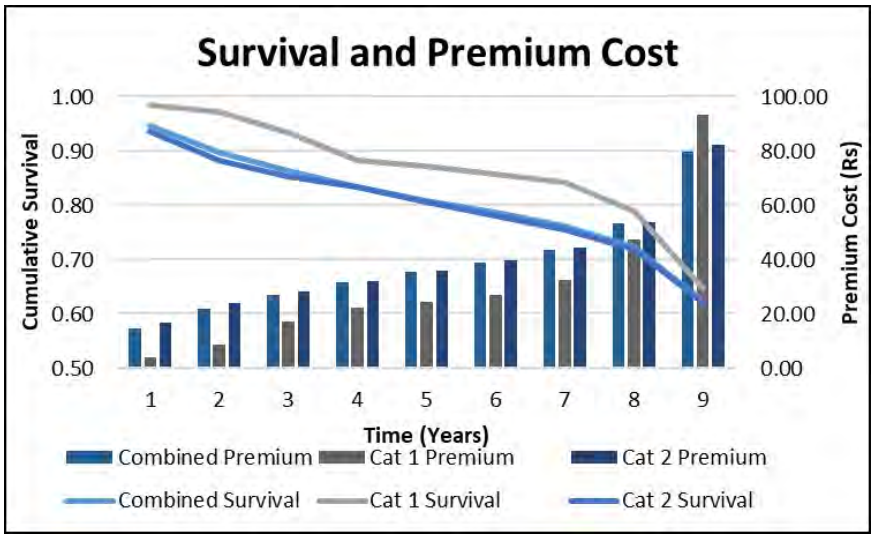


Figure 9: Comparison of survival estimates (primary-axis) and premium cost (secondary-axis) for all the HIV/AIDS patients

to the others. For a five-year difference the premium cost increases by 142.66%, 487.79% and 114.11% respectively in the three scenarios. So, it is recommended to opt for the insurance as early as possible because the premium cost keeps increasing over the time as the survival estimates decline.

Acknowledgements

We thank our colleagues for their valuable suggestions. We also, extend our thanks to the anonymous referees who suggested many improvements and thank the Chair Editor for his encouragement, guidance and counsel.

References

Bowers, N., Gerber, H., Hickman, J., Jones, D., and Nesbitt, C. (1997). *Actuarial Mathematics*. Society of Actuaries.

Burns, F., Edwards, S., Woods, J., Haidari, G., Calderon, Y., Leider, J., Morris, S., Tobin, R., Cartledge, J., and Brown, M. (2013). Acceptability, feasibility and costs of universal offer of rapid point of care testing for HIV in an acute admissions unit: results of the RAPID project. *HIV Medicine*, **14**, 10–14.

Collaborators, G. . H. et al. (2016). Estimates of global, regional, and national incidence, prevalence, and mortality of HIV, 1980–2015: The Global Burden of Disease Study 2015. *The Lancet HIV*, **3**, 361–387.

Grover, G., Vinit, P. K., and Sehgal, V. (2021). Estimation of premium cost for HIV/AIDS patients under ART. *International Journal of System Assurance Engineering and Management*, **12**, 77–83.

Grover, G., Vinit, P. K., and Thakur, A. K. (2018). Actuarial modeling of insurance premium for patients with acute lymphoblastic leukemia (ALL). *Journal Applied Quantitative Methods*, **13**, 1842–4562.

Haberman, S. and Pitacco, E. (1998). *Actuarial Models for Disability Insurance*. CRC Press.

Jain, N., Kumar, A., Nandraj, S., and Furtado, K. M. (2015). Nsso 71st Round: same data, multiple interpretations. *Economic and Political Weekly*, **50**, 84–87.

King Jr, J. T., Justice, A. C., Roberts, M. S., Chang, C.-C. H., and Fusco, J. S. (2003). Long-term HIV/AIDS survival estimation in the highly active antiretroviral therapy era. *Medical Decision Making*, **23**, 9–20.

Luptáková, I. D. and Bilíková, M. (2014). Actuarial modeling of life insurance using decrement models. *Journal of Applied Mathematics, Statistics and Informatics*, **10**, 81–91.

NACO (2019). India HIV estimates 2019: Report. Technical report, National AIDS Control Organization.

NACO and ICMR-NIMS (2017). India HIV Estimations 2017 Fact Sheet. Technical report, National AIDS Control Organization & ICMR-National Institute of Medical Statistics.

Norberg, R. (2000). *Basic Life Insurance Mathematics*. Lecture notes, Laboratory of Actuarial Mathematics, University of Copenhagen.

Ong, K., Thornton, A., Fisher, M., Hutt, R., Nicholson, S., Palfreeman, A., Perry, N., Stedman-Bryce, G., Wilkinson, P., Delpech, V., et al. (2016). Estimated cost per HIV infection diagnosed through routine HIV testing offered in acute general medical admission units and general practice settings in England. *HIV medicine*, **17**, 247–254.

Slud, E. V. (2001). *Actuarial Mathematics and Life-table Statistics*. Chapman and Hall/CRC.

UNAIDS (2021). Ending the AIDS Epidemic. Technical report, Global HIV Statistics.

WHO (2005). Interim WHO clinical staging of HIV/AIDS and HIV/AIDS case definitions for surveillance: African Region. Technical report, World Health Organization.

WHO (2016). The use of antiretroviral drugs for treating and preventing HIV infection. Technical report, World Health Organization.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 22, No. 1, 2024 (New Series), pp 137-147
http://www.ssca.org.in/journal

Integrated Redundant Reliability Model using k out of n Configuration with Integer Programming Approach

Srinivasa Rao Velampudi[1], Sridhar Akiri[2] and Pavankumar Subbara[3]
[1]*Department of Sciences and Humanities, Raghu Institute of Technology Visakhapatnam, Andhra Pradesh, India*
[2]*Department of Mathematics, GITAM School of Science GITAM (Deemed to be University), Visakhapatnam, Andhra Pradesh, India*
[3]*Department of Mathematics, GITAM School of Science GITAM (Deemed to be University), Bangalore Campus, Karnataka, India*

Received: 06 December 2022; Revised: 22 May 2023; Accepted: 22 June 2023

Abstract

Reliability engineering is a branch of systems engineering concerned with the dependability of machinery. Reliability is the degree to which a system or component continues to function as intended over time and under stress. In a 'k' out of 'n' systems, the total efficiency is greater than the efficiency of any individual component. Here, we propose an additional system, An Integrated Reliability Model (IRM) for the 'k' out of 'n' systems to take into account the factors' efficiencies, the number of factors in each stage, and the various constraints in order to maximize the system's efficiency. The authors used the above-cited integrated model for obtaining various components' reliability and efficiency in a Muffle Box Furnace machine by using Lagrangean methods to calculate the price-component, weight-component and volume-component associated with various configurations of the system, all in an effort to maximize overall system performance. To get a real-looking result in an integer space, we adopted the integer programming method and the dynamic programming technique.

Key words: IRM; Lagrangean approach; Stage efficiency; Integer programming; D.P. approach; Structure's efficiency.

AMS Subject Classifications: Primary: 90B25; Secondary: 90C39, 90C59

1. Introduction

Aggarwal *et al.* (1975) present a practical approach for resolving the redundancy-based, multi-criteria optimization issues that typically occur in the dependability design of engineering systems. Because it can take into consideration any combination of redundancy, limitations, and individual cost functions, it might resolve many design issues relating to dependability. Kuo and Prasad (2000) are well-known names when it comes to the secure and efficient creation of technical systems. They gave actual instances to show how multicriteria optimization issues may be utilised to successfully tackle redundancy optimization issues.

Corresponding Author: Sridhar Akiri
Email: sakiri@gitam.edu

The Structure's reliability can be improved by either placing superfluous units, applying the element of greater reliability or by adopting the two methods at a time and both of them use extra resources. Optimizing structure reliability, and conditions to resource availability viz. price-components, weight-components and volume-components are examined. In general, reliability is tested as an element of price-component; But, when tested with real-world problems, the invisible effect of other restraints such as weight-component, volume-components, etc., has a special effect on improving structural reliability.

A team of researchers under the direction of Sankaraiah *et al.* (2011) set out to look at how numerous restrictions affect system reliability. An integrated redundant reliability system is modelled and solved using a Lagrangian multiplier. This provides a real-valued response on the total number of components, the dependability of all system stages, and the dependability of individual stages. Sridhar *et al.* (2013) created a novel method for optimising a redundant IRM with several limitations. The method accounts for the k-out-of-n configuration system and enables the optimization to discover the unexpected effects of other constraints in addition to the cost constraint. A unique strategy for optimising a redundant IRM was developed by Sasikala *et al.* (2013), however it has several drawbacks. The technique takes into consideration the k-out-of-n configuration system and enables the optimization to find the unexpected impacts of additional constraints aside from the cost constraint.

The specific functionality of the over-reliability model with several limitations to optimize the recommended setup was examined to maximize the recommended setup. The problem examines the unknowns that is, various elements (Y_{cj}), the element reliability (r_{cj}), and the stage reliability (R_{RP}) at a specific point for disposing of multiple restraints to magnify the structure reliability (R_{RS}) that is described as a United Reliability Model (URM). In literature, United Reliability Models are enhanced by applying value restraints where there is a fixed association between price-component and reliability. A unique pattern of planned work is a deliberation of the weight-component and volume-component as supplementary restraints along with price-component to form and improve the superfluous reliability system for 'k' out of 'n' structure composition. The rest of the paper has been organized into five sections. In Section 2, we detail the Lagrangean analysis of the corresponding new mathematical function. In Section 3, we get an overview of the Muffle Box furnace's parts, including its price-component, weight-component, volume-component, stage, component, and structure reliability. Our rounded-off Lagrangean approach results are presented in Section 4. In Sections 5 and 6, we will present the Integer programming, results and make comparisons to the Lagrangean method (both without and with rounding off). Finally, the author concludes and makes some suggestions in Section 7.

1.1. Assumptions and notations

1. Each stage's elements are believed to be identical, i.e., all elements have the same level of reliability.

2. All elements are supposed to be statistically independent, meaning that their failure

has no bearing on the performance of other elements in the structure.

$$\begin{aligned}
R_{RS} &= \text{Reliability of structure} \\
R_{RP} &= \text{Reliability of stage }, 0 < R_{RP} < 1 \\
r_{cj} &= \text{Reliability of each component in stage } cj;\ 0 < r_{cj} < 1 \\
Y_{cj} &= \text{Number of components in stage } cj \\
PC_{cj} &= \text{Price component in stage } cj \\
WC_{cj} &= \text{Weight component in stage } cj \\
VC_{cj} &= \text{Volume component in stage } cj \\
P_{c0} &= \text{Greatest allowable complex for price component} \\
W_{c0} &= \text{Greatest allowable complex for weight component} \\
V_{c0} &= \text{Greatest allowable complex for volume component}
\end{aligned}$$

$P_{cj}; \alpha_{cj}; W_{cj}; \beta_{cj}; V_{cj}; \gamma_{cj}$ are constants.

2. Mathematical analysis

The efficiency of the system to the provided price-component function

$$R_{RS} = \sum_{i=1}^{n} B(m,i)(p)^i(1-p)^{(m-i)} \tag{1}$$

The following relationship between price-component and efficiency is used to calculate the price-component coefficient of each unit in Stage j.

$$r_{cj} = Cosh^{-1}\left[\frac{PC_{cj}}{P_{cj}}\right]^{\frac{1}{\alpha_{cj}}}$$

Therefore

$$PC_{cj} = P_{cj}Cosh[r_{cj}]^{\alpha_{cj}} \tag{2a}$$
$$WC_{cj} = W_{cj}Cosh[r_{cj}]^{\beta_{cj}} \tag{2b}$$
$$VC_{cj} = V_{cj}Cosh[r_{cj}]^{\gamma_{cj}} \tag{2c}$$

Since price-components are linear in Y_{cj},

$$\sum_{j=1}^{n} PC_{cj}Y_{cj} \leq P_{co} \tag{3a}$$

Similarly, weight-components and volume-components are also linear in Y_{cj}

$$\sum_{j=1}^{n} WC_{cj}Y_{cj} \leq W_{co} \tag{3b}$$
$$\sum_{j=1}^{n} VC_{cj}Y_{cj} \leq V_{co} \tag{3c}$$

Substituting (2a, 2b & 2c) in (3a, 3b & 3c) respectively, we get

$$\sum_{j=1}^{n} P_{cj} Cosh[r_{cj}]^{\alpha_{cj}}.\ Y_{cj} - P_{C0} \leq 0 \tag{4a}$$

$$\sum_{j=1}^{n} W_{cj} Cosh[r_{cj}]^{\beta_{cj}}.\ Y_{cj} - W_{C0} \leq 0 \tag{4b}$$

$$\sum_{j=1}^{n} V_{cj} Cosh[r_{cj}]^{\gamma_{cj}}.\ Y_{cj} - V_{C0} \leq 0 \tag{4c}$$

The transformed

$$Y_{cj} = \frac{lnR_{RP}}{lnr_{cj}} \tag{5}$$

where

$$R_{RP} = \sum_{k=2}^{n} B(Y_{cj}, k)(r_{cj})^{k}(1 - r_{cj})^{(cj-k)} \tag{6}$$

Subject to the constraints

$$\sum_{j=1}^{n} P_{cj} Cosh[r_{cj}]^{\alpha_{cj}}.\ \frac{lnR_{RP}}{lnr_{cj}} - P_{C0} \leq 0 \tag{7a}$$

$$\sum_{j=1}^{n} W_{cj} Cosh[r_{cj}]^{\beta_{cj}}.\ \frac{lnR_{RP}}{lnr_{cj}} - W_{C0} \leq 0 \tag{7b}$$

$$\sum_{j=1}^{n} V_{cj} Cosh[r_{cj}]^{\gamma_{cj}}.\ \frac{lnR_{RP}}{lnr_{cj}} - V_{C0} \leq 0 \tag{7c}$$

Positivity restrictions $Y_{cj} \geq 0$.

A Lagrangean function is defined as

$$\begin{aligned} L_G \;=\; & R_{RS} + \omega_1 \left[\sum_{j=1}^{n} P_{cj}.\ Cosh[r_{cj}]^{\alpha_{cj}}.\ \frac{lnR_{RP}}{lnr_{cj}} - P_{c0} \right] + \omega_2 \Big[\sum_{j=1}^{n} P_{cj}.\ Cosh[r_{cj}]^{\beta_{cj}}.\ \frac{lnR_{RP}}{lnr_{cj}} \\ & - W_{c0} \Big] + \omega_3 \left[\sum_{j=1}^{n} V_{cj}.\ Cosh[r_{cj}]^{\gamma_{cj}}.\ \frac{lnR_{RP}}{lnr_{cj}} - V_{c0} \right] \end{aligned} \tag{8}$$

The Lagrangean function can be used to find the ideal point and separating it by R_{RS}, r_{cj}, ω_1, ω_2 and ω_3.

$$\begin{aligned} \frac{\partial L_G}{\partial R_{RS}} \;=\; & 1 + \omega_1 \left[\sum_{j=1}^{n} P_{cj}.\ Cosh[r_{cj}]^{\alpha_{cj}}.\ \frac{1}{lnr_{cj}} \frac{1}{R_{RP}} \right] + \omega_2 \left[\sum_{j=1}^{n} W_{cj}.\ Cosh[r_{cj}]^{\beta_{cj}}.\ \frac{1}{lnr_{cj}} \frac{1}{R_{RP}} \right] \\ & + \omega_3 \left[\sum_{j=1}^{n} V_{cj}.\ Cosh[r_{cj}]^{\gamma_{cj}}.\ \frac{1}{lnr_{cj}} \frac{1}{R_{RP}} \right] \end{aligned} \tag{9}$$

$$\begin{aligned}\frac{\partial L_G}{\partial r_{cj}} &= \omega_1\left[\sum_{j=1}^{n} P_{cj}.\ Cosh[r_{cj}]^{\alpha_{cj}}.\ \frac{lnR_{RP}}{lnr_{cj}}\right]\left[\alpha_{cj}.\ Tanh(r_{cj}) - \frac{1}{r_{cj}.lnr_{cj}}\right] \\ &\quad +\omega_2\left[\sum_{j=1}^{n} W_{cj}.\ Cosh[r_{cj}]^{\beta_{cj}}.\ \frac{lnR_{RP}}{lnr_{cj}}\right]\left[\beta_{cj}.\ Tanh(r_{cj}) - \frac{1}{r_{cj}.lnr_{cj}}\right] \\ &\quad +\omega_3\left[\sum_{j=1}^{n} V_{cj}.\ Cosh[r_{cj}]^{\gamma_{cj}}.\ \frac{lnR_{RP}}{lnr_{cj}}\right]\left[\gamma_{cj}.\ Tanh(r_{cj}) - \frac{1}{r_{cj}.lnr_{cj}}\right] \qquad (10)\end{aligned}$$

$$\frac{\partial L_G}{\partial \omega_1} = \sum_{j=1}^{n} P_{cj}.\ Cosh[r_{cj}]^{\alpha_{cj}}.\ \frac{lnR_{RP}}{lnr_{cj}} - P_{c0} \qquad (11)$$

$$\frac{\partial L_G}{\partial \omega_2} = \sum_{j=1}^{n} W_{cj}.\ Cosh[r_{cj}]^{\beta_{cj}}.\ \frac{lnR_{RP}}{lnr_{cj}} - W_{c0} \qquad (12)$$

$$\frac{\partial L_G}{\partial \omega_3} = \sum_{j=1}^{n} V_{cj}.\ Cosh[r_{cj}]^{\gamma_{cj}}.\ \frac{lnR_{RP}}{lnr_{cj}} - V_{c0} \qquad (13)$$

where ω_1, ω_2 and ω_3 are Lagrangean multipliers.

Using the Lagrangean method, we can calculate the number of elements in each Stage (Y_{cj}), the best reliability of an individual element (r_{cj}), the reliability of an entire Stage ($\mathbf{R_{RP}}$), and the reliability of the entire structure (R_{RS}). When it comes to the price-component, weight-component, and volume-component, this method yields a true (valued) answer.

3. Case problem

To derive the multiple parameters of a given mechanical system using optimization techniques, where all the assumptions like price-component, weight-component, and volume-component are directly proportional to system reliability has been considered in this research work. The same logic may not be true in the case of electronic systems. Hence, the optimal element accuracy (r_{cj}), Stage reliability (R_{RP}), Number of elements in each Stage (Y_{cj}), and structure accuracy (R_{RS}) can be evaluated in any given mechanical system. In this work, an attempt has been made to evaluate the Structure accuracy of a special purpose of Muffle Box Furnace machine that is utilized in the laboratories for testing different materials.

The machine is used for the assembly of many components. But our case steady, we are considering 3 or 4 important components on the base of Muffle Box Furnace machine.

The muffle furnace is an essential laboratory testing instrument used for a wide variety of materials. The instrument is useful in the study of materials' properties and can be found in most laboratory settings. The tool is put to use in numerous heat-treating metallurgical procedures. Altering the molecular structure of a material is a common application of heat treatment. The machine's approximate price was Rs. 5000, which is considered a structure price, the weight of the machine is 156 kg, which is the volume of the structure, and the space occupied by the machine is $100cm^3$, which is the volume of the structure. To attract the authors from different cross sections, the authors attempted to use hypothetical numbers, which can be changed according to the environment. The Lagrangean approach of modelling and solving produces real-valued solutions for the number of elements, element reliability,

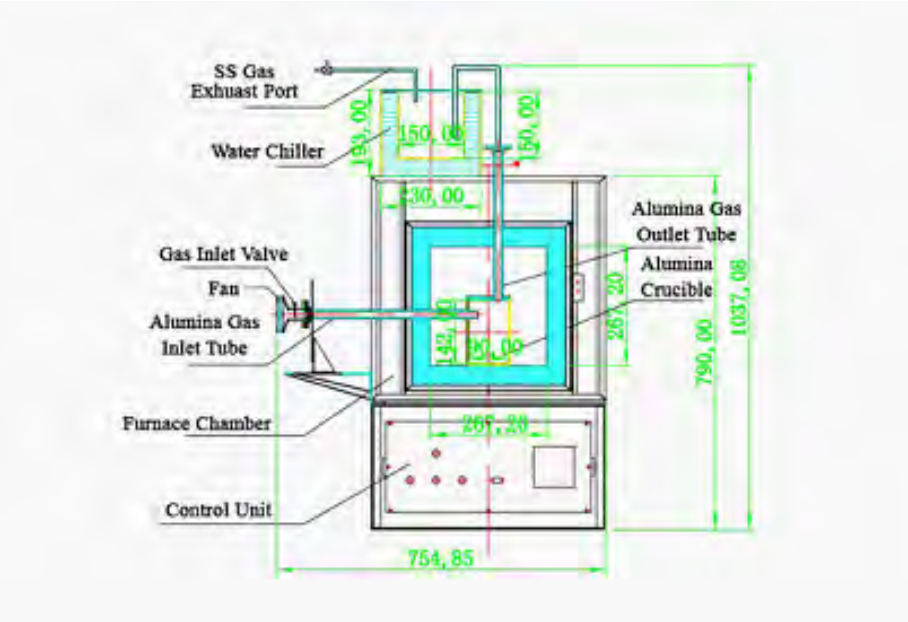


Figure 1: Schematic diagram of muffle box furnace

moment reliability, and therefore structure reliability, as shown by Sasikala *et al.* (2020).

3.1. Constants

The data required for the constants for the case problem are provided hereunder.

Table 1: The details of price-component, weight-component and volume-component for case problem

Phase	Price constants		Weight constants		Volume constants	
	P_{cj}	α_{cj}	W_{cj}	β_{cj}	V_{cj}	γ_{cj}
1	600	0.85	110	0.92	95	0.94
2	650	0.88	90	0.88	85	0.89
3	700	0.91	70	091	75	0.86

The efficiency of each factor, phase and number of factors in each stage, as well as the structural efficiency, are shown in the tables below.

3.2. The details of price-component, weight-component and volume-component

The price-components, weight-components, and volume-components related efficiency design and total price of the components is described in the table below.

Structure efficiency $= R_{CR} = 0.9664$

Table 2: The details of price-component, weight-component and volume-component by using Lagrangean approach

Phase	P_{cj}	α_{cj}	r_{cj}	$Log\ r_{cj}$	R_{RP}	$Log\ R_{RP}$	Y_{cj}	$PC_{cj} = P_{cj}Cosh[r_{cj}]^{\alpha_{cj}}$	$Y_{cj}.PC_{cj}$
1	600	0.85	0.8741	-0.0584	0.6777	-0.169	2.89	801.00	2314.89
2	650	0.88	0.8445	-0.0734	0.6487	-0.188	2.56	861.30	2204.93
3	700	0.91	0.8456	-0.0728	0.5461	-0.2627	3.61	937.60	3384.74
Total price of the components									7904.55
Phase	W_{cj}	β_{cj}	r_{cj}	$Log\ r_{cj}$	R_{RP}	$Log\ R_{RP}$	Y_{cj}	$WC_{cj} = W_{cj}Cosh[r_{cj}]^{\beta_{cj}}$	$Y_{cj}.WC_{cj}$
1	110	0.92	0.8741	-0.0584	0.6777	-0.169	2.89	150.48	434.89
2	90	0.88	0.8445	-0.0734	0.6487	-0.188	2.56	119.34	305.51
3	70	0.91	0.8456	-0.0.728	0.5461	-0.2627	3.61	93.73	338.37
Total weight of the components									1078.76
Phase	V_{cj}	γ_{cj}	r_{cj}	$Log\ r_{cj}$	R_{RP}	$Log\ R_{RP}$	Y_{cj}	$VC_{cj} = V_{cj}Cosh[r_{cj}]^{\gamma_{cj}}$	$Y_{cj}.VC_{cj}$
1	95	0.94	0.8741	-0.0584	0.6777	-0.169	2.89	130.8	378.01
2	85	0.89	0.8445	-0.0734	0.6487	-0.188	2.56	113.8	291.33
3	75	0.86	0.8456	-0.0728	0.5461	-0.2627	3.61	98.85	356.85
Total volume of the components									1026.19

4. Efficiency design with e_j rounding off

The acceptable outcomes for the price-component, weight-component, and volume components are listed in the tables, and the Y_{cj} values are summarised as integers (rounding the value of cj to the nearest integer) in the efficiency design. The information you seek can be determined by calculating the variance caused by the price-component, the weight-component, and the volume-component of the building's capacity (both before and after rounding Y_{cj} to the nearest integer).

4.1. Efficiency design concerning price-component, weight-component and volume-component with rounding off

Table 3: The details of price-component, weight-component and volume-component analysis by using rounding off approach

Phase	r_{cj}	R_{RP}	Y_{cj}	P_{cj}	$Y_{cj}.P_{cj}$	W_{cj}	$Y_{cj}.W_{cj}$	V_{cj}	$Y_{cj}.V_{cj}$
1	0.8741	0.6777	3	801	2403	150	450	131	393
2	0.8445	0.6487	3	861	2583	119	357	114	342
3	0.8456	0.5461	4	938	3752	94	376	99	396
Total price of the components					8738		1183		1131
Structure efficiency (R_{CR})					0.9687				

4.1.1. Variation in total price-component

$$= \frac{\text{Total price-component with rounding off - Total price-component without rounding off}}{\text{Total price without rounding off}} = 10.54\%$$

4.1.2. Variation in total weight-component

$$= \frac{\text{Total weight-component with rounding off - Total weight-component without rounding off}}{\text{Total weight-component without rounding off}} = 09.66\%$$

4.1.3. Variation in total price-component

$$= \frac{\text{Total volume-component with rounding off - Total volume-component without rounding off}}{\text{Total volume-component without rounding off}} = 10.21\%$$

4.1.4. Variation in structure efficiency

$$= \frac{\text{Structure efficiency with rounding off - Structure efficiency without rounding off}}{\text{Structure efficiency without rounding off}} = 01.24\%$$

Instead of using complex algorithms, the Lagrangian multiplier method provides a way to quickly find the best possible design. Naturally, this assumes that the component count at each Stage (Y_{cj}) is real. When Y_{cj} is rounded off to the nearest integer, it has a domino effect on all the other numbers in the reliability design, including the values of the reliability at each Stage ($\mathbf{R_{RP}}$) the reliability of the system as a whole ($\mathbf{R_{RS}}$), the total Price of each Stage, and the Price of the system as a whole. It is demonstrated in the examples how changing the way Y_{cj} is rounded can affect the reliability of a design. Integer programming can be used to counter this shortcoming.

5. Integer programming

To determine the number of components in each stage, stage reliabilities, and system reliability, integer programming requires the component reliabilities as input. The fundamental disadvantage of integer programming is that it cannot be utilised directly, i.e., without the input of the component reliabilities, even if it is beneficial for creating integrated reliability models. Therefore, integer programming may take the component reliabilities from the previous approach, the Lagrangian method, as input and output the stage reliabilities, system reliabilities, system reliabilities, stage reliabilities, and stage reliabilities. Integer programming allows you some flexibility to select the number of components in each step, the dependability of each stage, and the reliability of the entire system within the limits that are provided.

Pavankumar *et al.* (2020) look into how the many constraints listed above affect how the integrated reliability and optimization is formulated. Statistics are utilised with IRRCCS (Integrated Redundant Reliable Coherent Configuration system is considered). Sridhar *et al.* (2021), have provided a thorough examination, design, analysis, and optimization of a coherent redundant reliability design.The work of Sridhar *et al.* (2022) is applied to parallel-series systems where both technologies include parallel factors. A parallel approach can only function if all of its components are active at all times.To determine the optimum solution, Srinivasa Rao *et al.* (2022) recommended utilising an appropriate method based on Heuristic approach and included the redundancy strategy as a new decision variable.

Integer programming methodology as a whole has grown rapidly over the past half century, but linear integer programming has been its main focus. But there have been some promising theoretical and methodological developments in nonlinear integer programming in recent years. Because of these changes, nonlinear integer programming is now used in many areas of scientific computing. For example, it is a key criterion for choosing portfolios and managing risks.

The author used the LINGO Programme (created by Lindo Corporation, USA) to find

a decimal solution to the problem. Lingo is an all-inclusive Programme that streamlines the process of creating and solving linear, nonlinear, and integer optimization models. Lingo is an all-inclusive package that features a robust language for expressing optimization models, a comprehensive environment for creating and editing problems, and a collection of fast, in-built solvers.

An integrated reliability model for redundant systems with multiple constraints is established and optimized using integer programming for the considered function. By using the values of component dependability's (r_{cj}) and the number of components in each stage (Y_{cj}) as inputs for the application of integer programming, we can optimize the design in light of the case problem discussed in the previous section for the respective mathematical function (refer to equation 1). Since the values of Y_{cj} are integers, this method helps optimise the design and makes it easier to use in the real world.

6. Results

6.1. The details of price-component, weight component and volume-component constraint by using integer programming approach

The value-related efficiency design is described in the Table 4.

Table 4: The details of price-component, weight-component and volume-component constraint by using integer programming approach

Phase	P_{cj}	α_{cj}	r_{cj}	$Log\ r_{cj}$	R_{RP}	$Log\ R_{RP}$	Y_{cj}	$PC_{cj} = P_{cj}Cosh[r_{cj}]$	$Y_{cj}.PC_{cj}$
1	600	0.85	0.9982	-0.0008	0.9945	-0.0024	3	866	2598
2	650	0.88	0.9736	-0.0116	0.9229	-0.0348	3	936	2808
3	700	0.91	0.9891	-0.0048	0.9571	-0.0190	4	1031	4124
Final price									9530
Phase	W_{cj}	β_{cj}	r_{cj}	$Log\ r_{cj}$	R_{RP}	$Log\ R_{RP}$	Y_{cj}	$WC_{cj} = W_{cj}Cosh[r_{cj}]$	$Y_{cj}.WC_{cj}$
1	110	0.92	0.9982	-0.0008	0.9945	-0.0024	3	164	492
2	90	0.88	0.9736	-0.0116	0.9229	-0.0348	3	130	390
3	70	0.91	0.9891	-0.0048	0.9571	-0.0190	4	103	412
Final weight									1294
Phase	V_{cj}	γ_{cj}	r_{cj}	$Log\ r_{cj}$	R_{RP}	$Log\ R_{RP}$	Y_{cj}	$VC_{cj} = V_{cj}Cosh[r_{cj}]$	$Y_{cj}.VC_{cj}$
1	95	0.94	0.9982	-0.0008	0.9945	-0.0024	3	143	429
2	85	0.89	0.9736	-0.0116	0.9229	-0.0348	3	123	369
3	75	0.86	0.9891	-0.0048	0.9571	-0.0190	4	108	412
Final volume									1230
Structure efficiency (R_{SR})									0.9998

6.2. Comparison of optimization of integrated redundant reliability k out of n systems - LMM with rounding off and integer programming approach for price-component, weight-component and volume-component

7. Conclusion

This work proposes an integrated reliability model for a k out of n configuration system with many efficiency criteria. When the data are discovered to be in reals, the Lagrangean multiplier approach is used to compute the number of components (cj), component

Table 5: Results correlated LMM with rounding off approach and integer programming approach for price-component, weight-component and volume-component

		With rounding off				Integer programming			
Phase	Y_{cj}	r_{cj}	R_{RP}	PC_{cj}	$Y_{cj}.PC_{cj}$	r_{cj}	R_{RP}	PC_{cj}	$Y_{cj}.PC_{cj}$
1	3	0.8741	0.6777	801	2403	0.9982	0.9945	866	2598
2	3	0.8445	0.6487	861	2583	0.9736	0.9229	936	2808
3	4	0.8456	0.5461	938	3752	0.9891	0.9571	1031	4124
Total price		8738				9530			
Structure efficiency		Using with rounding off approach (R_{SR})			0.9987	Using integer programming approach (R_{SR})			0.9999
		With rounding off				Integer programming			
Phase	Y_{cj}	r_{cj}	R_{RP}	WC_{cj}	$Y_{cj}.WC_{cj}$	r_{cj}	R_{RP}	WC_{cj}	$Y_{cj}.WC_{cj}$
1	3	0.8741	0.6777	150	450	0.9982	0.9945	164	492
2	3	0.8445	0.6487	119	357	0.9736	0.9229	130	390
3	4	0.8456	0.5761	94	376	0.9891	0.9571	103	412
Total weight		1183				1294			
Structure efficiency		Using with rounding off approach (R_{SR})			0.9987	Using integer programming approach (R_{SR})			0.9999
		With rounding off				Integer programming			
Phase	Y_{cj}	r_{cj}	R_{RP}	VC_{cj}	$Y_{cj}.VC_{cj}$	r_{cj}	R_{RP}	VC_{cj}	$Y_{cj}.VC_{cj}$
1	3	0.8741	0.6777	131	393	0.9982	0.9945	143	429
2	3	0.8445	0.6487	114	342	0.9736	0.9229	123	369
3	4	0.8456	0.5461	99	396	0.9891	0.9571	108	432
Total volume		1131				1230			
Structure efficiency		Using with rounding off approach (R_{SR})			0.9664	Using Integer programming approach (R_{SR})			0.9998

efficiencies ($r_{cj} = 0.8741, 0.8445, 0.8456$), stage efficiencies ($R_{RP} = 0.6777, 0.6487, 0.5461$), and structure efficiency ($R_{RS} = 0.9664$). To obtain practical applicability, an integer way of programming approach is employed to construct an integer solution whereas component efficiencies ($r_{cj} = 0.9982, 0.9736, 0.9891$), stage efficiencies ($R_{RP} = 0.9945, 0.9229, 0.9571$), and structure efficiency ($R_{RS} = 0.9998$). using the inputs from the Lagrangean method. Finally, we observed that the price, weight, and volume components changed slightly, but the reliability of the stage and structure increased, resulting in increased system reliability.

The IRM model generated in this manner is quite valuable, particularly in real-world settings when a k from n configuration IRM with reliability engineer redundancy is required. In circumstances where the system value is low, the proposed model is especially valuable for the dependability design engineer to build high-quality and efficient materials.

In future study, the authors recommend utilizing a unique approach that limits the minimum and maximum component reliability values while maximizing system dependability using any of the current heuristic processes to build similar IRMs with redundancy.

References

Aggarwal, K. K., Misra, K. B., and Gupta, J. S. (1975). Reliability evaluation a comparative study of different techniques. *Microelectronics Reliability*, **14**, 49–56.

Kuo, W. and Prasad, V. R. (2000). An annotated overview of system-reliability optimization. *IEEE Transactions on Reliability*, **49**, 176–187.

Sankaraiah, G., Sarma, B. D., Umashankar, C., and Yadavalli, V. S. S. (2011). Designing, modelling and optimizing of an integrated reliability redundant system. *South African Journal of Industrial Engineering*, **22**, 100–106.

Sridhar, A., Pavankumar, S., Reddy, Y. R., Sankaraiah, G., and Umashankar, C. (2013). The k out of n redundant IRM optimization with multiple constraints. *IJRSAT*, **2**, 1–6.

Sasikala, P., Sridhar, A., Pavankumar, S., and Umashankar, C. (2013). Optimization of IRM-parallel-series redundant system. *International Journal of Engineering Research & Technology (IJERT)*, **2**, 1–6.

Kumar, S. P., Sridhar, A., and Sasikala, P. (2020). Optimization of integrated redundant reliability coherent systems - heuristic method. *Journal of Critical Reviews*, **7**, 2480–2491.

Sasikala, P., Sridhar, A., and Kumar, S. P. (2020). Optimization of IRM series-parallel and parallel-series redundant system. *Journal of Critical Reviews*, **7**, 1801–1811.

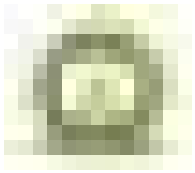
Akiri, S., Subbara, P. K., Sasikala, P., and Yadavalli, V. S. S. (2021). 9 Design and evaluation of coherent redundant system reliability. *In Systems Reliability Engineering*, De Gruyter, 137–152.

Akiri, S., Sasikala, P., Subbara, P. K., and Yadavalli, V. S. S. (2022). Design and evaluation of parallel-series IRM system. *In System Assurances*, Academic Press, 189–208

Srinivasa Rao, V., Akiri, S., and Subbara, P. K. (2022). k-out-of-n systems growth study focusing on redundant reliability systems by using heuristic programming approach. *Mathematical Statistician and Engineering Applications*, **71**, 5602–5613.

Srinivasa Rao, V., Akiri, S., and Subbara, P. K. (2023). A comprehensive case study on integrated redundant reliability model using k-out-of-n configuration. *Reliability: Theory & Applications*, **18**, 110–119.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 22, No. 1, 2024 (New Series), pp 149-169
http://www.ssca.org.in/journal



A Semi-Parametric Regression Hazards Model for Duration of Singlehood in North-East India

Lourembam Neroka Devi and Kshetrimayum Anand Singh
Department of Statistics
Manipur University,Canchipur, Imphal, Manipur, 795003, India

Received: 30 March 2023; Revised: 27 June 2023; Accepted: 30 June 2023

Abstract

Marriage is a life event which could change the qualitative status of an individual from single to married life. Studies on the duration of an individual's single status are of particular importance as it reflects the age pattern of marriage for a particular community or population. The median age at first marriage of India count to be 18.6 years for women and 24.5 years for men. This study aims the differentials and determinants of male and female singlehood durtation in North East States of India. Data from NFHS-4 are used to compute median duration of singlehood and its influential covariate are determined by using semi-parametric hazards model. Results show that the median duration of singlehood for North-East women and men are 21 years and 26 years respectively. Manipur women and men are recorded highest singlehood duration of 23 years and 27 years respectively. Findings shows that covariates such as place of residence, religion, ethnicity, wealth of the family and working status of women and men have significant effects on the duration of singlehood. As early marriages are expected to contribute more births it is important to increase the age at marriage of both men and women in order to reduce fertility.

Key words: Singlehood; NFHS-4; Median duration; Semi-parametric hazards model; North East India; Manipur.

Mathematics Subject Classification: 62G08, 91D20

1. Background

Singlehood is a term used for an individual who has never been married in his or her lifetime. Studies on singlehood for both men and women are equally important in a society. All the life course activities of an individual during his/her singlehood are determining factors for the individual's future shape. The status of single for an individual is defined in terms of their relationship to marriage. Marriage is a major life event where a change of status takes place in an individual's life course whoever male or female. In many societies, marriage is defined as the onset of socially accepted sexual activity and as such marriage is considered as an important proximate determinant of fertility by Bongaarts (1978). And also it is the onset

Corresponding Author:Lourembam Neroka Devi
Email: lourembamneroka@gmail.com

of making one family and society sooner or later. Marriage is a demographic event which could change the nomenclature of one's social recognition or status as an individual that is from single status (bachelor or spinster) to married life. There is differences or changes in the life course activities and routines between a never married individual and an ever married individual.

Stein (1975) suggest that U.S. census and surveys indicate the increased postponement of marriage led the growing number of singles. Singlehood as a positive choice have been made by adults who have chosen not to marry. Due to dissatisfaction with traditional marriage, a new lifestyle of being single throughout their life has been chosen. In addition, Stein (1975) reveals that more and more people are rejecting and postponing their marriage in favour of independence.

Many women in developing countries of the world are subject to early marriage. It is believed that many women in such countries have little to no chance to choose themselves to whom they should marry and at what age they would marry by Jensen and Thornton (2003). Hayase and Liaw (1997) claim that women who marry at an early age have a longer period of exposure to pregnancy and consequently led to high fertility level. Jensen and Thornton (2003) also reveals that early married women face many disadvantages in the field of education, status, autonomy and even including physical safety. They have less power on decision-making, and better experiences of domestic violence are reported from them.

Kumchulesi *et al.* (2011) suggest that many socio economic factors such as age of women, place of residence, region, economic status *etc.* have an effect on the age at first marriage. With rapid increase in the educational attainment, age at first marriage and age at first birth is also increased. In addition, Gangadharan and Maitra (2000) also found that education of husband significantly affects the time to conception.

The study from Weinberger (1987) found that early marriage occurs more often in the less educated women. Findings from the world fertility survey 1987 which include 38 countries around the world, shows that singulate mean age at marriage (SMAM) of women with seven or more years of education is almost four years higher than the women with no education. And Matlabi *et al.* (2013) also suggest that one of the method for reducing early child marriage is mandating girls stay in school. Early marriage is associated with early childbearing and also linked to various adverse health related outcomes for both mother and child. Such early child bearing is lowered by increasing longer duration of singlehood with subsequently slowing population growth.

Kumar (2016) studies provide that place of residence is a responsible variable for a wide variation in child marriage. The percentage of girls married before 18 years of age among all those got married 0-4 years before to census 2011 of India in rural areas was 21% while it is only half in urban area. Other studies from McLaughlin *et al.* (1993); Westoff and ORC Macro (2003) also show that rates of early marriage are higher in rural areas than in urban areas.

A longer duration of singlehood results in lowering childbearing experience with subsequently slowing population growth. Women who have higher level of intelligence, education and occupation are more likely to remain as a single for a longer period of time. Highly educated women want to live on their own way. However, studies from Spreitzer and Riley

(1974), in contrast to women, educated men, and those with higher occupational achievement want to get married sooner.

Lalmalsawmzauva *et al.* (2011) claim that, though the legal age at marriage for girls in India is fixed at 18 years many girls are married before reaching that particular age. The female age at marriage in India is not uniform in all states, districts, ethnic, caste, class and religious groups. The female of rural areas get married earlier than those of urban. Some states located to southern part (except Andhra Pradesh), North West and North East India have relatively higher mean age at marriage .

Although, till date marriage is universal in Indian context, there are certain shifts observed in the age at marriage. There is a consistence increasing trend in respect of mean and median age at marriage over cohort since 2005 to 2016 (NFHS-3 (2007) and NFHS-4 (2017)). Thus it becomes important to understand the current situation of marriage pattern in India in the light of policies aimed at increasing the age at marriage and the major contribution factors determining the change in median age at marriage in the last decade. Not much work had been done to model duration of singlehood exclusively for North East India. Thus, the present study attempt to analyze the differentials and determinants of singlehood duration of North Eastern States of India.

2. Data and methods

In view of literature reported on the age at marriage, the authors have an interest on the singlehood duration both for men and women the for whole North East India. The North Eastern region of India comprises of eight states *viz*. Assam, Arunachal, Manipur, Meghalaya, Mizoram, Nagaland, Sikkim and Tripura. These states have socio-economic and demographic characteristics, which are more or less different from the mainland population of India. Specifically, the economic activity in the region is quite different from the mainland as having little to no industrialization and mainly depends on agricultural activities. All states are dominated by tribal population except Assam where tribal population accounts for 12.5% only. The main religious groups in the region are Hindu, Muslim, Christians, Buddhists and some unrecognized local faiths still exist though fewer in number. The population of the region is sparse as compared to other parts of country and shares only 3.57% of whole population of India while the geographical area covers 7.5% of the total area of the country. All states except Meghalaya follow patrilineal norms while in Meghalaya there are ethnic groups who follow matrilineal norms. The present paper uses data from the National Family Health Survey-4 which provides information on various aspects of demographic analysis, reproductive health and nutrition for India. NFHS-4 (2017), 2015-16 (International Institute for Population Sciences (IIPS) and ICF, 2017) collected information from nationally representative samples regarding women, men, household and children. Interview was taken from 98702 women in the age group 15-49 years of the eight states of North Eastern region of India. Also, 14555 men in the age group 15-54 years were interviewed during the survey in the region. The duration of singlehood for these women and men were obtained from their age at marriage and current age. For those women or men who are never married, their duration of singlehood is obtained from their current age and is marked censored. Censored duration indicates that the event of interest (*i.e* marriage) does not occur to these women and men, whereas a complete duration indicates that the event has occurred at least once to the individuals. Out of total the women samples from North East India, 93321 women were

considered in the present study. Likewise, 14280 men in age group 15-54 were considered in the analysis.

2.1. Variables in the study

In any regression analysis one has to ascertain the outcome and the predictor variables. The outcome or dependent variable in the present study is the duration of singlehood. In the event history method duration of singlehood may be looked upon as the time to first marriage. In the literature, time to first marriage is the duration of the total time where an individual lives in the single state starting from the birth of individual. Several predictor variables are considered in the present study which are potential to influence the duration of singlehood. All the variables are categorical variables. These variables which are thought to influence the outcome variable are educational level of individuals, type of place of residence, religion, ethnicity, wealth of family, exposure to mass media and working status. At the community level, place of residence, religion and ethnicity are considered. At the household level wealth of family is considered and at the individual level educational level, exposure to mass media and working status are included as covariates. Table 1 gives the definition and categories of the predictor variables.

2.2. Methodology

The duration of singlehood (or time to first marriage) has been studied by way of survival analysis techniques using the non-parametric Kaplan-Meier, see Kaplan and Meier (1958) method and the semi-parametric Cox proportional hazards model, Cox (1972). As noted earlier duration of singlehood is the time an individual has got married for the first time starting from birth. In the present study, the event of interest is the first marriage. The Cox proportional hazards model is the most applied regression technique which addresses the risk of event time. Thus, the time to first marriage is fitted to the Cox model considering some potential covariates which are thought to explain the age of first marriage to estimate the relative risks. The median duration of singlehood is computed using the non-parametric K-M method.

Kaplan-Meier estimator of survival probability at time t is given by

$$\hat{S}(t) = \prod_{t_i<t} \frac{r_{t_i} - d_{t_i}}{r_{t_i}}$$

where r_{t_i} is the number of risk of experiencing the event at the time t_i, and d_{t_i} is the number of events at that time, with the convention that $\hat{S}(t) = 1$ if $t < t_i$.

Using Greenwood's formula for the variance of survival function

$$\hat{V}(\hat{S}(t)) = (\hat{S}(t))^2 \sum_{r_{t_i}} \frac{r_{t_i}}{r_{t_i} - d_{t_i}}$$

$$\hat{V}[\ln(\hat{S}(t)] = \sum_{r_{t_i}} \frac{r_{t_i}}{r_{t_i} - d_{t_i}}$$

The end point of $100(1 - \alpha)\%$ confidence interval for $S(t)$ on cumulative hazard or log-

Table 1: Variables and categories with sample size

Variables	Categories	Sample size	
		Women	Men
State	Arunachal Predesh	13667	2109
	Assam	27089	4138
	Manipur	12956	1856
	Meghalaya	8662	1220
	Mizoram	11115	1661
	Nagaland	10275	1567
	Sikkim	5114	860
	Tripura	4553	871
Type of place ofresidence	Urban	24754	3923
	Rural	68567	10357
Educational level	No Education	16418	1672
	Primary	13257	2142
	Secondary	54803	8629
	Higher	8843	1837
Religion	Hindu	34828	5512
	Muslim	10443	1500
	Christian	38789	5800
	Others	9266	1468
Ethnicity	SC	6707	1098
	ST	51621	7985
	OBC	12845	1993
	Others	14345	1933
Wealth of family	Poor	40363	6081
	Middle	24945	3891
	Rich	29613	4308
Working status	No	11417	3248
	Yes	3776	11032
Exposed to Media	No	16651	1378
	Yes	76670	12902

Working status and wealth of family are defined in the note.

survival scale is

$$\exp(\ln \hat{S}(t)) \pm z_{1-\frac{\alpha}{2}} \hat{se}(\ln \hat{S}(t)) \tag{1}$$

We have also computed the estimates of sample median and 95% confidence intervals for all the estimates. The median duration of singlehood is obtained at the time point at which $S(t)$ is less than 0.5. Using the test for Mantel (1966) Log rank test is used to compare the survival experience of duration of singlehood among the categories defined by socio-economic covariates. The regression model for the hazard function that addresses the

study goal is

$$h(t|x) = h_0(t).r(\beta' x) \tag{2}$$

Where $h(t|x)$ is the hazard function, $h_0(t)$ is the baseline hazard and β is the vector of regression parameters and x is the vector of explanatory covariates. Under the model in (2), the ratio of the hazard functions for two individuals (or group of individuals) with covariates x_1 and x_2 is

$$HR(t|x_1, x_2) = \frac{h(t|x_1)}{h(t|x_2)} = \frac{r(\beta' x_1)}{r(\beta' x_2)} \tag{3}$$

From (3), see that if the hazard ratio is easily interpreted then baseline hazard is of little importance. Cox (1972) proposed that the conditional hazard $h(t|x)$ be modelled as the product of $h_0(t)$ and an exponential function which is linear in x that is $r(\beta' x) = e^{\beta' x}$ so that

$$h(t|x) = h_0(t).e^{\beta' x} \tag{4}$$

Under the Cox model in (4), the hazard ratio

$$HR(t|x_1, x_2) = \exp(\beta' x_1 - x_2)$$

As the method in (4) forces the hazard ratio between two individuals to be constant over time, we call it proportional hazards model.

3. Results

The estimated median duration of singlehood along with 95% confidence interval (C.I.) using (1) for both women and men computed using the non-parametric Kaplan-Meier method is presented in Table 2. The median duration of singlehood for the whole North-East women and men are 21 years and 26 years respectively. Among the eight states Assam and Tripura have the shortest duration of singlehood estimated at 19 years each and Manipur have the longest duration estimated at 23 years for women. For men, Manipur and Nagaland have longest median duration for singlehood of 27 years and Arunachal, Meghalaya and Mizoram have the least median duration of 25 years. Those women and men who are living in urban area have longer median duration of singlehood as compared to their rural counterparts by two years. Women who are educated upto secondary or higher have longer singlehood duration than those women who are educated upto primary or illiterate. Women belonging to Christian and Others religious groups have longer median duration of singlehood. However, those men who are in Hindu religion have longest median duration as compared to the remaining groups. Results show that Muslim women and men have lowest median duration of 18 years and 25 years respectively. Those men who are belonging to SC, ST and OBC category have same median duration of 26 years and others category have the highest (27 years) median duration. Women belonging to SC category have shortest duration of singlehood. Median duration of singlehood for men living in poor and middle wealth categories increases successively by one year. Rich men tend to have longer duration of singlehood (28 years). Women from poor family have shortest singlehood duration among all wealth categories. Generally individuals from richest family have to stay longer in single status. Furthermore, working women have the longer length of singlehood duration than their non-working counterparts but it is contrast in men category. Exposure to mass media is also one of the significant covariates for the study of singlehood duration. Those women

Table 2: Median duration of singlehood for women and men and its p-values for testing significance of survivorship experience among categories

Variables	Categories	Median(95% C.I.)		Log rank test (p-value)
		Women	Men	
State	Overall	21(20.95,21.05)	26(25.85,26.16)	0.00
	Arunachal Pradesh	20(19.88,20.12)	25(24.61,25.38)	
	Assam	19(18.92,19.08)	26(25.71,26.28)	
	Manipur	23(22.83,23.16)	27(26.56,27.44)	
	Meghalaya	21(20.82,21.17)	25(24.57,25.42)	
	Mizoram	22(21.83,22.17)	25(24.57,25.42)	
	Nagaland	22(21.82,22.18)	27(26.53,27.47)	
	Sikkim	22(21.76,22.24)	26(25.37,26.65)	
	Tripura	19(18.83,19.17)	26(25.84,26.54)	
Type of place of residence	Urban	23(22.87,22.13)	27(26.86,27.31)	0.00
	Rural	20(19.94,20.06)	25(24.82,25.18)	
Educational level	No Education	18(17.91,18.08)	24(23.64,24.35)	0.00
	Primary	19(18.91,19.08)	24(23.67,24.32)	
	Secondary	21(20.94,21.06)	26(25.79,26.20)	
	Higher	28(27.76,28.23)	30(29.59,30.40)	
Religion	Hindu	21(20.92,21.08)	27(26.75,27.25)	0.00
	Muslim	18(17.90,18.10)	25(24.57,25.42)	
	Christian	22(21.94,22.08)	26(25.76,26.23)	
	Others	21(20.82,21.17)	25(24.53,25.46)	
Ethnicity	SC	20(19.82,20.17)	26(25.24,26.57)	0.00
	ST	21(20.92,21.07)	26(25.79,26.20)	
	OBC	21(20.86,21.13)	26(25.59,26.40)	
	Others	21(20.86,21.13)	27(26.54,27.45)	
Wealth of family	Poor	19(18.94,19.06)	25(24.79,25.20)	0.00
	Middle	21(20.90,21.09)	26(25.69,26.30)	
	Rich	23(22.88,23.11)	28(27.70,28.29)	
Working status	No	20(19.86,20.13)	29(28.17,29.80)	0.00
	Yes	22(21.69,22.31)	25(24.84,25.17)	
Exposed to Media	No	19(18.91,19.08)	25(24.62,25.37)	0.00
	Yes	21(20.94,21.05)	26(25.83,26.16)	

who are exposed to mass media have 2 years longer singlehood duration than women who are not exposed to mass media. The last column of Table 2 gives the p-values for testing the significant difference of the survival experience among the groups or categories defined by the socio-economic covariates. All covariates are significant at 5% level in the log rank test which in turn suggests that these covariates are important to influence singlehood duration and are potential candidates for the hazards regression model.

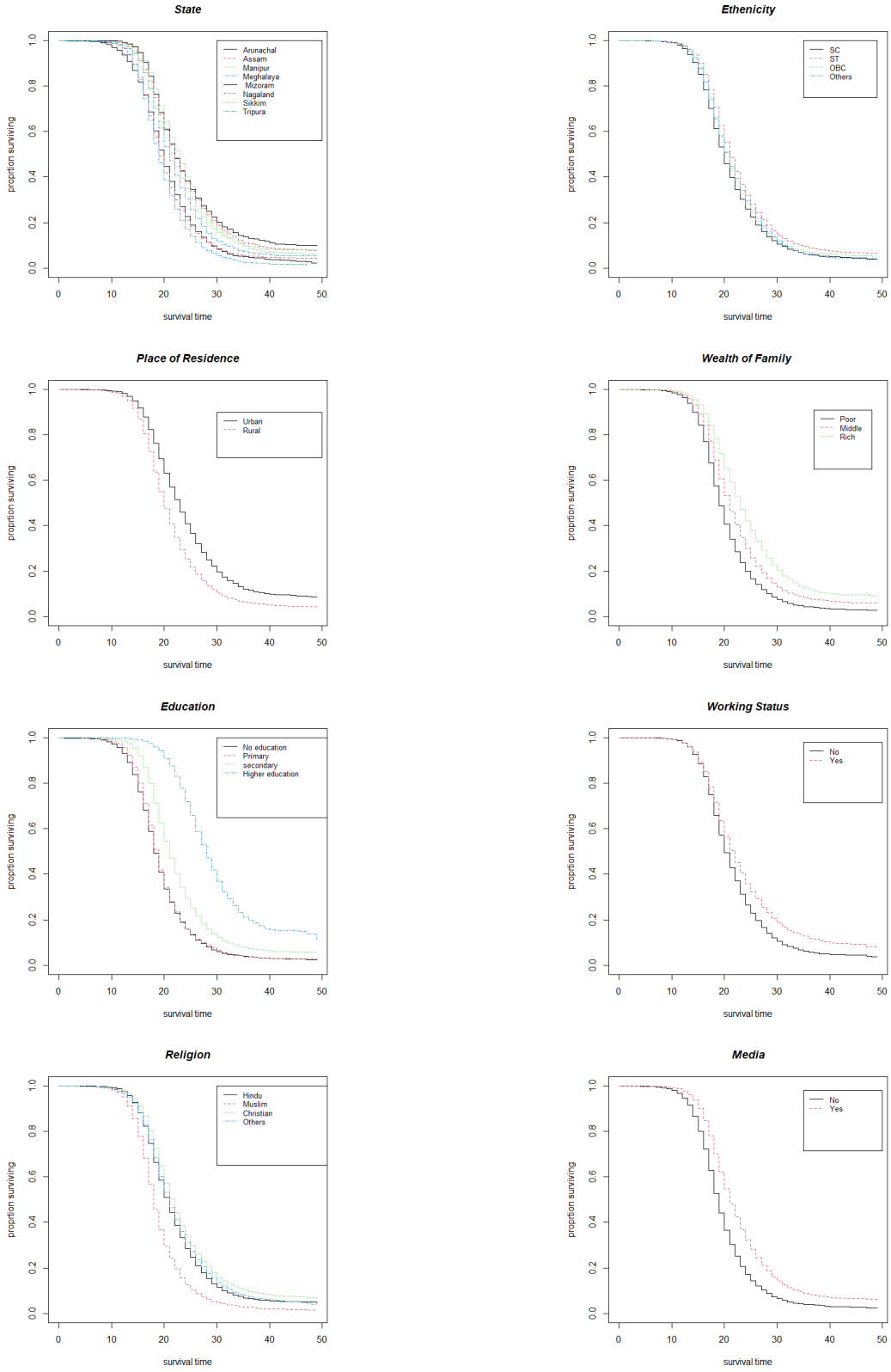


Figure 1: Survival curves by background characteristics(women)

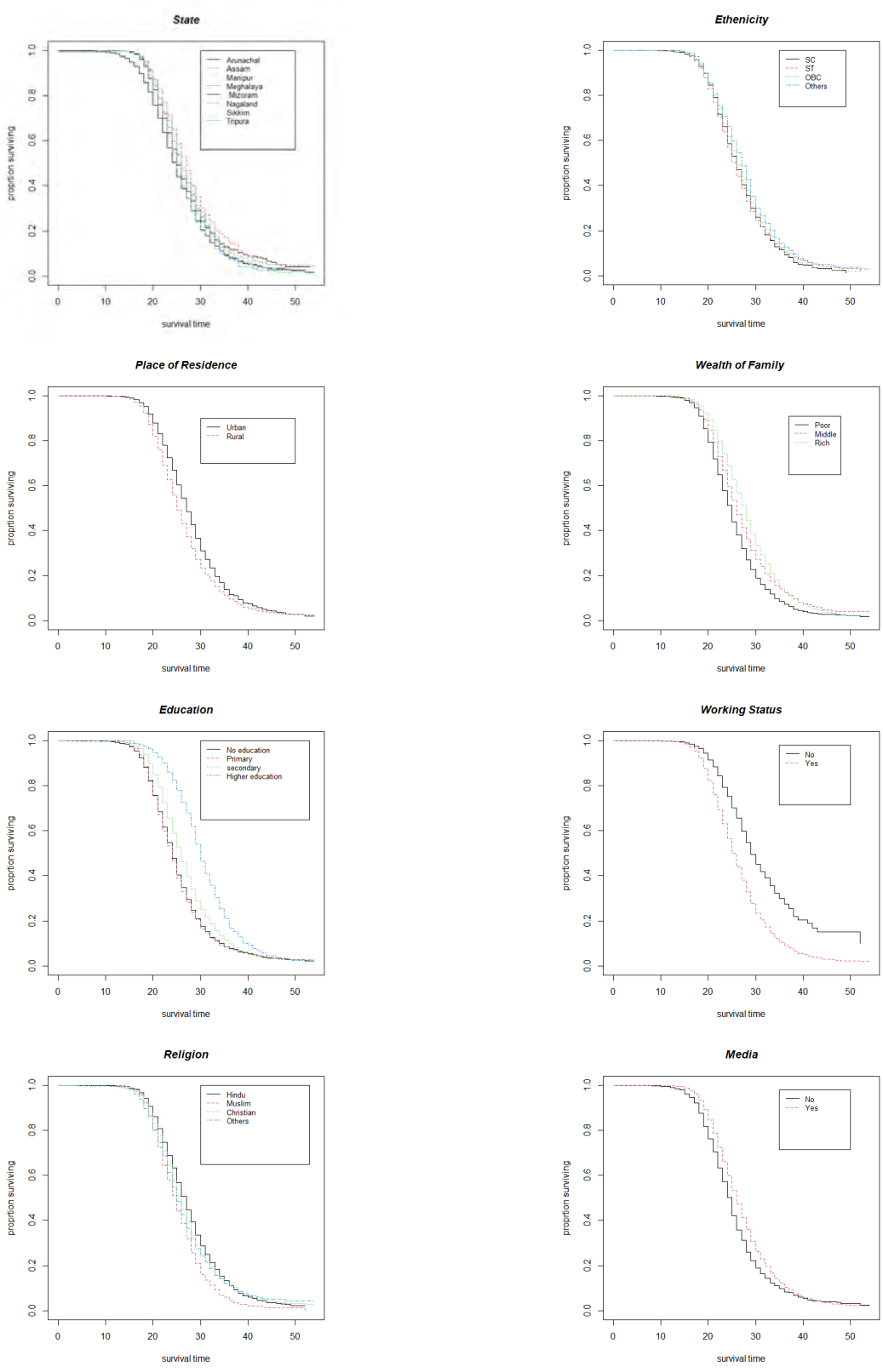


Figure 2: Survival curves by background characteristics(men)

3.1. Results of fitting the Cox hazards regression model

All covariates which are significant (at 5% level) in the bivariate analysis (Table 2) are potential covariates to include in the hazards model as explanatory variables. Consequently, two hazards models one for women and one for men are fitted with the significant covariates to regress the duration of singlehood. The results of fitting the two models are presented in Tables 3 and 4 which include the estimated coefficients, hazards ratio (indicating the reference category), standard error of the estimates and p-values for Wald test for testing the significance of the coefficients. The two models are checked and verified for violation of proportionality assumptions and leverage for influencing observations. Both the models pass the test for proportionality assumption and no influencing observation is present in the data. We discuss it in the next subsection model diagnostics.

3.2. Model diagnostics

Model based inferences depend completely on the fitted Statistical model. For these inferences to be valid in the real sense of the world, fitted model must provide an adequate summary of the data upon which it is based. A complete and thorough examination of the model's fit and adherence to model assumptions is just as important as careful model development. The methods for assessment of a fitted proportional hazards model essentially consists of i) methods for testing the assumption of proportional hazards and ii) subject specific diagnostic statistics that extend the notion of leverage and influence to the proportional hazards model.

A large number of tests of proportionality assumptions are found in the literature. However, works developed by Grambsch and Therneau (1994) and simulation work by Ng'andu (1997) have shown that an easily performed statistical test and an associated graph yield a powerful and effective method for examining the proportionality assumption. These are the two steps: 1) add the covariates by log time interaction to the model and assess their significance using partial likelihood ratio test, score test or Wald test, and 2) plot the scaled and smoothed scaled Schoenfeld residuals obtained from the model without interaction terms. The result of the two steps should support each other.

The plot for scaled Schoenfeld residuals for some of the covariates are shown in Annexure (Figure 3 and Figure 4) for female and male respectively. Results of statistical tests for proportionality assumptions are shown in Annexure (Table 5 and Table 6) for women and men respectively. We examine the p-values for Wald tests in the interaction terms which are all insignificant suggesting that the covariates have passed the proportionality tests. The graphical plots are more or less flat in all the covariates which support that they all have approximately zero slopes. Covariates with p-value significant at 5% level of significance are removed from the model fit as they may violate proportionality assumption. In the women data all covariates have insignificant p-values and they are all included in the main effects model. All covariates except education are insignificant at 5% level of significance in the men data. So, the final main effects model in the men data include the covariates state, place of residence, religion, ethnicity, wealth of family, working status and media.

It is also important to check outliers by visualising the deviance residuals to identify the influential subjects in the data. The plots of the deviance residuals are shown in Appendix (Figure 5). From the plots it is evident that there are no widely deviant observations in both

the women and men data.

Table 3: Estimated regression coefficients (β), hazards ratio (HR), standard error(SE), two tailed p-values for the proportional hazard model for women

Variables	Categories	β	HR	SE	p-value
State	Arunachal Pradesh(Ref)	-	-	-	-
	Assam	-0.13	0.88	0.04	0.001***
	Manipur	-0.34	0.71	0.04	0.000***
	Meghalaya	-0.17	0.84	0.045	0.000***
	Mizoram	-0.35	0.70	0.045	0.000***
	Nagaland	-0.35	0.70	0.04	0.000***
	Sikkim	-0.21	0.81	0.05	0.000***
	Tripura	0.18	1.20	0.05	0.000***
Type of place of residence	Urban(Ref)	-	-	-	-
	Rural	0.07	1.08	0.03	0.005***
Educational level	No Education(Ref)	-	-	-	-
	Primary	0.01	1.01	0.03	0.725
	Secondary	-0.38	0.69	0.03	0.000***
	Higher	-1.25	0.29	0.05	0.000***
Religion	Hindu(Ref)	-	-	-	-
	Muslim	0.35	1.41	0.05	0.000***
	Christian	0.06	1.06	0.04	0.111
	Others	-0.03	0.97	0.04	0.458
Ethnicity	SC(Ref)	-	-	-	-
	ST	-0.14	0.87	0.04	0.001***
	OBC	-0.06	0.94	0.04	0.151
	Others	-0.02	0.97	0.04	0.529
Wealth of family	Poor(Ref)	-	-	-	-
	Middle	-0.69	0.94	0.03	0.019**
	Rich	-0.09	0.92	0.03	0.006***
Working status	No(Ref)	-	-	-	-
	Yes	-0.08	0.93	0.02	0.001***
Media	No(Ref)	-	-	-	-
	Yes	-0.004	1.00	0.03	0.888

Ref=reference, ***=p <0.01, **=p <0.05.

3.3. Interpretation of fitted models

The popularity of a fitted regression hazards model is due to its ease in interpreting and understanding the hazards ratios which literally gives relative risk of experiencing the event of interest with respect to a reference category for a categorical covariate. The 4th column of Table 3 present the relative risk of first marriage for women and men respectively,

Table 4: Estimated regression coefficients (β), hazards ratio (HR), standard error(SE), two tailed p-values for the proportional hazard model for men

Variables	Categories	β	HR	SE	p-value
State	Arunachal Pradesh(Ref)	-	-	-	-
	Assam	-0.37	0.69	0.04	0.001***
	Manipur	-0.36	0.70	0.05	0.000***
	Meghalaya	-0.20	0.82	0.05	0.000***
	Mizoram	-0.18	0.83	0.05	0.000***
	Nagaland	-0.44	0.64	0.05	0.000***
	Sikkim	-0.05	0.95	0.05	0.334
	Tripura	-0.25	0.78	0.06	0.000***
Type of place of residence	Urban(Ref)	-	-	-	-
	Rural	0.05	1.05	0.03	0.078*
Educational level	No Education(Ref)	-	-	-	-
	Primary	0.01	1.01	0.03	0.725
	Secondary	-0.38	0.69	0.03	0.000***
	Higher	-1.25	0.29	0.05	0.000***
Religion	Hindu(Ref)	-	-	-	-
	Muslim	0.20	1.22	0.06	0.000***
	Christian	0.06	1.06	0.04	0.171
	Others	0.013	1.01	0.05	0.795
Ethnicity	SC(Ref)	-	-	-	-
	ST	-0.03	0.97	0.05	0.540
	OBC	-0.007	0.99	0.05	0.822
	Others	-0102	0.88	0.05	0.014**
Wealth of family	Poor(Ref)	-	-	-	-
	Middle	-0.24	0.79	0.03	0.000***
	Rich	-0.37	0.69	0.03	0.000***
Working status	No(Ref)	-	-	-	-
	Yes	0.65	1.91	0.04	0.000***
Media	No(Ref)	-	-	-	-
	Yes	0.005	1.01	0.04	0.884
Ref=reference,	***=p <0.01, **=p <0.05, *=p <0.1				

along with the regression coefficients and standard error of coefficients for different socio-economic covariates. In Table 3 and Table 4, the hazards ratios (HR) are shown for all the North East states (women) with Arunachal as the reference category state. From the p-values (Wald test)in the last column of Table 3, it is evident that the HR for all states are significant at 5% level. Manipur's HR of 0.71 reveals that women in Manipur have nearly 29% less risk of first marriage as compared to women in Arunachal Pradesh. Similarly, women in Nagaland have significantly lower risk (30%) of first marriage as compared to Arunachal women. However, women in Tripura marry earlier as the HR of 1.2 indicates that the risk of first marriage for women in Tripura is nearly 1.2 times that of Arunachal women.

As of now the insignificant HR's may not be interpreted as such. For men (Table 4) all the HRs for all the states except Sikkim are highly significant. Literally, the risks of marriage for men in these states are significantly lower than that of Arunachal men.

In the whole North East region, women who live in urban area have lower risk of age at marriage. Rural women have 1.08 times higher risk of first marriage as compared to urban women. Approximately, men in rural area have the same higher risk (5%) of first marriage as compared to urban men. Among women who are educated upto primary or no education the risk of marriage does not differ significantly. However, those women who have education upto secondary and higher have significantly lower risk of first marriage upto the tune of 31% and 70% respectively as compared women who have no education at all. Among the religious groups at the community level, the relative risk of first marriage for Muslim women is significantly higher than the Hindu women. Muslim women have 51% higher chance to marry earlier than the Hindu women. Similarly, for Muslim men have 22% more chance of first marriage as compared to men in the Hindu religion. Other categories of religion are not significant. Among the ethnic groups, ST category has HR = 0.97 which interpret that women of ST category have 3% less likely to marry as compared to women belonging to SC category. Other categories of ethnicity are not significant.

At the household level, women living in middle and rich family exhibit lower risks of marriage as compared to women in poor family. Women in middle and rich wealth quintiles are 6% and 8% less likely to marry as compared to women in poor wealth quintile respectively. Similarly, men belonging to middle and rich wealth quintiles are respectively 21% and 31% less likely to marry as compared to men belonging to poor wealth quintile.

At the individual level, working status of both women and men has significant effect on singlehood duration. Working women have less chance of marriage to the tune of 7% less as compared to women living with no working status. However, for men the result is just the reverse as working men have more chance to marry to the tune of 1.9 times more likely as compared to non-working men.

4. Discussion

Age at marriage is one of the significant life events for every individual. It signals the entry of each individual into the state of being married. This study attempted to investigate the median duration of singlehood for North east India using the NFHS-4 data. Cox proportional hazards model is fitted to assess the significant effect of various covariates on the singlehood duration.

First, it is observed that the duration of singlehood varies among groups of population identified by different covariates. In NFHS-4 (2017), 2015-16 (International Institute for Population Sciences (IIPS) and ICF, 2017) the median age at marriage for women and men in India is estimated to be 18.6 years and 24.5 years respectively. However, for North–east region the median age at first marriage for women is 21 years and 26 years for men. This indicates that the people in North east India are more likely to live in single status than the people in the rest of the country. North East region of India comprises of eight states with a different socio-cultural set up from the mainland India. From the results states with Christian as main religion like Arunachal, Meghalaya and Mizoram have least duration of singlehood in men category. Whereas Manipur and Nagaland show higher estimates of

median duration of singlehood than others in men category.

Rural women and men are more likely to get married at early ages which in turn indicate that urban people have longer median duration of singlehood as compared to rural people in the North east region. Educational level of individuals is one of the important determinant factors for the early marriage as many literatures have cited. North East women and men with higher education have longer duration of singlehood than others with low educational level which is in line with the findings of other studies. The chance of singlehood for women increases with increase in educational level.

For the whole country according to NFHS-4 (2017), 2015-16 (International Institute for Population Sciences (IIPS) and ICF, 2017) and NFHS-3 (2007), 2005-06 (International Institute for Population Sciences (IIPS) and Macro International, 2007) reports, Hindus and Muslims have similar median age at marriage for the whole country. However, in the North East region, Muslims are more likely to marry at early ages than the Hindus and Christians. Thus, Muslims have shorter duration of singlehood.

Present paper also explores the effect of ethnicity on the survival experience on the duration of singlehood. Schedule tribe women population has lower risk of marriage as compared to Schedule caste women population in the region less chance of marriage than the schedule caste women. Besides, men from others ethnicity groups are less likely to get married than those in other categories. Another important finding is that wealth of the family significantly affects the duration of singlehood. Women from poor family are more likely to marry earlier than others. In a similar manner,men fro m poor family have higher chance of marriage than men from richer families. Last but not the least; we assess the influence of working status on the singlehood duration. Working women are more likely to be in single state than non-working women. Interestingly, this phenomenon is just the reverse for men, which shows that men who are currently working have more risk of marrying earlier than their non- working counterparts. This is also in line with some of the findings in the literature. Exposure to mass media has no significant effect on the study of duration of singlehood in North East region.

5. Conclusion

Marriage is a major life event which basically marks the onset of married couples contributing to human reproduction. As such marriage is considered as an important proximate determinant of fertility for a country or a region. As early marriages are expected to contribute more births it is important to increase the age at marriage of both men and women in order to reduce fertility NFHS-3 (2007). The median age at marriage in India increases from NFHS-3 (2007) to NFHS-4 (2017) by two years for both men and women. In order to further improve the age at marriage, the policy makers have to give further attention to the socio-economic disparities of age at marriage in the country. Regional findings will be helpful in the present context and as such the findings in this paper could be helpful to policy and programme planners while addressing the issue of population control through improvement in age at marriage.

Note

1. According to NFHS-4 (2017), Wealth index is a measure of living standards based on households' ownership of items such as televisions to housing features such as drinking water sources. The population is divided into five equally sized groups based on index. The top 20% form the richest, and the bottom 20% the poorest quintile. In the present analysis the wealth index is condensed into 3 categories *viz* (1) poorer and poorest into poor, (2) middle and (3) richer and richest to rich.

2. Women who are currently working outside the home for earning are considered as working women. Such working women are categorized as "Yes" otherwise "No".

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

Bongaarts, J. (1978). A framework for analyzing the proximate determinants of fertility. *Population and Development Review*, **34**, 105–132.

Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B (Methodological)*, **34**, 187–202.

Gangadharan, L. and Maitra, P. (2000). *The Effect of Education on the Timing of Marriage and First Conception in Pakistan*. Monash University.

Grambsch, P. M. and Therneau, T. M. (1994). Proportional hazards tests and diagnostics based on weighted residuals. *Biometrika*, **81**, 515–526.

Hayase, Y. and Liaw, K.-L. (1997). Factors on polygamy in Sub-Saharan Africa: findings based on the demographic and health surveys. *The Developing Economies*, **35**, 293–327.

Jensen, R. and Thornton, R. (2003). Early female marriage in the developing world. *Gender & Development*, **11**, 9–19.

Kaplan, E. L. and Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*, **53**, 457–481.

Kumar, S. (2016). Measuring child marriage from census and large scale data systems in india. *Demogr India*, **45**, 59–76.

Kumchulesi, G., Palamuleni, M., and Kalule-Sabiti, I. (2011). Factors affecting age at first marriage in malawi. In *Sixth African Population Conference, Ouagadougou-Burkina Faso*, pages 5–9.

Lalmalsawmzauva, K., Nayak, D., and Vanlalvena, R. (2011). Interface between developments and female age at marriage in mizoram. *Transaction, Journal of the Institute of Indian Geographers*, **33**, 139–149.

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother Rep*, **50**, 163–170.

Matlabi, H., Rasouli, A., Behtash, H. H., Dastjerd, A. F., and Khazemi, B. (2013). Factors responsible for early and forced marriage in iran. *Science Journal of Public Health*, **1**, 227–229.

McLaughlin, D. K., Lichter, D. T., and Johnston, G. M. (1993). Some women marry young: Transitions to first marriage in metropolitan and nonmetropolitan areas. *Journal of Marriage and the Family*, **55**, 827–838.

NFHS-3 (2007). *National Family Health Survey (NFHS-3), 2005-06: India (2 v.+ suppl.)*, volume 1. International Institute for Population Sciences.

NFHS-4 (2017). *National Family Health Survey (NFHS-4), 2015-16: India*, volume 1. International Institute for Population Sciences.

Ng'andu, N. H. (1997). An empirical comparison of statistical tests for assessing the proportional hazards assumption of cox's model. *Statistics in Medicine*, **16**, 611–626.

Spreitzer, E. and Riley, L. E. (1974). Factors associated with singlehood. *Journal of Marriage and the Family*, **36**, 533–542.

Stein, P. J. (1975). Singlehood: An alternative to marriage. *Family Coordinator*, **24**, 489–503.

Weinberger, M. B. (1987). The relationship between women's education and fertility: selected findings from the world fertility surveys. *International Family Planning Perspectives*, **13**, 35–46.

Westoff, C. F. and ORC Macro, C. (2003). Trends in marriage and early childbearing in developing countries. **5**.

ANNEXURE

Table 5: Test for proportional hazards assumption(women)

Covariants	Categories	Chisq	p-value
State	Assam	0.120	0.729
	Manipur	0.517	0.472
	Meghalaya	0.630	0.428
	Mizoram	0.222	0.638
	Nagaland	0.268	0.604
	Sikkim	0.533	0.465
	Tripura	1.170	0.279
Type of place of residence	Rural	1.950	0.163
Educational level	Primary	2.600	0.107
	Secondary	0.001	0.969
	Higher	119	0.000
Religion	Muslim	1.880	0.171
	Christian	0.518	0.472
	Others	0.206	0.650
Ethnicity	ST	0.168	0.682
	OBC	3.060	0.081
	Others	1.290	0.257
Wealth of family	Middle	4.430	0.035
	Rich	15.600	0.000
Working status	Yes	16.600	0.000
Media	Yes	3.710	0.054
State:time	Assam	1.300	0.254
	Manipur	2.430	0.119
	Meghalaya	0.037	0.847
	Mizoram	1.420	0.234
	Nagaland	0.739	0.390
	Sikkim	0.365	0.546
	Tripura	0.075	0.784
Type of place of residence :time	Rural	0.592	0.442
Educational level :time	Primary	2.600	0.107
	Secondary	0.001	0.969
	Higher	119	0.000
Religion:time	Muslim	1.880	0.171
	Christian	0.518	0.472
	Others	0.206	0.650
Ethnicity :time	ST	0.168	0.682
	OBC	3.060	0.081
	Others	1.290	0.257
Wealth of family :time	Middle	0.757	0.384
	Rich	0.180	0.672
Working status:time	Yes	3.380	0.066
Media :time	Yes	0.483	0.487

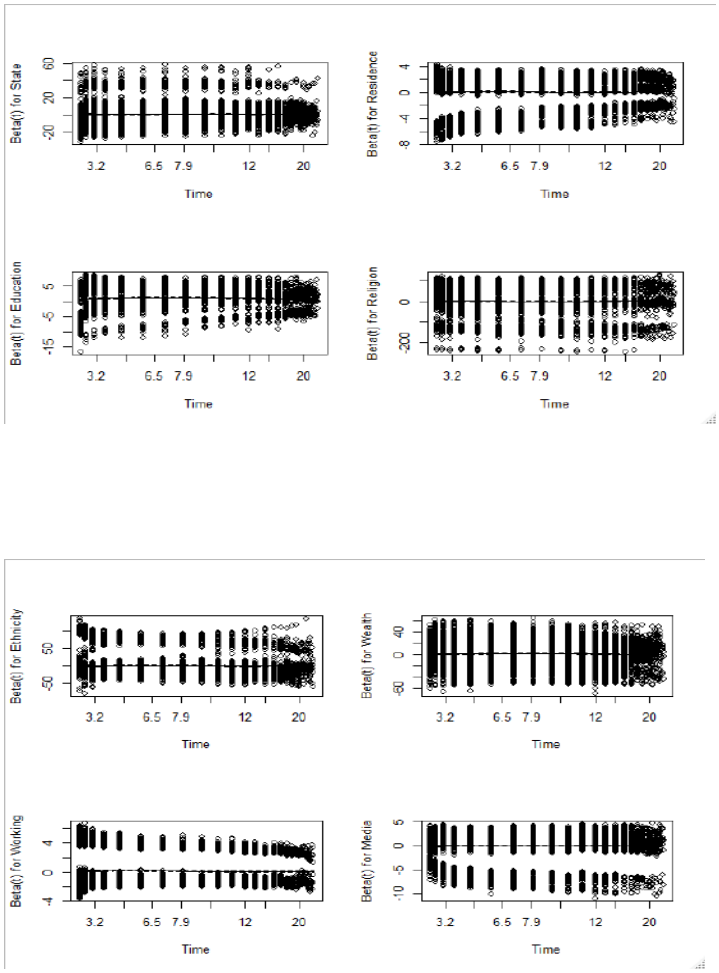


Figure 3: Plot of scaled Schoenfeld residuals and smoothed scaled Schoenfeld residuals for assessing proportionality assumptions in some covariates (women)

Table 6: Test for proportional hazards assumption(men)

Covariates	Categories	Chisq	p-value
State	Assam	0.800	0.371
	Manipur	1.450	0.229
	Meghalaya	8.640	0.003
	Mizoram	4.330	0.037
	Nagaland	0.137	0.711
	Sikkim	0.172	0.678
	Tripura	0.143	0.706
Type of place of residence	Rural	4.840	0.028
Educational level	Primary	6.400	0.011
	Secondary	1.510	0.219
	Higher	119.000	0.000
Religion	Muslim	0.096	0.757
	Christian	1.310	0.253
	Others	0.051	0.822
Ethnicity	ST	0.713	0.398
	OBC	0.263	0.608
	Others	0.000	0.922
Wealth of family	Middle	0.080	0.777
	Rich	0.732	0.392
Working status	Yes	1.400	0.237
Media	Yes	0.035	0.852
State:time	Assam	1.720	0.190
	Manipur	0.657	0.418
	Meghalaya	8.920	0.003
	Mizoram	6.340	0.012
	Nagaland	0.807	0.369
	Sikkim	0.102	0.749
	Tripura	0.196	0.658
Type of place of residence :time	Rural	1.890	0.169
Educational level :time	Primary	7.010	0.008
	Secondary	7.480	0.006
	Higher	91.700	0.000
Religion:time	Muslim	0.151	0.698
	Christian	0.645	0.422
	Others	0.123	0.725
Ethnicity :time	ST	0.183	0.669
	OBC	0.137	0.711
	Others	0.082	0.775
Wealth of family :time	Middle	0.070	0.792
	Rich	0.711	0.399
Working status:time	Yes	0.676	0.411
Media :time	Yes	0.013	0.911

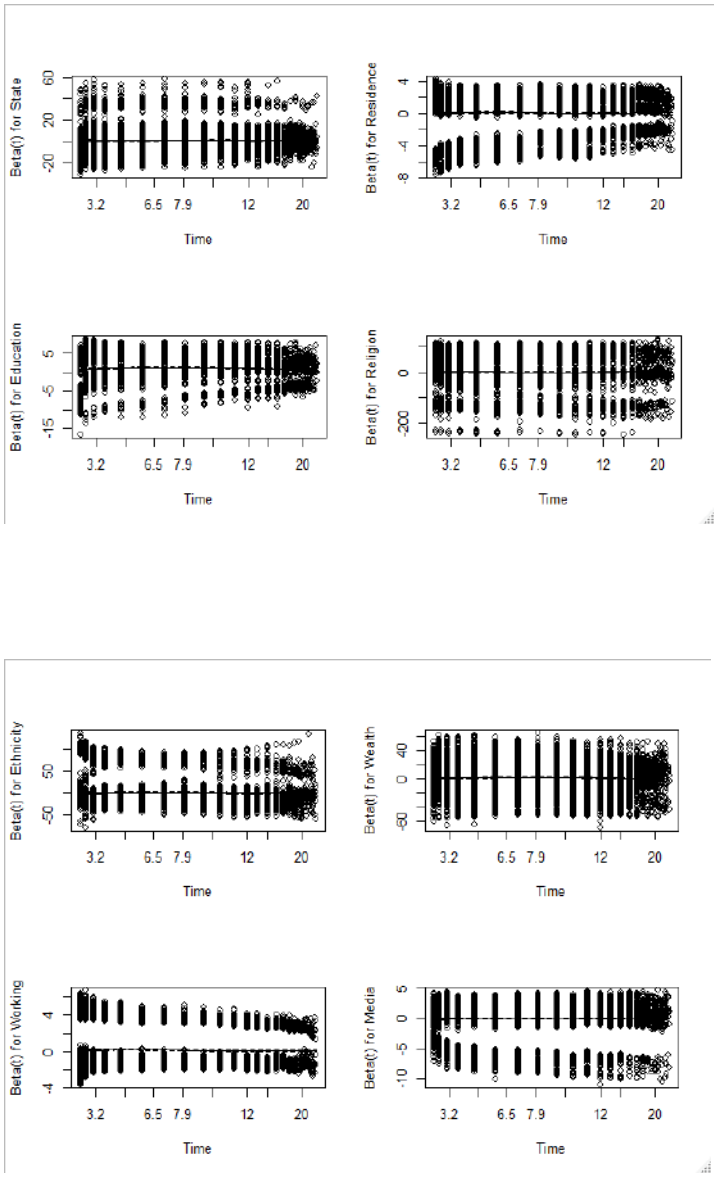


Figure 4: Plot of scaled Schoenfeld residuals and smoothed scaled Schoenfeld residuals for assessing proportionality assumptions in some covariates (men)

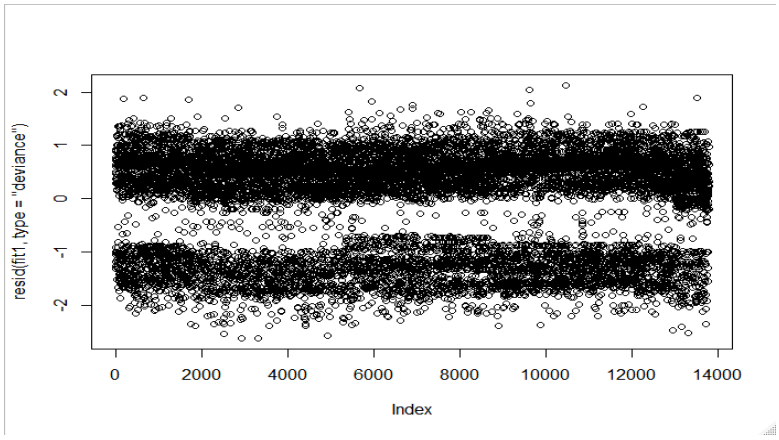


Figure 5a: Deviance residuals for women covariates

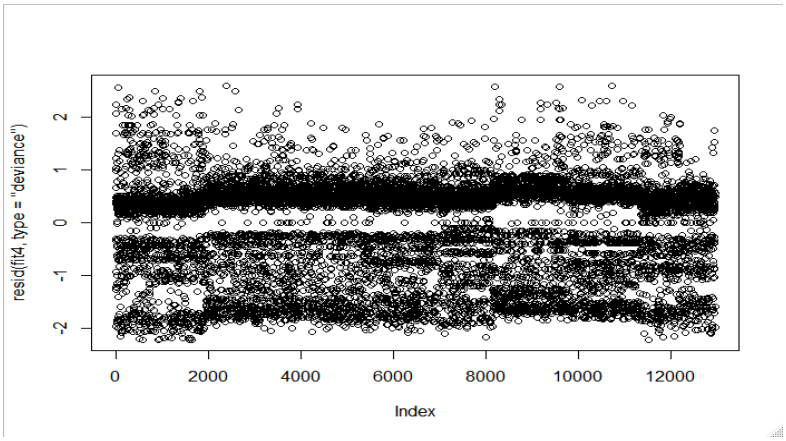
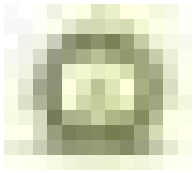


Figure 5b: Deviance residuals for men covariates

Figure 5: Deviance residuals

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 22, No. 1, 2024 (New Series), pp 171–192
http://www.ssca.org.in/journal



A Multi-Criteria Decision-Making Approach to Compare the Maternal Healthcare Status of Indian States: An Application of Data Science

Sangeeta Goala[1], Supahi Mahanta[2] and Dibyojyoti Bhattacharjee[1]
[1]*Department of Statistics, Assam University, Silchar, Cachar, Assam*
[2]*Department of Agricultural Statistics, Assam Agricultural University, Jorhat, Assam*

Received: 30 March 2023; Revised: 31 May 2023; Accepted: 14 July 2023

Abstract

The health system of a nation influences the well-being of its citizens. Maternal health is about the contentment of women throughout pregnancy, childbirth, and the postpartum period. In a country with millions of people like India, there are still goals in the area of maternal healthcare that need to be met despite widespread concern by the authorities. Spatial quantification of maternal health is necessary to identify the regions of immediate concern. In light of the methods and the variables used- the result of the quantification techniques produces a range of possible outcomes. The paper builds Composite Indicators based on some parameters of maternal healthcare, using different weighting methods, namely- TOPSIS, Iyengar-Sudarshan, Principal Component Analysis, Data Envelopment Analysis, and Ordered Weighted Average. Eventually, the most robust weighting technique is identified. The study finds Lakshadweep, Kerala, and Goa have better maternal healthcare, while Bihar, Arunachal Pradesh, and Nagaland are poorly positioned.

Key words: Demography; Composite index; TOPSIS; Principal component analysis; Data envelopment analysis; Ordered weighted average; Robustness.

AMS Subject Classifications: 90B50, 91B42

1. Introduction

Health, education, and income are the essential aspects of human development Swain and Mohanty (2010). A healthy society results from a community's access to quality healthcare services. The WHO rightly emphasized that the main objective of a healthcare system is to deliver better health services appropriately WHO (2000). However, the performance of a health system in achieving its objectives is measured by the actual health outcomes. The primary health system undertakes several interventions for promotive and preventive care of mother and child, along with curative and referral services Mishra (2001).

The healthcare services have two divisions, as detailed in the seventh schedule of part XI of the Indian Constitution, which deals with dividing powers between the central and

Corresponding Author: Sangeeta Goala
Email: goala.sangeeta@gmail.com

state governments. Some of the health services are under the Concurrent List [1] like All India Institutes of Medical Sciences (AIIMS), controlled by the central government under the jurisdiction of the Ministry of Health and Family Welfare (MoHFW), whereas Government Medical Colleges, Civil Hospitals, Community Health Centers (CHC), Primary Health Centers (PHC) and Sub Centers are under the control of Directorate of Health Services by state government comes under State List. 'The matter of public health belongs to state list and maternity benefits is under concurrent list so a state-wise variation in maternal healthcare is expected' Chakraborty and Bhattacharjee (2017). For example, *Janani Suraksha Yojana* (JSY) is a centrally sponsored scheme but the state governments implement it through civil hospitals, health centers, *etc.* The extent of implementation of such schemes varies depending upon the quality of governance at the state level.

Maternal Health refers to women's health during pregnancy, childbirth, and postpartum. It encompasses the healthcare dimensions of family planning, preconception, prenatal and postnatal care to reduce maternal morbidity and mortality Chakraborty and Bhattacharjee (2017).

With the safe motherhood initiative by the UN in the 1980s, India initiated the Reproductive and Child Health Policies in 1997, followed by the National Population Policy in 2000, the National Health Policy in 2002, and then the National Rural Health Mission (NRHM) in 2005 based on Global health commitments for Millennium Development Goals (MDGs) to enhance access to high-quality healthcare for women Mali (2018). It is estimated that about 21 million women benefited from this scheme between April 2005 to August 2009 Jain (2010). In the last three years, 28.223 million mothers benefitted from JSY, with an expenditure of 46.23 billion. In India, institutional delivery has increased to 78 percent Janani Suraksha Yojana (2017). According to Sunaina (2018) the fifth Millennium Development Goal (MDG) called for lowering the Maternal Mortality Ratio (MMR) by at least three-fourths by 2015, from 437 to 109 per 100,000 live births. The achievement by 2015 was 167, according to the country report for the MDG 2015 MDG (2015).

Recognizing the need for improved maternal health, the government of India came up with different schemes like the cash assistance program *Janani Suraksha Yojana* (JSY) in April 2005 to encourage institutional deliveries by providing cash incentives to pregnant women and *Accredited Social Health Activists* (ASHA) and thus to reduce the MMR, especially among the states with high maternal mortality. The *Janani Shishu Suraksha Karayakaram* (JSSK) provided free medical services such as nutritional supplements, antenatal check-ups, medical transportation, free admission to hospitals, *etc.*, during the period of pregnancy and with limited prenatal and post-natal healthcare through public healthcare institutions. According to Mali (2018) the *Indira Gandhi Matritva Sahyog Yojana* (IGMSY) provided cash to pregnant women to make up for the loss of income they experienced during pregnancy, subject to age and parity requirements. Another program, the *Pradhan Mantri Surakshit Matritva Abhiyan* (PMSMA) provides guaranteed, comprehensive, and high-quality antenatal care at no cost to all pregnant women on the ninth day of every month. PMSMA assures pregnant women a minimum package of antenatal care services in their *second/third* trimesters at assigned government health centers NHP (2016).

[1] A list of activities that both state and central government look after. Public Health is one such activity.

Pandey and Singh (2018) measured the frequency with which women utilize pregnancy and child health services using data from the National Family Health Survey (NFHS-III). Their work serves as an illustration of Andersen's Behavioural Model of healthcare usage. It was discovered that the quintile of home wealth and the mother's educational level were accurate indicators of the use of maternal healthcare services.

Obviously, better maternal healthcare namely, prenatal health, including antenatal check-ups, neo-natal tetanus protection, and pregnancy registration in suitable health centers shall lead to a decrease in maternal mortality. This study's primary goal is to investigate the maternal healthcare condition in different Indian states and Union Territories (UTs) to identify the maternal health services that need immediate attention. This shall help the government in policy-making to achieve uniform national growth.

Since many demographers are continually working on various issues relating to the health sector, there is a wealth of literature concerning maternal health in India. Blum and Fargues (1990) created a mechanism to predict maternal mortality when cause-of-death is insufficient. They provided two strategies: one based upon an extrapolation by smoothing the observed profile of deaths among women, which yields lower estimates, and another on the age-specific mortality ratio of men and women. By processing a few life tables, one can quickly determine the number, age pattern, and trend of maternal mortality regardless of the method. Bhat (2002) derived estimates of maternal mortality for India using the sisterhood technique and a regression method that took into account sex differences in adult mortality and compared those values to the values of the estimates from different sources.

Research on the disparity in maternal healthcare facilities in different regions is undertaken periodically using different approaches, leading to the classification of alternatives are abundant in the literature. Authors like Iyengar and Sudarshan (1982), Ram and Shekhar (2006), Mohanty and Ram (2001) developed different multivariate ranking techniques using various parameters to rank the districts/states of India. These studies mainly focused to calculate a single index and a conclusion is made based on the value of the index. In this study, some methods of computing maternal healthcare are considered, and the authors tried to reach a unique solution that gives the most robust result.

Robustness signifies the insensitivity of a result to minor deviations from the assumptions Huber and Ronchetti (2009). "In a broad informal sense, robust statistics is a body of knowledge, partly formalized into theories of robustness, relating to deviations from idealized assumptions in statistics" Hampel *et al.* (1986). For robust composite indices, minor changes in the values of the participating variables in the index should not change the values in ranking. Robustness analysis is required to limit the possibility of getting meaningless Composite Indicators. This kind of analysis can enhance the final results' accuracy, credibility, and interpretability OECD (2008).

Many studies are carried out by converging values from different relevant parameters into a single index using several different methods of aggregation and weighting Chakraborty and Bhattacharjee (2017), Iyengar and Sudarshan (1982), Mohanty and Ram (2001). Gang *et al.* (2012) ranked the alternatives using various Multi-Criteria Decision Making (MCDM) methods and later used Spearman's rank correlation coefficient to generate the final ranking to resolve the inconsistency. A significant value of Spearman's rank correlation coefficient indicates a good agreement between a given MCDM method with other MCDM methods.

In this study, different weighting techniques are visited to build Composite Indicators using parameters associated with maternal healthcare. This paper distinguishes itself from the previously mentioned papers by employing multiple weighting methods to construct Composite Indicators, rather than creating a single index. (Gang *et al.*, 2012, pp.198) claims that - applying various MCDM methods to a sorting problem is beneficial because the ranking agreed by several methods is more trustful than a single method. However, it is necessary to check the robustness of the various methods used for ranking the regions to identify the most reliable method of ranking.

In this paper, the researchers identified some parameters that influenced maternal healthcare. These parameters were then combined into a single index using different weighting methods, forming different indices. However, it should be noted that no single approach can be superior in all aspects, and the selection of the optimal method depends on its compatibility and robustness. Combining some scattered statistical tools, the researcher's aim is to determine the robust Composite Indicator (MCDM method) from the competing approaches.

2. Objectives of the study

Based on the research gap identified and the issues raised in the above discussion, this paper intends to attain the following objectives :

- Formulate Composite Indicators to measure the extent of maternal healthcare status of the states/UTs of India combining all the maternal healthcare parameters.
- Identifying the most robust composite index amongst the competing methods of weighting.
- Ranking the states/UTs according to maternal health care services and accordingly identifying the state-wise level of maternal healthcare attainments.

3. Data source

The study uses secondary data from 36 Indian states and union territories that can be found in factsheets for the National Family Health Survey (NFHS-4) (`http://rchiips.org/nfhs/factsheet_NFHS-4.shtml`). The following parameters for evaluating maternal healthcare are identified from the aforementioned data source:

- P_1 = Mothers who had an antenatal check-up in the first trimester (%)
- P_2 = Mothers who had at least 4 antenatal care visits (%)
- P_3 = Mothers whose last birth was protected against neonatal tetanus (%)
- P_4 = Mothers who consumed iron folic acid for 100 days or more when they were pregnant (%)
- P_5 = Mothers who had full antenatal care (%)

- P_6 = Registered pregnancies for which the mother received Mother and Child Protection (MCP) card (%)
- P_7 = Mothers who received postnatal care from a doctor/ nurse/ LHV/ ANM/ midwife/ other health personnel within 2 days of delivery (%)
- P_8 = Mothers who received financial assistance under *Janani Suraksha Yojana* (JSY) for births delivered in an institution (%)

4. Different steps of composite indicator building

A Composite Indicator (CI) is a multidimensional concept calculated based on two or more single indicators on the basis of an underlying model. It compares spatial performance and is increasingly recognized as a useful tool in policy analysis and public communication OECD (2008). Defining a composite index is an integral part of MCDM problem which looks into selecting, ranking, and evaluating a finite set of alternatives (in this case states/UTs) Singh and Pant (2021). A brief description of the various steps involved in building a composite index is provided in the subsequent Sub-sections.

4.1. Normalization of parameters

The first step of Composite Indicator (CI) building invites the normalization of the variables. By converting the data to pure, dimensionless numbers, the data collected for the variables under consideration are normalized to bring the indicators to the same standard. Although there are other normalizing methods, in the current work re-scaling approach is utilized which is commonly coined as the max-min approach of indicators. To know in detail about different normalization techniques, one may refer to (OECD, 2008, pp.29–32).

Let, x_{ij} represents the value of the i^{th} state of the j^{th} parameter. The normalized decision matrix y_{ij} is calculated as

$$y_{ij} = \frac{x_{ij} - \min\limits_{i}(x_{ij})}{\max\limits_{i}(x_{ij}) - \min\limits_{i}(x_{ij})}, \quad i = 1, 2, \ldots, n \quad \text{and} \quad j = 1, 2, \ldots, m \tag{1}$$

The normalization technique is fixed throughout the study only the weighing techniques are changed. The weights of the normalized parameters are computed using five different weighting techniques, then the weights and normalized score of a given state are combined as a sum-product (linear aggregation) to attain the value of the composite index of the state. As five weighting techniques are used so for each state five composite index values are obtained, one for each weighing technique.

4.2. Weighting of the indicators

As weights quantify the relative importance of the different factors in the composite index and also control the dominance of the parameters with higher variance, so different popular methods of weighting are identified from the available literature and different sets of weights w_j for each of the weighting methods are computed. The values of the Composite

Indicators are calculated using the method of linear aggregation. A brief description of the various weighting techniques is provided below:

(i) Technique for order preference by similarity to ideal solution

The Technique for Order Preference by Similarity to Ideal Solution (TOPSIS) approach, developed by Hwang and Yoon (1981) is a mechanism for ranking alternatives based on a variety of factors by minimizing the distance to the ideal solution and maximizing the distance to the negative-ideal solution.

Let, x_{ij} represent the value of the i^{th} state of the j^{th} parameter; $L(i, IDR)$ and $L(i, NIDR)$ are two components of an ideal solution and negative-ideal solution, then

$$L(i, IDR) = \sqrt{\frac{\sum_{j=1}^{m} \left(x_{ij} - \max_{i}(x_{ij})\right)^2 w_j}{\sum_{j=1}^{m} x_{ij}^2}} \tag{2}$$

$$L(i, NIDR) = \sqrt{\frac{\sum_{j=1}^{m} \left(x_{ij} - \min_{i}(x_{ij})\right)^2 w_j}{\sum_{j=1}^{m} x_{ij}^2}} \tag{3}$$

The weight (w_j) is calculated using Shannon's entropy Wu *et al.* (2011), Chakraborty and Bhattacharjee (2017)

$$w_j = \frac{1 - \phi_j}{\sum_{j}(1 - \phi_j)}; \quad 0 < \phi_j < 1 \quad \text{and} \quad \sum_{j=1}^{m} w_j = 1 \tag{4}$$

The entropy of the j^{th} parameter is given by

$$\phi_j = -\sum_{i} \frac{p_{ij} ln(p_{ij})}{ln(n)} \tag{5}$$

$$\text{where,} \quad p_{ij} = \frac{d_{ij}}{\sum_{j} d_{ij}} \quad \text{and} \quad n = \text{total no of state/UT;}$$

$$d_{ij} = \frac{x_{ij}}{\max_{j}(x_{ij})} \quad \text{in case of positive indicators}$$

$$\text{and} \quad d_{ij} = \frac{x_{ij}}{\min_{j}(x_{ij})} \quad \text{in case of negative indicators}$$

Composite Indicator (CI_{TOP}) for the $TOPSIS$ method is given by:

$$CI_{TOP} = \frac{L(i, IDR)}{L(i, IDR) + L(i, NIDR)} \tag{6}$$

(ii) Iyengar-Sudarshan method

Iyengar and Sudarshan (1982) proposed a weighting technique where the weights vary inversely proportional to the variation in the respective variables. Here the weights act as variance stabilizers of the participating parameters.

Let, y_{ij} represents the normalized value of the i^{th} state of the j^{th} parameter, and w_j represents the weights of the j^{th} parameter, then,

$$w_j = \frac{k}{\sqrt{\underset{i}{var}(y_{ij})}} \quad \text{with} \quad \sum_j w_j = 1 \quad \text{and} \quad 0 < w_j < 1 \tag{7}$$

$$\text{and} \quad k = \left[\sum_{j=1}^{m} \frac{1}{\sqrt{var(y_{ij})}}\right] \tag{8}$$

These weights stabilize the variance of the normalized parameters and prevent any one of the variables from dominating the composite index. The choice of the weights in this manner would ensure that large variation in any one of the indicators would not unduly dominate the contribution of the rest of the indicators and distort the inter-state comparisons Bhattacharjee and Wang (2011).

Composite Indicator (CI_{IS}) for the $Iyengar - Sudarshan$ method is given by:

$$CI_{IS} = \sum_{j=1}^{m} w_j y_{ij} \tag{9}$$

(iii) Principal component analysis

The eigenvalues indicate the proportion of each variable's variance that can be explained by the primary component. The Eigenvalues of the parameters for maternal healthcare can be obtained using the *Principal Component Analysis* (PCA) method.

The term *Principal Component Analysis* (PCA) refers to a technique that employs complex mathematical principles to reduce a large number of variables that could be associated with one another into a smaller set. It rotates the data point cluster to highlight the maximum variance. Additionally, because the input variables are grouped in a particular way using the Principal component analysis, the least important variables can be eliminated while the most useful ones can be retained.

$$w_j = \frac{\text{Individual Eigen values}}{\text{Sum of all Eigen values}} \tag{10}$$

Composite Indicator (CI_{PCA}) for the *Principal Component Analysis* method is given by:

$$CI_{PCA} = \sum_{j=1}^{m} w_j y_{ij} \tag{11}$$

(iv) Data envelopment analysis

Data Envelopment Analysis (DEA) is a mathematical programming technique presented by Charnes *et al.* (1978). Its application has been focused mainly on efficiency assessment. An efficiency frontier that could be used as a benchmark to compare countries' relative performance is estimated using linear programming tools by DEA. This requires the development of a benchmark (the frontier) and the estimation of the distance between nations in a multi-faceted system OECD (2008).

The weighted composite index for the i^{th} state is given by,

$$CI_{DEA} = \frac{\sum_{j=1}^{m} w_j y_{ij}}{\sum_{j=1}^{m} w_j} \tag{12}$$

The weights are to be selected in such a way that CI is *maximized* for the i^{th} state. Thus, the objective function is

$$\text{M}aximize \quad \text{C}I_{DEA} = \frac{\sum_{j=1}^{m} w_j y_{ij}}{\sum_{j=1}^{m} w_j} \tag{13}$$

$$\text{Constrained by the following relations} \quad \sum_{j=1}^{m} w_j = 1 \quad ; \quad a < w_j < b \quad \forall j$$

The values of a and b are fixed for a particular problem and it depends on the value of the number of parameters (m) in the composite index.

(v) Ordered weighted average

Yager (1988) introduced the concept of *Ordered Weighted Average* (OWA). The main objective of the technique is to determine the weights of the different components participating in the formation of the composite index.

The weighted composite index for the i^{th} state is given by,

$$CI_{OWA} = \sum_{j=1}^{m} w_j y_{ij} \tag{14}$$

where y_{ij} is the i^{th} largest observation of the normalized matrix and w_j are the corresponding weights with the ordered values of the component y_{ij} such that

$$w_j > 0, \quad \sum_{j=1}^{m} w_j = 1, \forall j$$

Accordingly, some OWA operators are defined as the entropy function explaining the dispersion in the weights,

$$Disp(w_j) = -\sum_{j=1}^{m} w_j ln(w_j) \tag{15}$$

and some OWA operators are called the orness and are defined as,

$$\alpha = Orness(w_j) = \frac{1}{n-1}\sum_{j=1}^{m}(n-j)w_j \tag{16}$$

The OWA weights are determined using the linear programming method minimax disparity rule proposed by Wang and Parkan (2005). The objective function is

$$\text{M}inimize \quad \delta \tag{17}$$

$$\text{such that} \quad \sum_{j=1}^{m}\Big(\frac{n-j}{n-1}\Big)w_j = \alpha, \quad \text{where} \quad \alpha \in [0,1], w_j > 0 \tag{18}$$

$$\sum_{j=1}^{m} w_j = 1, \forall j \quad \text{and} \quad |w_j - w_{j+1}| \leqslant \delta, \quad j = 1, 2, \ldots, j-1 \tag{19}$$

The ultimate value of the composite index shall lie between the highest and the lowest value of the participating components in the formation of the index.

5. Process of testing robustness of composite indicators

As such, no fixed method is available to check the robustness of composite indices. With the help of the available resources from the literature, combining some scattered statistical tools an algorithm is developed to identify the robust Composite Indicator/MCDM method from a set of competing approaches. The algorithm is provided in Table A.1 of the paper. However, before introducing the method one needs to know about some statistical tests detailed below:

The process of measuring the robustness of the ranks of the same set of subjects (alternatives) obtained from different processes discussed here is improvised notation-wise over the method explained by Saisana *et al.* (2005) and Saltelli *et al.* (2008). Here, R_{ij} denotes the rank of the i^{th} state/UT (alternative) obtained from the j^{th} method. Let there be n states/UTs and m competing methods.

5.1. Inter-rater agreement of subjective judgment

The method proposed by Tinsley and Weiss (1975) looks into the agreement in ranks of a common group of subjects provided by different raters. The method can be hired and designed into the current setup to compare the agreement of the rating (in this case the ranking) of states/UTs obtained from different weighting methods. From the ranks of the n competing states and m methods S and S_1 are computed, where,

$$S = mn\frac{(n^2-1)}{12}, \quad S_1 = \frac{1}{m}\sum_{i=1}^{n}\Big[(\sum_{j=1}^{m}R_{ij}) - (\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{m}R_{ij})\Big]^2 \text{and} \quad S_2 = S - S_1$$

The corresponding test statistic as defined by Tinsley and Weiss (1975) is

$$F = \frac{S_1/n-1}{S_2/nm-n} \sim F_(n-1, nm-n) \tag{20}$$

It can be applied to test the null-hypothesis of independence amongst the raters *i.e.*, the different methods in this case related to the ranking of states are independent of each other. The agreement between the raters takes place in case the null hypothesis is rejected.

5.2. Rank correlation matrix

The Rank correlation coefficient measures the degree of similarity between two rank sets of the same groups of subjects. A high value of the Rank correlation coefficient implies a better agreement between two rank sets and vice versa. The Rank correlation coefficient (R_{ij}) is defined as

$$R_{ij} = 1 - \frac{6 \sum d_i jk^2}{n(n-1)} \tag{21}$$

where n is the number of alternatives (states/UTs) and d_{ijk} is the difference between ranks of k^{th} state in the i^{th} and j^{th} weighting technique.

The correlation between the various methods of ranking can be checked with the help of the rank correlation matrix. It gives the pairwise comparison of rank correlation with every method.

5.3. Distance between ranks across different methods

The aggregate absolute difference in rank of the i^{th} state across all the different methods is given by,

$$\sum_{j'=1, j' \neq j}^{m} \left| R_{ij} - R_{ij'} \right| \tag{22}$$

gives the sum of the absolute difference of the rank of the i^{th} state obtained through the j^{th} method with the rank of the same i^{th} state obtained from all the other $(j')^{th}$ methods *i.e.*, $j' = 1, 2, \ldots, j-1, j+1, \ldots, m$. More precisely, in (22) j is fixed but. j' is varying from 1 to m (assuming there are m methods) but, $j \neq j'$

Subsequently, the average $\bar{R}_j$, aggregating the difference of ranks of all the states across all the competing methods is defined as

$$\bar{R}_j = \frac{1}{n(m-1)} \sum_{i=1}^{n} \sum_{j'=1, j' \neq j}^{m} \left| R_{ij} - R_{ij'} \right| \tag{23}$$

The ideal situation when all the techniques are equally robust is that the value of $\bar{R}_j$ should be very close to $0 \, \forall \, j$ and accordingly the $Avg(\bar{R}_j)$, with average taken over all the different methods shall be close to 0 too. Accordingly, the one-sided Studentized t is proposed for testing the null-hypothesis that H_0 : Avg $\bar{R}_j = 0$ against the alternative hypothesis H_1 : Avg $\bar{R}_j > 0$, taking the values of $\bar{R}_j$ as $j = 1, 2, \ldots, m$ as the values of the test variable. Accepting the null hypothesis takes us to the conclusion that the methods are equally robust.

Another single measure for overall robustness is defined in OECD (2008) as,

$$\bar{R} = \sum_{j=1}^{m} \bar{R}_j = \frac{1}{nm(m-1)} \sum_{j=1}^{n} \sum_{i=1}^{n} \sum_{j'=1, j' \neq j}^{m} \left| R_{ij} - R_{ij'} \right| \tag{24}$$

High values of $\bar{R}$ indicate the need for robustness check or sensitivity analysis of the competing methods (OECD, 2008, pp.117–118). However, the existing body of literature is yet to define any test for the statistic $\bar{R}$ to be utilized to test the hypothesis H_0: $\bar{R} = 0$ against the alternative $H_1 : \bar{R} > 0$.

5.4. Kendall Tau distance

The Kendall Tau rank distance Kendall (1938) is a statistic originally used to compute the dissimilarity between two rank sets and can be extended to find the robust composite index out of the competing composite indices. Considering two rank-sets of the same set of subjects (alternatives) say τ_1 and τ_2, the distance is defined as

$$\begin{aligned} K(\tau_1, \tau_2) = & \left| \{(i, i') : i < i', [\tau_1(i) < \tau_1(i') \wedge \tau_2(i) > \tau_2(i')] \right. \\ & \left. \vee [\tau_1(i) > \tau_1(i') \wedge \tau_2(i) < \tau_2(i')] \} \right| \end{aligned} \tag{25}$$

where, $\tau_1(i)$ and $\tau_2(i')$ are the ranking of the i^{th} and $(i')^{th}$ subject in the rank set τ_1 and τ_2 respectively. The expression in (25) is summarized as

$$K(\tau_1, \tau_2) = \sum_{\substack{\{i,i'\} \in P \\ i < i'}} \hat{K}_{i,i'}(\tau_1, \tau_2) \tag{26}$$

where, P is the set of unordered pairs of all the distant subjects in the rank set τ_1 and τ_2 respectively, and

$\hat{K}_{i,i'}(\tau_1, \tau_2) = 0$ if the i^{th} and $(i')^{th}$ subject are in the same order in both the rank set τ_1 and τ_2
$= 1$, otherwise

Thus, the statistic $K(\tau_1, \tau_2)$ is a measure of the distance between the rank set τ_1 and τ_2 for the same set of subjects. Accordingly, the statistic for an aggregate distance of a rank set τ_j and τ_j with all other rank set $\tau_{j'}$ is defined as

$$K_j = \sum_{\substack{j'=1 \\ j' \neq j}}^{m} K(\tau_j, \tau_{j'}), \;\; j = 1, 2, \ldots, m \tag{27}$$

The rank set τ_j shall be considered as the most robust set of ranks of the subjects (alternatives) over any other rank set $\tau_{j'}$ if

$$K_j < K_{j'} \;\; \forall \;\; j' \;\; (\text{but} \neq j) \tag{28}$$

Eventually, the j^{th} method of ranking stands out to be the most robust technique of ranking the subjects given the dataset.

6. Analysis and result

As described in Section 4.1. the normalized values of the parameters for the different states/UTs are computed. The normalized values of the parameters are used to calculate the composite index with the use of five weighting techniques *viz*; TOPSIS, Iyengar-Sudarshan, Principal Component Analysis, Data Envelopment Analysis, and Ordered Weighted Average. The aggregation of the normalized score was done using the linear aggregation method. The normalization and aggregation method remains the same throughout the study only the weighting techniques varied. Based on the methods discussed in Section 4.2. above, 5 sets of composite indices along with their ranks are obtained and are given in Table A.2. From Table A.2, it can be seen that Kerala and Lakshadweep are the two states that are on the top list of ranking for all the methods. However, for other states, it can be seen that there is heterogeneity of ranking especially for ranks obtained by the TOPSIS method which has significant dissimilarity in ranking in comparison to the other four methods. Moving down the table one can find a lack of consensus on ranks among the different methods. This shows the relevance of the current study. Different method generates different ranking so one needs to check for the robustness of different competing approaches to reach a unique set of ranks.

The next step is to use the proposed algorithm (*c.f.* Table A.1) to check the robustness of the methods. All the steps of the algorithm are implemented in R-software.

Initially, the test proposed by Tinsley and Weiss (1975) is applied to look for consistency in the rankings obtained from several methods under

Null hypothesis H_0: There is independence in ranks obtained from different methods.

Alternative hypothesis H_1: the ranking methods are in agreement with each other.

After performing the test in R-software we have found that $Cal.F(42.3184) > Crit.F(1.5050)$ at 5% level of significance. Accordingly, the null hypothesis is rejected and it is concluded that there is an agreement between the ranking methods.

As agreement could be found between the rankings one can skip steps III and IV of the algorithm (*c.f.* Table A.1) and directly can move towards step V to check:

Null hypothesis H_0: All the ranking methods are equally robust.

Alternative hypothesis H_1: All the ranking methods are not equally robust.

From Table 1 we find that the p-value is $0.0024 < 0.05$, so, we shall conclude that all the methods are not equally robust. Hence, the next task is to determine the most robust ranking method out of the methods used (*c.f.* Table A.1). Kendall's distance measures the pairwise disagreement between two rankings. The lesser the distance, the more robust the method is. A detailed discussion on the same is available in Section 5.4.

Table 1: Result for robustness check

$\overline{R}_1$	$\overline{R}_2$	$\overline{R}_3$	$\overline{R}_4$	$\overline{R}_5$	$\overline{R}$	t-value	p-value
4.6389	2.2917	4.0278	2.1389	2	3.0389	5.646494	0.0024

Table 2 confirms that method *Ordered Weighted Average* (OWA) has the least score (190), *i.e.*, the lowest distance with all other methods. Hence, the composite index based on the OWA method of weighting is considered to be the most robust method for the said data. Accordingly, the composite index values obtained from OWA method are shown in Figure 1.

Table 2: Kendall's distance measurement

Sl no	Methods	Distance between rankings (S)	Rank
1	TOPSIS	422	5
2	I-S	215	3
3	PCA	398	4
4	DEA	199	2
5	OWA	190	1

*one with minimum score is ranked 1

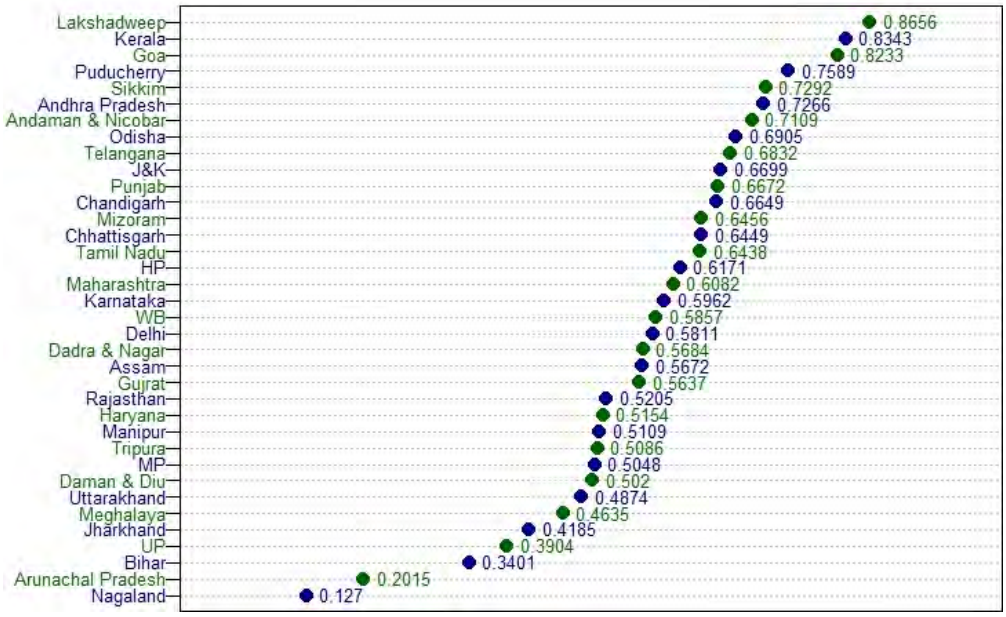


Figure 1: CI values of all the states/UTs using OWA method

7. Discussion and conclusion

Using several healthcare indicators, the article aimed to compare the situation of maternal healthcare in India's various states and UTs using the MCDM technique. In order to determine the most reliable weighting approach for merging multiple healthcare indicators, the study suggests a methodology incorporating a number of dispersed statistical methods. The states and UTs were rated in terms of their maternal healthcare facilities using the *Ordered Weighted Average* (OWA) technique, as this weighting tool posted itself as the most robust of the various weighting techniques. The states and UTs that do poorly in terms

of maternal healthcare are identified, and pertinent discussion concerning those states is conducted.

According to the aforementioned survey, the top five performing states/UTs are Lakshadweep, Kerala, Goa, Puducherry, and Sikkim, whereas the bottom five states or UTs are Jharkhand, Uttar Pradesh, Bihar, Arunachal Pradesh, and Nagaland.

As per census 2011 data, it has been observed that the female literacy rate of all the top 5 states/UTs, namely Puducherry (81.2), Goa (81.8), Lakshadweep (88.3), Kerala (92), is very high (above 81 percent) except Sikkim (76.4 percent). Similarly, for the bottom five states, Jharkhand (56.2), Uttar Pradesh (59.3), Bihar (53.3), and Arunachal Pradesh (59.7), the female literacy rate is below 60 percent; the exception lies with Nagaland, where the female literacy is (76.7) percent which is same as Sikkim Ministry of Statistics and Program Implementation (2017). Thus, female literacy might be one of the major contributors to better maternal healthcare status as literature has evidence of the positive and significant influence of mother's schooling on maternal care utilization Govindasamy and Ramesh (1997). Nagaland ranks lowest in maternal healthcare despite having the same literacy rate as that of Sikkim. This might be due to Nagaland's remote location, restricted access to healthcare, and lack of proper medical facilities.

Among the top-ranked states/ UTs, Puducherry has a compact geographical area with a high literacy rate and better health facilities both in the public and private sectors. As per Kayaroganam *et al.* (2016) 78.6 percent of the mother avail full ANC. Paul and Chouhan (2020) suggested that Maternal mortality can be prevented by regular ANC visits, supervised deliveries, and postnatal care (PNC). According to Census 2011, in Puducherry, 87 percent of mothers consult an Obstetrician and prefer delivery at government hospitals, while 56 percent have good knowledge of nutrition during pregnancy. Additionally, 99 percent of mothers have adequate knowledge of breastfeeding and its benefits Ramaiah *et al.* (2022).

Goa, a coastal state with mountains on its western border, covers an area of 3702 square kilometers and has a population of 1.46 million. The state excels in protecting mothers against neonatal tetanus, with a rate of 96.2 percent surpassing the national average of 89 percent. Iron deficiency anemia among mothers is a major threat to safe motherhood and to the health and survival of infants, but Goa has achieved a consumption rate of 67.4 percent for Iron and Folic Acid (IFA), surpassing the national average of 30.3 percent Dehury *et al.* (2017).

Kerala, a state in South India is home to more than 33 million people "with females enjoying higher status compared to other states" Gupta and Mani (2022). This state is known for its remarkable achievements in education and health Mukherjee (2010). As per the report of the Department of Health, Govt of Kerala, the state has 1280 numbers of modern medicine Institutions including Hospitals, Community Health Centers (CHC), Public Health Center (PHC) *etc.*, with 38004 numbers of Beds with a population bed ratio of 879 per person Saritha (2018). As per the 2011 census, 92 percent of the women of Kerala are literate. The government's efforts to improve healthcare have positively impacted the state's socioeconomic development. Easy access to adequate medical facilities and the availability of such facilities to its stakeholders are significant contributors to maternal health issues. Access to quality obstetric care is a priority to prevent complications during and after delivery.

Sikkim is a mountainous state, with *Kangchenjunga* having the highest peak in India. It is one of the states of North-east India with a population of 6 million. Numerous studies concur that women's empowerment improves access to health and wellness. 95.3 percent of the women in Sikkim participate in household decision-making. Proper nutrition is crucial for women's health; inadequate nutrition can lead to anemia and health issues Dehury *et al.* (2017).

As derived, Nagaland has one of the worst maternal healthcare, preceded by Arunachal Pradesh, Bihar, Uttar Pradesh, and Jharkhand. Nagaland's maternal care and child immunization indicators are significantly below the national average Chakraborty and Bhattacharjee (2017). Geographic isolation, limited healthcare access, high medical costs, and inadequate health facilities contribute to the challenges faced by women in Nagaland. They often prefer home births assisted by traditional midwives to reduce expenses. A lack of encouragement to participate in seminars and awareness programs on maternal healthcare is observed, particularly in rural areas. Traditional midwives support childbirth and reduce family costs Humtso and Soundari (2019).

Arunachal Pradesh, the neighboring state of Nagaland, has the second lowest position in maternal healthcare status. Singh *et al.* (2009) points out that approximately 50 percent of women in Arunachal Pradesh do not make any prenatal visits. Prenatal visits are crucial for recognizing pregnancy complications and knowing when to seek emergency obstetric care, reducing the risk of maternal death.

The neighboring states Bihar, Uttar Pradesh, and Jharkhand located in eastern India, rank among the bottom five states. Under utilization of professional assistance during delivery may contribute to the poor conditions of maternal healthcare. Singh *et al.* (2009) stated that nearly 75 percent of women still give birth without any medical assistance in Uttar Pradesh and Bihar. Despite the fact that 56 percent of women are aware of the requirement for three ANC check-ups, 49 percent do not adhere to it because they are unaware of the hazards associated with pregnancy without these check-ups. Many people think that unless there are difficulties, a normal pregnancy doesn't need three ANC checks. Only 22 percent of pregnant women are advised to have a minimum of three ANC visits Khan *et al.* (2014).

Despite its abundant resources, Jharkhand faces issues in maternal healthcare, with high maternal mortality and low utilization of prenatal and secure delivery services IIPS (2010). In 2009, the maternal mortality rate in Jharkhand was 261 per 100,000 live births, higher than the national average of 212 Ogala *et al.* (2012). As per the guidelines developed by the Ministry of Health and Family Welfare (2010) and WHO (2006), complete ANC is one of the key factors of maternal healthcare utilization. Only 9 percent of women in Jharkhand used complete ANC services during 2007-2008, compared to 18.8 percent nationally. Socioeconomic disparities, caste, and media exposure influence the utilization of ANC services Kavitha and Audinarayan (1997), Pandey *et al.* (2004). Complete ANC services were provided to approximately 19 percent from other social groups, compared to 7 percent of SC and 6 percent of ST married women. All ANC services were utilized by 10 percent of Hindus and 27 percent of urban women Singh and Chaturvedi (2015). Women exposed to mass media were 65 percent more likely to use all ANC services Gupta *et al.* (2016). The coal mine industry in Jharkhand also contributes to the problem of an irregular visit to

health centers with women not receiving paid leave and facing occupational hazards Dubey (2016), Bhanumathi (2002). It is essential to address these issues for improved maternity healthcare.

Governments must prioritize maternal healthcare as an investment in society. India needs a clear healthcare vision, emphasizing immunization, maternity care, primary health centers, committed doctors, and support staff. Public awareness campaigns can reduce disparities in ANC utilization. Spreading knowledge about prenatal screening and highlighting government health programs are essential.

Although the study is exploratory and indicative, it provides comparative information on the status of Indian states/UTs in terms of maternal health care. Policymakers should focus on lower-ranked states/UTs as they shall show a higher convergence rate. To identify such spatial black spots, further research is needed at the district level in such low-performing states. Various normalization and aggregation techniques can benefit the quantification of maternal healthcare. Incorporating other demographic and socioeconomic variables can generate an advanced composite score. Examining further the disadvantaged districts may offer focused insights and potential solutions. Timely ANC check-ups, preparedness for delivery, postnatal care, and family planning are crucial for improving maternal health Khan *et al.* (2014). This approach of identifying a reliable MCDM method from among the many choices available can also be applied to several other MCDM-related activities in other knowledge domains.

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

Bhanumathi, K. (2002). The status of women affected by mining in India. In *Tunnel Vision: Women, Mining and Communities*.

Bhat, P. N. (2002). Maternal mortality in India: an update. *Studies in Family Planning*, **33**, 227–236.

Bhattacharjee, D. and Wang, J. (2011). Assessment of facility deprivation in the households of the north eastern states of India. *Demography India*, **40**, 35–54.

Blum, A. and Fargues, P. (1990). Rapid estimations of maternal mortality in countries with defective data: an application to Bamako (1974–85) and other developing countries. *Population Studies*, **44**, 155–171.

Chakraborty, A. and Bhattacharjee, D. (2017). Maternal health: which states are more caring? *Demography India*, **46**, 82–96.

Charnes, A., Cooper, W., and Rhodes, E. (1978). Measuring the efficiency of decision making units. *European Journal of Operational Research*, **2**, 429–444.

Dehury, R. K., Samal, J., Desouza, N. V., and Dehury, P. (2017). Status of women's health in Goa and Sikkim: a comparative analysis of state fact sheets of NFHS-3 and 4. *International Journal of Medicine and Public Health*, **7**, 196–202.

Dubey, K. (2016). *Socio Economic Impact Study of Mining and Mining Polices on the Livelihoods of Local Population in the Vindhyan Region of Uttar Pradesh*. Technical Report, NITI Aayog.

Gang, K., Lu, Y., Peng, Y., and Shi, Y. (2012). Evaluation of classification algorithms using MCDM and rank correlation. *International Journal of Information Technology & Decision Making*, **11**, 197–225.

Govindasamy, P. and Ramesh, B. M. (1997). *Maternal Education and the Utilization of Maternal and Child Health Services in India*. Technical Report, International Institute for Population Sciences and Macro International Inc.

Gupta, A., Kumar, P., and Dorcas, O. A. (2016). Decomposing the socio-economic inequalities in utilization of full antenatal care in Jharkhand state, India. *International Journal of Population Studies*, **2**, 92–106.

Gupta, A. and Mani, S. S. (2022). Assessing mortality registration in Kerala: the MARANAM study. *Genus*, **78**.

Hampel, F., Rousseeuw, P., and Stahel, W. (1986). *Robust Statistics: The Approach Based on Influence Functions*. John Wiley & Sons, New York.

Huber, P. J. and Ronchetti, E. M. (2009). *Robust Statistics*. Wiley, Hoboken, 2nd edition.

Humtso, M. Y. and Soundari, M. H. (2019). Maternal health care practices of Lotha Naga tribal women in India. *PEOPLE: International Journal of Social Sciences*, **5**, 738–755.

Hwang, C. and Yoon, K. (1981). *Multiple Attribute Decision Making Methods and Applications*. Springer.

IIPS (2010). *District Level Household and Facility Survey (DLHS-3), 2007-08: Jharkhand*. Technical Report, International Institute for Population Sciences.

Iyengar, N. and Sudarshan, P. (1982). A method of classifying regions from multivariate data. *Economic and Political Weekly*, **17**, 2047–2052.

Jain, A. K. (2010). Janani suraksha yojana and the maternal mortality ratio. *Economic and Political Weekly*, **45**, 15–16.

Janani Suraksha Yojana (2017). National health portal of India. Retrieved August 7, 2017, from `https://www.nhp.gov.in/janani-suraksha-yojana-jsy-_pg`.

Kavitha, N. and Audinarayan, N. (1997). Utilisation and determinates of selected maternal and child health care in rural areas of Tamil Nadu. *Health and Population: Perspectives and Issues*, **20**, 112–125.

Kayaroganam, R., Saya, G. K., and Kar, S. S. (2016). Utilization of maternal health services among janani suraksha yojana beneficiaries in Puducherry, India. *International Journal of Advanced Medical and Health Research*, **3**, 73–77.

Kendall, M. (1938). A new measure of rank correlation. *Biometrika*, **30**.

Khan, M. E., Agarwal, P. K., Hazra, A., Dixit, A., Bhatnagar, I., Ahmad, J., and Ahmad, D. (2014). *Maternal and Newborn Health Behaviors in Rural Uttar Pradesh: Findings from Learning Phase Baseline Survey 2013*. Population Council, New Delhi.

Mali, N. V. (2018). A comparative assessment of maternal health and maternal health policies in India and the US: need to transition from a biomedical model to a biopsychosocial model for maternal health policies. *Journal of Health and Human Services Administration*, **40**, 462–498.

MDG (2015). *Millennium Development Goals (MDG) India Country Report 2015*. Technical Report, MoSPI.

Ministry of Statistics and Program Implementation (2017). *Women & Men in India-2017*. Technical Report, MoSPI.

Mishra, R. K. (2001). *Utilisation of Maternal Child Health and Family Planning Services Through Primary Health Centres in Rural Rajasthan*. Ph.D. Thesis, Jawaharlal Nehru University.

Mohanty, S. and Ram, F. (2001). *District at a Glance: India*. Mimeograph, IIPS, Mumbai.

Mukherjee, S. (2010). State of health and health care in West Bengal: a study in comparative perspective with Kerala and Tamil Nadu. *Demography India*, **39**, 259–276.

NHP (2016). Pradhan mantri surakshit matritva abhiyan. Retrieved October 13, 2022, from `https://pmsma.nhp.gov.in/about-scheme/`.

OECD (2008). *Handbook on Constructing Composite Indicators: Methodology and User Guide*. OECD, Paris.

Ogala, V., Avan, B., and Roy, I. (2012). Women's Level of Satisfaction with Maternal Health Services in Jharkhand. Technical Report, United States Agency International Development. Retrieved from `http://pdf.usaid.gov/pdf_docs/PA00JDGP.pdf`.

Pandey, A., Roy, N., Sahu, D., and Acharya, R. (2004). Maternal health care services observations from Chhattisgarh, Jharkhand and Uttaranchal. *Economic and Political Weekly*, **39**, 713–720.

Pandey, K. and Singh, R. (2018). Accessing maternal and child health care services utilization: an application of andersen's behaviour model. *Elixir International Journal*, **118**, 50646–50655.

Paul, P. and Chouhan, P. (2020). Socio-demographic factors influencing utilization of maternal health care services in India. *Clinical Epidemiology and Global Health*, **8**, 666–670.

Ram, F. and Shekhar, C. (2006). *Ranking and Mapping of Districts Based on Socio-economic and Demographic Indicators*. Technical Report, International Institute of Population Sciences, Mumbai.

Ramaiah, C. K., Raju, S., Chennupati, D., and Prakash, G. S. (2022). Maternal healthcare information requirements of first-time mothers in Puducherry: a survey.

Saisana, M., Saltelli, A., and Tarantola, S. (2005). Uncertainty and sensitivity analysis techniques as tools for the quality assessment of composite indicators. *Journal of the Royal Statistical Society Series A: Statistics in Society*, **168**, 307–323.

Saltelli, A., Ratto, M., Andres, T., Campolongo, F., Cariboni, J., Gatelli, D., Saisana, M., and Tarantola, S. (2008). *Global Sensitivity Analysis: The Primer*. John Wiley & Sons.

Saritha, R. (2018). *Health at a Glance 2018*. Technical Report, Department of Health Services. Kerala: Directorate of Health Services. Retrieved January 07, 2022, from `https://dhs.kerala.gov.in/wp-content/uploads/2020/03/health_25022019.pdf`.

Singh, K. J. and Chaturvedi, H. K. (2015). Factors influencing the skilled birth attendant utilization of non-institutional delivery in empowered action group states, India. *American International Journal of Research in Humanities, Arts and Social Sciences*, **12**, 54–62.

Singh, M. and Pant, M. (2021). A review of selected weighing methods in MCDM with a case study. *International Journal of Systems Assurance Engineering and Management*, **12**, 126–144.

Singh, S., Ramez, L., Ram, U., Moore, A. M., and Audam, S. (2009). Barriers to Safe Motherhood in India. Technical Report, IIPS.

Sunaina, P. (2018). *Inequalities in Maternal Health Care in Kerala*. Ph.D. Thesis, Mahatma Gandhi University.

Swain, A. and Mohanty, B. (2010). Socio-demographic disparities in Orissa-maternal and child health and welfare perspectives. *Demography India*, **39**, 129–131.

Tinsley, H. E. and Weiss, D. J. (1975). Inter-rater reliability and agreement of subjective judgments. *Journal of Counseling Psychology*, **22**, 358–376.

Wang, Y.-M. and Parkan, C. (2005). A minimax disparity approach for obtaining OWA operator weights. *Information Sciences*, **175**, 20–29.

WHO (2000). World health report 2000 health systems: improving performance. Retrieved January 07, 2022, from `https://apps.who.int/iris/bitstream/handle/10665/42281/WHR_2000-eng.pdf?sequence=1&isAllowed=y`.

Wu, J., Sun, J., Liang, L., and Zha, Y. (2011). Determination of weights for ultimate cross efficiency using shannon entropy. *Expert Systems with Applications*, **38**, 5162–5165.

Yager, R. R. (1988). On ordered weighted averaging aggregation operators in multicriteria decision making. *IEEE Transactions on Systems, Man, and Cybernetics*, **18**, 183–190.

Appendix A

Table A.1: Algorithm for robustness check

Step I	Compute rank of all the subjects (alternatives) based on the composite index developed from all the methods i.e., rank of the n-subjects for the m-different methods.
Step II	Perform test for Inter-rater agreement of subjective judgment (different methods in this exercise) as defined in Tinsley and Weiss (1975). Null hypothesis: H_0: There is independence in ranks (of subjects) obtained from different methods tested against the alternative H_1: The ranking methods are in agreement to each other If H_0 is rejected go to Step V else go to step III.
Step III	Rank correlation Matrix taking all the methods in pairs are computed
Step IV	Look for insignificant rank correlations if any. Identify the method(s) which is (are) insignificant from the other methods and drop it from further analysis. Repeat step II else GO TO Step V
Step V	Compute $\bar{R}_j$ (Eqn. 23) for each of the methods ($j = 1, 2, \ldots, m$). Test if Avg ($\bar{R}_j$) is significantly greater than 0. If Avg ($\bar{R}_j$) is significantly close to zero conclude that "All the methods are equally robust"- GOTO Step VIII ELSE – "All methods are not equally robust and there is a need of Robustness study"- GOTO Step VI
Step VI	Compute Kendall Tau distance of a method (j, say) with all the other methods (except j) and add the distances. Call it K_j Compute K_j for all the methods ($j = 1, 2, \ldots, m$). The method with a minimum value of K_j is the most robust technique.

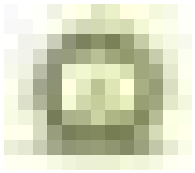
Table A.2: Rank & CI values of states/UTs of maternal healthcare parameters

State/UTs	Methods				
	TOPSIS	**I-S**	**PCA**	**DEA**	**OWA**
Andaman & Nicobar	13 (0.5200)	7 (0.7277)	10 (0.7178)	7 (0.7179)	7 (0.7109)
Andhra Pradesh	11 (0.5410)	6 (0.7439)	4 (0.8158)	5 (0.7315)	6 (0.7266)
Arunachal Pradesh	36 (0.1836)	35 (0.2381)	35 (0.1505)	35 (0.2089)	35 (0.2015)
Assam	12 (0.5264)	23 (0.5857)	28 (0.469)	21 (0.576)	22 (0.5672)
Bihar	32 (0.3582)	34 (0.3578)	34 (0.1955)	34 (0.352)	34 (0.3401)
Chandigarh	20 (0.4566)	11 (0.6908)	14 (0.6513)	11 (0.6686)	12 (0.6649)
Chhattisgarh	10 (0.5607)	15 (0.6574)	15 (0.6473)	13 (0.6526)	14 (0.6449)
Daman & Diu	25 (0.4237)	27 (0.5257)	18 (0.6178)	29 (0.5088)	29 (0.502)
Delhi	21 (0.4428)	20 (0.5983)	20 (0.6034)	20 (0.5863)	20 (0.5811)
Dadra & Nagar Haveli	26 (0.4148)	21 (0.5902)	19 (0.6039)	22 (0.5752)	21 (0.5684)
Goa	7 (0.5828)	3 (0.8373)	3 (0.8811)	3 (0.8280)	3 (0.8233)
Gujarat	28 (0.4032)	22 (0.5859)	12 (0.6774)	23 (0.5703)	23 (0.5637)
Haryana	34 (0.3256)	24 (0.5454)	25 (0.5465)	25 (0.5219)	25 (0.5154)
Himachal Pradesh	16 (0.4678)	16 (0.6411)	13 (0.6579)	16 (0.6218)	16 (0.6171)
Jammu & Kashmir	8 (0.5719)	12 (0.6797)	8 (0.7225)	10 (0.6746)	10 (0.6699)
Jharkhand	33 (0.336)	32 (0.442)	32 (0.3954)	32 (0.4282)	32 (0.4185)
Karnataka	18 (0.466)	18 (0.6136)	17 (0.6208)	18 (0.6011)	18 (0.5962)
Kerala	2 (0.6353)	2 (0.8383)	1 (0.9744)	2 (0.8373)	2 (0.8343)
Lakshadweep	1 (0.6453)	1 (0.8789)	2 (0.9404)	1 (0.8669)	1 (0.8656)
Madhya Pradesh	19 (0.4596)	28 (0.5256)	31 (0.4198)	28 (0.5147)	28 (0.5048)
Maharashtra	23 (0.4262)	17 (0.6294)	16 (0.6439)	17 (0.609)	17 (0.6082)

Table A.2: Continued

State/UTs	Method				
	TOPSIS	**I-S**	**PCA**	**DEA**	**OWA**
Manipur	15 (0.4728)	31 (0.4866)	11 (0.6982)	26 (0.5179)	26 (0.5109)
Meghalaya	29 (0.3916)	30 (0.4895)	29 (0.4259)	31 (0.4624)	31 (0.4635)
Mizoram	4 (0.6189)	13 (0.6634)	22 (0.5931)	15 (0.6485)	13 (0.6456)
Nagaland	35 (0.2122)	36 (0.151)	36 (0.0122)	36 (0.1333)	36 (0.127)
Odisha	6 (0.6034)	8 (0.7034)	21 (0.6032)	8 (0.6959)	8 (0.6905)
Punjab	17 (0.4667)	10 (0.6914)	9 (0.7194)	12 (0.6682)	11 (0.6672)
Puducherry	3 (0.6215)	4 (0.7758)	6 (0.7918)	4 (0.7644)	4 (0.7589)
Rajasthan	22 (0.4359)	25 (0.5442)	25 (0.5076)	24 (0.5251)	24 (0.5205)
Sikkim	9 (0.5644)	5 (0.7487)	7 (0.7624)	6 (0.7291)	5 (0.7292)
Tamil Nadu	5 (0.6103)	14 (0.6609)	24 (0.5866)	14 (0.6508)	15 (0.6438)
Telangana	14 (0.5074)	9 (0.7015)	5 (0.7948)	9 (0.6882)	9 (0.6832)
Tripura	31 (0.3582)	26 (0.5259)	23 (0.5923)	27 (0.5165)	27 (0.5086)
Uttar Pradesh	30 (0.3605)	33 (0.4084)	33 (0.313)	33 (0.3999)	33 (0.3904)
Uttarakhand	27 (0.4082)	29 (0.5129)	30 (0.4206)	30 (0.497)	30 (0.4874)
West Bengal	24 (0.4257)	19 (0.6058)	26 (0.5434)	19 (0.5912)	19 (0.5857)

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 22, No. 1, 2024 (New Series), pp 193–217
http://www.ssca.org.in/journal



A Rank-based Test of Independence of Covariate and Error in Nonparametric Regression with Missing Completely at Random Response Situation

Sthitadhi Das and Saran Ishika Maiti
Department of Statistics
Visva Bharati, Santiniketan, India

Received: 17 January 2023; Revised: 18 July 2023; Accepted: 22 July 2023

Abstract

In the context of nonparametric regression, statistical relationship between the covariate and the random error is a matter of interest. For a traditional nonparametric regression model $Y = g(X) + \epsilon$ where Y is the response, X the covariate, ϵ the random error and $g(\cdot)$ a suitably chosen smooth function, null hypothesis may be framed as the independence of X and ϵ against all possible alternatives citing dependence between them. It may be of further concern, whether for an incomplete data set with several missing observations, such rank based testing of independence can be performed. For example, some observations on Y are unreported whereas the covariate X has complete data. On this structure of missingness completely at random (MCAR) situation, process of rank based testing on independence between X and ϵ may be thought of. This article delineates such testing techniques, based on Kendall's τ or Bergsma's (2014) τ^* and Blum *et al.* (1961) distance based test statistics, in order to develop consistent test procedures against a sequence of contiguous alternatives. The asymptotic powers of these test statistics are further studied through the finite sample simulation study, choosing different levels of missingness percentage. Finally, a real data analysis presents a comparative testimony of those proposed test statistics.

Key words: Asymptotic power; Contiguous alternative; Distance covariance; Kendall's τ; Missing completely at random; Nonparametric regression model; Local linear smoothing.

AMS Subject Classifications: 62G08, 62G30

1. Introduction

For a quite substantial period of time in statistics literature, missing data context continues to be a live topic. The impact of missing data on quantitative research can be serious, heading to biased estimates of parameters, loss of information, increased standard errors and debilitated the generalizability of findings. Usually, most statistical processes are designed for complete data. In the presence of missing values, failing to edit the incomplete

Corresponding Author: Saran Ishika Maiti
Email: saranishika.maiti@visva-bharati.ac.in

data into "complete" one can turn the data statistically unsuitable. Particularly, statistical inference process experiences a huge toll in presence of missingness. Thus, as a default approach, one may delete those missing observations before going to conduct the necessary analysis using statistical methods. Most inevitable drawback of such listwise deletion is that a large fraction of sample might get trimmed causing severe loss to statistical power. Some articles by Anderson(1957), Wilks(1932), Afifi and Elashoff(1966), Hartley and Hocking(1971) discussed the problem of listwise deletion where each value of data set is equally likely to be missing.

In regression set up, missing scenario mostly occurs in response variable Y where some of the observations in Y are not available. The chance mechanism of this missingness may be independent of X and Y or may depend fully on the covariate, X. The first case is termed as missing completely at random (MCAR) while the second type of missingness is missing at random (MAR) (Little and Rubin (2014)). Mathematically speaking, in regression set-up, missingness can be interpreted via a triplet (X_i, Y_i, δ_i) for a set of n observations on (X, Y). At a given point X_i, the response Y_i is either observed or missing. The indicator variable δ takes the value 1 or 0 according as the value of Y is reported or not. Clearly for MCAR, $\text{Prob}[\delta = 1/X, Y] = p$ (a constant) while for MAR $Prob[\delta = 1/X, Y] = P[\delta = 1/X] = p(X)$ (a function of X). We shall proceed with an MCAR data to test the association in the context of nonparametric regression further.

Suppose in nonparametric regression model $Y = g(X) + \epsilon$ with g being the unknown regression function and ϵ the error, missingness at random occurs in Y. Instead of complete deletion of those unavailable (X, Y) observations, imputation techniques may be used where substitutes for missing values are looked for. In contrast to imputing certain global estimates such as mean/median of available Y figures, it may be worthwhile to opt for some other imputation alternatives based on nonparametric regression estimation, like local linear smoothing, kernel density estimation *etc.* (Chung *et al.* (1993), Cheng (1994)), thereafter examining the impact of missingness on their performances. One may note that downside of imputation technique is to produce underestimates of standard errors, which leads in turn to inflated test statistics.

In nonparametric regression, a fundamental assumption is homoscedasticity, *i.e.* $E(\epsilon^2/X = x) = \sigma^2 > 0$. However even for homoscedastic model, inference based on unknown regression function $g(x)$ may be unconvincing, for instance in isotonic mean/median regression model, confidence interval for the regression function at a given point will be wrong even if the homoscedasticity holds. In such cases, it is safer to assume the independence between X and ϵ. This issue of checking the independence against all possible alternatives, has been addressed in the literature by Einmahl *et al.*(2008), Neumeyer(2009), Hlavka *et al.* (2011), Dhar *et al.*(2018). Most of the test statistics proposed are distance based except the rank based test statistic by Bergsma (2014), followed by Dhar *et al.* (2018), Das *et al.*(2022) where the test statistic is constructed on the sign function of second/third order differences of neighbouring quadruplet of responses.

The present article is evolved on the adoption of such rank based test statistic to investigate the independence of ϵ and X in nonprametric regression when the data has MCAR in Y. At the first stage, the missing places are imputed by the regression estimator through Nadaraya- Watson estimation and local linear smoothing technique respectively. Thereafter,

filling those unregistered Y values we try to form rank based test statistic following the roadmap by Bergsma (2014). We also investigate the asymptotic theory of those test statistics under null and contiguous alternative (Lehmann and Romano, 2005).

The rest of the article is organized as follows. Section 2 describes original regression model and the transformed imputed model. Section 3 provides the methodologies to estimate the regression function $g(.)$ using various estimation techniques. In section 4, test statistics are constructed based on the newly obtained bivariate observations X and Y. The asymptotic local powers of the test statistics under contiguous alternatives are computed in Section 5. Section 6 includes a real data study. A precise conclusion is presented in section 6. Appendix 1 contains derivation of technical details while appendix 2 contains numerical results of asymptotic power study.

2. Regression setting

Let the nonparametric regression model to be considered as $Y = g(X) + \epsilon$. Consider the following imcomplete data: (X_i, Y_i, δ_i), $i = 1, 2 \cdots, n$ where $\delta = 1$ if Y_i is observed otherwise $\delta_0 = 0$ if Y_i is missing. Also, $Prob(\delta = 1/X, Y) = Prob(\delta = 1/X) = p\,(0 < p < 1)$ where p being a fixed constant, *i.e.*, missingness is MCAR type. Let there be k bivariate observations assuming missingness on Y and the remaining $(n - k)$ pairs are complete. Suppose (X_i', Y_i') denote the i-th complete observation of (X, Y), $i = 1, 2, \cdots, (n - k)$. A nonparametric sub-model can be formulated on these **complete pairs** as

$$Y' = g_1(X') + \epsilon' \tag{1}$$

with the assumptions on error ϵ' similar to the assumptions, already drawn on error ϵ of the original model, as $E(\epsilon'|X' = x') = 0\ \forall\ x'$ and $E({\epsilon'}^2|X' = x') = \sigma^2(x')$ where $\sigma^2(x') > 0$. The regression function $g_1(\cdot)$ is the **first step regression function.** Its nonparametric estimator may be treated as a naive alternative against the estimator of $g(X)$ in the original model. After deducing the estimator of $g_1(\cdot)$ as $\hat{g}_1(\cdot)$, the missing observations on Y will be filled up by $\hat{g}_1(\cdot)$ at the values of the covariate X corresponding to the missing responses. These fillers are known as **imputed responses**. Thus, by imputing the missing values of Y, the complete data set (X^*, Y^*) of size n can be re-framed as follows.

$$Y_i^* = \begin{cases} Y_i' & \text{when } \delta = 1 \\ \hat{g}_1(X_i), & \text{when } \ \delta = 0\,;\ i = 1, 2, \cdots, n \end{cases}$$

Then, the following regression model is proposed on the hence completed bivariate data (X^*, Y^*).

$$Y^* = g_2(X^*) + \epsilon^* \tag{2}$$

where X^* being the covariate and ϵ^* being the error. Finally, $g_2(X^*)$ is estimated using the conventional methods like Nadaraya-Watson (NW) estimation and local linear smoothing method respectively.

3. Estimation of regression functions

3.1. Estimation using Nadaraya-Watson method

The first step regression function $g_1(.)$ in (1) can be estimated using Nadaraya Watson (NW) estimation process at $X' = x'$ as

$$\hat{g}_1(x') = \frac{\sum_{i=1}^{n} k\left(\frac{X_i' - x'}{h}\right) Y_i'}{\sum_{i=1}^{n} k\left(\frac{X_i' - x'}{h}\right)} \tag{3}$$

where $k(\cdot)$ is the kernel density function and h is the bandwidth satisfying $h \to 0$ with $nh \to \infty$ where $n \to \infty$. A variety of kernel functions are possible to be chosen but for practical and theoretical considerations we choose a very common one, Epanichnikov kernel $k(u)$, where $k(u) = .75(1-u^2).I(|u| \leq 1)$. This parabolic shape kernel enjoys some optimality properties.

The second stage estimator of the regression function $g_2(X^*)$ in (4) is also deduced in a similar manner.

$$\hat{g}_2(x^*) = \frac{\sum_{i=1}^{n} k\left(\frac{X_i^* - x^*}{h}\right) Y_i^*}{\sum_{i=1}^{n} k\left(\frac{X_i^* - x^*}{h}\right)} \tag{4}$$

Further, proposition of some test statistics are made.

3.2. Estimation using local linear smoothing (LLS)

In addressing the same issue, another alternative approach against NW estimation can be the technique of local linear smoothing (Chu *et al.*, 1995). This method begins with the minimization of the local weighted least squares based on all bivariate observations, *i.e.* minimization of the following expression.

$$\sum_{i=1}^{n} [Y_i - r_0 - r_1(x - X_i)]^2 \, k\left(\frac{x - X_i}{h}\right) \delta_i \tag{5}$$

As per the notation stated in section 2, specifically for non missing pairs of observations (X', Y') the above expression of minimization can be re-framed as minimization of

$$\sum_{i=1}^{n-k} [Y_i' - r_0 - r_1(x' - X_i')]^2 \, k\left(\frac{x' - X_i'}{h}\right) \tag{6}$$

The minimization yields the solutions of the constants r_0 and r_1. (5) gives

$$\hat{r}_0 = \frac{\sum_{i=1}^{n} (M_2 - (x - X_i)M_1) \; k\left(\frac{x - X_i}{h}\right) \delta_i \, Y_i}{\sum_{i=1}^{n} [M_2 - (x - X_i)M_1] \; k\left(\frac{x - X_i}{h}\right) \delta_i} \tag{7}$$

where $M_l = \sum_{i=1}^{n}(x - X_i)^l \; k\left(\frac{x - X_i}{h}\right)\delta_i, \; l = 1, 2$. Clearly, for non-missing pairs of observations (X', Y') (6) would be reshaped as

$$\hat{r}_0 = \frac{\sum_{i=1}^{n-k} [M_2' - (x' - X_i')M_1'] \; k\left(\frac{x' - X_i'}{h}\right) Y_i}{\sum_{i=1}^{n-k} [M_2' - (x' - X_i')M_1'] \; k\left(\frac{x' - X_i'}{h}\right)} \tag{8}$$

where $M_l' = \sum_{i=1}^{n-k}(x' - X_i')^l \; k\left(\frac{x' - X_i'}{h}\right), \; l = 1, 2$. The least square estimate $\hat{r}_1$ of r_1 can be deduced in a similar way from (5) or (6) which is simply

$$\hat{r}_1 = \frac{\sum_{i=1}^{n} (x' - X_i') \, k\left(\frac{x - X_i}{h}\right) \delta_i \, Y_i - \hat{r}_0 M_1'}{M_2'}.$$

Next, by the first order Taylor's expansion, $g(X_i)$ can be expanded in the neighbourhood of x as

$$g(X_i) = \; g(x) - (x - X_i)g^{(1)}(x) \tag{9}$$

where $g^{(1)}(x)$ is the first order derivative of $g(x)$. Hence the response Y_i can be approximated as $\{g(x) - (x - X_i)g^{(1)}(x) + \epsilon_i\}$, $i = 1, \ldots, n$. Synonymously, under non missing set up Y_i' may be approximated as $\{g(x') - (x' - X_i')g^{(1)}(x') + \epsilon_i'\}$, $i = 1, \ldots, n$. Then substituting Y_i' in (3), we obtain

$$\begin{aligned} \hat{g}_1(x') &= \frac{\sum_{i=1}^{n} k\left(\frac{X_i' - x'}{h}\right) \{\hat{r}_0 + \hat{r}_1 \, (x' - X_i')\}}{\sum_{i=1}^{n} k\left(\frac{X_i' - x'}{h}\right)} \\ &= \hat{r}_0 - h\hat{r}_1 \frac{\sum_{i=1}^{n} \left(\frac{X_i' - x'}{h}\right) k\left(\frac{X_i' - x'}{h}\right)}{\sum_{i=1}^{n} k\left(\frac{X_i' - x'}{h}\right)} \end{aligned}$$

which approaches to $\hat{r}_0$ mentioned in (7) for relatively small bandwidth h such that $h \to 0$. Noticeably, the estimator $\hat{r}_1$ is not of use when $h \to 0$. Denote $\beta_i' = M_2' - (x' - X_i')M_1' \; \forall$ $i = 1, \ldots, n$. Then the estimate of $g_1(x)$ will be

$$\hat{g}_1(x') = \frac{\sum_{i=1}^{n} \beta_i' Y_i'}{\sum_{i=1}^{n} \beta_i'} \tag{10}$$

or a slightly modified estimator $\hat{g}_1(x') = \dfrac{\sum_{i=1}^{n} \beta_i' Y_i'}{\sum_{i=1}^{n} \beta_i' + n^{-2}}$ where n^{-2} is added to the denominator to avoid the situation of $\sum_{i=1}^{n} \beta_i' \approx 0$. This $\hat{g}_1(x')$ is called *simplified local linear smoother* (SLLS) of $g_1(x')$.

As we mentioned in the introduction, deletion of incomplete pairs may cause loss of information in data analysis. Hence the technique of refilling the missing observations or imputation would be thought of. $\hat{g}_1(x')$ can be treated as the imputed estimator for those k missing responses at the values of corresponding X. Subsequently, the estimator $\hat{g}_2(\cdot)$ is to be derived on the basis of complete bivariate observations (X, Y), denoted as (X^*, Y^*) after the imputation process.

Thus, in this concocted data $X^* = X$ and $Y_i^* = \delta_i\, Y_i' + (1 - \delta_i)\, \hat{g}_1(X_i')$.

Minimizing $\sum_{i=1}^{n} [Y_i^* - s_0 - s_1(x^* - X_i^*)]^2\, k\left(\dfrac{x^* - X_i^*}{h}\right)$ with respect to the linear constants s_0 and s_1 following the same arguments already proposed in (5) and (6),

$$\hat{s}_0 = \frac{\sum_{i=1}^{n} (M_2^* - (x^* - X_i^*)M_1^*)\; k\left(\dfrac{x^* - X_i^*}{h}\right) \delta_i\, Y_i^*}{\sum_{i=1}^{n} (M_2^* - (x^* - X_i^*)M_1^*)\; k\left(\dfrac{x^* - X_i^*}{h}\right)} \tag{11}$$

where

$$M_l^* = \sum_{i=1}^{n} (x^* - X_i^*)^l\; k\left(\frac{x^* - X_i^*}{h}\right),\; l = 1, 2.$$

and $\hat{s}_1$ be the solution of s_1.

Ultimately, using the same logic as projected in (10), the final estimator $\hat{g}_2(\cdot)$ at $X^* = x^*$ is derived as

$$\hat{g}_2(x^*) = \frac{\sum_{i=1}^{n} \beta_i^* Y_i^*}{\sum_{i=1}^{n} \beta_i^*} \tag{12}$$

where $\beta_i^* = M_2^* - (x^* - X_i^*)M_1^*\ \forall\ i = 1, \ldots, n$. Alternatively, (11) can be written as $\hat{g}_2(x^*) = \dfrac{\sum_{i=1}^{n} \beta_i^* Y_i^*}{\sum_{i=1}^{n} \beta_i^* + n^{-2}}$ in order to avoid the possibility of the inflation of $\hat{g}_2(x^*)$. This estimator $\hat{g}_2(\cdot)$ is called the *imputed local linear smoother* (**ILLS**) of $g(x)$.

4. Relevant test statistics

In order to test $H_0 : X^* \perp\!\!\!\perp \epsilon^*$ ($\perp\!\!\!\perp$ means independence) we consider a sequence of contiguous alternatives, say H_n, that converges to H_0 as $n \to \infty$. In this case, the sequence of contiguous alternative H_n, indicating to the dependence between X^* and ϵ^*, has the following expression

$$H_n : \ F_{n;X^*,\epsilon^*}(x^*, e^*) = (1 - \frac{\gamma}{\sqrt{n}})\, G_{X^*}(x^*)\, H_{\epsilon^*}(e^*) + \frac{\gamma}{\sqrt{n}}\, K_{X^*,\epsilon^*}(x^*, e^*) \tag{13}$$

where $F_{n;X^*,\epsilon^*}(\cdot,\cdot)$ denote the joint CDF of (X^*, ϵ^*) under H_n while, $H_{\epsilon^*}(\cdot)$ and $G_{X^*}(\cdot)$ are the marginal CDFs of ϵ^* and X^* respectively and $K_{X^*,\epsilon^*}(\cdot,\cdot)$ is the proper joint distribution function of (X^*, ϵ^*). $\gamma > 0$ is the mixing constant for $F_0(\cdot,\cdot)$ and $K_{X^*,\epsilon^*}(\cdot,\cdot)$ where $F_0(x^*, e^*) = G_{X^*}(x^*)\, H_{\epsilon^*}(e^*)$ is the joint CDF of (X^*, ϵ^*) under H_0. First we generate a bivariate sample $\{(x_1^*, e_1^*), \ldots, (x_n^*, e_n^*)\}$ of size n from $F_0(x^*, e^*)$ under H_0. Then, using the regression model $Y^* = g(X^*) + \epsilon^*$ we obtain the bivariate observations $(x_1^*, y_1^*), \ldots, (x_n^*, y_n^*)$ can be from the joint distribution function of (X^*, Y^*). Taking the ordered observations on X^* as $x_{(1)}^*, \ldots, x_{(n)}^*$ and then the corresponding Y^*-values as $y_{(1)}^*, \ldots, y_{(n)}^*$ ($y_{(i)}^*$'s termed as *induced ordered statistics*), we achieve the ordered set $\{(x_{(1)}^*, y_{(1)}^*), \ldots, (x_{(n)}^*, y_{(n)}^*)\}$. The related errors are $\epsilon_{(1)}^*, \ldots, \epsilon_{(n)}^*$ which are also viewed as the induced ordered values of $\epsilon_1^*, \ldots, \epsilon_n^*$. The second order differences of these induced ordered observations $y_{(i)}^*$'s, $i = 1, \ldots, n$ are defined as $y_{(i)}^{*(2)} := y_{(i+1)}^* - 2y_{(i)}^* + y_{(i-1)}^*$ with the marginal considerations as $y_{(0)}^* = y_{(1)}^*$, $y_{(n+1)}^* = y_{(n)}^*$, resulting two threshold figures as $y_{(1)}^{*(2)} = y_{(2)}^* - y_{(1)}^*$ and $y_{(n)}^{*(2)} = y_{(n-1)}^* - y_{(n)}^*$. Based on the these bivariate observation $(x_{(i)}^*, y_{(i)}^{*(2)})$s for $i = 1, \ldots, n$, the following test statistics (Dhar *et al.* (2018)) are proposed as

$$T_{n,1} = \frac{1}{\binom{n}{2}} \sum_{1 \le i < j \le n} sign\{(x_{(i)}^* - x_{(j)}^*)(y_{(i)}^{*(2)} - y_{(j)}^{*(2)})\} \tag{14}$$

$$T_{n,2} = \frac{1}{\binom{n}{4}} \sum_{1 \le i < j \le n} a(x_{(i)}^*, x_{(j)}^*, x_{(k)}^*, x_{(l)}^*) a(y_{(i)}^{*(2)}, y_{(j)}^{*(2)}, y_{(k)}^{*(2)}, y_{(l)}^{*(2)}) \tag{15}$$

$$T_{n,3} = \frac{1}{\binom{n}{4}} \sum_{1 \le i < j \le n} \frac{1}{4} h(x_{(i)}^*, x_{(j)}^*, x_{(k)}^*, x_{(l)}^*) h(y_{(i)}^{*(2)}, y_{(j)}^{*(2)}, y_{(k)}^{*(2)}, y_{(l)}^{*(2)}) \tag{16}$$

where $sign(t) = \frac{t}{|t|}$ if $t \neq 0$ or 0 otherwise, $h(p, q, r, s) = \{|p - q| + |r - s| - |p - r| - |q - s|\}$; $p, q, r, s \in \mathbb{R}$ and $a(p, q, r, s) = sign\{|p-q|+|r-s|-|p-r|-|q-s|\}$. (14) is the sample version of Kendall's tau statistics between X^* and $Y^{*(2)}$ while (15) is the sample statistic in favour to τ^* which is an extended version of Kendall's tau by Bergsma *et al.* (2014). In contrast, (16) is the sample counterpart of the distance based measure D introduced by Blum-Kiefer-Rosenblatt (1961).

For the sake of readers' interest the population versions of the aforementioned test statistics for unordered observations on (X^*, Y^*), $i = 1, 2, 3$ are too presented herewith.

$$T_1 = E[sign(X_1^* - X_2^*)(Y_1^* - Y_3^*)]$$

$$T_2 = E[a(X_1^*, X_2^*, X_3^*, X_4^*) a(Y_1^*, Y_2^*, Y_3^*, Y_4^*)]$$

$$T_3 = E[\frac{1}{4}h(X_1^*, X_2^*, X_3^*, X_4^*)h(Y_1^*, Y_2^*, Y_3^*, Y_4^*)].$$

To check $H_0 : X^* \perp\!\!\!\perp \epsilon^*$ is analogous of checking $H_0 : X^* \perp\!\!\!\perp f(\epsilon^*)$ for any proper function $f(\cdot)$. Let us assume the form of the function as $f(\epsilon^*) = \epsilon^{*(2)} = \epsilon_{(i+1)} - 2\epsilon_{(i)} + \epsilon_{(i-1)}$, $i = 1, \ldots, n$, the second order difference of ϵ^*. Thus modified H_0 is $H_0 : X^* \perp\!\!\!\perp \epsilon^{*(2)}$. Since ϵ_i's are unobservable, so is $\epsilon^{*(2)}$. Thus instead of $\epsilon^{*(2)}$ we may judiciously approximate it by $Y^{*(2)}$ provided the function $g(.)$ is sufficiently smooth. Thus H_0 can further be modified to $H_0 : X^* \perp\!\!\!\perp Y^{*(2)}$. Evidently, independence of X^* and ϵ^* implies and implied by $T_k = 0$ for $k = 1, 2, 3$. So, $H_0 : X^* \perp\!\!\!\perp Y^{*(2)}$ implies $T_k = 0$, $k = 1, 2, 3$ and vice versa. Therefore their sample representatives, *viz.*, $T_{n,k}$ for $k = 1, 2, 3$ would be regarded as the desired test statistics to carry out the test of independence.

To kick-start the test process it is reasonable to approximate $T_{n,k}((x_{(1)}^*, e_{(1)}^{*(2)}), \ldots, (x_{(4)}^*, e_{(4)}^{*(2)}))$ by $T_{n,k}((x_{(1)}^*, y_{(1)}^{*(2)}), \ldots, (x_{(4)}^*, y_{(4)}^{*(2)}))$ for $k = 1, 2, 3$, as due to smoothness of $g(.)$, $y^{*(2)}$ would enable to sweep out the effect of g for large n. In fact, any function sorting out the effect of $g(\cdot)$ can be chosen instead of $y_{(i)}^{*(2)}$. For instance, the test statistic based on first order differences of Y^* may be applicable also for testing homoscedasticity of errors against all possible alternatives, which coincides with any traditional nonparametric test of homoscedasticity [see the discussion in Einmahl *et al.*, 2008]. Under H_0 the critical regions can be determined by the test statistics $T_{n,i}$'s $(i = 1, 2, 3)$ as $\omega_{n,i} : T_{n,i} > c_{\alpha,i}$, $i = 1, 2, 3$, where $\alpha \in (0, 1)$ is the level of significance satisfying $P_{H_0}[T_{n,i} > c_{\alpha,i}] = \alpha$ and $c_{\alpha,i}$ is the α-th critical point of the limiting distribution of $T_{n,i}$ under H_0. To study the statistical powers of all $T_{n,i}$'s under H_n for different values of γ, we have to ascertain their limiting distributions.

5. Study on asymptotic powers of the test statistics

It can be shown that the proposed test statistics $T_{n,1}$, $T_{n,2}$ and $T_{n,3}$ are all degenerate U-statistics. In order to study their asymptotic powers we would use various asymptotic properties such as consistency, efficiency, limiting law related to degenerate U statistic. Hence, the order of degeneracy of $T_{n,i}$ for each $i = 1, 2, 3$ is derived hereafter so that their asymptotic distributions under H_0 and H_n can be established.

5.1. Contiguity

For two arbitrary sequences of probability measures, say P_n and Q_n, the definition of contiguity of P_n and Q_n on the sequence of measurable spaces $(\chi_n, \mathcal{A}_n)$ is stated from Le Cam (1960a).

Definition 1: For an arbitrary sequence of events $A_n \in \mathcal{A}_n$, if $P_n(A_n) \longrightarrow 0 \implies Q_n(A_n) \longrightarrow 0$ for sufficiently large sample size n, then Q_n is concluded as contiguous with respect to P_n. It is symbolically expressed as $P_n \triangleleft Q_n$.

To detect whether $P_n \triangleleft Q_n$ holds, the theory of local asymptotic normality (LAN) needs to be expounded. Le Cam's first lemma describes the asymptotic Gaussian nature of the quantity $\log \frac{dQ_n}{dP_n}$ under the probability measure P_n (p.253, Hajek *et al.*, 1999)

Lemma 1: Let $l_n = \frac{dQ_n}{dP_n}$ be a sequence of likelihood ratios corresponding to P_n and Q_n.

Define G_n to be the sequence of distribution functions of l_n. Furthermore, G_n converges to another distribution function G such that

$$\int_0^\infty v\, dG(v) = 1.$$

Then, $P_n \triangleleft Q_n$.

Corollary 1 below delves out an useful consequence of Lemma 1 .

Corollary 0.1: $\log l_n \overset{P_n}{\sim} N(-\frac{1}{2}\theta, \theta)$ implies that Q_n is contiguous with respect to P_n.

The proof of Corollary 1 can be derived using Lemma 1 (for details see Van Der Vaart (2002)). To derive the asymptotic distributions of $T_{n,1}$, $T_{n,2}$ and $T_{n,3}$ using Le Cam's first lemma under contiguous alternatives H_n we assume

Assumption 1: $f_{X^*,\epsilon^*}(x^*, e^*) > 0$ for all x^* and e^*, where f_{X^*,ϵ^*} is the joint PDF of (X^*, ϵ^*).

Assumption 2: $E_{F_{X^*,\epsilon^*}}(\frac{k_{X^*,\epsilon^*}(x^*,e^*)}{f_{X^*,\epsilon^*}(x^*,e^*)} - 1)^2 < \infty$ where $k_{X^*,\epsilon^*}(\cdot,\cdot)$ is the joint proper PDF of (X^*, ϵ^*).

Theorem 1: Under Assumption 1 and Assumption 2, H_n is a sequence of contiguous alternatives.

The formal proof of Theorem 1 is provided in Appendix 1. Next, we explore out the limiting laws of an U-statistic with certain order of degeneracy so that limiting distributions of $T_{n,i}$'s under both hypotheses can be intuited further.

Definition 2: (U statistic) Suppose $\psi(z_1, \ldots, z_m)$ be a real-valued measurable function. Based on a sample $\{Z_1, \ldots, Z_n\}$ from $F_Z(\cdot) \in \mathcal{F}$, $m \leq n$, a U-statistic with kernel ψ is defined as

$$U_n \equiv U_n(\psi) \;=\; \frac{1}{\binom{n}{m}} \sum_{1 \leq i_1 < \ldots < i_m \leq n} \psi(Z_{i_1}, \ldots, Z_{i_m}). \tag{17}$$

U_n is an unbiased estimator of population parameter θ. Also, U_n attains the minimum variance among all other unbiased estimators of θ.
Let us define a sequence of functions related to ψ. For $c = 0, 1, \cdots, m$, let

$\psi_c(z_1, \ldots, z_c) = E[\psi(z_1, \ldots, z_c, Z_{c+1}, \ldots, Z_m)]$ where $X_{c+1}, \cdots, X_n$ are i.i.d. Clearly, $E\psi_c(z_1, \ldots, z_c) = \theta$.

Denote, $\psi_c^*(z_1, \ldots, z_c) = \psi_c(z_1, \ldots, z_c) - E[\psi_c(z_1, \ldots, z_c)]$ and $\xi_c = var[\psi_c^*(z_1, \ldots, z_c)]$, $0 \leq c \leq m$.

Under this notation, the degeneracy of U statistic of order m is defined as follows.

Definition 3: (Order of degeneracy) The order of degeneracy of a U statistic is p if $\xi_0 = \ldots = \xi_p = 0$ and $\xi_{p+1} > 0$.

Here p is the order of degeneracy for the associated kernel $\psi(.)$ and the corresponding U-statistic U_n as well. Some useful theorems, provided by Lee (1990), are pertinent in the context of variance of U_n.

Theorem 2: (i) $\psi_c(z_1, \ldots, z_c) = E[\psi_d(z_1, \ldots, z_c, Z_{c+1}, \ldots, Z_d)]$ for $1 \leq c < d \leq m$.

(ii) $E[\psi_c(Z_1, \ldots, Z_c)] = E[\psi(Z_1, \ldots, Z_m)]$.

Theorem 3: $\xi_c = cov(\psi(N_1), \psi(N_2))$ with N_1, N_2 being the subsets of $\mathcal{C}_{m,n}$, $c = 1, \ldots, m$ each with m number of elements.

Theorem 4: The variance of U_n based on kernel ψ of degree m is

$$Var(U_n) = \binom{n}{m}^{-1} \sum_{c=1}^{m} \binom{m}{c} \binom{n-m}{m-c} \xi_c \tag{18}$$

The asymptotic distribution of $\sqrt{n}(U_n - \theta)$ for large n is normal with mean 0 and variance $m^2\,\xi_1$ (Serfling, 1980) . Unfortunately, in degenerate situation the asymptotic distribution of U_n is no longer normally distributed. Also, it can be explained that $\sqrt{n}(U_n - \theta)$ does not converge to a random variable with degenerate distribution function. If the kernel ψ possesses order of degeneracy p, then the asymptotic distribution of $n^{\frac{d+1}{2}}(U_n - \theta)$ converges to a nonnormal distribution as n increases. The following theorem from Serfling (1980) unveils on the pattern of distribution when $p = 1$(*i.e.* order of degeneracy 1).

Theorem 5: Let $\tilde{\psi}_2(z_1, z_2) = E[\psi(Z_1, Z_2, Z_3, \ldots, Z_m)|Z_1 = z_1, Z_2 = z_2]$, and $\xi_2 = Var[\tilde{\psi}_2(z_1, z_2)]$. If $\xi_1 = 0 < \xi_2$ and $E[\psi^2(Z_1, \ldots, Z_m)] < \infty$, then for some real constants λ_1, λ_2, ... and *iid* $N(0, 1)$ random variables Γ_1, Γ_2, ...,

$$n(U_n - \theta) \xrightarrow{L} Y \tag{19}$$

where $Y \sim \binom{m}{2} \sum_{i=1}^{\infty} \lambda_i(\Gamma_i^2 - 1)$, $m \geq 2$.

The asymptotic non-Gaussian distribution of degenerate U-statistic may also be explicated through obtaining the variance of a symmetric and positive definite quadratic kernel $W(Z_1, Z_2)$ with order of degeneracy 1 where Z_1, Z_2 are i.i.d. random variables. The kernel $W(Z_1, Z_2)$ can be expanded as

$$W(z_1, z_2) = \sum_{k=1}^{\infty} \lambda_k \phi_k(z_1) \phi_k(z_2)$$

where λ_k's are the eigenvalues with corresponding eigenfunctions $\phi_k(z)$'s satisfying

$$\int_{-\infty}^{\infty} W(z, Z_2) \phi_k(Z_2) d\, Z_2 = \lambda_k \phi_k(z).$$

In contiguous set up, the distribution of degenerate U statistic can be deduced (Gregory, 1977). Let $Q_{n,1}$ be the sequence of probability measures with $Q_n = Q_{n,1} \times \ldots \times Q_{n,1}$ (n times). P_0 is the probability measure under H_0 with $P_n = P_0 \times \ldots \times P_0$ (n times). Further suppose, Q_n is contiguous with respect to P_n. Then, the following theorem asserts the limiting distribution of an U-statistic T_n under the probability measure Q_n.

Theorem 6: (Gregory, 1977) Suppose the Radon-Nikodym derivative $dQ_{n,1}/dP_0 = 1 + n^{-\frac{1}{2}}\, h_n$ holds for some sequence $\{h_n\}$ in $L_2(\chi, \mathcal{A})$ that converges to $h \in L_2$. Then, for an U-statistic T_n with order of degeneracy 1,

$$\lim_{n\to\infty} Q_{n,1}\{T_n \le x\} = P\left(\sum_{k=1}^{\infty} \lambda_k\{(\Gamma_k + a_k)^2 - 1\} \le x\right) \tag{20}$$

where $a_k = \int h\, \phi_k\, dP_0$ and $\Gamma_1, \Gamma_2, \ldots$ are iid $N(0,1)$ random variables.

The asymptotic distributions for $T_{n,2}$ and $T_{n,3}$ under H_0 and H_n are easily obtainable using Theorem 6.

Generally speaking, let us define an operator E on $L_2(\chi, \mathcal{A})$ for $\tilde{\psi}_2(z_1, z_2)$ associated with the kernel ψ as

$$E\, g(z) = \int_{-\infty}^{\infty} \tilde{\psi}_2(z, y) g(y) d(F(y)),\ z \in \mathbb{R},\ g \in L_2 \tag{21}$$

and corresponding to E the eigenvalues $\lambda_1, \lambda_2, \ldots$ satisfy $E\, g = \lambda\, g$. Hence one can conclude that $\tilde{\psi}_2(z_1, z_2) = \sum_{k=1}^{\infty} \lambda_k g_k(z_1) g_k(z_2)$ with being orthonormal sequence g_k's satisfying $E[g_k(Z_1) g_l(Z_2)] = 1$ if $k = l$ and 0 if $k \neq l$. Here g_k's are the eigenfunctions corresponding to λ_k's of the transformation

$$E[\tilde{\psi}_2(z, Z_1) g_k(Z_1)] = \lambda_k\, g_k(z) \tag{22}$$

and in L_2,

$$\sum_{k=1}^{n} \lambda_k g_k(Z_1) g_k(Z_2) \xrightarrow{q.m.} \tilde{\psi}_2(Z_1, Z_2). \tag{23}$$

5.2. Limiting distributions of $T_{n,1}$, $T_{n,2}$ and $T_{n,3}$

These test statistics are constructed by the spacings function formed from the distribution function of X^* *i.e.* $G_{X^*}(\cdot)$. Regarding consistency of the test statistics under H_0, we prefer to mention below an important result related to the expectation of an ordered uniform spacing due to Bairamov *et al.* (2010).

Result 1: For $r \ge 1$ and $n \to \infty$,

$$E(V_{(n+2-r)}) \sim \frac{\log n}{n} \longrightarrow 0 \tag{24}$$

where $V_{(s)}$ is the s^{th} order statistic among $\{V_{(1)}, \ldots, V_{(n)}\}$ based on the uniform spacings $V_i = U_{(i)} - U_{(i-1)}$'s $\forall\ i = 1, \ldots, n$. $V_{(s)}$ is also called the s^{th} ordered uniform spacing, $1 \le s \le n$. $U_{(i)}$ is the i^{th} order statistic based on $\{U_1, \ldots, U_n\}$ obtained from $Uniform(a, b)$ distribution, $a < b$, $1 \le i \le n$.

Along with Assumptions 1 and 2, let us further assume

Assumption 3: $X_1^*, \ldots, X_n^*$ (as defined earlier) are *i.i.d.* random variables with distribution function G_{X^*}.

Assumption 4: $Y_1^*, \ldots, Y_n^*$ (as defined earlier) are obtained from the model $Y_i^* = g(X_i^*)+\epsilon_i^*$, $i = 1, \ldots, n$, with $g(\cdot)$ having bounded derivative, ϵ^* having bounded probability density function and $E(\epsilon_i^*|X_i^*) = 0 \; \forall \; i = 1, \ldots, n$.

Based on Assumption 1-4, we develop the following theorems (Theorem 7, 8 and 9) regarding the limiting properties of $T_{n,i}$'s, $i = 1, 2, 3$. In each theorem, part (i) detects the order of degeneracy attached to each of $T_{n,i}$'s, $i = 1, 2, 3$. Part (ii) and part (iv) are directly followed from (i), describing the limiting distributions of $T_{n,1}$, $T_{n,2}$ and $T_{n,3}$. Part (ii) establishes the consistency of each of the test statistics. Suppose $\epsilon^{*(2)}$ has the CDF $H^*_{\epsilon^{*(2)}}(\cdot)$.

Theorem 7: (i) $T_{n,1}$ has kernel of order of degeneracy 0.

(ii) $T_{n,1} \xrightarrow{P} 0$ under H_0.

(iii) Under H_0, $\sqrt{n}(T_{n,1} - E(T_{n,1})) \xrightarrow{L} N(0, 4\xi_1)$.

(iv) Under H_n, $\sqrt{n}(T_{n,1} - E(T_{n,1})) \xrightarrow{L} N(\mu_1, 4\xi_1)$, where

$$\mu_1 = 2\gamma \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}[2\int_{-\infty}^{x^*}\int_{-\infty}^{y^*} dG_{X^*}(u^*)dH^*_{\epsilon^{*(2)}}(v^*) + 2\int_{x^*}^{\infty}\int_{y^*}^{\infty} dG_{X^*}(u^*)dH^*_{\epsilon^{*(2)}}(v^*)]dK_{X^*,\epsilon^*}(x^*, y^*) \tag{25}$$

and,

$$\xi_1 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}[2\int_{-\infty}^{x^*}\int_{-\infty}^{y^*} dG_{X^*}(u^*)dH^*_{\epsilon^{*(2)}}(v^*) + 2\int_{x^*}^{\infty}\int_{y^*}^{\infty} dG_{X^*}(u^*)dH^*_{\epsilon^{*(2)}}(v^*)]^2 dG_{X^*}(x^*)dH_{\epsilon^*}(y^*). \tag{26}$$

Theorem 8: (i) $T_{n,2}$ has kernel of order of degeneracy 1.

(ii) $T_{n,2} \xrightarrow{P} 0$ under H_0.

(iii) The asymptotic distribution for $T_{n,2}$ under H_0 is given by

$$n(T_{n,2} - E(T_{n,2})) \xrightarrow{L} \sum_{k=1}^{\infty} \lambda_k\{\Gamma_k^2 - 1\}$$

where $\Gamma_1, \Gamma_2, \ldots$ are *iid* $N(0, 1)$ random variables, λ_k's are the eigenvalues associated with

$$\begin{aligned} l(x, y) \;=\; & E[sign\{|X^*_{(1)} - X^*_{(2)}| + |X^*_{(3)} - X^*_{(4)}| - |X^*_{(1)} - X^*_{(3)}| - |X^*_{(2)} - X^*_{(4)}|\} \\ & \times sign\{|Y^{*(2)}_{(1)} - Y^{*(2)}_{(2)}| + |Y^{*(2)}_{(3)} - Y^{*(2)}_{(4)}| - |Y^{*(2)}_{(1)} - Y^{*(2)}_{(3)}| - |Y^{*(2)}_{(2)} - Y^{*(2)}_{(4)}|\} \\ & |X^*_{(1)} = x^*, Y^{*(2)}_{(1)} = y^*]. \end{aligned}$$

(iv) The asymptotic distribution for $T_{n,2}$ under H_n is given by

$$n(T_{n,2} - E(T_{n,2})) \xrightarrow{L} \sum_{k=1}^{\infty} \lambda_k \{(\Gamma_k + a_k)^2 - 1\} \tag{27}$$

where $\Gamma_1, \Gamma_2, \ldots$ are *iid* $N(0,1)$ random variables, λ_k's are the eigenvalues associated with $l(x^*, y^*)$ given in (*iii*). The quantities a_k's are defined as

$$a_k = \int h\, f_k(x^*) f_k(y^*)\, dG_{X^*}(x^*) dH^*_{\epsilon^{*(2)}}(y^*). \tag{28}$$

where f_k's are the eigenfunctions corresponding to λ_k's, $k = 1, 2, \ldots$.

Theorem 9: (i) $T_{n,3}$ has kernel of order of degeneracy 1.

(ii) $T_{n,3} \xrightarrow{P} 0$ under H_0.

(iii) The asymptotic distribution for $T_{n,3}$ under H_0 is given by

$$n(T_{n,3} - E(T_{n,3})) \xrightarrow{L} \sum_{k=1}^{\infty} \lambda^*_k \{\Gamma^{*^2}_k - 1\}$$

where $\Gamma^*_1, \Gamma^*_2, \ldots$ are *iid* $N(0,1)$ random variables, λ^*_k's are the eigenvalues associated with

$$\begin{aligned} l^*(x^*, y^*) \;=\; & E[\{|X^*_{(1)} - X^*_{(2)}| + |X^*_{(3)} - X^*_{(4)}| - |X^*_{(1)} - X^*_{(3)}| - |X^*_{(2)} - X^*_{(4)}|\} \\ & \times \{|Y^{*(2)}_{(1)} - Y^{*(2)}_{(2)}| + |Y^{*(2)}_{(3)} - Y^{*(2)}_{(4)}| - |Y^{*(2)}_{(1)} - Y^{*(2)}_{(3)}| - |Y^{*(2)}_{(2)} - Y^{*(2)}_{(4)}|\} \\ & |X^*_{(1)} = x^*, Y^{*(2)}_{(1)} = y^*]. \end{aligned}$$

(iv) The asymptotic distribution for $T_{n,3}$ under H_n is given by

$$n(T_{n,3} - E(T_{n,3})) \xrightarrow{L} \sum_{k=1}^{\infty} \lambda^*_k \{(\Gamma^*_k + a^*_k)^2 - 1\} \tag{29}$$

where $\Gamma^*_1, \Gamma^*_2, \ldots$ are *iid* $N(0,1)$ random variables, λ^*_k's are the eigenvalues associated with $l^*(x^*, y^*)$ given in (*iii*). The quantities a^*_k's are defined as

$$a^*_k = \int h\, f^*_k(x^*) f^*_k(y^*)\, dG_{X^*}(x^*) dH^*_{\epsilon^{*(2)}}(y^*). \tag{30}$$

where f^*_k's are the eigenfunctions corresponding to λ^*_k's, $k = 1, 2, \ldots$.

Proofs of all three theorems are furnished in Appendix 1.

5.3. Examples on asymptotic power calculation

To check on the performance of asymptotic power curves of $T_{n,1}$, $T_{n,2}$ and $T_{n,3}$ with respect to different values of the mixing constant γ introduced in (13) we consider the values of γ from 0 to 10. We investigate on power against the H_0 in reference with these three statistics when the different percentage of missingness occurs in Y values under missing at random (MCAR) structure. All those missing values are refilled by NW estimation process as well as local linear smoothing (ILLS) as elaborately discussed in Section 3. Thereafter, the power functions for $T_{n,1}$, $T_{n,2}$ and $T_{n,3}$ are found for the imputed set of (X^*, Y^*) under $n = 100$. We generate such 500 sets of bootstap sample.

Let us pick up a couple of examples from Einmahl *et al.*(2008) where the conditional distributions of the error ϵ^* for given value of the covariate X^*, along with the joint proper distribution of (X^*, ϵ^*) are proposed. Epanechnikov kernel is used as the kernel function in the expression of the test statistics. Note that for each of the examples under consideration, the null model is taken as independent bivariate normal, *i.e.*, $f_{X^*,\epsilon^*}(.,.) = \frac{1}{2\pi} e^{-\frac{\epsilon^{*2}+x^{*2}}{2}}$. Since under H_0, $F_{X^*,\epsilon^*}(.,.) = G_{X^*}(.)H_{\epsilon^*}(.)$, μ_1 and ξ_1 in (25) and (26) are theoretically found out using the integral of standard normal variable. The rest of the results related to $T_{n,2}$ and $T_{n,3}$ are deduced by approximating infinite sum of weighted chi-square by finite one (taking upto the tenth term of (27) and (29)).

Example 1: $k_{X^*,\epsilon^*}(x^*, e^*)$ is such that $(\epsilon^*|X^* = x^*) \sim N(0, \frac{1+5x^*}{100})$ with $X^* \sim N(0, 1)$.

Example 2: $k_{X^*,\epsilon^*}(x^*, e^*)$ is such that $(\epsilon^*|X^* = x^*) \stackrel{\mathcal{D}}{=} Cauchy(0, x^{*2})$ with $X^* \sim N(0, 1)$.

Percentages of missingness are chosen as 5%, 10% and 20% respectively. For each example, power curves of three statistics under complete data (without missing value) and other three missing proportion cases are drawn (a total of eight figures). The red line denotes the power curve of $T_{n,1}$, whereas the green and blue lines denote the power curves of $T_{n,2}$ and $T_{n,3}$ respectively. Due to space constraint, the power curves obtained only through LLS imputation technique in $n = 100$ are provided here. Appendix 2 contains the detailed and comparative tables of power calculation derived by both NW estimation and ILS technique taking sample size 100 with bootsrap size 500.

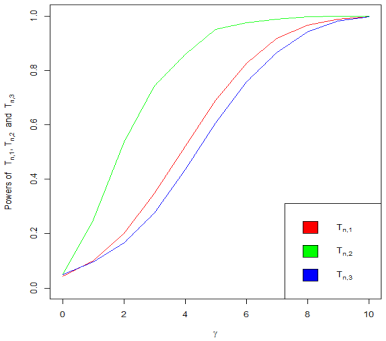


Figure 1: Power for Example 1 against γ in no missing setup

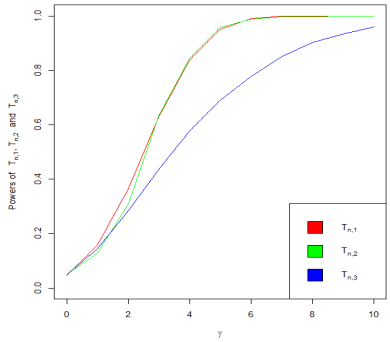


Figure 2: Power for Example 1 against γin 5% MCAR setup

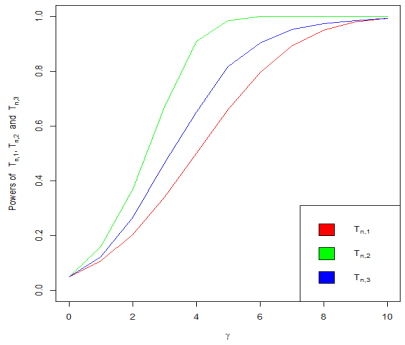


Figure 3: Power for Example 1 against γ in 10% MCAR setup

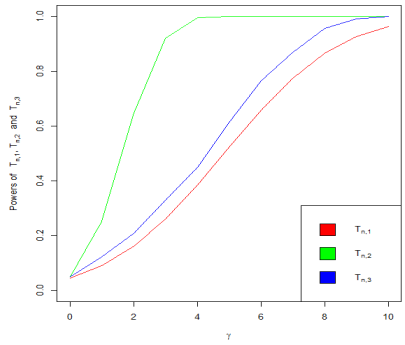


Figure 4: Power for Example 1 against γ in 20% MCAR setup

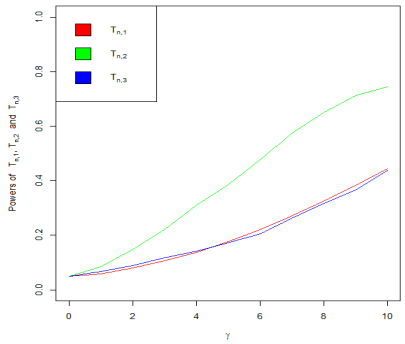


Figure 5: Power for Example 2 against γ in no missing setup

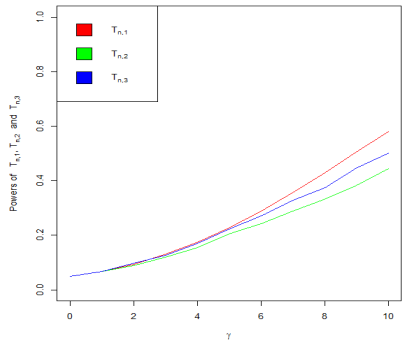


Figure 6: Power for Example 2 against γ in 5% MCAR setup

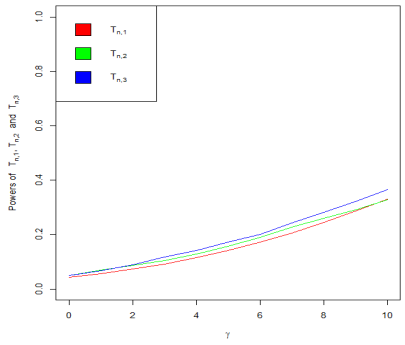


Figure 7: Power for Example 2 against γ in 10% MCAR setup

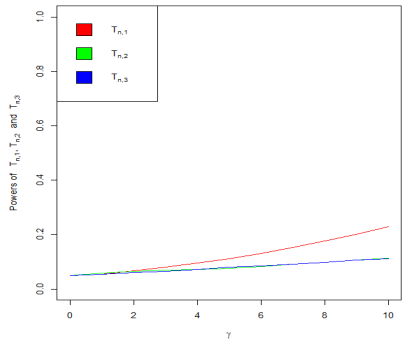


Figure 8: Power for Example 2 against γ in 20% MCAR setup

Although for no missing case power exerted by $T_{n,2}$ performs better across the mixing constant γ, in presence of missingness its power gets deteriorated as compared with the power by Kendall's tau, *i.e.* $T_{n,1}$. In contrast, power by distance based measure $T_{n,3}$ behaves not so well for all choices of missingness. Imputation done by local linear smoothing also does not change the scenario. In applying rank based test when observations on Y are missing does not guarantee the universal superiority in power. The more the counts in bivariate pairing

in test statistic; lesser will be the power with the increase of missingness. Since in $T_{n,2}$ four bivariate pairs are in use, impact of missingness hits it more sharply than $T_{n,1}$. Plausible imputation can not improve the downfall as well.

Additionally, under normally distributed alternative the power exerted by all three statistics are quite reasonable and closer to 1. In contrast, Example 2 dealing with Cauchy alternatives experiences poorer power performance. Cauchy distribution being a heavy tailed distribution might be a good indicator of how sensitive the tests are to departures from normality, *i.e.* in presence of extreme observations. Although in no missing case the proposed $T_{n,2}$ holds its superiority, it fails to hold that in missing cases. In fact more the missingness worse the power comes out.
The entire simulation exercise is performed by R 4.0.5.

6. Real data analysis

In this segment of real data analysis, we choose out Abalone Data collected by the Department of Primary Industry and Fisheries, Tasmania. The data is available online in UCI Machine Learning Repository Data Set page (https://archive.ics.uci.edu/ml/datasets/Abalone).

The primary objective of this zoological data is to predict the age of abalone (a common species of marine gastropod molluscs, mainly inhabited in warm seas) from different physical measurements. This data consists of 4177 observations each having 10 qualitative and quantitative characters. Among those there are 9 independent characters, based on the physical measurements – viz, sex (nominal), length (in mm) for longest shell measurement, diameter (in mm) perpendicular to length, height (in mm) with meat in shell, whole weight (in grams) of abalone, shucked weight (in grams) *i.e.* weight of meat, viscera weight (in grams) *i.e.* gut weight (after bleeding), shell weight (in grams) after being dried, rings (integer) and one dependent variable — age (in years).

In our study, we pick up a single nonparametric regressor, *viz.*, shell weight after being dried (X in grams) and the regressand, *viz.* age (Y in years). For the sake of preciseness, we select first 100 observations instead of the whole. As a preliminary exploratory analysis, let us highlight the scatter plot on age against scaled shell weights below. The plot projects positive association with weakly linear tendency.

In order to incite readers' interest, the group of histograms (Figure 1) on underlying distributions of the response variable Y for complete case as well as for of several percentage of missingness is provided. In this figure, the missing observations are imputed by Nadaraya Watson estimator. Also the kernel density inlay is curved over each histogram. The underlying distribution is mildly right skewed which remains almost same not only in complete case but also in imputed distributions under 5%,10% and 20% missingness. Therefore imputation does not trigger any significant change in the underlying distribution.

To test the independence of X and ϵ we carry out bootstrap tests on 200 resamples having 100 sample observations in each set. At first, the observed values of the test statistics under the null hypothesis are obtained for the fixed sample size 100. Suppose the b^{th} resample of $T_{n,k}$ be $T_{n,k}^{\ b}$, $b = 1, \ldots, 200$, $k = 1, 2, 3$. The estimated p-value of $T_{n,k}$ is computed as $\frac{\#\{T_{n,k}^{\ b} > T_{n,k}^{*}\}}{200}$, $b = 1, \ldots, 200$, $k = 1, 2, 3$ where $T_{n,k}^{*}$ is the observed value of $T_{n,k}$ under H_0.

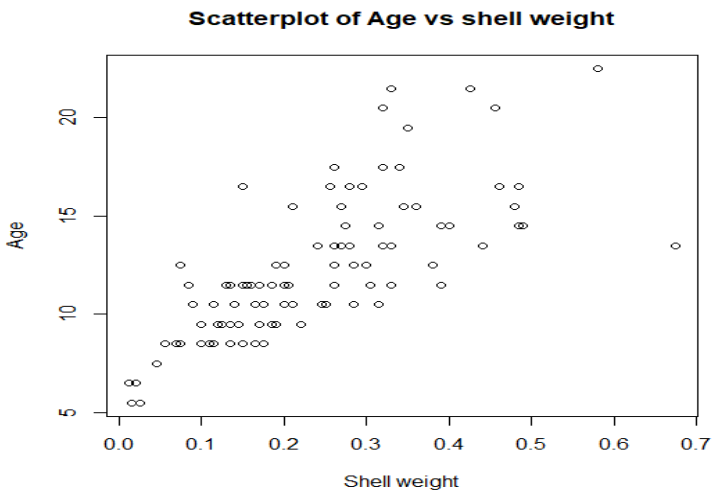


Figure 9: Scatter diagram

The same is repeated for (i) complete case (with 100 observations in each bootstrap set); (ii) 5% randomly missing observations, (iii) 10% randomly missing observations and (iv) 20% randomly missing observations. In each of the missing scenario, the missing observations are imputed by NW estimation as well as ILLS estimation and p-value is reported accordingly. Higher the p-value stronger is the evidence in favour of H_0. Tacitly speaking, for this data, under missingness, each p-value indicates preference towards H_0.

Table 1: Table showing p-values of $T_{n,1}$, $T_{n,2}$ and $T_{n,3}$ under missingness estimated by N-W & ILLS imputation respectively

Statistic	Complete case	p-values					
		N-W			ILLS		
		5%	10%	20%	5%	10%	20%
$T_{n,1}$	0.575	0.375	0.480	0.415	0.490	0.515	0.635
$T_{n,2}$	0.680	0.940	0.930	0.900	0.980	0.989	0.890
$T_{n,3}$	0.660	0.900	0.920	0.880	0.980	0.999	0.905

7. Conclusion

In this article we have investigated the performance of three statistics– two rank based and one distance based, in the presence of MCAR missingness of observations. These tests are consistent. Powers are calculated under contiguous alternatives. For complete case situation $T_{n,2}$ shows best staging over $T_{n,1}$ and $T_{n,3}$ in both Gaussian and the heavy tailed distribution Cauchy but $T_{n,2}$ is not robust enough in presence of constant proportion of missingness. Specifically for non Gaussian alternative, missingness yields poor power exerted by $T_{n,2}$ and $T_{n,3}$ as compared to that by $T_{n,1}$. On the other hand, estimation of missing responses by imputated local linear smoothing (ILLS) method may yield a better power over that deduced by Nadaraya Watson (N-W) method, still those results are not convincing enough for non Gaussian distribution. Therefore, applying a rank based test statistic in testing of independence under nonparametric regression set up in presence of missingness would not

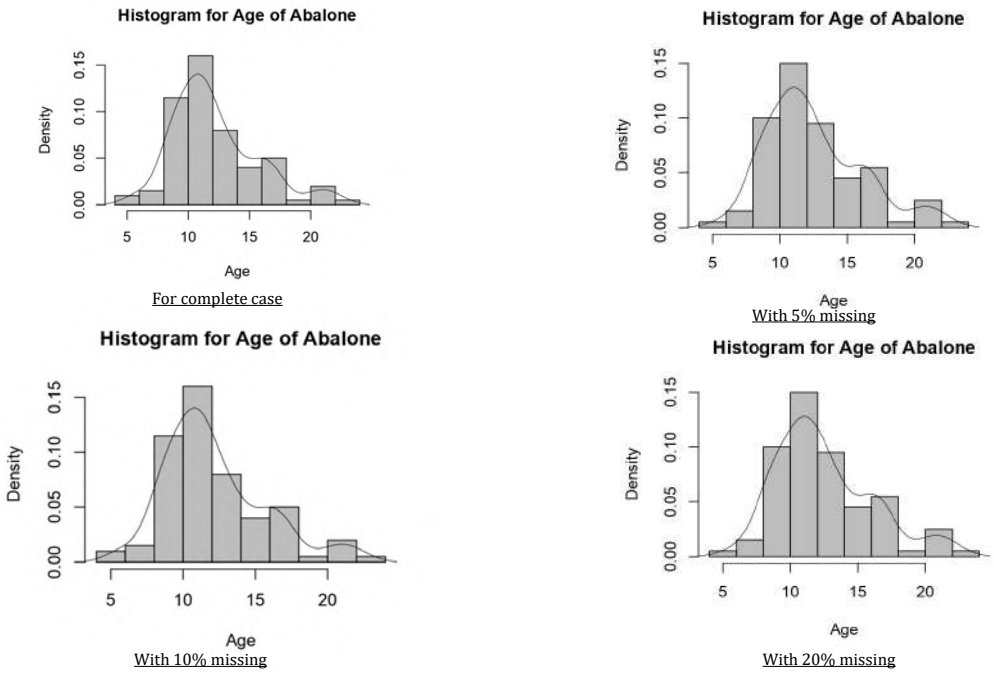


Figure 10: Histograms and regression curve in-lays for complete and missing cases

add substantial amount of power. In order to deal with such a situation few other distance based measure on distribution functions, *e.g.* Kolmogorov-Smirnov or Cramer-Von-Mises might be given a thought. It is to be noted that Alvo *et al.* (1995) proposed a new class of measures of rank correlation which are formed on a notion of distance between incomplete rankings. This approach utilizes the information on the positions of the actual observations relative to the string of incomplete observations. This mechanism would compensate for missing values and may be used as consistent test statistic in same context too.

In missing situation the strongest assumption that is commonly made is that the data are missing completely at random (MCAR) as probability that any variable is missing can not depend on any other variable in the model of interests. But for most data sets, the MCAR assumption is unlikely to be precisely specified, specially in design data. In those cases, a much weaker assumption, missing at random (MAR) is more common in practice. In MAR, the missingness of response depends on another observed variable. Therefore, effectivity of $T_{n,2}$ may be more worth investigating subject under MAR situation as compared with the performance by $T_{n,1}$ and $T_{n,3}$, considering a certain probability distribution of missingness.

Acknowledgments

Both authors are thankful to the anonymous referee for the valuable suggestions. The corresponding author is indebted to Dr. Ujjwal Das, Associate Professor, Indian Institute of Management, Udaipur, India for sharing his initial idea of this problem. Also, the corre-

sponding author like to thank the Chair Editor for his meticulous proof reading which led to substantial improvement of this paper.

References

Afifi, A. A. and Elashoff (1966). R. M. Missing observations in multivariate statistics:I.review of literature. *Journal of the American Statistical Association*, **61**, 595-604.

Alvo, M.and Cabilio, P. (1995). Rank correlation methods for missing data. *The Canadian Journal of Statistics*, **13**, 345-358.

Bairamov, I., Berred, A., and Stepanov, A. (2010). Limit results for ordered uniform spacings. *Statistical Papers*, **51**, 227-240.

Bergsma, W. and Dassios, A. (2014). A consistent test of independence based on a sign covariance related to Kendall's tau. *Bernoulli*, **20**, 1006-1028.

Blum, J. R., Kiefer, J., and Rosenblatt (1961). Distribution free tests of independence based on the sample distribution function. *Annals of Mathematical Statistics*, **32**, 485-498.

Cheng, P. E. (1994). Nonparametric estimation of mean functionals with data missing at random. *Journal of the American Statistical Association*, **89**, 81-87.

Chu, C. K. and Cheng, P. E. (1995). Nonparametric regression estimation with missing data. *Journal of Statistical Planning and Inference*, **48**, 85-99.

Das, S., Halder, S., and Maiti, S.I. (2022). An extended approach to test of independence between error and covariates under nonparametric regression model. *Thailand Statistician*, **21**, 19-36.

Dhar, S. S., Bergsma, W., and Dassios, A (2018). Testing Independence of Covariates and Errors in Non-parametric Regression. *Scandinavian Journal of Statistics*, **45**, 421-443.

Einmahl, J. H. and Van Keilegom, I.(2008). Tests for independence in nonparametric regression. *Statistica Sinica*, **18**, 601-615.

Einmahl, J. H. and Van Keilegom, I.(2008). Specification tests in nonparametric regression. *Journal of Econometrics*, **143**, 88-102.

Gregory, G. G. (1977). Large sample theory for U-statistics and tests of fit. *The Annals of Statistics*, **5**, 110-123.

Hajek J., Sidak Z., and Sen P. K. (1999). *Theory of Rank Tests*. Academic Press.

Hartley, H. O. and Hocking R. R. (1971). The analysis of Incomplete Data. *Biometrics*, **27**, 783-823.

Hlavka Z., Huskova M., and Meintanis S. G. (2011). Tests for independence in nonparametric heteroscedastic regression models. *Journal of Multivariate Analysis*, **102**, 816-27.

Lee A. J. (1990). *U-Statistics: Theory and Practice*. Marcel Dekker: New York and Basel.

Lehmann E. L. and Romano, J. P. (2005). *Testing of Statistical Hypotheses, 3rd Ed.*. Springerlink.

Little, R. J. and Rubin, D. B. (2019). *Statistical Analysis with Missing Data*. 3rd Edition. John Wiley & Sons Inc.

Neumeyer, N. (2009). Testing independence in nonparametric regression. *Journal of Multivariate Analysis*, **100**, 1551-1566.

Serfling, R. J. (1981). *Approximation Theorems of Mathematical Statistics*. John-Wiley & Sons, Inc.

Van Der Vaart A. W. (1998). *Asymptotic Statistics*. Cambridge University Press.

ANNEXURE

Appendix 1

Proof of Theorem 1

The expansion of $\log L_n$ takes the form as follows

$$\begin{aligned}\log L_n &= \log \prod_{i=1}^{n} \frac{f_{n;X^*,\epsilon^*}(x_i^*, e_i^*)}{f_{X^*,\epsilon^*}(x_i^*, e_i^*)} \\ &= \log \prod_{i=1}^{n} \Big\{ \frac{(1-\frac{\gamma}{\sqrt{n}}) f_{X^*,\epsilon^*}(x_i^*, e_i^*) + \frac{\gamma}{\sqrt{n}} k_{X^*,\epsilon^*}(x_i^*, e_i^*)}{f_{X^*,\epsilon^*}(x_i^*, e_i^*)} \Big\} \\ &= \sum_{i=1}^{n} \log \Big\{ \frac{(1-\frac{\gamma}{\sqrt{n}}) f_{X^*,\epsilon^*}(x_i^*, e_i^*) + \frac{\gamma}{\sqrt{n}} k_{X^*,\epsilon}(x_i^*, e_i^*)}{f_{X^*,\epsilon^*}(x_i^*, e_i^*)} \Big\}.\end{aligned}$$

With the aid of Taylor's expansion of $\log(1+r)$, $r > -1$ as well as the weak law of large numbers, $\log L_n$ is further expanded as

$$\sum_{i=1}^{n} \frac{\gamma}{\sqrt{n}} \left(\frac{k_{X^*,\epsilon^*}(x_i^*, e_i^*)}{f_{X^*,\epsilon^*}(x_i^*, e_i^*)} - 1 \right) - \frac{\gamma^2}{2n} \sum_{i=1}^{n} \left(\frac{k_{X^*,\epsilon^*}(x_i^*, e_i^*)}{f_{X^*,\epsilon^*}(x_i^*, e_i^*)} - 1 \right)^2 + O_P(n^{-1/2}). \tag{31}$$

Then,

$$\log L_n - \textstyle\sum_{i=1}^{n} \frac{\gamma}{\sqrt{n}} \left(\frac{k_{X^*,\epsilon^*}(x_i^*, e_i^*)}{f_{X^*,\epsilon^*}(x_i^*, e_i^*)} - 1 \right) + \frac{\gamma^2}{2n} \sum_{i=1}^{n} \left(\frac{k_{X^*,\epsilon^*}(x_i^*, e_i^*)}{f_{X^*,\epsilon^*}(x_i^*, e_i^*)} - 1 \right)^2 = O_P(n^{-1/2}) \longrightarrow 0 \ \ as \ \ n \to \infty.$$

Define a sequence of random variables W_n as $\displaystyle\sum_{i=1}^{n} \frac{\gamma}{\sqrt{n}} \left(\frac{k_{X^*,\epsilon^*}(x_i^*, e_i^*)}{f_{X^*,\epsilon^*}(x_i^*, e_i^*)} - 1 \right)$. With the help of Lindeberg's condition, the asymptotic distribution of W_n is developed as $\frac{W_n - E(W_n)}{\sqrt{Var(W_n)}} \xrightarrow{L} N(0,1)$ under H_0, where

$$E_{H_0}(W_n) = \sum_{i=1}^{n} \frac{\gamma}{\sqrt{n}} E_{H_0} \left(\frac{k_{X^*,\epsilon}(x_i^*, e_i^*)}{f_{X^*,\epsilon^*}(x_i^*, e_i^*)} - 1 \right) = 0$$

and $Var_{H_0}(W_n) = \dfrac{\gamma^2}{n} \displaystyle\sum_{i=1}^{n} E_{H_0} \left(\frac{k_{X^*,\epsilon^*}(x_i^*, e_i^*)}{f_{X^*,\epsilon^*}(x_i^*, e_i^*)} - 1 \right)^2 = \gamma^2 E_{H_0} \left(\frac{k_{X^*,\epsilon^*}}{f_{X^*,\epsilon^*}} - 1 \right)^2$. Hence under H_0,

$$W_n \xrightarrow{L} N\left(0, \gamma^2 E_{H_0} \left(\frac{k_{X^*,\epsilon^*}}{f_{X^*,\epsilon^*}} - 1 \right)^2 \right).$$

Another sequence of random variables $V_n = \dfrac{\gamma^2}{2n} \displaystyle\sum_{i=1}^{n} \left(\frac{k_{X^*,\epsilon^*}(x_i^*, e_i^*)}{f_{X^*,\epsilon^*}(x_i^*, e_i^*)} - 1 \right)^2$ weakly converges to $\frac{\gamma^2}{2} E_{H_0} \left(\frac{k_{X^*,\epsilon^*}}{f_{X^*,\epsilon^*}} - 1 \right)^2$. So, $\log L_n - W_n + V_n = o_p(1)$. Slutsky's theorem further ensures that

the limiting distribution of the sequence of random variables $M_n = W_n - V_n$ converges to a random variable M such that

$$M \sim N\left(-\frac{1}{2}\gamma^2 E_{H_0}\left(\frac{k}{f}-1\right)^2, \gamma^2 E_{H_0}\left(\frac{k}{f}-1\right)^2\right). \tag{32}$$

Summing up all, one can conclude that $\log L_n - M_n = o_p(1)$, *i.e.* $\log L_n$ has the limiting distribution which is identical with that of limiting distribution of M_n, *i.e.* $N(-\frac{1}{2}\sigma, \sigma)$ where $\sigma = \gamma^2 E_{H_0}\left(\frac{k}{f}-1\right)^2$. Thereafter, the Corollary 5.1 of lemma 5.1 is sufficient enough in establishing the fact that H_n is a contiguous sequence of alternatives due to asymptotic normality of $\log L_n$. Notationally, contiguity can be expressed as $F_{X^*,\epsilon^*} \triangleleft F_{n;X^*,\epsilon^*}$.

Proof of Theorem 7

(i) Suppose the kernel of $T_{n,1}$ is denoted by $\psi((X^*_{(1)}, Y^{*(2)}_{(1)}), (X^*_{(2)}, Y^{*(2)}_{(2)}))$. One can simplify its form as

$$\begin{aligned}\psi_1(x^*, y^*) &= E[\psi((X^*_{(1)}, Y^{*(2)}_{(1)}), (X^*_{(2)}, Y^{*(2)}_{(2)}))|X^*_{(1)} = x^*, Y^{*(2)}_{(1)} = y^*] \\ &= E[sign\{(X^*_{(1)} - X^*_{(2)})(Y^{*(2)}_{(1)} - Y^{*(2)}_{(2)})\}|X^*_{(1)} = x^*, Y^{*(2)}_{(1)} = y^*] \\ &= 2P[(X^*_{(1)} - X^*_{(2)})(Y^{*(2)}_{(1)} - Y^{*(2)}_{(2)}) > 0|X^*_{(1)} = x^*, Y^{*(2)}_{(1)} = y^*] - 1.\end{aligned}$$

Now under H_0 one can determine that

$$E_{(X^*_{(1)}, Y^{*(2)}_{(1)})}[\psi_1(X^*_{(1)}, Y^{*(2)}_{(1)})] = E_{(X^*_{(1)}, Y^{*(2)}_{(1)}), (X^*_{(2)}, Y^{*(2)}_{(2)})}[\psi((X^*_{(1)}, Y^{*(2)}_{(1)}), (X^*_{(2)}, Y^{*(2)}_{(2)}))] = 0.$$

Then, $\xi_1 = Var[\psi_1(X^*_{(1)}, Y^{*(2)}_{(1)})] = E[\psi_1^2(X^*_{(1)}, Y^{*(2)}_{(1)})] > 0$, where $Y^{*(2)}$ is approximately identically distributed with $\epsilon^{*(2)}$. Therefore, $\xi_0 = 0$ and $\xi_1 > 0$ is enough to conclude that ψ has order of degeneracy 0.

(ii) From Theorem 4 it is clear that the variance of $T_{n,1}$ gets approximated as $\frac{4\xi_1}{n}$ for large n, and $E[sign\{(X^*_{(i)} - X^*_{(j)})(Y^{*(2)}_{(i)} - Y^{*(2)}_{(j)})\}] = 0 \; \forall \; 1 \le i < j \le n$ as $P[(X^*_{(i)} - X^*_{(j)})(Y^{*(2)}_{(i)} - Y^{*(2)}_{(j)}) > 0] = P[(X^*_{(i)} - X^*_{(j)})(Y^{*(2)}_{(i)} - Y^{*(2)}_{(j)}) < 0]$ under H_0. One may conclude that $T_{n,1} \xrightarrow{P} 0$ as $E(T_{n,1}) = 0$ and $var(T_{n,1}) \to 0$ for $n \to \infty$ under H_0.

(iii) Deducing the asymptotic variance in Theorem 4 when $n \to \infty$, we derive the asymptotic distribution of $\sqrt{n}(T_{n,1} - E(T_{n,1}))$ under H_0. To prove this part of the theorem, any standard textbook on nonparametric inference would suffice.

(iv) Directed from the Le Cam's third lemma (Dhar *et al.* (2018)) the asymptotic distribution of $(\sqrt{n}(T_{n,1} - E(T_{n,1})), \log L_n)$ converges to $N_2\left(\begin{pmatrix} 0 \\ -\frac{\theta}{2} \end{pmatrix}, \begin{pmatrix} 4\xi_1 & \tau \\ \tau & \theta \end{pmatrix}\right)$, $\theta > 0$ under H_0. Then it is easy to determine the limiting distribution of $\sqrt{n}(T_{n,1} - E(T_{n,1}))$ under H_n as $N(0 + \tau, 4\xi_1)$ *i.e.* $N(\tau, 4\xi_1)$. Hence $\tau = \lim_{n\to\infty} cov_{H_0}(\sqrt{n}(T_{n,1} - E(T_{n,1})), \log L_n)$ which can be finally derived as

$$2\gamma \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}[2\int_{-\infty}^{x^*}\int_{-\infty}^{y^*} dG_{X^*}(u^*)dH_{\epsilon^*}(v^*) + 2\int_{x^*}^{\infty}\int_{y^*}^{\infty} dG_{X^*}(u^*)dH_{\epsilon^*}(v^*) - 1]dK_{X^*,\epsilon^*}(x^*, y^*).$$

Proof of Theorem 8

(i) The simplification of the kernel of $T_{n,2}$ is done as

$$\begin{aligned}
& a(X^*_{(1)}, X^*_{(2)}, X^*_{(3)}, X^*_{(4)})a(Y^*_{(1)}, Y^*_{(2)}, Y^*_{(3)}, Y^*_{(4)}) \\
= \; & 2I(|X^*_{(1)} - X^*_{(2)}| + |X^*_{(3)} - X^*_{(4)}| - |X^*_{(1)} - X^*_{(3)}| - |X^*_{(2)} - X^*_{(4)}| > 0, \\
& |Y^{*(2)}_{(1)} - Y^{*(2)}_{(2)}| + |Y^{*(2)}_{(3)} - Y^{*(2)}_{(4)}| - |Y^{*(2)}_{(1)} - Y^{*(2)}_{(3)}| - |Y^{*(2)}_{(2)} - Y^{*(2)}_{(4)}| > 0) + \\
& 2I(|X^*_{(1)} - X^*_{(2)}| + |X^*_{(3)} - X^*_{(4)}| - |X^*_{(1)} - X^*_{(3)}| - |X^*_{(2)} - X^*_{(4)}| < 0, \\
& |Y^{*(2)}_{(1)} - Y^{*(2)}_{(2)}| + |Y^{*(2)}_{(3)} - Y^{*(2)}_{(4)}| - |Y^{*(2)}_{(1)} - Y^{*(2)}_{(3)}| - |Y^{*(2)}_{(2)} - Y^{*(2)}_{(4)}| < 0) - 1 \\
= \; & 2P(|Y^{*(2)}_{(1)} - Y^{*(2)}_{(2)}| + |Y^{*(2)}_{(3)} - Y^{*(2)}_{(4)}| - |Y^{*(2)}_{(1)} - Y^{*(2)}_{(3)}| - |Y^{*(2)}_{(2)} - Y^{*(2)}_{(4)}| < 0) - 1 \\
= \; & \tilde{a}((X^*_{(1)}, Y^{*(2)}_{(1)}), (X^*_{(2)}, Y^{*(2)}_{(2)}), (X^*_{(3)}, Y^{*(2)}_{(3)}), (X^*_{(4)}, Y^{*(2)}_{(4)})) \qquad (33)
\end{aligned}$$

where $I(\cdot)$ is an indicator function. Now define, for $c = 0, \ldots, 4$,

$$\begin{aligned}
& \tilde{a}_c((x^*_{(1)}, y^{*(2)}_{(1)}), \ldots, (x^*_{(c)}, y^{*(2)}_{(c)})) \\
= \; & E[\tilde{a}((x^*_{(1)}, y^{*(2)}_{(1)}), \ldots, (x^*_{(c)}, y^{*(2)}_{(c)}), (X^*_{(c+1)}, Y^{*(2)}_{(c+1)}), \ldots, (X^*_{(4)}, Y^{*(2)}_{(4)}))]
\end{aligned}$$

and, $\xi_c = Var[\tilde{a}_c((X_{(1)}, Y^{*(2)}_{(1)}), \ldots, (X_{(c)}, Y^{*(2)}_{(c)}))]$.
In equation (33), $|Y^{*(2)}_{(1)} - Y^{*(2)}_{(3)}|$ and $|Y^{*(2)}_{(2)} - Y^{*(2)}_{(4)}|$ can be written into following two inequalities as $|Y^{*(2)}_{(1)} - Y^{*(2)}_{(3)}| \le |Y^{*(2)}_{(1)} - Y^{*(2)}_{(2)}| + |Y^{*(2)}_{(2)} - Y^{*(2)}_{(3)}|$ and
$|Y^{*(2)}_{(2)} - Y^{*(2)}_{(4)}| \le |Y^{*(2)}_{(2)} - Y^{*(2)}_{(3)}| + |Y^{*(2)}_{(3)} - Y^{*(2)}_{(4)}|$. Then,
$P(Y^{*(2)}_{(2)} > Y^{*(2)}_{(3)}, Y^{*(2)}_{(1)} > Y^{*(2)}_{(4)})$
$= P(Y^{*(2)}_{(2)} > Y^{*(2)}_{(3)}, Y^{*(2)}_{(1)} > Y^{*(2)}_{(4)}, Y^{*(2)}_{(3)} > Y^{*(2)}_{(1)}) + P(Y^{*(2)}_{(2)} > Y^{*(2)}_{(3)}, Y^{*(2)}_{(1)} > Y^{*(2)}_{(4)}, Y^{*(2)}_{(3)} \le Y^{*(2)}_{(1)}) = \frac{1}{4!} \times 6 = \frac{1}{4}$. Similarly, $P(Y^{*(2)}_{(2)} > Y^{*(2)}_{(3)}, Y^{*(2)}_{(1)} \le Y^{*(2)}_{(4)})$ is calculated as $\frac{1}{4}$.
Then, $P(Y^{*(2)}_{(2)} < Y^{*(2)}_{(3)}) = \frac{1}{2} = P(Y^{*(2)}_{(2)} > Y^{*(2)}_{(3)})$.
Finally we obtain $2P(|Y^{*(2)}_{(1)} - Y^{*(2)}_{(2)}| + |Y^{*(2)}_{(3)} - Y^{*(2)}_{(4)}| - |Y^{*(2)}_{(1)} - Y^{*(2)}_{(3)}| - |Y^{*(2)}_{(2)} - Y^{*(2)}_{(4)}| < 0) = 2 \min\left(\frac{1}{2}, \frac{1}{2}\right) = 1$. Therefore,
$E[\tilde{a}((X^*_{(1)}, Y^{*(2)}_{(1)}), (X^*_{(2)}, Y^{*(2)}_{(2)}), (X^*_{(3)}, Y^{*(2)}_{(3)}), (X^*_{(4)}, Y^{*(2)}_{(4)}))] = 0$.
On the other hand, the expression of ξ_1 is same as
$cov[\tilde{a}((X^*_{(1)}, Y^{*(2)}_{(1)}), (X^*_{(2)}, Y^{*(2)}_{(2)}), (X^*_{(3)}, Y^{*(2)}_{(3)}), (X^*_{(4)}, Y^{*(2)}_{(4)}))]$ which equals
$\{1 + 4P[Y^{*(2)}_{(2)} > Y^{*(2)}_{(3)}, Y^{*(2)}_{(5)} > Y^{*(2)}_{(6)}] - 2P[Y^{*(2)}_{(2)} > Y^{*(2)}_{(3)}] - 2P[Y^{*(2)}_{(5)} > Y^{*(2)}_{(6)}]\}$.
For four distinct numbers (i_1, i_2, i_3, i_4) with $1 \le i_1 \ne i_2 \ne i_3 \ne i_4 \ne 7$ it is easy to verify that
$P[Y^{*(2)}_{(i_1)} > Y^{*(2)}_{(i_2)} > Y^{*(2)}_{(i_3)} > Y^{*(2)}_{(i_4)}] = \frac{6}{4!} = \frac{1}{4}$ and furthermore $P[Y^{*(2)}_{(i_1)} > Y^{*(2)}_{(i_2)}] = \frac{1}{2}$.
Then $\xi_1 = 1 + 4 \cdot \frac{1}{4} - 2 \cdot \frac{1}{2} - 2 \cdot \frac{1}{2} = 0$.
Consequently, the computation of ξ_2 becomes necessary to verify whether it is equal to 0 or not. ξ_2 is evaluated further as $\{1 + 4P[Y^{*(2)}_{(2)} > Y^{*(2)}_{(3)}, Y^{*(2)}_{(2)} > Y^{*(2)}_{(5)}] - 2P[Y^{*(2)}_{(2)} > Y^{*(2)}_{(3)}] - 2P[Y^{*(2)}_{(2)} > Y^{*(2)}_{(5)}]\}$ which equals $4 \times 5 \times 6 \times P[Y^{*(2)}_{(2)} > Y^{*(2)}_{(3)} > Y^{*(2)}_{(5)} > Y^{*(2)}_{(6)} > Y^{*(2)}_{(4)} > Y^{*(2)}_{(1)}] + 4 \times 5 \times 6 \times P[Y^{*(2)}_{(2)} > Y^{*(2)}_{(5)} > Y^{*(2)}_{(3)} > Y^{*(2)}_{(6)} > Y^{*(2)}_{(4)} > Y^{*(2)}_{(1)}] =$

$2 \times \frac{4\times5\times6}{6!} = \frac{1}{3} > 0$. So $\xi_2 > 0$, which naturally implies that the order of degeneracy of $T_{n,2}$ is 1.

(ii) It is to be noted that $(|X^*_{(i)} - X^*_{(j)}| + |X^*_{(k)} - X^*_{(l)}| - |X^*_{(i)} - X^*_{(k)}| - |X^*_{(j)} - X^*_{(l)}|)(|Y^{*(2)}_{(i)} - Y^{*(2)}_{(j)}| + |Y^{*(2)}_{(k)} - Y^{*(2)}_{(l)}| - |Y^{*(2)}_{(i)} - Y^{*(2)}_{(k)}| - |Y^{*(2)}_{(j)} - Y^{*(2)}_{(l)}|) = O_p(\frac{\log n}{n}), 1 \le i < j < k < l \le n$ by Result 5.1 originally introduced by Bairamov *et al.* (2010).
The distribution function of $(|\epsilon^{*(2)}_{(i)} - \epsilon^{*(2)}_{(j)}| + |\epsilon^{*(2)}_{(k)} - \epsilon^{*(2)}_{(l)}| - |\epsilon^{*(2)}_{(i)} - \epsilon^{*(2)}_{(k)}| - |\epsilon^{*(2)}_{(j)} - \epsilon^{*(2)}_{(l)}|)$ is $\int_{-\infty}^{\infty} \left\{ H_{\epsilon^*}\left(y^* + \frac{t}{2}\right) - H_{\epsilon^*}\left(y^* - \frac{t}{2}\right) \right\} dH_{\epsilon^*}(y^*)$, denoted by $H^*_{\epsilon^{*(2)}}(t)$. Also the distribution function of $\epsilon^{*(2)}$ is approximately equal to the distribution function of $Y^{*(2)}$. One can derive that $a(X^*_{(i)}, X^*_{(j)}, X^*_{(k)}, X^*_{(l)})a(Y^{*(2)}_{(i)}, Y^{*(2)}_{(j)}, Y^{*(2)}_{(k)}, Y^{*(2)}_{(l)}) \longrightarrow 0$ in probability for $1 \le i < j < k < l \le n$ under H_0. Consequently a final conclusion becomes inevitable that $T_{n,2} \xrightarrow{P} 0$ as $n \to \infty$.

(iii) Due to Serfling (1981)'s theorem on the asymptotic distribution of a degenerate U-statistic presented by Theorem 5, it is quite straightforward to derive the limiting distributional form of $n(T_{n,2} - E(T_{n,2}))$ under H_0.

(iv) To furnish the elaborate proof regarding the asymptotic distribution of $n(T_{n,2} - E(T_{n,2}))$ under H_n, Theorem 6 by Gregory (1977) is required.

Proof of Theorem 9

In similar way to the proof of Theorem 8, Theorem 9 can also be proved.

Appendix 2

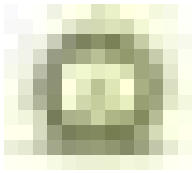
Table 2: Powers of $T_{n,1}$, $T_{n,2}$ and $T_{n,3}$ for Example 1 for complete and missing cases using N-W and ILLS imputation

γ	Powers of test statistics in MCAR setup using NW estimation												Powers of test statistics in MCAR setup using ILLS								
	No missing			5% missing			10% missing			20% missing			5% missing			10% missing			20% missing		
	Power of $T_{n,1}$	Power of $T_{n,2}$	Power of $T_{n,3}$	Power of $T_{n,1}$	Power of $T_{n,2}$	Power of $T_{n,3}$	Power of $T_{n,1}$	Power of $T_{n,2}$	Power of $T_{n,3}$	Power of $T_{n,1}$	Power of $T_{n,2}$	Power of $T_{n,3}$	Power of $T_{n,1}$	Power of $T_{n,2}$	Power of $T_{n,3}$	Power of $T_{n,1}$	Power of $T_{n,2}$	Power of $T_{n,3}$	Power of $T_{n,1}$	Power of $T_{n,2}$	Power of $T_{n,3}$
0	0.042	0.05	0.05	0.053	0.05	0.05	0.05	0.05	0.05	0.054	0.05	0.05	0.047	0.05	0.05	0.05	0.05	0.05	0.046	0.05	0.05
1	0.1	0.25	0.095	0.103	0.06	0.188	0.156	0.064	0.12	0.157	0.087	0.087	0.157	0.129	0.143	0.108	0.16	0.123	0.091	0.252	0.124
2	0.201	0.537	0.166	0.18	0.085	0.474	0.355	0.1	0.249	0.343	0.143	0.148	0.366	0.309	0.284	0.205	0.371	0.268	0.161	0.647	0.209
3	0.347	0.742	0.276	0.286	0.139	0.796	0.604	0.148	0.468	0.579	0.206	0.241	0.628	0.632	0.436	0.341	0.668	0.465	0.261	0.92	0.33
4	0.521	0.859	0.436	0.415	0.23	0.969	0.816	0.238	0.695	0.789	0.273	0.344	0.84	0.845	0.579	0.501	0.909	0.652	0.386	0.995	0.45
5	0.691	0.95	0.606	0.554	0.356	0.997	0.938	0.381	0.877	0.92	0.325	0.501	0.951	0.957	0.691	0.661	0.984	0.818	0.523	1	0.613
6	0.827	0.975	0.759	0.686	0.483	1	0.985	0.5	0.968	0.978	0.359	0.66	0.99	0.989	0.777	0.796	1	0.904	0.658	1	0.766
7	0.918	0.988	0.866	0.798	0.628	1	0.998	0.63	0.996	0.995	0.391	0.758	0.999	0.998	0.85	0.892	1	0.953	0.775	1	0.869
8	0.967	0.996	0.943	0.882	0.758	1	1	0.753	1	0.999	0.411	0.84	1	0.998	0.902	0.951	1	0.974	0.866	1	0.957
9	0.989	0.999	0.982	0.938	0.852	1	1	0.851	1	1	0.421	0.898	1	1	0.934	0.981	1	0.986	0.927	1	0.991
10	0.997	0.999	0.996	0.97	0.924	1	1	0.913	1	1	0.42	0.948	1	1	0.96	0.993	1	0.994	0.964	1	1

Table 3: Powers of $T_{n,1}$, $T_{n,2}$ and $T_{n,3}$ for Example 2 for complete and missing cases using N-W and ILLS imputation

γ	Powers of test statistics in MCAR setup using N-W estimation												Powers of test statistics in MCAR setup using ILLS								
	No missing			5% missing			10% missing			20% missing			5% missing			10% missing			20% missing		
	Power of $T_{n,1}$	Power of $T_{n,2}$	Power of $T_{n,3}$	Power of $T_{n,1}$	Power of $T_{n,2}$	Power of $T_{n,3}$	Power of $T_{n,1}$	Power of $T_{n,2}$	Power of $T_{n,3}$	Power of $T_{n,1}$	Power of $T_{n,2}$	Power of $T_{n,3}$	Power of $T_{n,1}$	Power of $T_{n,2}$	Power of $T_{n,3}$	Power of $T_{n,1}$	Power of $T_{n,2}$	Power of $T_{n,3}$	Power of $T_{n,1}$	Power of $T_{n,2}$	Power of $T_{n,3}$
0	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.05	0.044	0.05	0.05	0.05	0.05	0.05
1	0.059	0.084	0.066	0.063	0.083	0.067	0.085	0.065	0.061	0.065	0.058	0.051	0.066	0.067	0.068	0.057	0.069	0.067	0.055	0.058	0.054
2	0.08	0.149	0.09	0.085	0.148	0.088	0.146	0.073	0.077	0.087	0.06	0.051	0.094	0.09	0.097	0.073	0.086	0.089	0.066	0.064	0.06
3	0.106	0.221	0.117	0.112	0.206	0.104	0.231	0.08	0.093	0.115	0.067	0.055	0.13	0.119	0.126	0.092	0.104	0.118	0.08	0.069	0.064
4	0.138	0.31	0.142	0.145	0.29	0.124	0.338	0.103	0.113	0.149	0.069	0.056	0.175	0.154	0.17	0.115	0.128	0.141	0.095	0.072	0.072
5	0.176	0.386	0.173	0.184	0.396	0.151	0.46	0.121	0.132	0.189	0.079	0.056	0.228	0.205	0.222	0.141	0.157	0.173	0.112	0.075	0.08
6	0.22	0.478	0.206	0.229	0.488	0.182	0.586	0.146	0.15	0.235	0.09	0.057	0.289	0.243	0.271	0.172	0.189	0.201	0.131	0.083	0.085
7	0.27	0.573	0.263	0.28	0.57	0.211	0.704	0.181	0.18	0.287	0.106	0.063	0.357	0.289	0.327	0.206	0.227	0.242	0.152	0.091	0.091
8	0.325	0.651	0.316	0.336	0.64	0.244	0.803	0.194	0.207	0.343	0.122	0.064	0.43	0.332	0.374	0.244	0.259	0.283	0.176	0.098	0.098
9	0.383	0.712	0.365	0.395	0.692	0.287	0.879	0.213	0.245	0.403	0.136	0.069	0.506	0.382	0.446	0.286	0.29	0.321	0.201	0.106	0.107
10	0.445	0.745	0.437	0.458	0.72	0.325	0.932	0.238	0.288	0.465	0.146	0.072	0.581	0.444	0.501	0.331	0.327	0.365	0.229	0.113	0.112

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 22, No. 1, 2024 (New Series), pp 219–228
http://www.ssca.org.in/journal



Partially Accelerated Reliability Demonstration Tests For A Parallel System With Weibull Distributed Components Under Periodic Inspection

Preeti Wanti Srivastava and Satya Rani
Department of Operational Research, University of Delhi, Delhi 110007, India

Received: 13 February 2023; Revised: 30 June 2023; Accepted: 26 July 2023

Abstract

A Reliability Demonstration Test (RDT) demonstrates whether a product has met a certain reliability requirement with a specific confidence. This paper deals with construction of RDTs for a two-component parallel system subject to constant-stress partially accelerated life testing (CSPALT) using periodic mode of inspection and Weibull life distribution. The data from periodic inspection consist of number of failures of systems due to each component in each inspection period. In CSPALT the test specimens are allocated to two test chambers with test items running at normal operating condition in one and at accelerated condition in the other till the termination of the experiment. The optimal test plan consists in obtaining optimal number of allocations in each test chamber and optimal inspections points. RDTs based on optimal test plan are carried out for mean lives of components as well as the system. A numerical example is presented to illustrate the method developed.

Key words: Reliability demonstration tests; Partially accelerated life tests; Periodic inspection; Two-component parallel system; Weibull life distribution; D-optimality criterion.

1. Introduction

Accelerated life tests (ALTs) facilitate bringing about early failures in highly reliable items lasting for several years and hence obtaining reliability information about them in timely manner. This in turn helps the manufacturer to sustain in competitive market where technology is constantly changing with change in consumers' tastes. The book by Nelson (2009) gives a detailed account of Accelerated Tests. The data from a periodic inspection referred to as "grouped data" or "interval data" comprises number of failures in each inspection period. In contrast to continuous inspection wherein exact failure times of the test units are observed, periodic inspection requires less testing effort and is administratively convenient as compared to continuous inspection. In the literature the periodic inspection has been used by many authors, for example, Kulldorff (1961); Ehrenfeld (1962); Nelson (1977); Archer (1982); Flygare *et al.* (1985); Meeker (1986); Yum and Choi (1989); Seo and Yum (1991); Ahmad *et al.* (1994); Islam and Ahmad (1994); Ahmad and Islam (1996); Ahmad *et al.* (2006). A PALT is modelled using an acceleration factor (AF) and a life distribution,

Corresponding Author: Preeti Wanti Srivastava
Email: preetisrivastava.saxena@gmail.com

where AF is defined as the ratio of a reliability measure, say mean life, at use condition to that at accelerated condition. AF= k (say) means that the unit under consideration runs k times longer at normal operating condition than at accelerated condition. In CSPALT the test specimens are allocated to two test chambers with test items running at normal operating condition in one and at accelerated condition in the other till the termination of the experiment. ALTs have been studied extensively in the literature see for example, Srivastava (2017) and Chen *et al.* (2018).

The theory of testing statistical hypotheses provides the tools for reliability demonstration. If either the life distribution or its parameters are unknown, then the problem of reliability demonstration is that of obtaining suitable data and using them to test the null hypothesis that $R(t_0) \geq R_0$ against the alternative that $R(t_0) < R_0$, where t_0 is the specified time point and R_0 is desired reliability. We wish to test whether the reliability of the device at age t_0, $R(t_0)$ satisfies the requirement that $R(t_0) \geq R_0$. Nelson (1977) has provided the optimal demonstration tests with grouped inspection data from an exponential distribution and has also explained how to use the results for a Weibull distribution with known shape parameter. Optimal demonstration tests with grouped inspection data for logistic, log-logistic, normal/Gaussian, and log-normal distributions have been obtained Wei and Bau (1987).

The present paper deals with formulation of reliability demonstration tests for a two-component parallel system subject to CSPALT using periodic mode of inspection and Weibull life distribution. The Weibull life distribution incorporates various failure rates-increasing, decreasing and constant and is therefore of importance in industries manufacturing electronic and mechanical components. It adequately fits the life of several types of capacitors and resistors, such as electrolytic aluminium and tantalum capacitors and carbon film resistors (Yang, 2007; Shaw, 1987).

2. Notation

δ	Weibull shape parameter
μ_1	Weibull scale parameter for component 1
μ_2	Weibull scale parameter for component 2
λ_1	Exponential Scale parameter for component 1
λ_2	Exponential scale parameter for component 2
A	Acceleration Factor, $A > 1$
$R(t)$	Reliability function
n	Total number of two-component parallel systems
w_{1j}, w_{A1j}	The number of systems failing due to component 1 in $(t_{j-1}, t_j]$, $j = 1, 2, \ldots k+1$ in chamber 1 and chamber 2, respectively
w_{2j}, w_{A2j}	The number of systems failing due to component 2 in $(t_{j-1}, t_j]$, $j = 1, 2, \ldots k+1$ in chamber 1 and chamber 2, respectively
P_{1j}, P_{A1j}	The probability of failure of a system due to component 1 in $(t_{j-1}, t_j]$, $j = 1, 2, \ldots k+1$ in chamber 1 and chamber 2, respectively
P_{2j}, P_{A2j}	The probability of failure of a system due to component 2 in $(t_{j-1}, t_j]$, $j = 1, 2, \ldots k+1$ in chamber 1 and chamber 2, respectively
$N(0,1)$	Standard normal distribution

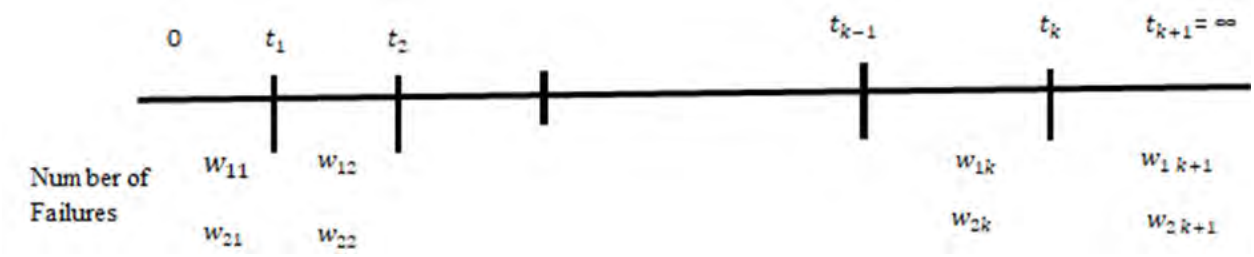


Figure 1: Structure of periodic inspection in chamber 1

Figure 2: Structure of periodic inspection in Chamber 2

3. Formulation of likelihood function

n independent parallel systems each with two independent components are put to test under CSPALT. Out of n systems n_1 systems are put to test in test chamber 1 where they are run under normal operating condition, and n_2 systems are put to test in test chamber 2 where they are run at accelerated condition. The systems are examined for failures periodically at optimally spaced inspection t_1, t_2, ... , t_{k+1}. Let, $t_0 = 0$ and $t_{k+1} = \infty$. Define ρ as the proportion of units that are allocated in chamber 1, and $1-\rho$ as the proportion of units that are allocated in chamber 2.

The structures of periodic inspection of systems at chamber 1 and chamber 2 are displayed in Figure 1 and Figure 2, respectively.

Assume that the lifetimes of test units are iid as Weibull with shape parameter δ known and scale parameter μ unknown. That is, the pdf, cdf, and reliability function of lifetime T at normal operating condition are:

$$f(t) = \mu\delta(\mu t)^{\delta-1}e^{-(\mu t)^{\delta}},\ t \geq 0\ , \tag{1}$$

$$F(t) = 1 - e^{-(\mu t)^{\delta}},\ t \geq 0\ , \tag{2}$$

and

$$\overline{F}(t) = e^{-(\mu t)^{\delta}},\ t \geq 0\ , \tag{3}$$

respectively, and the pdf, cdf, and reliability function of lifetime T at accelerated operating condition are given as

$$f(t) = A\mu\delta(\mu t)^{\delta-1}e^{-A(\mu t)^{\delta}},\ t \geq 0\ , \tag{4}$$

$$F(t) = 1 - e^{-A(\mu t)^{\delta}},\ t \geq 0\ , \tag{5}$$

and

$$\overline{F}(t) = e^{-A(\mu t)^{\delta}},\ t \geq 0\ , \tag{6}$$

Assuming transformation times $y_j = {t_j}^{\delta},\ j = 1, \ldots, k+1$, are iid from an exponential distribution with failure rate $\lambda = \mu^{\delta}$.

So the Weibull distribution reduces to the exponential distribution with pdf, cdf, and reliability function under accelerated condition as:

$$f(y) = A\lambda e^{-A\lambda y},\ y > 0\ , \tag{7}$$

$$F(y) = 1 - e^{-A\lambda y},\ y > 0\ , \tag{8}$$

and

$$\overline{F}(y) = e^{-A\lambda y},\ y > 0\ , \tag{9}$$

respectively. At normal operating condition, $A = 1$ in the above equations. The probability of failure of a system due to component 1 under normal operating condition in $(y_{j-1}, y_j]\,, j = 1, 2, \ldots k+1$,

$$P_{1j} = \int_{y_{j-1}}^{y_j} F_2(y)\, f_1(y)\, dy\ ,$$

giving

$$P_{1j} = e^{-\lambda_1 y_{j-1}} - e^{-\lambda_1 y_j} + \frac{\lambda_1 \left(e^{-(\lambda_1+\lambda_2)y_j} - e^{-(\lambda_1+\lambda_2)y_{j-1}}\right)}{\lambda_1 + \lambda_2},\ for\ j = 1, 2, \ldots,\ k+1. \tag{10}$$

The probability of failure of a system due to component 2 under normal operating condition in $(y_{j-1}, y_j]\,, j = 1, 2, \ldots k+1$,

$$P_{2j} = \int_{y_{j-1}}^{y_j} F_1(y)\, f_2(y)\, dy\ ,$$

giving

$$P_{2j} = e^{-\lambda_2 y_{j-1}} - e^{-\lambda_2 y_j} + \frac{\lambda_2 \left(e^{-(\lambda_1+\lambda_2)y_j} - e^{-(\lambda_1+\lambda_2)y_{j-1}}\right)}{\lambda_1 + \lambda_2},\ for\ j = 1, 2, \ldots,\ k+1. \tag{11}$$

The probability of failure of a system due to component 1 under accelerated condition in $(y_{j-1}, y_j]\,, j = 1, 2, \ldots k+1$,

$$P_{A1j} = \int_{y_{j-1}}^{y_j} F_2(y)\, f_1(y)\, dy\ ,$$

giving

$$P_{A1j} = e^{-A\lambda_1 y_{j-1}} - e^{-A\lambda_1 y_j} + \frac{\lambda_1 \left(e^{-A(\lambda_1+\lambda_2)y_j} - e^{-A(\lambda_1+\lambda_2)y_{j-1}}\right)}{\lambda_1 + \lambda_2}, for\ j = 1, 2, \ldots,\ k+1. \tag{12}$$

The probability of failure of a system due to component 2 under accelerated condition in $(y_{j-1}, y_j]\,, j = 1, 2, \ldots k+1$,

$$P_{A2j} = \int_{y_{j-1}}^{y_j} F_1(y)\, f_2(y)\, dy\ ,$$

giving

$$P_{A2j} = e^{-A\lambda_2 y_{j-1}} - e^{-A\lambda_2 y_j} + \frac{\lambda_2 \left(e^{-A(\lambda_1+\lambda_2)y_j} - e^{-A(\lambda_1+\lambda_2)y_{j-1}}\right)}{\lambda_1+\lambda_2},\ for\ j = 1, 2, \ldots,\ k{+}1. \quad (13)$$

At stress level, the grouped data $w_{ij}, j = 1, 2, \ldots, k+1$ are multinomially distributed with parameters n_i and $P_{ij}, j = 1, 2, \ldots, k+1$. The likelihood function of parallel system for independent components is then given by

$$L\,(\lambda_1,\ \lambda_2,\ A) = \ L_1 L_2, \quad (14)$$

where L_1 is the likelihood corresponding to systems' failures in chamber 1 (normal operating condition) is:

$$\mathrm{L}_1 = n_1! \left[\prod_{j=1}^{k+1} (w_{1j} + w_{2j})!\right]^{-1} \left[\prod_{j=1}^{k+1} P_{1j}^{w_{1j}} P_{2j}^{w_{2j}}\right],$$

L_2 is the likelihood systems' failures in chamber 2 (accelerated operating condition) is:

$$L_2 = n_2! \left[\prod_{j=1}^{k+1} (\ w_{A1j} + w_{A2j})!\right]^{-1} \left[\prod_{j=1}^{k+1} P_{A1j}^{w_{A1j}} P_{A2j}^{w_{A2j}}\right],$$

$$L = L_1 L_2\ ,$$

From properties of Multinomial Distribution,

$w_{1j} + w_{2j} = n_{1j}$, and, $w_{A1j} + w_{A2j} = n_{2j}$,

$\sum_{j=1}^{k+1} (P_{1j} + (P_{2j}) = 1$, and, $\sum_{j=1}^{k+1} (P_{A1j} + P_{A2j}) = 1\ ,$

Thus, the log-likelihood function is a function of unknown parameters $\lambda_1,\ \lambda_2,\ and\ A$ given as:

$$lnL(\lambda_1,\ \lambda_2,\ A) = lnL_1 + lnL_2.$$

$$lnL(\lambda_1,\ \lambda_2,\ A)$$

$$= \ln \left\{ \left[n_1! \left(\prod_{j=1}^{k+1} (w_{1j} + w_{2j})! \right)^{-1} \left(\prod_{j=1}^{k+1} P_{1j}^{w_{1j}} P_{2j}^{w_{2j}} \right) \right] \cdot \right.$$

$$\left. \left[n_2! \left(\prod_{j=1}^{k+1} (\ w_{A1j} + w_{A2j})! \right)^{-1} \left(\prod_{j=1}^{k+1} P_{A1j}^{w_{A1j}} P_{A2j}^{w_{A2j}} \right) \right] \right\}$$

$$= ln\,(n_1!)\ + ln\,(n_2!)\ - \sum_{j=1}^{k+1} \ln(w_{1j} + w_{2j})! - \sum_{j=1}^{k+1} \ln\,(w_{A1j} + w_{A2j})\ !$$

$$+ \sum_{j=1}^{k+1} w_{1j} \ln(P_{1j}) + \sum_{j=1}^{k+1} w_{2j} \ln\,(P_{2j})\ + \sum_{j=1}^{k+1} w_{A1j} \ln(P_{A1j}) + \sum_{j=1}^{k+1} w_{A2j} \ln(P_{A2j})$$

$$= \sum_{j=1}^{k+1} w_{1j} \ln(P_{1j}) + \sum_{j=1}^{k+1} w_{2j} \ln(P_{2j}) + \sum_{j=1}^{k+1} w_{A1j} \ln(P_{A1j}) + \sum_{j=1}^{k+1} w_{A2j} \ln(P_{A2j}) + C, \qquad (15)$$

where C is a constant independent of parameters. Maximum Likelihood (ML Estimates of $\lambda_1,\ \lambda_2,\ and\ A$ are obtained by maximizing $lnL\,(\lambda_1,\ \lambda_2, A)$ using NMaximize option of Mathematica 10 software package.

If $\widehat{\lambda}$ is the ML estimate for $\lambda = \mu^{\delta}$ in the transformed problem, then $\widehat{\mu} = \widehat{\lambda}^{1/\delta}$ is the ML estimate for μ.

4. Fisher information matrix

The Fisher Information Matrix (Nelson, 1977) is given by,

$$F = n\rho F_1 + n\,(1-\rho)\,F_2,$$

where,

$$F_i = \begin{bmatrix} E\left[-\frac{\partial^2 lnL_i}{\partial {\lambda_1}^2}\right] & E\left[-\frac{\partial^2 lnL_i}{\partial \lambda_1 \partial \lambda_2}\right] & E\left[-\frac{\partial^2 lnL_i}{\partial \lambda_1 \partial A}\right] \\ E\left[-\frac{\partial^2 lnL_i}{\partial \lambda_1 \partial \lambda_2}\right] & E\left[-\frac{\partial^2 lnL_i}{\partial {\lambda_2}^2}\right] & E\left[-\frac{\partial^2 lnL_i}{\partial \lambda_2 \partial A}\right] \\ E\left[-\frac{\partial^2 lnL_i}{\partial \lambda_1 \partial A}\right] & E\left[-\frac{\partial^2 lnL_i}{\partial \lambda_2 \partial A}\right] & E\left[-\frac{\partial^2 lnL_i}{\partial A^2}\right] \end{bmatrix},\ i = 1, 2. \qquad (16)$$

5. Optimization problem

The optimum plan consists in determining optimum allocation ρ and optimal inspection times using D-optimality which consists in maximizing the determinant of Fisher information matrix which is the same as the reciprocal of the asymptotic variance-covariance matrix. The volume of the asymptotic joint confidence region of parameters, say, (μ, δ) is proportional to the square root of the determinant of the inverse of the Fisher information matrix, $|F^{-1}|^{1/2}$, at a fixed confidence level. In other words, it is inversely proportional to $|F|^{1/2}$. Consequently, a smaller value of the determinant would correspond to a higher (joint) precision of the estimators of μ, δ. The D-optimality criterion is therefore preferred to other optimality criteria existing in the literature such as A-optimality criterion, C-optimality or variance-optimality criterion. Thus, the optimization problem for determining optimal allocation and two inspection points $y1$ and $y2$ with $y3$ specified is:

Maximize $|F|$

$$s.t. 0 < \rho < 1, 0 < y_1 < y_2 < y_3. \qquad (17)$$

Using transformed problem $t_j = {y_j}^{\delta}$, j =1, 2, 3, we get inspection points t_1 and t_2 with t_3 specified.

6. Reliability demonstration testing

In the present section, reliability demonstration testing for the mean life of the components and the system comprising these components has been presented. The acceptance of the null hypothesis in Section 6.1 and Section 6.2 corresponds to a demonstration of mean life of at least the specified value with confidence $100((1-\alpha_1)\%)$, where α_1 is the probability of committing Type-I error.

6.1. Reliability demonstration for components

For component $i, i = 1, 2$, the objective is to test:

$$H_{0i} : \frac{1}{\mu_i} \geq \frac{1}{\mu_{i0}}, versus H_{1i} : \frac{1}{\mu_i} < \frac{1}{\mu_{i0}}, \tag{18}$$

where $\mu_i = {\lambda_i}^{1/\delta}$ and $\mu_{i0} = {\lambda_{i0}}^{1/\delta}$ (18) is equivalent to testing

$H_{0i} : \frac{1}{\mu_i^{\delta}} \geq \frac{1}{\mu_{i0}^{\delta}}$, versus $H_{1i} : \frac{1}{\mu_i^{\delta}} < \frac{1}{\mu_{i0}^{\delta}}$,

or,

$H_{0i} : \frac{1}{\lambda_i} \geq \frac{1}{\lambda_{i0}}$, versus $H_{1i} : \frac{1}{\lambda_i} < \frac{1}{\lambda_{i0}}$,

where $\frac{1}{\lambda_{i0}}$ is a specified mean life of component i, and $\frac{1}{\widehat{\lambda}_i}$ is estimated value of $\frac{1}{\lambda_i}$. Under H_{0i}, the test statistic ((Nelson, 1977))

$$T_i = \frac{\frac{1}{\widehat{\lambda}_i} - \frac{1}{\lambda_{i0}}}{\sqrt{est.var\left(1/\ \widehat{\lambda}_i\right)}} \sim N(0,1) \ as \ n \ \to \infty \ . \tag{19}$$

6.2. Reliability demonstration test for system

The time to failure of a two-component parallel structure is not Weibull distributed, even if both components have Weibull distributed times to failure.

The MTTF (mean time to failure) of the 2-component parallel system is,

$$MTTF = \int_0^{\infty} R(t)\, dt = \frac{\Gamma\left(\frac{1}{\delta}\right)}{\delta}\left(\frac{1}{{\lambda_1}^{1/\delta}} + \frac{1}{{\lambda_2}^{1/\delta}} - \frac{1}{(\lambda_1+\lambda_2)^{1/\delta}}\right).$$

Under H_{03}: MTTF$\geq MTTF_0$ versus H_{13} : MTTF $< MTTF_0$ the test statistic (Nelson, 1977),

$$T_s = \frac{(Est.MTTF - MTTF_0)}{\sqrt{Est.variance\ of\ Est.MTTF}} \sim N(0,1) \ as \ n \to \infty, \tag{20}$$

where

$$\text{Est.}MTTF = \frac{\Gamma\left(\frac{1}{\delta}\right)}{\delta}\left(\frac{1}{\widehat{\lambda}_1^{1/\delta}} + \frac{1}{\widehat{\lambda}_2^{1/\delta}} - \frac{1}{\left(\widehat{\lambda}_1+\widehat{\lambda}_2\right)^{1/\delta}}\right) = h(say).$$

$$h_1 = \frac{dh}{d\lambda_1},\ h_2 = \frac{dh}{d\lambda_2},\ h_3 = \frac{dh}{dA}\ ,$$

$$h = \begin{bmatrix} h_1 \\ h_2 \\ h_3 \end{bmatrix}, h^T = [h_1\ h_2\ h_3]$$

$\text{Est.}variance\ of\ Est.MTTF = h^T F^{-1} h.$

7. Numerical example

Let the total number of inspection k = 5, $\lambda_1 = 0.1$ and $\lambda_2 = 0.5$, inspection times $y_0 = 0,\ y_5 = 30,\ y_6 = \infty$, acceleration factor $A = 1.7$, for $n = 35$ items put to test. The optimal allocation $\rho = 0.521892$, and times inspection times $y_1 = 2$,$y_2 = 4$,$y_3 = 8,$ and $y_4 = 16$. The simulated data is depicted in Table 1.1.

Table 1: Simulated data

Intervals	Chamber 1		Chamber 2	
	Component 1	Component 2	Component 1	Component 2
(0, 2]	3	1	0	3
(2, 4]	3	0	2	5
(4, 8]	2	3	0	2
(8,16]	3	0	0	4
(16,30]	2	0	0	1
(30, ∞)	1	0	0	0

7.1. Hypothesis testing problem for component 1

$H_{01} : \frac{1}{\lambda_1} \geq 10$ versus $H_{11} : \frac{1}{\lambda_1} < 10$

Under H_{01}, the test statistic:

$$T_1 = -0.786602$$

Thus accept H_{01} at 5% level of significance.

7.2. Hypothesis testing Problem for component 2

$H_{02} : \frac{1}{\lambda_2} \geq 2$ versus $H_{12} : \frac{1}{\lambda_2} < 2$

Under H_{02}, the test statistic:

$$T_2 = 2.71621$$

Thus, accept H_{02} at 5% level of significance.

7.3. Hypothesis testing problem for the 2-component parallel system

Under H_{03}: MTTF≥ 5 versus H_{13} : MTTF < 5 the test statistic,

$$T_s = 1.04019$$

Thus, accept H_{03} at 5% level of significance. Thus, components as well as the system meet the specified reliability requirements.

8. Conclusion

The paper deals with reliability demonstration tests for a two-component parallel system with subject to CSPALT under periodic inspection using Weibull life distribution. The optimal plan consists in determining optimal allocation and optimal inspection times using D-optimality criterion. The method proposed is illustrated using a numerical example.

The future scope of RDTs under normal operating or accelerated conditions is vast and still unexploited.

These tests can be also constructed for two-component parallel systems with dependent components. RDTs can also be formulated for other reliability systems such as series-parallel, parallel-series, and k-out-of-n system, etc. Conducting RDTs for small sample size for various reliability systems is still an open problem. Parametric approach has been used in the present paper. The tests can also be formulated using Non-parametric and Bayesian approaches.

Acknowledgements

The authors are grateful to the reviewer(s) for the valuable comments.

References

Ahmad, N. and Islam, A. (1996). Optimal accelerated life test designs for burr type XII distributions under periodic inspection and type I censoring. *Naval Research Logistics (NRL)*, **43**, 1049–1077.

Ahmad, N., Islam, A., Kumar, R., and Tuteja, R. (1994). Optimal design of accelerated life test plans under periodic inspection and type I censoring: the case of rayleigh failure law. *South African Statistical Journal*, **28**, 93–101.

Ahmad, N., Islam, A., and Salam, A. (2006). Analysis of optimal accelerated life test plans for periodic inspection: the case of exponentiated weibull failure model. *International Journal of Quality & Reliability Management*, **23**, 1019–1046.

Archer, N. P. (1982). Maximam likelihood estimation with welbull models when the data are grouped. *Communications in Statistics-Theory and Methods*, **11**, 199–207.

Chen, W. H., Gao, L., Pan, J., Qian, P., and He, Q. C. (2018). Design of accelerated life test plans overview and prospect. *Chinese Journal of Mechanical Engineering*, **31**, 1–15.

Ehrenfeld, S. (1962). Some experimental design problems in attribute life testing. *Journal of the American Statistical Association*, **57**, 668–679.

Flygare, M. E., Austin, J. A., and Buckwalter, R. M. (1985). Maximum likelihood estimation for the 2-parameter weibull distribution based on interval-data. *IEEE Transactions on Reliability*, **34**, 57–59.

Islam, A. and Ahmad, N. (1994). Optimal design of accelerated life tests for the weibull distribution under periodic inspection and type I censoring. *Microelectronics Reliability*, **34**, 1459–1468.

Kulldorff, G. (1961). *Contributions to the Theory of Estimation from Grouped and Partially Grouped Samples*. Almqvist and Wiksell.

Meeker, W. Q. (1986). Planning life tests in which units are inspected for failure. *IEEE Transactions on Reliability*, **35**, 571–578.

Nelson, W. (1977). Optimum demonstration tests with grouped inspection data from an exponential distribution. *IEEE Transactions on Reliability*, **26**, 226–231.

Nelson, W. B. (2009). *Accelerated Testing: Statistical Models, Test Plans, and Data Analysis*. John Wiley & Sons.

Seo, S. K. and Yum, B. J. (1991). Accelerated life test plans under intermittent inspection and type-I censoring: The case of weibull failure distribution. *Naval Research Logistics*, **38**, 1–22.

Shaw, M. (1987). Weibull analysis of component failure data from accelerated testing. *Reliability Engineering*, **19**, 237–243.

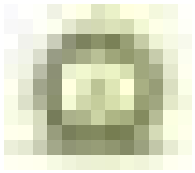
Srivastava, P. W. (2017). *Optimum Accelerated Life Testing Models with Time-varying Stresses*. World Scientific.

Wei, D. and Bau, J. J. (1987). Some optimal designs for grouped data in reliability demonstration tests. *IEEE Transactions on Reliability*, **36**, 600–604.

Yang, G. (2007). *Life Cycle Reliability Engineering*. John Wiley & Sons.

Yum, B.-J. and Choi, S. C. (1989). Optimal design of accelerated life tests under periodic inspection. *Naval Research Logistics*, **36**, 779–795.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 22, No. 1, 2024 (New Series), pp 229–240
http://www.ssca.org.in/journal



Inference on Stress-Strength Reliability for Lomax Exponential Distribution

Parameshwar V. Pandit and Kavitha N.
Department of Statistics, Bangalore University, Bengaluru-560056

Received: 01 June 2023; Revised: 05 July 2023; Accepted: 28 July 2023

Abstract

In this paper, the reliability estimation of single component stress-strength model is studied with strength(X) and stress(Y) of the component follow Lomax exponential distribution. The maximum likelihood and Bayesian estimation methods are applied to derive estimators of reliability. The Bayesian estimators for reliability are constructed under different loss functions such as squared error and linex loss functions with non-informative and gamma priors using Lindley's approximation technique. The simulation experiment is conducted to estimate the mean squared error of the estimators which enable the comparison of different estimators. The construction of asymptotic confidence interval of reliability is also constructed. The real data analysis is done to illustrate the developed procedures.

Key words: Lomax exponential distribution (LED); Stress-strength reliability; maximum likelihood estimation; Bayesian inference; Lindley's approximation technique.

1. Introduction

In the recent years there has been growing interest in defining new generators for univariate continuous distributions by introducing one or more additional shape parameters to the baseline distribution. Some well-known generators are beta-G and gamma-G due to Eugene and Famoye (2002) and Zografos and Balakrishnan (2009), respectively. Torabi and Montazeri (2014) introduced the logistic-G family. Recently, Cordeiro and Pescim (2014) studied a new family of distributions based on the Lomax distribution. The probability density function (pdf) and cumulative distribution function (cdf) of Lomax-G family with two additional parameters α and β are given by

$$f(x) = \alpha\beta^{\alpha} g(x) \left([1 - G(x)] \{\beta - \log [1 - G(x)]\}^{\alpha+1} \right)^{-1}, \; x > 0, \; \alpha, \; \beta > 0$$

and

$$F(x) = 1 - \beta^{\alpha} (\beta - \log [1 - G(x)])^{-\alpha}, \; x > 0, \; \alpha, \; \beta > 0$$

where g(x) and G(x) are the pdf and cdf of parent distribution. The parameters α and β are the shape and scale parameters of the distribution, respectively.

Corresponding Author: Parameshwar V. Pandit
Email: panditpv12@gmail.com

In this paper, the estimation of stress-strength reliability is considered when X and Y are independently distributed Lomax exponential distribution (LED) which was introduced by Ieren and Kuhe (2018) which is a new generalization of exponential distribution. The LED is constructed by Cordeiro and Pescim (2014) using Lomax-G family. The general form of cumulative distribution function (cdf) and probability density function (pdf) of Lomax G-family with baseline distribution G is given below:
The cumulative distribution function (cdf) and probability density function (pdf) are given by

$$f(x) = \alpha\beta^{\alpha} g(x) \left([1 - G(x)] \{\beta - \log [1 - G(x)]\}^{\alpha+1} \right)^{-1}, \; x > 0, \; \alpha, \; \beta > 0$$

and

$$F(x) = 1 - \beta^{\alpha} (\beta - \log [1 - G(x)])^{-\alpha}, \; x > 0, \; \alpha, \; \beta > 0$$

where g(x) and G(x) are the pdf and cdf of baseline distribution. The parameters α and β are the shape and scale parameters of the distribution, respectively.
In this paper, the problem of estimation of stress-strength reliability is considered when X and Y are independently distributed Lomax exponential distribution (LED) due to Ieren and Kuhe (2018) with exponential distribution as baseline distribution. Then, the cdf and pdf of LED are given by

$$F(x) = 1 - \beta^{\alpha} (\beta + \lambda x)^{-\alpha}, \; \alpha > 0, \; x > 0, \; \beta > 0, \; \lambda > 0$$

and

$$f(x) = \alpha\lambda\beta^{\alpha} (\beta + \lambda x)^{-(\alpha+1)}, \; \alpha > 0, \; x > 0, \; \beta > 0, \; \lambda > 0$$

It is denoted by LED(α, β, λ).
Some particular cases of Lomax exponential distribution are as given below:

1. If λ=1 and β=1, then LED is Pareto type-II distribution.
2. LED is Lomax standard exponential distribution, when λ=1.
3. When β=1, LED is generalized Pareto distribution.

The main focus of the paper is to study the problem of estimating stress-strength reliability when stress and strength variables follow LED. In the literature several authors have studied the estimation of stress-strength reliability for different life time distributions. Awad and Gharraf (1986) considered the estimation of R for Burr distribution. Mokhlis (2005) and Panahi and Asadi (2011) estimated the stress-strength reliability for Burr type-III and Lomax distributions respectively. Abravesh and Mostafaiy (2019) studied the classical and Bayesian estimation of stress-strength reliability based on type II censored sample from Pareto distribution.
The rest of the paper is organised as below. Section 2 deals with derivation of stress-strength reliability when strength X and stress Y follow LED. Maximum likelihood estimation of R and its asymptotic confidence intervals are given in Section 3. In Section 4, the Bayesian estimator of R is presented. The real data analysis is considered in Section 5. Section 6 contains a simulation study and the conclusions are presented in Section 7.

2. Stress-strength reliability

Let X and Y be two independent random variables having LED(α_1, β, λ) and LED(α_2, β, λ) respectively.
Then, the stress-strength reliability is given by

$$\begin{aligned} R &= P(X > Y) \\ &= \int_0^\infty F(x) f(x) dx \\ &= \frac{\alpha_2}{\alpha_1 + \alpha_2} \end{aligned} \tag{1}$$

3. Maximum likelihood estimation (MLE) of reliability

Let $\underline{X} = (X_1, X_2, ..., X_n)$ and $\underline{Y} = (Y_1, Y_2, ..., Y_m)$ be independent random samples from LED(α_1, β, λ) and LED(α_2, β, λ), respectively. Then the likelihood function of α_1, α_2, β and λ given $\left(\underline{x}, \underline{y}\right)$ is

$$L\left(\alpha_1, \alpha_2, \beta, \lambda | \underline{x}, \underline{y}\right) = \prod_{i=1}^{n} \alpha_1 \beta^{\alpha_1} \lambda (\beta + \lambda x_i)^{-(\alpha_1+1)} \prod_{j=1}^{m} \alpha_2 \beta^{\alpha_2} \lambda (\beta + \lambda y_j)^{-(\alpha_2+1)} \tag{2}$$

and the log-likelihood function is

$$\begin{aligned} \log L = n \log \alpha_1 + m \log \alpha_2 + (n\alpha_1 + m\alpha_2) \log \beta + (n+m) \log \lambda - (\alpha_1 + 1) \textstyle\sum_{i=1}^{n} \log (\beta + \lambda x_i) - \\ (\alpha_2 + 1) \textstyle\sum_{j=1}^{m} \log (\beta + \lambda y_j) \end{aligned} \tag{3}$$

The likelihood equations are

$$\frac{n}{\alpha_1} + n \log \beta - \sum_{i=1}^{n} \log (\beta + \lambda x_i) = 0 \tag{4}$$

$$\frac{m}{\alpha_2} + m \log \beta - \sum_{j=1}^{m} \log (\beta + \lambda y_j) = 0 \tag{5}$$

$$\frac{n\alpha_1 + m\alpha_2}{\beta} - (\alpha_1 + 1) \sum_{i=1}^{n} \left(\frac{1}{(\beta + \lambda x_i)} \right) - (\alpha_2 + 1) \sum_{j=1}^{m} \left(\frac{1}{(\beta + \lambda y_j)} \right) = 0 \tag{6}$$

and

$$\frac{n+m}{\lambda} - (\alpha_1 + 1) \sum_{i=1}^{n} \left(\frac{x_i}{(\beta + \lambda x_i)} \right) - (\alpha_2 + 1) \sum_{j=1}^{m} \left(\frac{y_j}{(\beta + \lambda y_j)} \right) = 0 \tag{7}$$

The above equations do not yield solution in closed form. Hence, a popular iterative technique, namely, Newton Raphson technique is used.
Using the invariance property of MLE, the MLE of R is given by

$$\hat{R} = \frac{\hat{\alpha}_2}{\hat{\alpha}_1 + \hat{\alpha}_2}.$$

3.1. Asymptotic distribution of R

Under general regularity conditions, the asymptotic distribution of $\left(\underline{\hat{\theta}} - \underline{\theta}\right)$ is multivariate $N_{p+4}\left(\underline{0},\, I\left(\underline{\theta}\right)^{-1}\right)$ distribution, where $I\left(\underline{\theta}\right)$ is the expected information matrix and $\underline{\theta} = [\alpha_1, \alpha_2, \beta, \lambda]^{\intercal}$. Here, $I\left(\underline{\theta}\right)^{-1}$can be approximated by the inverse of observed information matrix $I\left(\underline{\hat{\theta}}\right)^{-1}$ evaluated at $\hat{\theta}$. This distribution is used to construct the 100(1-α)% confidence interval for each parameters.
Asymptotic confidence intervals of the parameters are given by

$$\alpha_1 \pm Z_{\frac{\alpha}{2}}\sqrt{I_{11}^{-1}},\ \alpha_2 \pm Z_{\frac{\alpha}{2}}\sqrt{I_{22}^{-1}},$$

$$\beta \pm Z_{\frac{\alpha}{2}}\sqrt{I_{33}^{-1}}\ and\ \lambda \pm Z_{\frac{\alpha}{2}}\sqrt{I_{44}^{-1}}$$

The asymptotic confidence interval of R is

$$\hat{R} \pm Z_{\frac{\alpha}{2}}\sqrt{AV\left(\hat{R}\right)},$$

where

$$AV\left(\hat{R}\right) = \left[\frac{\partial R}{\partial \alpha_1}\right]^2 I_{11}^{-1} + \left[\frac{\partial R}{\partial \alpha_2}\right]^2 I_{22}^{-1}.$$

4. Bayesian estimation of R

In this section, the Bayesian estimation of R under different loss functions and priors is presented. The non-informative and gamma priors are considered to obtain Bayes estimator of R. The prior distribution of α_1, α_2, β and λ are gamma (c_1, d_1), gamma (c_2, d_2), gamma (c_3, d_3) and gamma (c_4, d_4), respectively.
The joint prior distribution of α_1, α_2, β and λ is given by

$$g_1\left(\alpha_1,\, \alpha_2,\, \beta,\, \lambda\right) = g\left(\alpha_1\right)\, g\left(\alpha_2\right)\, g\left(\beta\right)\, g\left(\lambda\right) \tag{8}$$

where

$$g\left(\alpha_1\right) = \frac{d_1^{c_1}}{\Gamma c_1}\exp\left(-d_1\alpha_1\right)\, \alpha_1^{c_1-1},\ \alpha_1 > 0,\ c_1,\ d_1 > 0,$$

$$g\left(\alpha_2\right) = \frac{d_2^{c_2}}{\Gamma c_2}\exp\left(-d_2\alpha_2\right)\alpha_2^{c_2-1},\ \alpha_2 > 0,\ c_2,\ d_2 > 0,$$

$$g\left(\beta\right) = \frac{d_3^{c_3}}{\Gamma c_3}\exp\left(-d_3\beta\right)\beta^{c_3-1},\ \beta > 0,\ c_3,\ d_3 > 0$$

and

$$g\left(\lambda\right) = \frac{d_4^{c_4}}{\Gamma c_4}\exp\left(-d_4\lambda\right)\, \lambda^{c_4-1},\ \lambda > 0,\ c_4,\ d_4 > 0$$

If, $c_1 = c_2 = c_3 = c_4 = d_1 = d_2 = d_3 = d_4 = 0$, then it reduces to non-informative prior.
The posterior distribution is

$$\pi\left(\alpha_1,\, \alpha_2,\, \beta,\, \lambda\right) = \frac{L\left(\alpha_1,\, \alpha_2,\, \beta,\, \lambda\right)\, g\left(\alpha_1,\, \alpha_2,\, \beta,\, \lambda\right)}{\int_0^\infty \int_0^\infty \int_0^\infty \int_0^\infty L\left(\alpha_1,\, \alpha_2,\, \beta,\, \lambda\right)\, g\left(\alpha_1,\, \alpha_2,\, \beta,\, \lambda\right)\, d\alpha_1\, d\alpha_2\, d\beta\, d\lambda}$$

$$\pi\left(\alpha_1,\,\alpha_2,\,\beta,\,\lambda\right)=\frac{G}{\int_0^\infty\int_0^\infty\int_0^\infty\int_0^\infty G\,d\alpha_1\,d\alpha_2\,d\beta\,d\lambda} \tag{9}$$

where

$$G=\exp\left(-d_1\alpha_1\right)\alpha_1^{n+c_1-1}\exp\left(-d_2\alpha_2\right)\alpha_2^{m+c_2-1}\exp\left(-d_3\beta\right)\exp\left(-d_4\lambda\right)\beta^{n\alpha_1+m\alpha_2+c_3-1}$$
$$\lambda^{n+m+c_4-1}\prod_{i=1}^{n}\left(\beta+\lambda x_i\right)^{-(\alpha_1+1)}\prod_{j=1}^{m}\left(\beta+\lambda y_j\right)^{-(\alpha_2+1)}$$

The posterior distribution of R is non-tractable. Hence, the Lindley's approximation technique is used to derive Bayes estimator of R.
For four parameter case, the Lindley's approximation to Bayes estimator of R under squared error loss function is given by

$$\begin{aligned}\hat{R}_S=u+&\left(u_1a_1+u_2a_2+u_3a_3+u_4a_4+a_5+a_6\right)+\\&\tfrac{1}{2}\left[A\left(u_1\sigma_{11}+u_2\sigma_{12}+u_3\sigma_{13}+u_4\sigma_{14}\right)\right]+\\&\tfrac{1}{2}\left[B\left(u_1\sigma_{21}+u_2\sigma_{22}+u_3\sigma_{23}+u_4\sigma_{24}\right)\right]+\\&\tfrac{1}{2}\left[C\left(u_1\sigma_{31}+u_2\sigma_{32}+u_3\sigma_{33}+u_4\sigma_{34}\right)\right]+\\&\tfrac{1}{2}\left[D\left(u_1\sigma_{41}+u_2\sigma_{42}+u_3\sigma_{43}+u_4\sigma_{44}\right)\right],\end{aligned} \tag{10}$$

where,$u=\hat{R}$, u_i, $i=1,\,2,\,3,\,4$ and u_{ij}, $i,\,j=1,\,2,\,3,\,4$ are the first and second order derivatives of R, σ_{ij}, $i,\,j=1,\,2,\,3,\,4$ is the (i, j)th element in the inverse of the matrix [-L_{ij}], L_{ij} and L_{ijk} are the second and third order derivatives of log-likelihood function and ρ_i, $i=1,\,2,\,3,\,4$ is the first order differentiation of log of prior with respect to α_1, α_2, β and λ.
Here,

$$a_i=\rho_1\sigma_{i1}+\rho_2\sigma_{i2}+\rho_3\sigma_{i3}+\rho_4\sigma_{i4},\ i=1,\,2,\,3,\,4.\,,$$

$$a_5=u_{12}\sigma_{12}+u_{13}\sigma_{13}+u_{14}\sigma_{14}+u_{23}\sigma_{23}+u_{24}\sigma_{24}\,,$$

$$a_6=\frac{1}{2}\left(u_{11}\sigma_{11}+u_{22}\sigma_{22}+u_{33}\sigma_{33}+u_{44}\sigma_{44}\right),$$

$$\begin{aligned}A=&\sigma_{11}L_{111}+2\sigma_{12}L_{121}+2\sigma_{13}L_{131}+2\sigma_{14}L_{141}+2\sigma_{23}L_{231}+2\sigma_{24}L_{241}+\\&\sigma_{22}L_{221}+\sigma_{33}L_{331}+2\sigma_{34}L_{341}+\sigma_{44}L_{441}\,,\end{aligned}$$

$$\begin{aligned}B=&\sigma_{11}L_{112}+2\sigma_{12}L_{122}+2\sigma_{13}L_{132}+2\sigma_{14}L_{142}+2\sigma_{23}L_{232}+2\sigma_{24}L_{242}+\\&\sigma_{22}L_{222}+\sigma_{33}L_{332}+2\sigma_{34}L_{342}+\sigma_{44}L_{442}\,,\end{aligned}$$

$$\begin{aligned}C=&\sigma_{11}L_{113}+2\sigma_{12}L_{123}+2\sigma_{13}L_{133}+2\sigma_{14}L_{143}+2\sigma_{23}L_{233}+2\sigma_{24}L_{243}+\\&\sigma_{22}L_{223}+\sigma_{33}L_{333}+2\sigma_{34}L_{343}+\sigma_{44}L_{443}\,,\end{aligned}$$

and

$$\begin{aligned}D=&\sigma_{11}L_{114}+2\sigma_{12}L_{124}+2\sigma_{13}L_{134}+2\sigma_{14}L_{144}+2\sigma_{23}L_{234}+2\sigma_{24}L_{244}+\\&\sigma_{22}L_{224}+\sigma_{33}L_{334}+2\sigma_{34}L_{344}+\sigma_{44}L_{444}\,.\end{aligned}$$

According to our case,

$$\begin{aligned}\hat{R}_S=u+&\left(u_1a_1+u_2a_2+a_6\right)+\tfrac{1}{2}\left[A\left(u_1\sigma_{11}\right)+B\left(u_2\sigma_{22}\right)\right]+\\&\tfrac{1}{2}\left[C\left(u_1\sigma_{31}+u_2\sigma_{32}\right)+D\left(u_1\sigma_{41}+u_2\sigma_{42}\right)\right],\end{aligned} \tag{11}$$

where

$$a_1=\rho_1\sigma_{11}+\rho_3\sigma_{13}+\rho_4\sigma_{14},\ a_2=\rho_2\sigma_{22}+\rho_3\sigma_{23}+\rho_4\sigma_{24}$$

$$a_5=+u_{14}\sigma_{14}+u_{23}\sigma_{23}+u_{24}\sigma_{24}\,,$$

$$a_6 = \frac{1}{2}\left(u_{11}\sigma_{11} + u_{22}\sigma_{22} + u_{33}\sigma_{33} + u_{44}\sigma_{44}\right),$$

$$A = \sigma_{11}L_{111} + \sigma_{22}L_{221} + \sigma_{33}L_{331} + 2\sigma_{34}L_{341} + \sigma_{44}L_{441},$$

$$B = \sigma_{22}L_{222} + \sigma_{33}L_{332} + 2\sigma_{34}L_{342} + \sigma_{44}L_{442},$$

$$C = 2\sigma_{13}L_{133} + 2\sigma_{14}L_{143} + 2\sigma_{23}L_{233} + 2\sigma_{24}L_{243} + \sigma_{33}L_{333} + 2\sigma_{34}L_{343} + \sigma_{44}L_{443},$$

$$D = 2\sigma_{13}L_{134} + 2\sigma_{14}L_{144} + 2\sigma_{23}L_{234} + 2\sigma_{24}L_{244} + \sigma_{33}L_{334} + 2\sigma_{34}L_{344} + \sigma_{44}L_{444},$$

$$u = \frac{\hat{\alpha}_2}{\hat{\alpha}_1 + \hat{\alpha}_2},\ u_1 = \frac{-\hat{\alpha}_2}{\left(\hat{\alpha}_1 + \hat{\alpha}_2\right)^2},\ u_2 = \frac{\hat{\alpha}_1}{\left(\hat{\alpha}_1 + \hat{\alpha}_2\right)^2},\ u_3 = u_4 = 0,$$

$$u_{11} = \frac{2\hat{\alpha}_2}{\left(\hat{\alpha}_1 + \hat{\alpha}_2\right)^3},\ u_{22} = \frac{-2\hat{\alpha}_1}{\left(\hat{\alpha}_1 + \hat{\alpha}_2\right)^3},\ u_{12} = u_{21} = \frac{\hat{\alpha}_2 - \hat{\alpha}_1}{\left(\hat{\alpha}_1 + \hat{\alpha}_2\right)^3},$$

$$\rho_1 = \frac{c_1 - 1}{\hat{\alpha}_1} - d_1,\ \rho_2 = \frac{c_2 - 1}{\hat{\alpha}_2} - d_2,\ \rho_3 = \frac{c_3 - 1}{\hat{\beta}} - d_3,\ \rho_4 = \frac{c_4 - 1}{\hat{\lambda}} - d_4,$$

$$L_{11} = -\frac{n}{\hat{\alpha}_1^2},\ L_{22} = -\frac{m}{\hat{\alpha}_2^2},$$

$$L_{13} = L_{31} = -\sum_{i=1}^{n}\left(\frac{1}{\hat{\beta} + \hat{\lambda}x_i}\right) + \frac{n}{\hat{\beta}},$$

$$L_{23} = L_{32} = -\sum_{j=1}^{m}\left(\frac{1}{\hat{\beta} + \hat{\lambda}y_j}\right) + \frac{m}{\hat{\beta}},$$

$$L_{14} = L_{41} = -\sum_{i=1}^{n}\left(\frac{x_i}{\hat{\beta} + \hat{\lambda}x_i}\right),$$

$$L_{24} = L_{42} = -\sum_{j=1}^{m}\left(\frac{y_j}{\hat{\beta} + \hat{\lambda}y_j}\right),$$

$$L_{33} = -\frac{(n\hat{\alpha}_1 + m\hat{\alpha}_2)}{\hat{\beta}^2} + (\hat{\alpha}_1 + 1)\sum_{i=1}^{n}\left(\frac{1}{\left(\hat{\beta} + \hat{\lambda}x_i\right)^2}\right) + (\hat{\alpha_2} + 1)\sum_{j=1}^{m}\left(\frac{1}{\left(\hat{\beta} + \hat{\lambda}y_j\right)^2}\right),$$

$$L_{44} = -\frac{(n+m)}{\hat{\lambda}^2} + (\hat{\alpha}_1 + 1)\sum_{i=1}^{n}\left(\frac{x_i^2}{\left(\hat{\beta} + \hat{\lambda}x_i\right)^2}\right) + (\hat{\alpha}_2 + 1)\sum_{j=1}^{m}\left(\frac{y_j^2}{\left(\hat{\beta} + \hat{\lambda}y_j\right)^2}\right),$$

$$L_{34} = L_{43} = (\hat{\alpha}_1 + 1)\sum_{i=1}^{n}\left(\frac{x_i}{\left(\hat{\beta} + \hat{\lambda}x_i\right)^2}\right) + (\hat{\alpha}_2 + 1)\sum_{j=1}^{m}\left(\frac{y_j}{\left(\hat{\beta} + \hat{\lambda}y_j\right)^2}\right),$$

$$L_{111} = \frac{2n}{\hat{\alpha}_1^2},\ L_{222} = \frac{2m}{\alpha_2^2},$$

$$L_{134} = L_{413} = L_{314} = L_{341} = L_{431} = L_{143} = \sum_{i=1}^{n}\left(\frac{x_i}{\left(\hat{\beta} + \hat{\lambda}x_i\right)^2}\right),$$

$$L_{244} = L_{424} = L_{442} = \sum_{j=1}^{m} \left(\frac{y_j^2}{\left(\hat{\beta} + \hat{\lambda} y_j\right)^2} \right),$$

$$L_{234} = L_{423} = L_{324} = L_{342} = L_{432} = L_{243} = \sum_{j=1}^{m} \left(\frac{y_j}{\left(\hat{\beta} + \hat{\lambda} y_j\right)^2} \right),$$

$$L_{144} = L_{414} = L_{441} = \sum_{i=1}^{n} \left(\frac{x_i^2}{\left(\hat{\beta} + \hat{\lambda} x_i\right)^2} \right),$$

$$L_{334} = L_{343} = L_{433} = -(\hat{\alpha}_1 + 1) \sum_{i=1}^{n} \left(\frac{x_i}{\left(\hat{\beta} + \hat{\lambda} x_i\right)^3} \right) - (\hat{\alpha}_2 + 1) \sum_{j=1}^{m} \left(\frac{y_j}{\left(\hat{\beta} + \hat{\lambda} y_j\right)^3} \right),$$

$$L_{344} = L_{434} = L_{443} = -(\hat{\alpha}_1 + 1) \sum_{i=1}^{n} \left(\frac{x_i^2}{\left(\hat{\beta} + \hat{\lambda} x_i\right)^3} \right) - (\hat{\alpha}_2 + 1) \sum_{j=1}^{m} \left(\frac{y_j^2}{\left(\hat{\beta} + \hat{\lambda} y_j\right)^3} \right),$$

$$L_{111} = \frac{2n}{\hat{\alpha}_1^2},\ L_{222} = \frac{2m}{\hat{\alpha}_2^2},$$

$$L_{134} = L_{413} = L_{314} = L_{341} = L_{431} = L_{143} = \sum_{i=1}^{n} \left(\frac{x_i}{\left(\hat{\beta} + \hat{\lambda} x_i\right)^2} \right),$$

$$L_{244} = L_{424} = L_{442} = \sum_{j=1}^{m} \left(\frac{y_j^2}{\left(\hat{\beta} + \hat{\lambda} y_j\right)^2} \right),$$

$$L_{234} = L_{423} = L_{324} = L_{342} = L_{432} = L_{243} = \sum_{j=1}^{m} \left(\frac{y_j}{\left(\hat{\beta} + \hat{\lambda} y_j\right)^2} \right),$$

$$L_{144} = L_{414} = L_{441} = \sum_{i=1}^{n} \left(\frac{x_i^2}{\left(\hat{\beta} + \hat{\lambda} x_i\right)^2} \right),$$

$$L_{334} = L_{343} = L_{433} = -(\hat{\alpha}_1 + 1) \sum_{i=1}^{n} \left(\frac{x_i}{\left(\hat{\beta} + \hat{\lambda} x_i\right)^3} \right) - (\hat{\alpha}_2 + 1) \sum_{j=1}^{m} \left(\frac{y_j}{\left(\hat{\beta} + \hat{\lambda} y_j\right)^3} \right),$$

$$L_{344} = L_{434} = L_{443} = -(\hat{\alpha}_1 + 1) \sum_{i=1}^{n} \left(\frac{x_i^2}{\left(\hat{\beta} + \hat{\lambda} x_i\right)^3} \right) - (\hat{\alpha}_2 + 1) \sum_{j=1}^{m} \left(\frac{y_j^2}{\left(\hat{\beta} + \hat{\lambda} y_j\right)^3} \right),$$

$$L_{133} = L_{313} = L_{331} = \sum_{i=1}^{n} \left(\frac{1}{\left(\hat{\beta} + \hat{\lambda} x_i\right)^2} \right) - \frac{n}{\hat{\beta}^2},$$

$$L_{233} = L_{323} = L_{332} = \sum_{j=1}^{m} \left(\frac{1}{\left(\hat{\beta} + \hat{\lambda} y_j\right)^2} \right) - \frac{m}{\hat{\beta}^2},$$

$$L_{333} = \frac{2\left(n\hat{\alpha}_1 + m\hat{\alpha}_2\right)}{\hat{\beta}^3} - 2\left(\hat{\alpha}_1 + 1\right) \sum_{i=1}^{n} \left(\frac{1}{\left(\hat{\beta} + \hat{\lambda} x_i\right)^3} \right) - 2\left(\hat{\alpha}_2 + 1\right) \sum_{j=1}^{m} \left(\frac{1}{\left(\hat{\beta} + \hat{\lambda} y_j\right)^3} \right),$$

and

$$L_{444} = \frac{2\left(n+m\right)}{\hat{\lambda}^3} - \left(\hat{\alpha}_1 + 1\right) \sum_{i=1}^{n} \left(\frac{2x_i^3}{\left(\hat{\beta} + \hat{\lambda} x_i\right)^3} \right) - \left(\hat{\alpha}_2 + 1\right) \sum_{j=1}^{m} \left(\frac{2y_j^3}{\left(\hat{\beta} + \hat{\lambda} y_j\right)^3} \right),$$

Under linex loss function,

$$\hat{R}_L = -\frac{1}{\delta} \log \left(\begin{array}{l} v + \left(v_1 a_1 + v_2 a_2 + a_6\right) + \frac{1}{2}\left[A\left(v_1\sigma_{11}\right) + B\left(v_2\sigma_{22}\right)\right] + \\ \frac{1}{2}\left[C\left(v_1\sigma_{31} + v_2\sigma_{32}\right) + D\left(v_1\sigma_{41} + v_2\sigma_{42}\right)\right] \end{array} \right), \tag{12}$$

where

$$v = \exp\left(-\delta \frac{\hat{\alpha}_2}{\hat{\alpha}_1 + \hat{\alpha}_2}\right),\ v_1 = \delta \exp\left(-\delta \frac{\hat{\alpha}_2}{\hat{\alpha}_1 + \hat{\alpha}_2}\right) \frac{\hat{\alpha}_2}{\left(\hat{\alpha}_1 + \hat{\alpha}_2\right)^2},$$

$$v_2 = -\delta \exp\left(-\delta \frac{\hat{\alpha}_2}{\hat{\alpha}_1 + \hat{\alpha}_2}\right) \frac{\hat{\alpha}_1}{\left(\hat{\alpha}_1 + \hat{\alpha}_2\right)^2},$$

$$v_{11} = -\delta\hat{\alpha}_2 \exp\left(-\delta \frac{\hat{\alpha}_2}{\hat{\alpha}_1 + \hat{\alpha}_2}\right) \left[\frac{\hat{\alpha}_2\left(\delta + 2\right) + 2\hat{\alpha}_1}{\left(\hat{\alpha}_1 + \hat{\alpha}_2\right)^4}\right],$$

$$v_{12} = v_{21} = \delta \exp\left(-\delta \frac{\hat{\alpha}_2}{\hat{\alpha}_1 + \hat{\alpha}_2}\right) \left[\frac{\hat{\alpha}_1^2 - \hat{\alpha}_2^2 - \delta\hat{\alpha}_1\hat{\alpha}_2}{\left(\hat{\alpha}_1 + \hat{\alpha}_2\right)^4}\right]$$

and

$$v_{22} = \delta\hat{\alpha}_1 \exp\left(-\delta \frac{\hat{\alpha}_2}{\hat{\alpha}_1 + \hat{\alpha}_2}\right) \left[\frac{\hat{\alpha}_1\left(\delta + 2\right) + 2\hat{\alpha}_2}{\left(\hat{\alpha}_1 + \hat{\alpha}_2\right)^4}\right].$$

5. Real data analysis

In this section, two real data sets are analysed to illustrate the proposed estimation methods. These data sets are initially used by Nelson (1982). The data sets represent times to breakdown of an insulating fluid between electrodes at different voltage. The failure times (in minutes) for an insulating fluid between two electrodes subject to a voltage of 34kV and 36kV are presented as data set 1 and data set 2, respectively.
Data set 1 (X): 0.19, 0.78, 0.96, 1.31, 2.78, 3.16, 4.15, 4.67, 4.85, 6.50, 7.35, 8.01, 8.27, 12.06, 31.75, 32.52, 33.91, 36.71, 72.89.
Data set 2 (Y): 0.35, 0.59, 0.96, 0.99, 1.69, 1.97, 2.07, 2.58, 2.71, 2.90, 3.67, 3.99, 5.35, 13.77, 25.50.

To check the fitness for the two data sets, -logL, Akaike information criteria (AIC), Bayesian information criteria (BIC), Akaike information criteria corrected (AICc), Kolmogrov-smirnov (K-S) and Anderson-Darling (A-D) statistics with corresponding p-values are computed and the results for both data sets are given in the Tables below.

Table 1: Estimates of the parameters with corresponding standard error and the values of -logL, AIC, AICc, BIC, K-S and A-D statistics for different distributions for data set 1

Name of the distribution	Estimates of parameters	-logL	AIC	AICc	BIC	$K-S$ $(p-value)$	$A-D$ $(p-value)$
Lomax Exponential	$\alpha = 2.0302(1.6632)$ $\beta = 5.1863(11.6778)$ $\lambda = 0.3101(0.8155)$	68.4234	140.8468	142.4468	143.683	0.12654 (0.5029)	0.32203 (0.9198)
Weibull	$\alpha = 0.7956(0.1561)$ $\beta = 0.1752(0.0380)$	69.1296	142.2592	143.0092	144.1481	0.1613 (0.336)	0.3918 (0.8552)
Exponentiated exponential	$\alpha = 0.6825(0.1941)$ $\beta = 0.0535(0.0180)$	69.3980	142.796	143.546	144.684	0.1886 (0.2292)	0.5057 (0.7388)

Table 2: Estimates of the parameters with corresponding standard error and the values of -logL, AIC, AICc, BIC, K-S and A-D statistics for different distributions for data set 2

Name of the distribution	Estimates of parameters	-logL	AIC	AICc	BIC	$K-S$ $(p-value)$	$A-D$ $(p-value)$
Lomax Exponential	$\alpha = 3.0369(2.889)$ $\beta = 8.5039(3.124)$ $\lambda = 0.8991(1.182)$	36.9792	79.9583	81.5584	77.794	0.14339 (0.4938)	0.45395 (0.7915)
Weibull	$\alpha = 0.8891(0.1635)$ $\beta = 0.2738(0.1151)$	38.0125	81.3828	83.5646	80.7989	0.1917 (0.2941)	0.6271 (0.6199)
Exponentiated exponential	$\alpha = 1.9271(0.5985)$ $\beta = 0.3875(0.0954)$	41.4606	86.9212	87.6712	88.810	0.1979 (0.2724)	1.3637 (0.2125)

The estimate of reliability using MLE is 0.7042. Bayes estimates under different priors and loss functions are presented in Table 3.

Table 3: MLE and the Bayes estimates under different loss functions with different priors.

		Bayes estimaes								
		Non informative prior			Gamma prior					
	MLE	$\hat{R}_S$	$\hat{R}_L$	$\hat{R}_{L1}$	Prior 1			Prior 2		
					$\hat{R}_S$	$\hat{R}_L$	$\hat{R}_{L1}$	$\hat{R}_S$	$\hat{R}_L$	$\hat{R}_{L1}$
λ unknown	0.7042	0.7248	0.7247	0.7248	0.7312	0.7312	0.7312	0.7489	0.7591	0.7467
λ known	0.7046	0.7253	0.7252	0.7253	0.7343	0.7344	0.7343	0.8528	0.8583	0.8477

$\hat{R}_{L1}$ refers to linex loss function with loss parameter $\delta = -0.5$.

6. Simulation study

A simulation study of 10000 observations is conducted by generating samples of different sizes such as (n, m) = (5, 5), (5, 10), (10, 10), (10, 15), (20, 20), (20, 25), (30, 30), (30, 35) and (40, 40). The true values of R which are considered under simulation study are 0.57142 and 0.47058. The parameter values of the prior distribution for squared error and linex loss functions are $c_1 = 1$, $d_1 = 0.8$, $c_2 = 2$, $d_2 = 0.4$, $c_3 = 5$, $d_3 = 0.2$, $c_4 = 1$, $d_4 = 2$ (prior1) and $c_1 = 4$, $d_1 = 3$, $c_2 = 3$, $d_2 = 0.9$, $c_3 = 5$, $d_3 = 5$, $c_4 = 1$ and $d_4 = 2$ (prior 2). The values of loss parameters under linex loss function are 0.5 and -0.5. The proposed

estimators are compared using mean squared error (MSE) criteria. The MLEs and Bayes estimates with corresponding MSEs are given in the Tables given in annexure.

7. Conclusions

The estimation of stress-strength reliability (R) is considered, when stress and strength variables follow LED. The maximum likelihood and Bayesian estimation methods are used to estimate stress-strength reliability. MLEs are derived. Bayes estimators under different loss functions such as squared error and linex loss functions with gamma and non-informative priors are obtained. The Lindley's approximation technique is used to approximate the Bayes estimator of R. The real data analysis is conducted to illustrate the developed estimation procedures. A simulation experiment is conducted to study the performance of estimators which are derived in the paper and it reveals that Bayes estimator with non-informative prior is better when compared to MLEs. However, the Bayes estimator with gamma prior is better than that the non-informative prior. The gamma prior under linex loss function is better than the squared error loss function. Specially, the linex with loss parameter -0.5 is better than all.

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.

References

Abravesh, A. Ganji, M. and Mostafaiy, B. (2019). Classical and bayesian estimation of stress-strength reliability in type ii censored pareto distributions. *Communication in Statistics - Simulation and Computation*, **48**, 2333–2358.

Awad, A. and Gharraf, K. (1986). Estimation of p(y<x) for burr distribution. *Communications in Statistics - Simulation and Computation*, **15**, 389–402.

Cordeiro, G. M., O. E. M. M. P. B. V. and Pescim, R. R. (2014). The lomax generator of distributions: properties, minification process and regression model. *Applied Mathematics and Computation*, **247**, 465–486.

Eugene, N. Lee, C. and Famoye, F. (2002). Beta-normal distribution and its applications. *Statistics & Probability Letters*, **31**, 497–512.

Ieren, T. G. and Kuhe, D. A. (2018). On the properties and applications of lomax exponential distribution. *Technometrics*, **1**, 1–13.

Mokhlis, N. A. (2005). Reliability of a stress-strength reliability model with burr type iii distributions. *Communication in Statistics-Theory and methods*, **34**, 1643–1657.

Nelson (1982). *Applied Life Data Analysis*. John Wiley and Sons Inc.: Newyork.

Panahi, H. and Asadi, S. (2011). Inference of stress strength model for lomax distribution. *International Journal of Mathematical Sciences*, **5**, 937–940.

Torabi, H. and Montazeri, N. H. (2014). The logistic-uniform distribution and its application. *Communication in Statistics-Simulation And Computation*, **43**, 2551–2569.

Zografos, K. and Balakrishnan, N. (2009). On families of beta and generalized gamma-generated distributions and associated inference. *Statistical Methodology*, **6**, 344–362.

ANNEXURE

Table 4: The maximum likelihood and Bayes estimates of R with corresponding MSEs, when $\mathrm{R} = 0.47058$**,** $\alpha_1 = 4.5$, $\alpha_2 = 4$, $\beta = 0.25$, $\lambda = 1$, $c_1 = 1$, $d_1 = 0.8, c_2 = 2$, $d_2 = 0.4$, $c_3 = 5$, $d_3 = 0.2$, $c_4 = 1$ *and* $d_4 = 2$

Sample Size n, m	$\hat{R}$	Bayes estimates					
		Non-informative prior			Gamma prior		
		$\hat{R}_S$	$\hat{R}_L$	$\hat{R}_{L1}$	$\hat{R}_S$	$\hat{R}_L$	$\hat{R}_{L1}$
	0.475778 (0.00935)	0.476434 (0.00921)	0.476434 (0.00921)	0.476432 (0.00921)	0.469151 (0.01326)	0.469537 (0.01319)	0.468764 (0.01333)
	0.484687 (0.00761)	0.484751 (0.00759)	0.484751 (0.00758)	0.484752 (0.00756)	0.498398 (0.00568)	0.498469 (0.00567)	0.498328 (0.00568)
	0.509822 (0.00386)	0.518294 (0.00289)	0.518835 (0.00283)	0.517758 (0.00294)	0.627643 (0.00342)	0.631826 (0.00392)	0.623797 (0.00298)
	0.491192 (0.00646)	0.49118 (0.00645)	0.49119 (0.00646)	0.49128 (0.00648)	0.502418 (0.00482)	0.502452 (0.004812)	0.502386 (0.00482)
	0.499716 (0.00516)	0.502127 (0.004826)	0.501969 (0.004848)	0.501969 (0.00484)	0.5360198 (0.001302)	0.536513 (0.00127)	0.535539 (0.001336)
	0.493795 (0.00604)	0.493784 (0.006042)	0.493785 (0.006043)	0.493784 (0.006042)	0.502259 (0.00481)	0.502278 (0.00485)	0.502241 (0.004810)
	0.498063 (0.00539)	0.499178 (0.005233)	0.499253 (0.005222)	0.499104 (0.005243)	0.517388 (0.002941)	0.513875 (0.002923)	0.517221 (0.002959)
	0.495227 (0.005814)	0.495219 (0.005815)	0.495217 (0.005816)	0.494567 (0.005834)	0.501950 (0.002838)	0.501961 (0.002837)	501939 (0.002840)
	0.497783 (0.005431)	0.498424 (0.005337)	0.498467 (0.005331)	0.49838 (0.005343)	0.510512 (0.002721)	0.510596 (0.002711)	0.510428 (0.002731)

Table 5: The maximum likelihood and Bayes estimates of R with corresponding MSEs, when $\mathrm{R} = 0.47058$, $\alpha_1 = 4.5$, $\alpha_2 = 4$, $\beta = 0.25$, $\lambda = 1$, $c_1 = 4$, $d_1 = 3$,$c_2 = 3$, $d_2 = 0.9$, $c_3 = 5$, $d_3 = 5$, $c_4 = 1$ *and* $d_4 = 2$

Bayes estimator						
Gamma prior						
Samplesize n, m	$\hat{R}_S$	$MSE\left(\hat{R}_S\right)$	$\hat{R}_L$	$MSE\left(\hat{R}_L\right)$	$\hat{R}_{L1}$	$MSE\left(\hat{R}_{L1}\right)$
5, 5	0.497594	(0.001201)	0.497486	(0.001202)	0.497486	(0.001199)
10,10	0.496694	(0.0006817)	0.496698	(0.0006819)	0.4966903	(0.0006815)
20, 20	0.495962	(0.0006442)	0.495967	(0.0006445)	0.4959559	(0.0006439)
30, 30	0.493617	(0.000531)	0.493629	(0.000532)	0.493605	(0.0005305)
40, 40	0.491194	(0.000427)	0.491219	(0.000429)	0.49117	(0.000426)
50, 50	0.490183	(0.000419)	0.491835	(0.000422)	0.490092	(0.000418)

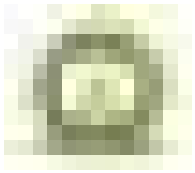
Table 6: The maximum likelihood and Bayes estimates of R with corresponding MSEs, when $\mathrm{R} = 0.57142$, $\alpha_1 = 1.5$, $\alpha_2 = 2$, $\beta = 0.05$, $\lambda = 1$, $c_1 = 1.5$, $d_1 = 0.8$, $c_2 = 2$, $d_2 = 0.4$, $c_3 = 5$, $d_3 = 0.2$, $c_4 = 1$ *and* $d_4 = 2$

Sample Size n, m	$\hat{R}$	Bayes estimates					
		Non-informative prior			Gamma prior		
		$\hat{R}_S$	$\hat{R}_L$	$\hat{R}_{L1}$	$\hat{R}_S$	$\hat{R}_L$	$\hat{R}_{L1}$
	0.512716	0.508545	0.608551	0.608536	0.608231	0.608375	0.608084
	(0.008113)	(0.001532)	(0.001531)	(0.001532)	(0.001415)	(0.001414)	(0.001416)
	0.598296	0.594557	0.594561	0.594554	0.592429	0.592497	0.592359
	(0.001438)	(0.001177)	(0.001178)	(0.001176)	(0.000554)	(0.000556)	(0.000552)
	0.584903	0.582614	0.582616	0.582613	0.584997	0.585012	0.584972
	(0.001183)	(0.001032)	(0.001032)	(0.001031)	(0.000508)	(0.000509)	(0.000507)
	0.585307	0.584972	0.581633	0.581278	0.580119	0.580272	0.581965
	(0.000615)	(0.000230)	(0.000226)	(0.000234)	(0.000139)	(0.000136)	(0.000123)
	0.579466	0.57843	0.571844	0.571843	0.566447	0.56646	0.466435
	(0.001084)	(0.00098)	(0.000973)	(0.000981)	(0.000118)	(0.000125)	(0.000111)
	0.578644	0.573740	0.573672	0.573808	0.572867	0.573873	0.573863
	(0.000789)	(0.000538)	(0.000536)	(0.000542)	(0.000110)	(0.000111)	(0.000109)
	0.576447	0.571814	0.572679	0.565506	0.569978	0.569784	0.569989
	(0.000671)	(0.000411)	(0.000407)	(0.000415)	(0.000040)	(0.000039)	(0.000040)

Table 7: The maximum likelihood and Bayes estimates of R with corresponding MSEs, when $\mathrm{R} = 0.57142$, $\alpha_1 = 1.5$, $\alpha_2 = 2$, $\beta = 0.05$, $\lambda = 1$, $c_1 = 4$, $d_1 = 3$,$c_2 = 3$, $d_2 = 0.9$, $c_3 = 5$, $d_3 = 5$, $c_4 = 1$ *and* $d_4 = 2$

Bayes estimator						
Gamma prior						
Samplesize n, m	$\hat{R}_S$	$MSE\left(\hat{R}_S\right)$	$\hat{R}_L$	$MSE\left(\hat{R}_L\right)$	$\hat{R}_{L1}$	$MSE\left(\hat{R}_{L1}\right)$
5, 5	0.614573	(0.001506)	0.614682	(0.001492)	0.614325	(0.001485)
10,10	0.591334	(0.0008835)	0.596698	(0.0008829)	0.5966903	(0.0008821)
20, 20	0.583426	(0.0007566)	0.582466	(0.0007545)	0.581345	(0.0007536)
30, 30	0.581215	(0.0005104)	0.581103	(0.0005102)	0.579996	(0.0005009)
40, 40	0.579855	(0.0004551)	0.578847	(0.0004545)	0.577634	(0.0004495)
50, 50	0.574673	(0.0002486)	0.573321	(0.0002465)	0.572589	(0.0002355)

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 22, No. 1, 2024 (New Series), pp 241–261
http://www.ssca.org.in/journal



Bayesian Inference for Glioma Data Using Generalized Burr X-G (GBX-G) Family with R and Stan

Devashish, Shazia Farhin, and Athar Ali Khan
Department of Statistics and Operations Research
Aligarh Muslim University, Aligarh - 202002, India

Received: 17 April 2023; Revised: 29 August 2023; Accepted: 03 September 2023

Abstract

Bayesian modeling of generalized distributions is currently highly appreciated due to the impressive growth in computational capabilities and software accessibility. This work attempts to fit the Bayesian inference methods for the generalized Burr X-G (GBX-G) family. On the basis of the GBX-G family, three distributions— the generalized Burr X-Weibull, the generalized Burr X-Exponential and the generalized Burr X-Lomax are analysed and fitted to censored survival data of malignant glioma patients using the probabilistic programming language STAN. In order to apply censored mechanisms throughout using STAN, codes are developed. Finally, a comparison has been made between the models through the use of Watanabe Akaike information criteria and leave one out cross-validation information criteria and conclusion has been given regarding the Bayesian model fitting of the glioma dataset.

Key words: Generalized Burr X-G (GBX-G) family; Bayesian survival modelling; Censored data; MCMC; LOOIC and WAIC methods.

1. Introduction

The time until an event occurs is the outcome variable of interest in a group of statistical techniques for data analysis called survival analysis. In the literature, there are many models that may be used to analyse lifetime data. Burr X (BX) distribution and it's generalization has been a part of research interest for survival data analysis for a long period of time. Burr (1942) introduced the Burr X (BX) distribution and later Yousof *et al.* (2017) defined Burr X-G (BX-G) family of distributions and also discussed it's application in analysing lifetime data. Also several other extended forms of the Burr X-G family were studied such as the transmuted Burr X-G (TBX-G) family and the truncated Burr X-G family of distributions that have been proposed and discussed by Al-Babtain *et al.* (2021) and Bantan *et al.* (2021) respectively. Apart from this Akhtar and Khan (2014) has conducted Bayesian analysis of generalized log-Burr family using R.

Based on the Burr X (BX) distribution, Aldahlan *et al.* (2021) created a new class of continuous distributions known as the generalised Burr X-G (GBX-G) family, studied its

Corresponding Author: Devashish
Email: devashishstatistics@gmail.com

mathematical properties, such as explicit expressions for the quantile and generating functions, ordinary and incomplete moments, order statistics, *etc.*, and provided its applications to real data sets. The GBX-G family's versatility in accommodating various forms of the hazard rate function (for details see Aldahlan *et al.* (2021)) turns into the driving force behind our work. The three GBX-G family-based models have been taken into consideration namely generalized Burr X-Exponential (GBXEx) Model, generalized Burr X-Weibull (GBXW) Model, and generalized Burr X-Lomax (GBXLx) Model in order to fit a real censored survival data named **glioma** which was initially discussed by Grana *et al.* (2002), under the Bayesian setup.

The comprehensive Bayesian inference-supporting probabilistic programming language **STAN** Carpenter *et al.* (2017) in R R Core Team (2021) is used to fit the aforementioned models. In Bayesian analysis, the computer language STAN is most typically used as a Hamiltonian Monte Carlo (HMC) sampler Duane *et al.* (1987); Brooks *et al.* (2011). Statistical models are defined using STAN. For Bayesian analysis, STAN predominantly uses the No-U-Turn sampler (NUTS) Hoffman *et al.* (2014) to obtain posterior simulation. We have also assessed and chosen the most appropriate model for the glioma data using the Watanabe-Akaike information criteria, or widely applicable information criteria (WAIC) and the Leave-One-Out information criteria (LOOIC). LOOIC and WAIC are two techniques for assessing the accuracy of pointwise out-of-sample predictions using a fitted Bayesian model and the log-likelihood evaluated at the posterior simulations of the parameter values, see Vehtari *et al.* (2017).
The article is structured as follows:
1. Explanation of PDF, CDF and hazard function for GBX-G family and all three models GBXEx, GBXW, and GBXLx of it (Section 2).
2. Explanation of the glioma data set and it's structure for STAN (Section 3).
3. Analysis under Bayesian approach by providing Likelihood, prior and posterior for all three models (Section 4).
4. Implementation and model fitting using STAN (Section 4.5).
5. Numeric as well as graphic results and interpretation of Bayesian analysis for the glioma data set (Section 4.8 - Section 4.12).
6. Bayesian Model comparison for the glioma data set (Section 5).
7. Conclusion (Section 6).

2. Generalized Burr X-G (GBX-G) family

A continuous random variable T is said to have the GBX-G family that is T $\sim$ GBX-G(α, β, η), if it has following probability density function (PDF), cumulative distribution function (CDF), survival function, and hazard function respectively, see (Aldahlan *et al.*, 2021)-

$$f_T(t,\alpha,\beta,\eta) = \frac{2\alpha\beta g(t,\eta)G(t,\eta)^{2\alpha-1}}{[1-G(t,\eta)^{\alpha}]^3} exp(-[\frac{G(t,\eta)^{\alpha}}{1-G(t,\eta)^{\alpha}}]^2) \times(1-exp(-[\frac{G(t,\eta)^{\alpha}}{1-G(t,\eta)^{\alpha}}]^2))^{\beta-1} \tag{1}$$

$$F_T(t,\alpha,\beta,\eta) = (1-exp(-[\frac{G(t,\eta)^{\alpha}}{1-G(t,\eta)^{\alpha}}]^2))^{\beta} \tag{2}$$

$$S_T(t,\alpha,\beta,\eta) = 1-(1-exp(-[\frac{G(t,\eta)^\alpha}{1-G(t,\eta)^\alpha}]^2))^\beta \tag{3}$$

$$h_T(t,\alpha,\beta,\eta) = \frac{f_T(t,\alpha,\beta,\eta)}{S_T(t,\alpha,\beta,\eta)} \tag{4}$$

Where α and β are positive shape parameters and η is parameter vector.

2.1. Generalized Burr X-exponential (GBXEx) model

Let the PDF $g(t) = \lambda e^{-\lambda t}$, for t > 0, of the exponential distribution with scale parameter λ, λ>0. Then, the probability density function (PDF), cumulative distribution function (CDF), survival function of the GBXEx model becomes

$$\begin{aligned} f(t) = \frac{2\alpha\beta\lambda e^{-\lambda t}(1-e^{-\lambda t})^{2\alpha-1}}{[1-(1-e^{-\lambda t})^\alpha]^3} exp(-[\frac{(1-e^{-\lambda t})^\alpha}{1-(1-e^{-\lambda t})^\alpha}]^2) \\ \times(1-exp(-[\frac{(1-e^{-\lambda t})^\alpha}{1-(1-e^{-\lambda t})^\alpha}]^2))^{\beta-1} \end{aligned} \tag{5}$$

$$F(t) = (1-exp(-[\frac{(1-e^{-\lambda t})^\alpha}{1-(1-e^{-\lambda t})^\alpha}]^2))^\beta \tag{6}$$

$$S(t) = 1-(1-exp(-[\frac{(1-e^{-\lambda t})^\alpha}{1-(1-e^{-\lambda t})^\alpha}]^2))^\beta \tag{7}$$

Also, using above expressions the hazard function can be obtained as-

$$h(t) = \frac{f(t)}{S(t)} \tag{8}$$

Here the random variable T will be denoted as T $\sim$ GBXEx(α, β, λ).

Random number generation - For random number generation of time variable from any survival model we will equate Survival function, S(t) to u, where U is a Uniform(0,1) variate and solve this equation for the value of t. Gelman *et al.* (2013) explained this method to generate the random numbers. Farhin *et al.* (2022) used this method recently.
The random number generation from GBXEx model is obtained by -

$$t = \frac{-1}{\lambda} log(1-(\frac{(-log(1-(1-u)^{1/\beta}))^{1/2}}{1+(-log(1-(1-u)^{1/\beta}))^{1/2}})^{1/\alpha}) \tag{9}$$

2.2. Generalized Burr X-Weibull (GBXW) model

Let the PDF $g(t) = a\lambda x^{\lambda-1}e^{-ax^\lambda}$, for t > 0, of the Weibull distribution with parameters λ and a, λ>0, a>0. Thus, the probability density function (PDF), cumulative distribution function (CDF), survival function of the GBXW model becomes

$$\begin{aligned} f(t) = \frac{2\alpha\beta a\lambda t^{\lambda-1}e^{-at^\lambda}(1-e^{-at^\lambda})^{2\alpha-1}}{[1-(1-e^{-at^\lambda})^\alpha]^3} exp(-[\frac{(1-e^{-at^\lambda})^\alpha}{1-(1-e^{-at^\lambda})^\alpha}]^2) \\ \times(1-exp(-[\frac{(1-e^{-at^\lambda})^\alpha}{1-(1-e^{-at^\lambda})^\alpha}]^2))^{\beta-1} \end{aligned} \tag{10}$$

$$F(t) = (1 - exp(-[\frac{(1 - e^{-at^{\lambda}})^{\alpha}}{1 - (1 - e^{-at^{\lambda}})^{\alpha}}]^2))^{\beta} \tag{11}$$

$$S(t) = 1 - (1 - exp(-[\frac{(1 - e^{-at^{\lambda}})^{\alpha}}{1 - (1 - e^{-at^{\lambda}})^{\alpha}}]^2))^{\beta} \tag{12}$$

Also, using above expressions the hazard function can be obtained as-

$$h(t) = \frac{f(t)}{S(t)} \tag{13}$$

Here the random variable T will be denoted as T $\sim$ GBXW(α, β, a, λ).

Also, the generation of random numbers from GBXW model is obtained by -

$$t = \lambda \times (-log(1 - (\frac{(-log(1 - (1 - u)^{1/\beta}))^{1/2}}{1 + (-log(1 - (1 - u)^{1/\beta}))^{1/2}})^{1/\alpha}))^{1/a} \tag{14}$$

2.3. Generalized Burr X-Lomax (GBXLx) model

Let the PDF $g(t) = a/\lambda[1 + t/\lambda]^{-a-1}$, for t $>$ 0, of the Lomax distribution with parameters λ and a, λ>0, a>0. Thus, the probability density function (PDF), cumulative distribution function (CDF), survival function of the GBXLx model is given by

$$f(t) = \frac{2\alpha\beta a/\lambda[1 + t/\lambda]^{-a-1}(1 - [1 + t/\lambda]^{-a})^{2\alpha-1}}{[1 - (1 - [1 + t/\lambda]^{-a})^{\alpha}]^3} exp(-[\frac{(1 - [1 + t/\lambda]^{-a})^{\alpha}}{1 - (1 - [1 + t/\lambda]^{-a})^{\alpha}}]^2) \times (1 - exp(-[\frac{(1 - [1 + t/\lambda]^{-a})^{\alpha}}{1 - (1 - [1 + t/\lambda]^{-a})^{\alpha}}]^2))^{\beta-1} \tag{15}$$

$$F(t) = (1 - exp(-[\frac{(1 - [1 + t/\lambda]^{-a})^{\alpha}}{1 - (1 - [1 + t/\lambda]^{-a})^{\alpha}}]^2))^{\beta} \tag{16}$$

$$S(t) = 1 - (1 - exp(-[\frac{(1 - [1 + t/\lambda]^{-a})^{\alpha}}{1 - (1 - [1 + t/\lambda]^{-a})^{\alpha}}]^2))^{\beta} \tag{17}$$

Also, using above expressions the hazard function can be obtained as-

$$h(t) = \frac{f(t)}{S(t)} \tag{18}$$

Here the random variable T will be denoted as T $\sim$ GBXLx(α, β, a, λ). The generation of random numbers from GBXLx model is obtained by-

$$t = \lambda \times ((1 - (\frac{((1 - (1 - u)^{1/\beta}))^{1/2}}{1 + (-log(1 - (1 - u)^{1/\beta}))^{1/2}})^{1/\alpha})^{-1/a} - 1) \tag{19}$$

3. Data: malignant glioma pilot study

On malignant glioma patients receiving pretargeted adjuvant radioimmunotherapy with yttrium-90-biotin, Grana *et al.* (2002) did a non-randomized pilot research and evaluated overall survival and the time to relapse. In this study, 37 high-grade glioma patients, 17 with grade III glioma and 20 with glioblastoma (GBM) were enrolled in a controlled open non-randomized study. Among them, 19 patients were treated with adjuvant radioimmunotherapy (RIT) and 18 were represented as the Control group. The survival time for each patient alongwith other helpful information such as gender, histology, age, *etc.* had been recorded. There are 14 censored observations out of 37 in the dataset.
This complete data set can be accessed through the R R Core Team (2021) package **coin** Zeileis *et al.* (2008) with the name **glioma**.
The discription of variables of glioma data set are given below:
no.: patient number.
age: patient age in years.
sex: a factor indicating patient's gender with levels "M" for Male and "F" for Female.
histology: a factor with levels "Grade3" (grade III glioma) and "GBM" (grade IV or glioblastoma).
group: a factor with levels "RIT"(radioimmunotherapy) and "Control".
event: censoring status indicator: FALSE for right-censored values and TRUE otherwise.
time: survival time in months.

3.1. Data creation for computation in stan

The model matrix x, a number of predictors M, and details of the censoring and response variable are needed for data production. N is the number of observations that are stated. Censoring is considered, with 0 being censored values and 1 denoting uncensored values. Finally, a listed form of data named 'datg' is created by combining all of these operations.

```
library(coin)
library(survival)
data("glioma")
glioma
help(glioma)
head(glioma)
y=glioma$time
x1=glioma$age
x2=as.numeric(glioma$sex)
#0=Female, 1=Male
x2=as.numeric(x2==2)
x3=as.numeric(glioma$histology)
#0=GBM, 1=Grade3
x3=as.numeric(x3==2)
x4=as.numeric(glioma$group)
#0=Control, 1=RIT
x4=as.numeric(x4==2)
```

```
#0=censored, 1=observed
censor=as.numeric(glioma$event)
x=cbind(1,x1,x2,x3,x4)
N=nrow(x)
M=ncol(x)
datg=list(y=y,censor=censor,x=x,N=N,M=M)
```

4. Analysis using Bayesian mechanism

In Bayesian analysis, in accordance with Bayes Theorem, we look for the posterior distribution which is the exact parameter distribution, by combining likelihood or data with the prior distribution of the parameter. The likelihood of the data and the prior distribution or the prior belief about the model's parameters must be established before the Bayesian regression model can be built.

4.1. Likelihood

Following Collett (2015), The right censored data can be formulated using the following joint likelihood function-

$$L = \prod_{i=1}^{n} h(t_i)^{\gamma_i} S(t_i) \tag{20}$$

And, the log-likelihood can be re-written as an alternative to the above form as-

$$logL = \sum_{i=1}^{n} (\gamma_i (logh(t_i) + logS(t_i))) \tag{21}$$

Here γ_i is the censoring indicator such that $\gamma = 1$ if the observation is not censored and $\gamma = 0$ if the observation is censored. To obtain the likelihood of GBXEx, GBXW and GBXLx survival models, the survival function $S(t_i)$ and the hazard function $h(t_i)$ of GBXEx, GBXW and GBXLx models respectively can be sustituted in the equation 20.

4.2. Modeling information

Following Lawless (2011), We have introduced covariates using the log link function in order to construct a regression model.

$$log(\lambda_i) = b_1 + b_2 x_{i1} + b_3 x_{i2} + b_4 x_{i3} + b_5 x_{i4} \tag{22}$$

$$\lambda_i = exp(b_1 + b_2 x_{i1} + b_3 x_{i2} + b_4 x_{i3} + b_5 x_{i4}) \tag{23}$$

$$\lambda_i = exp(x_i b) \tag{24}$$

Here $b = [b_1, b_2, b_3, b_4, b_5]$ are regression coefficients and x_i's are covariates of the data set discussed in Section 3. In particular, b_1 is the intercept, b_2 is the coefficient of covariate (x_1) of Age, b_3 is the coefficient of covariate (x_2) of Sex, b_4 is the coefficient of covariate (x_3) of Histology and b_5 is the coefficient of covariate (x_4) of Group.
In **STAN** the *transformed parameter* block of the stan model code contains above regression model. These stan codes are discussed in Section 4.5.

4.3. Prior

A prior probability distribution for parameters of the model needs to be specified before building the Bayesian regression model. In this article, for shape and scale parameters, we have chosen a half-Cauchy prior, and for the regression coefficient, a regularizing prior. We have opted for the Normal prior with mean 0, and standard deviation 5 for regression coefficient as a regularizing prior. Regularizing prior reduces the rate of learning from the data and prevents a model from becoming overexcited by it. Notably, it reduces the overfitting of the model to the data.
The half-Cauchy distribution with scale parameter 25, used as a noninformative prior distribution for shape parameter. Taking 25 as the value of the scale parameter, the half-Cauchy distribution becomes almost flat. Gelman (2006) support the use of half-Cauchy or uniform prior for the regression coefficents. Khan and Khan (2018) and Farhin and Khan (2023) explained the use of Gaussian prior for the regression coefficient and half-Cauchy prior for the shape as well as the scale in detail.

4.4. Posterior

Here the Bayes Theorem is used to obtain the joint posterior distribution of parameter $\theta = (\alpha, \beta, a, b) = (\alpha, \beta, a, b_0, b_1, ..., b_p)$ given the data as

$$P(\theta|t, X) \propto L(t|\theta, X)P(\theta) \tag{25}$$

Taking X as the matrix of covariates and assuming the parameters as independent, we have

$$P(\alpha, \beta, a, b|t, X) \propto P(t|\alpha, \beta, a, b, X)P(\alpha)P(\beta)P(a)P(b) \tag{26}$$

Hence the joint posterior distribution of GBXEx Model, GBXW Model and GBXLx Model can be obtained by sustituting priors and the likelihood of the corresponding models in Equation 26.

4.4.1. Posterior density of GBXEx model

$$\begin{aligned}
&P(\alpha, \beta, b|t, X) \propto P(t|\alpha, \beta, b, X)P(\alpha)P(\beta)P(b)\\
&\propto \prod_{i=1}^{n} \left\{ \frac{2\alpha\beta e^{x_i b} e^{-e^{x_i b}t}(1 - e^{-e^{x_i b}t_i})^{2\alpha-1}}{[1 - (1 - e^{-e^{x_i b}t_i})^{\alpha}]^3} exp(-[\frac{(1 - e^{-e^{x_i b}t_i})^{\alpha}}{1 - (1 - e^{-e^{x_i b}t_i})^{\alpha}}]^2) \right\}^{\gamma_i}\\
&\times \left\{ (1 - exp(-[\frac{(1 - e^{-e^{x_i b}t_i})^{\alpha}}{1 - (1 - e^{-e^{x_i b}t_i})^{\alpha}}]^2))^{\beta-1} \right\}^{\gamma_i}\\
&\times \left\{ 1 - (1 - exp(-[\frac{(1 - e^{-e^{x_i b}t_i})^{\alpha}}{1 - (1 - e^{-e^{x_i b}t_i})^{\alpha}}]^2))^{\beta} \right\}^{1-\gamma_i}\\
&\times \frac{2 \times 25}{\pi(\alpha^2 + 25^2)} \times \frac{2 \times 25}{\pi(\beta^2 + 25^2)} \times \prod_{j=0}^{J} \frac{1}{5\sqrt{2\pi}} exp\left(-\frac{1}{2 \times 25} b_j^2\right)
\end{aligned} \tag{27}$$

4.4.2. Posterior density of GBXW model

$$
\begin{aligned}
&P(\alpha,\beta,a,b|t,X) \propto P(t|\alpha,\beta,a,b,X)P(\alpha)P(\beta)P(a)P(b)\\
&\propto \prod_{i=1}^{n}\left\{\frac{2\alpha\beta a e^{x_i b}t_i^{e^{x_i b}-1}e^{-at_i^{e^{x_i b}}}(1-e^{-at_i^{e^{x_i b}}})^{2\alpha-1}}{[1-(1-e^{-at_i^{e^{x_i b}}})^{\alpha}]^3}exp(-[\frac{(1-e^{-at_i^{e^{x_i b}}})^{\alpha}}{1-(1-e^{-at_i^{e^{x_i b}}})^{\alpha}}]^2)\right\}^{\gamma_i}\\
&\times\left\{(1-exp(-[\frac{(1-e^{-at_i^{e^{x_i b}}})^{\alpha}}{1-(1-e^{-at_i^{e^{x_i b}}})^{\alpha}}]^2))^{\beta-1}\right\}^{\gamma_i}\\
&\times\left\{1-(1-exp(-[\frac{(1-e^{-at_i^{e^{x_i b}}})^{\alpha}}{1-(1-e^{-at_i^{e^{x_i b}}})^{\alpha}}]^2))^{\beta}\right\}^{1-\gamma_i}\\
&\times\frac{2\times 25}{\pi(\alpha^2+25^2)}\times\frac{2\times 25}{\pi(\beta^2+25^2)}\times\frac{2\times 25}{\pi(a^2+25^2)}\times\prod_{j=0}^{J}\frac{1}{5\sqrt{2\pi}}exp\left(-\frac{1}{2\times 25}b_j^2\right) \qquad (28)
\end{aligned}
$$

4.4.3. Posterior density of GBXLx model

$$
\begin{aligned}
&P(\alpha,\beta,a,b|t,X) \propto P(t|\alpha,\beta,a,b,X)P(\alpha)P(\beta)P(a)P(b)\\
&\propto \prod_{i=1}^{n}\left\{\frac{2\alpha\beta a/e^{x_i b}[1+t_i/e^{x_i b}]^{-a-1}(1-[1+t_i/e^{x_i b}]^{-a})^{2\alpha-1}}{[1-(1-[1+t_i/e^{x_i b}]^{-a})\alpha]^3}\right\}^{\gamma_i}\\
&\times\left\{exp(-[\frac{(1-[1+t_i/e^{x_i b}]^{-a})^{\alpha}}{1-(1-[1+t_i/e^{x_i b}]^{-a})^{\alpha}}]^2)\times(1-exp(-[\frac{(1-[1+t_i/e^{x_i b}]^{-a})^{\alpha}}{1-(1-[1+t_i/e^{x_i b}]^{-a})^{\alpha}}]^2))^{\beta-1}\right\}^{\gamma_i}\\
&\times\left\{1-(1-exp(-[\frac{(1-[1+t_i/e^{x_i b}]^{-a})^{\alpha}}{1-(1-[1+t_i/e^{x_i b}]^{-a})^{\alpha}}]^2))^{\beta}\right\}^{1-\gamma_i}\\
&\times\frac{2\times 25}{\pi(\alpha^2+25^2)}\times\frac{2\times 25}{\pi(\beta^2+25^2)}\times\frac{2\times 25}{\pi(a^2+25^2)}\times\prod_{j=0}^{J}\frac{1}{5\sqrt{2\pi}}exp\left(-\frac{1}{2\times 25}b_j^2\right) \qquad (29)
\end{aligned}
$$

Now, to find marginal posterior distribution we need to solve a a high-dimensional integral over all model parameters. Since it is difficult to derive the normalised joint posterior distribution and the marginal distributions of the parameters analytically, we approximate these integrals using Markov chain Monte Carlo (MCMC) methods. Thus, the estimates and other pertinent findings are achieved using the MCMC simulation approach with the aid of STAN.

4.5. Implementation using stan

STAN incorporates the use of the no-U-turn sampler (NUTS), an adaptive variant of Hamiltonian Monte Carlo (HMC) sampling, to efficiently simulate from the posterior distribution. HMC, which extends the capabilities of the Metropolis algorithm, is particularly advantageous in high-dimensional models due to its improved effectiveness and speed. While

Bayesian inference using Gibbs sampling (BUGS) is a commonly used approach, it often faces challenges when confronted with large datasets or complex models, leading to lengthy computation times or even failure to provide solutions. STAN, on the other hand, excels in handling such scenarios, offering faster computations and requiring a reduced number of iterations to achieve convergence when compared to BUGS (See Ashraf-Ul-Alam and Khan (2021) and Gelman *et al.* (2013)).
In R, to execute the STAN code, the package `rstan` is necessary. A Stan programme has six code blocks that are used for Bayesian modelling. Each block accommodates a list of instructions for distinct tasks. These blocks are - Data block, Transformed data block, Parameter block, Transformed parameter block, Model block, and Generated quantities block. In the Appendix, the stan codes with all these blocks for GBXEx, GBXW, and GBXLx models are provided.

4.6. Fitting the model using stan

The package `rstan` has a function named *stan* which is used to fit all the models based on GBX-G family. STAN uses C^{++} compiler for sampling from the posterior distrbution of the model parameters. All necessory codes for the numeric and graphical illustrations are provided in upcoming sub sections.

4.6.1. Bayesian data visualization: key plots for analysis

Graphical summary is an improtant part for analysis to assess the model convergence and communicate posterior related findings effectively. In this study, four plots are used namely- Traceplot, Caterpiller plot, Posterior predictive density plot and Posterior density plot. The traceplot provides a visual assessment of Markov Chain Monte Carlo (MCMC) convergence and it can be seen by comparing several Markov chains in a single plot. The Caterpiller plot shows the credible intervals or the quantiles for various parameters of the model and can be used to see the statistical significance of various coeffcients of the model. Posterior predictive checks (PPD plots) allow us to evaluate how well the model fits the observed data by comparing the density of the predicted values produced using the posterior predictive distribution of the specified model to the observed data. The posterior density plot is a graphical representation of the posterior distribution of a parameter and is constructed using simulated draws of the parameter from the posterior distribution. The R packages `Bayesplot` Gabry *et al.* (2019) and `ggplot2` Wickham and Wickham (2016) with `rstan` are used to create these plots in this paper.

4.7. Running the GBXEx model using stan

```
#calling rstan package
require(rstan)
#fitting the model
GBXE=stan(model_code = MGXE,data=datg,iter=4000,chains = 4)
print(GBXE)
```

4.8. Output summary and interpretaion

The results of Bayesian model fitting of GBXEx model are provided in Table 1. Also graphs are provided for summaries of posterior density and model convergence. The coeffcients b[2] of age (x1) and b[3] of sex (x2) are negative which shows that the older patients tends to have less survival probability then younger ones and the chances of survival for female are greater than male. The coeffcient b[5] of group is positive which indicates the chances of survival for patients who recieved the radioimmunotherapy (RIT) are greater than patients of control group which is a clear indication of positive impact of radioimmunotherapy on glioma patients. We can also see that the coeffcient b[4] of histology is positive which indicates patients with Grade 3 glioma have higher survival rates than those with glioblastoma (GBM). Also, after observing the estimates summary, it can be observed that the 95% credible intervals of b[4] and b[5] do not contain a value of zero, so the effect of coefficients of histology (GBM and Grade3) and group (Control and RIT) is statistically significant. Additionally, the summary table conatins the posterior estimates mean and se_mean, the standard deviation (sd), and the credible interval. Apart from this, the numerical summary table also contain the n_eff that is the effective number of samples which is a measure of the number of independent samples from the posterior distribution and the Rhat or the potential scale reduction factor, see (Gelman *et al.*, 2013), which is a quantitative criterion to assess convergence to the target distribution. In general n_eff > 100 and Rhat < 1.1 is acceptable for appropriate parameter estimates and model convergence, see (Gelman *et al.*, 2013). We can discern that the Rhat values for all parameters of the GBXEx model fall within an acceptable range, signifying successful convergence of the Markov chains to the desired distribution. The effective sample size is also reasonable. Using the **Bayesplot** package Gabry *et al.* (2019) , posterior predictive density (PPD) charts are created to visually assess the model. Posterior predictive density charts (Figure 2) depicts that the GBXEx model is consistent with the current data. Trace plots are also provided (Figure 1) to indicate the convergence of MCMC algorithm. Also, Figure 1 displays the caterpillar plot, wherein a vertical line appears at the value zero. Notably, the credible intervals of coefficients b[4] and b[5] lie on the right-hand side of the line, indicating that these intervals do not encompass the value of zero. This finding serves as evidence of the statistical significance of these coefficients.

Table 1: Posterior estimates results of GBXEx model parameters

Parameters	mean	se_mean	sd	2.5%	25%	50%	97.5%	n_eff	Rhat
b[1]	3.691	0.057	1.203	2.246	2.882	3.354	6.925	439	1.005
b[2]	-0.013	0.000	0.009	-0.031	-0.019	-0.013	0.006	899	1.002
b[3]	-0.284	0.008	0.245	-0.759	-0.444	-0.291	0.227	875	1.000
b[4]	0.962	0.010	0.307	0.431	0.758	0.940	1.627	967	1.003
b[5]	1.209	0.007	0.247	0.789	1.039	1.183	1.756	1418	1.001
alpha	4.652	0.155	4.223	0.122	0.972	3.576	14.582	747	1.004
beta	1.255	0.133	2.628	0.105	0.183	0.303	10.175	391	1.009

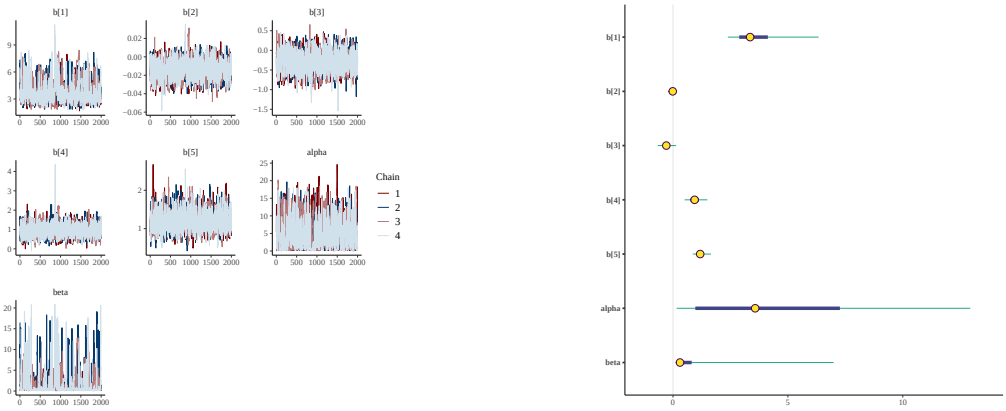


Figure 1: (i) Traceplot for GBXEx model, four chains were displayed in two separate runs; combining the two chains successfully indicates that MCMC algorithm has converged to the target joint posterior distribution. (ii) Caterpillar plot for GBXEx model.

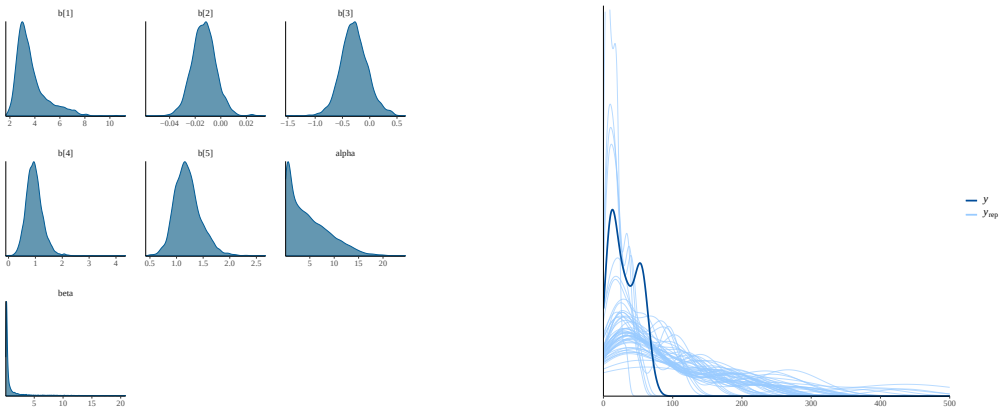


Figure 2: (i) Posterior density plot for GBXEx model. (ii) Posterior predictive density plot of the GBXEx model to assess the convergence of model. The GBXEx's posterior predictive density adequately fits the data, according to the PPD plot.

4.9. Running the GBXW model using stan

```
GBXW=stan(model_code = MGXW,data=datg,iter=4000,chains = 4, init = "random")
print(GBXW)
```

4.10. Output summary and interpretaion

The results of Bayesian model fitting of GBXW model are provided in Table 2. The Rhat values for model parameters are < 1.1, which depicts that the Markov chain converges to the desired distribution. And, the n_eff is more than 100 for all parameters of the model. The PPD chart (Figure 4) of the GBXW model indicates a good fit of the posterior predictive density with the data. It can also be seen that the coeffcients b[2] of age (x1) and b[3] of sex (x2) are negative and the coeffcients b[4] of histology (x3) and b[5] of group (x4) are positive. From the numeric summary of posterior estimates (Table 2) and the caterpillar plot (Figure 3), it is observed that the 95% credible intervals do not encompass the value

of zero for the coefficients of histology and group, which serves as evidence of the statistical significance of these coefficients.

Table 2: Posterior estimates results of GBXW model parameters

Parameters	mean	se_mean	sd	2.5%	25%	50%	97.5%	n_eff	Rhat
b[1]	0.500	0.112	3.827	-7.926	-1.967	0.971	6.993	1177	1.007
b[2]	-0.013	0.000	0.010	-0.032	-0.019	-0.013	0.007	2105	1.000
b[3]	-0.162	0.007	0.277	-0.696	-0.348	-0.166	0.372	1781	1.001
b[4]	0.966	0.006	0.302	0.417	0.763	0.948	1.612	2861	1.000
b[5]	1.208	0.005	0.254	0.756	1.036	1.190	1.760	2775	1.001
alpha	12.997	0.948	41.793	0.057	1.319	4.608	69.750	1945	1.003
beta	7.361	0.567	20.603	0.134	0.470	1.461	53.536	1319	1.003
a	0.859	0.072	1.671	0.086	0.168	0.292	6.352	539	1.010

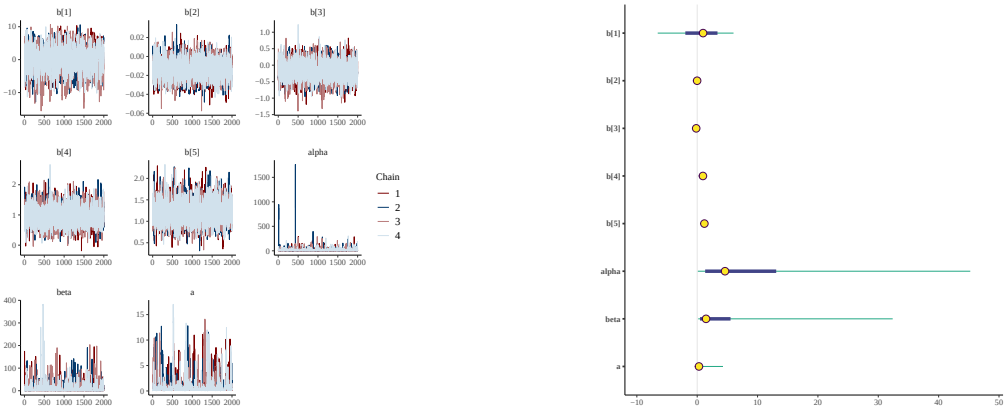


Figure 3: (i) Traceplot for GBXW model, four chains were displayed in two separate runs; combining the two chains successfully indicates that MCMC algorithm has converged to the target joint posterior distribution. (ii) Caterpillar plot for GBXW model.

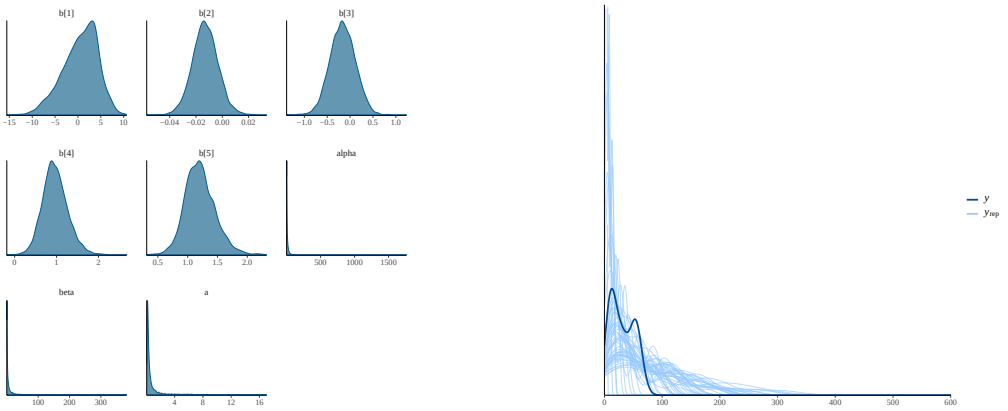


Figure 4: (i) Posterior density plot for GBXW model. (ii) Posterior predictive density plot of the GBXW model to assess the convergence of model. The GBXW's posterior predictive density adequately fits the data, according to the PPD plot.

4.11. Running the GBXLx model using stan

```
GBXL=stan(model_code = MGXL,data=datg,iter=4000,chains = 4)
print(GBXL)
```

4.12. Output summary and interpretaion

The results of Bayesian model fitting of GBXLx model are provided in Table 3. The Rhat values for model parameters are < 1.1, which depicts that the Markov chain converges to the desired distribution. And, the n_eff is more than 100 for all parameters of the model. The PPD chart (Figure 6) for the GBXLx model indicates a good fit of the posterior predictive density with the data. It can also be seen that the coeffcients b[2] of age (x1) and b[3] of sex (x2) are negative and the coeffcients b[4] of histology (x3) and b[5] of group (x4) are positive. From the numeric summary of posterior estimates (Table 3) and the caterpillar plot (Figure 5), it is observed that the 95% credible intervals do not encompass the value of zero for the coefficients of histology and group, which serves as evidence of the statistical significance of these coefficients.

Table 3: Posterior estimates results of GBXLx model parameters

Parameters	mean	se_mean	sd	2.5%	25%	50%	97.5%	n_eff	Rhat
b[1]	3.916	0.251	4.042	-4.775	1.203	5.100	9.957	259	1.016
b[2]	-0.013	0.000	0.009	-0.031	-0.019	-0.014	0.004	1453	1.002
b[3]	-0.203	0.009	0.254	-0.692	-0.372	-0.204	0.304	815	1.004
b[4]	0.966	0.007	0.286	0.459	0.770	0.951	1.594	1541	1.005
b[5]	1.194	0.006	0.241	0.773	1.034	1.173	1.723	1655	1.002
alpha	12.687	1.891	45.047	0.129	1.830	5.370	63.206	568	1.007
beta	2.067	0.228	5.376	0.118	0.221	0.440	15.137	556	1.004
a	26.023	2.067	95.486	0.274	1.062	6.787	152.409	2134	1.001

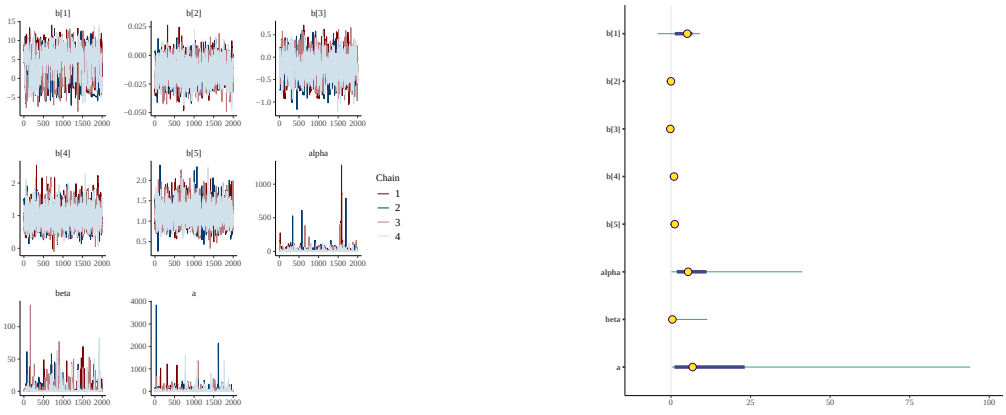


Figure 5: (i) Traceplot for GBXLx model, four chains were displayed in two separate runs; combining the two chains successfully indicates that MCMC algorithm has converged to the target joint posterior distribution. (ii) Caterpillar plot for GBXLx model.

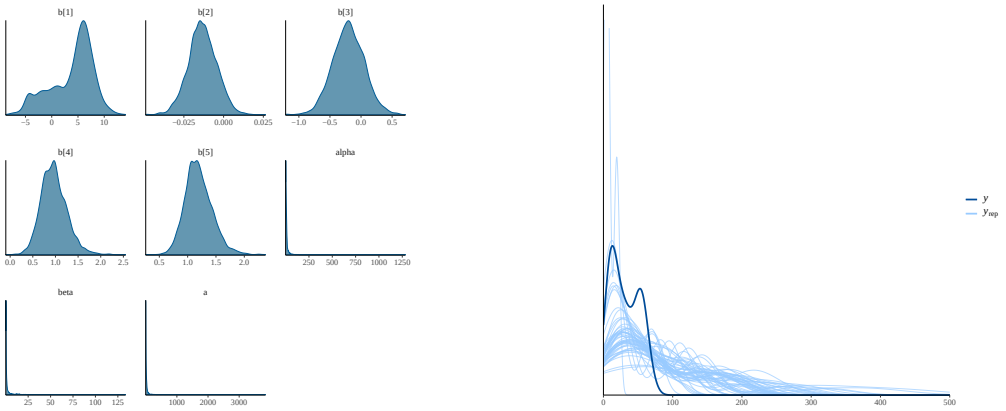


Figure 6: (i) Posterior density plot for GBXLx model. (ii) Posterior predictive density plot of the GBXLx model to assess the convergence of model. The GBXLx's posterior predictive density adequately fits the data, according to the PPD plot.

5. Model comparison with Bayesian criteria

For the purpose of comparing the fitted models, we use criteria for model evaluation and selection like the Leave One Out cross-validation Information Criteria (LOOIC) and the Watanabe Akaike Information Criteria (WAIC) both of which are methods for estimating pointwise out-of-sample prediction accuracy from a fitted Bayesian model Watanabe and Opper (2010); Vehtari *et al.* (2018). There are simpler estimates of predictive accuracy such as Akaike Information Criterion (AIC) and Deviance Information Criterion (DIC) but LOOIC and WAIC are better as instead of using only point estimates both LOOIC and WAIC use the pointwise log-likelihood of the full Bayesian posterior distribution. LOOIC and WAIC are more advantageous as they account for model complexity more effectively and offer fully Bayesian model comparison (See Vehtari *et al.* (2017)). LOOIC and WAIC quantifies the predictive accuracy of a model by estimating the expected log pointwise predictive density and in Stan, the generated quantities block computes these values. After fitting the model through STAN, the LOOIC and WAIC values are obtained by utilizing an R package `loo` (see Vehtari *et al.* (2018)) by calculating the log-likelihood assessed using posterior simulations of the parameters. A better model fit is indicated by a lower value for these selection criterias. Recently Ashraf-Ul-Alam and Khan (2021) and AbuJarad *et al.* (2022) used LOOIC and WAIC as the basis of comparision of the Bayesian survival models.

Table 4: LOOIC and WAIC values for GBXEx, GBXLx, and GBXW models

Model	LOOIC	WAIC
GBXEx	190.9	190.2
GBXLx	191.8	191.2
GBXW	192.8	192.4

We can observe from Table 4 that the GBXEx model's LOOIC and WAIC values are the lowest of the three, demonstrating that it exhibits superior performance as a survival model when compared to the other models applied to the glioma data.

6. Conclusion

In present study, Bayesian paradigm was applied to the analysis of a censored survival data using the GBXG family. The `Rstan` package of R is used to implement the simulation and analytical approximation techniques. The covariates Histology and Group are significant, and Markov chains for all models converge to the target distribution. The GBXEx model stands out as the most suitable option for fitting the glioma data, as evidenced by thorough comparisons of posterior predictive density plots, LOOIC, and WAIC. Additionally, patients who received radioimmunotherapy (RIT) had higher survival rates than individuals in the control group. Compared to patients with glioblastoma (GBM), patients with Grade 3 glioma had higher survival chances.

Acknowledgements

We sincerely appreciate the Editors' advice and support. We are very appreciative of the reviewer's insightful remarks and recommendations to generously add numerous helpful references.

References

AbuJarad, M. H., AbuJarad, E. S., and Khan, A. A. (2022). Bayesian survival analysis of type I general exponential distributions. *Annals of Data Science*, **9**, 347–367.

Akhtar, M. T. and Khan, A. A. (2014). Bayesian analysis of generalized log-Burr family with R. *SpringerPlus*, **3**, 1–10.

Al-Babtain, A. A., Elbatal, I., Al-Mofleh, H., Gemeay, A. M., Afify, A. Z., and Sarg, A. M. (2021). The flexible Burr XG family: properties, inference, and applications in engineering science. *Symmetry*, **13**, 474.

Aldahlan, M. A., Khalil, M. G., and Afify, A. Z. (2021). A new generalized family of distributions for lifetime data. *Journal of Modern Applied Statistical Methods*, **19**, 6.

Ashraf-Ul-Alam, M. and Khan, A. A. (2021). Generalized Topp-Leone-Weibull AFT modelling: a Bayesian analysis with MCMC tools using R and stan. *Austrian Journal of Statistics*, **50**, 52–76.

Bantan, R. A., Chesneau, C., Jamal, F., Elbatal, I., and Elgarhy, M. (2021). The truncated burr XG family of distributions: properties and applications to actuarial and financial data. *Entropy*, **23**, 1088.

Brooks, S., Gelman, A., Jones, G., and Meng, X.-L. (2011). *Handbook of Markov Chain Monte Carlo*. CRC press.

Burr, I. W. (1942). Cumulative frequency functions. *The Annals of Mathematical Statistics*, **13**, 215–232.

Carpenter, B., Gelman, A., Hoffman, M. D., Lee, D., Goodrich, B., Betancourt, M., Brubaker, M. A., Guo, J., Li, P., and Riddell, A. (2017). Stan: a probabilistic programming language. *Journal of Statistical Software*, **76**.

Collett, D. (2015). *Modelling Survival Data in Medical Research*. CRC press, third edition.

Duane, S., Kennedy, A. D., Pendleton, B. J., and Roweth, D. (1987). Hybrid monte carlo. *Physics Letters B*, **195**, 216–222.

Farhin, S. and Khan, A. A. (2023). Bayesian survival analysis of Rayleigh-X family with time varying covariate. *Applied Mathematics E-Notes*, **23**, 124–145.

Farhin, S., Yousuf, F., and Khan, A. A. (2022). Bayesian survival modeling of Marshal Olkin generalized-G family with random effects using R and stan. *Reliability: Theory & Applications*, **17**, 422–440.

Gabry, J., Simpson, D., Vehtari, A., Betancourt, M., and Gelman, A. (2019). Visualization in Bayesian workflow. *Journal of the Royal Statistical Society Series A: Statistics in Society*, **182**, 389–402.

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper). *Bayesian Analysis*, **1**, 515–534.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A., and Rubin, D. B. (2013). *Bayesian Data Analysis*. CRC press.

Grana, C., Chinol, M., Robertson, C., Mazzetta, C., Bartolomei, M., De Cicco, C., Fiorenza, M., Gatti, M., Caliceti, P., and Paganelli, G. (2002). Pretargeted adjuvant radioimmunotherapy with yttrium-90-biotin in malignant glioma patients: a pilot study. *British Journal of Cancer*, **86**, 207–212.

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.

Khan, N. and Khan, A. A. (2018). Bayesian analysis of Topp-Leone generalized exponential distribution. *Austrian Journal of Statistics*, **47**, 1–15.

Lawless, J. F. (2011). *Statistical Models and Methods for Lifetime Data*. John Wiley & Sons.

R Core Team (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Vehtari, A., Gabry, J., Yao, Y., and Gelman, A. (2018). loo: Efficient leave-one-out cross-validation and WAIC for Bayesian models. *R Package Version*, **2**, 1003.

Vehtari, A., Gelman, A., and Gabry, J. (2017). Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, **27**, 1413–1432.

Watanabe, S. and Opper, M. (2010). Asymptotic equivalence of Bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, **11**.

Wickham, H. and Wickham, H. (2016). *Data Analysis*. Springer.

Yousof, H. M., Afify, A. Z., Hamedani, G., and Aryal, G. R. (2017). The burr X generator of distributions for lifetime data. *Journal of Statistical Theory and Applications*, **16**, 288–305.

Zeileis, A., Wiel, M. A., Hornik, K., and Hothorn, T. (2008). Implementing a class of permutation tests: the coin package. *Journal of Statistical Software*, **28**, 1–23.

APPENDIX

A. Stan code for GBXE model

```
MGXE="functions{
 real gbxe_lpdf(real t, real alpha, real beta, real lambda){
  real log_fe;
  log_fe=log(2)+log(alpha)+log(beta)+exponential_lpdf(t|lambda)+
  (2*alpha-1)*exponential_lcdf(t|lambda)-3*log
  (1-(exponential_cdf(t,lambda))^alpha)-
  ((exponential_cdf(t,lambda))^alpha/(1-(exponential_cdf(t,lambda))
  ^alpha))^2 + (beta-1)*log(1-exp(-((exponential_cdf(t,lambda))^alpha
  /(1-(exponential_cdf(t,lambda))^alpha))^2));
  return log_fe;
 }
 real gbxe_lccdf(real t, real alpha, real beta, real lambda){
  real log_ccfe;
  log_ccfe=log(1-(1-exp(-((exponential_cdf(t,lambda))^alpha/
  (1-(exponential_cdf(t,lambda))^alpha))^2))^beta);
  return log_ccfe;
 }
 real surv_gbxe_lpdf(vector t, vector d, real alpha,
 real beta, vector lambda){vector[num_elements(t)] llk_gbxe;
  real prob;
  for(i in 1:num_elements(t)){
   llk_gbxe[i]=log_mix(d[i],gbxe_lpdf(t[i]|alpha,beta,lambda[i]),
   gbxe_lccdf(t[i]|alpha,beta,lambda[i]));
  }
  prob=sum(llk_gbxe);
  return prob;
 }
}
data{
 int N;
 vector<lower=0>[N] y;
 vector<lower=0,upper=1>[N] censor;
 int M;
 matrix[N,M] x;
}
parameters{
 vector[M] b;
 real<lower=0> alpha;
 real<lower=0> beta;
}
transformed parameters{
 vector[N] linpred;
 vector<lower=0>[N] lambda;
 linpred=x*b;
```

```
 for(i in 1:N){
  lambda[i]=exp(-linpred[i]);
 }
}
model{
 //priors
 target+=cauchy_lpdf(alpha|0,25)- 1 * cauchy_lccdf(0|0,25);
 target+=cauchy_lpdf(beta|0,25)- 1 * cauchy_lccdf(0|0,25);
 target+=normal_lpdf(b|0,5);
 //liklihood
 target+=surv_gbxe_lpdf(y|censor,alpha,beta,lambda);
}
generated quantities{
 vector[N] log_lik;
 vector[N] yrepgbxe;
 for(n in 1:N) log_lik[n]=log_mix(censor[n],
 gbxe_lpdf(y[n]|alpha,beta,lambda[n]),
 gbxe_lccdf(y[n]|alpha,beta,lambda[n]));
 {real u;
  u=uniform_rng(0,1);
  for(n in 1:N) yrepgbxe[n]=-1/lambda[n]*
  log(1-(((-log(1-(1-u)^(1/beta)))^(1/2))/
  (1+(-log(1-(1-u)^(1/beta)))^(1/2)))^(1/alpha));
 }
}"
```

B. Stan code for GBXW model

```
MGXW="functions{
 real gbxw_lpdf(real t, real alpha, real beta,real a, real lambda){
  real log_fw;
  log_fw=log(2)+log(alpha)+log(beta)+weibull_lpdf(t|a,lambda)+(2*alpha-1)
  *weibull_lcdf(t|a,lambda)-3*log(1-(weibull_cdf(t,a,lambda))^alpha)-
  ((weibull_cdf(t,a,lambda))^alpha/(1-(weibull_cdf(t,a,lambda))^alpha))^2
  +(beta-1)*log(1-exp(-((weibull_cdf(t,a,lambda))^alpha/
  (1-(weibull_cdf(t,a,lambda))^alpha))^2));
  return log_fw;
 }
 real gbxw_lccdf(real t, real alpha, real beta,real a, real lambda){
  real log_ccfw;
  log_ccfw=log(1-(1-exp(-((weibull_cdf(t,a,lambda))^alpha/
  (1-(weibull_cdf(t,a,lambda))^alpha))^2))^beta);
  return log_ccfw;
 }
 real surv_gbxw_lpdf(vector t, vector d, real alpha, real beta,real a,
 vector lambda){vector[num_elements(t)] llk_gbxw;
```

```
  real prob;
  for(i in 1:num_elements(t)){
   llk_gbxw[i]=log_mix(d[i],gbxw_lpdf(t[i]|alpha,beta,a,lambda[i]),
   gbxw_lccdf(t[i]|alpha,beta,a,lambda[i]));
  }
  prob=sum(llk_gbxw);
  return prob;
 }
}
data{
 int N;
 vector<lower=0>[N] y;
 vector<lower=0,upper=1>[N] censor;
 int M;
 matrix[N,M] x;
}
parameters{
 vector[M] b;
 real<lower=0> alpha;
 real<lower=0> beta;
 real<lower=0> a;
}
transformed parameters{
 vector[N] linpred;
 vector<lower=0>[N] lambda;
 linpred=x*b;
 for(i in 1:N){
  lambda[i]=exp(linpred[i]);
 }
}
model{
 //priors
 target+=cauchy_lpdf(alpha|0,25)- 1 * cauchy_lccdf(0|0,25);
 target+=cauchy_lpdf(beta|0,25)- 1 * cauchy_lccdf(0|0,25);
 target+=cauchy_lpdf(a|0,25)- 1 * cauchy_lccdf(0|0,25);
 target+=normal_lpdf(b|0,5);
 //liklihood
 target+=surv_gbxw_lpdf(y|censor,alpha,beta,a,lambda);
}
generated quantities{
 vector[N] log_lik;
 vector[N] yrepgbxw;
 for(n in 1:N) log_lik[n]=log_mix(censor[n],
 gbxw_lpdf(y[n]|alpha,beta,a,lambda[n]),
 gbxw_lccdf(y[n]|alpha,beta,a,lambda[n]));
 {real u;
  u=uniform_rng(0,1);
```

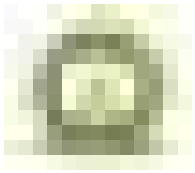
```
  for(n in 1:N) yrepgbxw[n]=lambda[n]*
  (-log(1-(((-log(1-(1-u)^(1/beta)))^(1/2))/
  (1+(-log(1-(1-u)^(1/beta)))^(1/2)))^(1/alpha)))^(1/a);
 }
}"
```

C. Stan code for GBXLx model

```
MGXL="functions{
 real gbxl_lpdf(real t, real alpha, real beta,real a, real lambda){
  real log_fl;
  log_fl=log(2)+log(alpha)+log(beta)+pareto_type_2_lpdf(t|0,lambda,a)
  +(2*alpha-1)*pareto_type_2_lcdf(t|0,lambda,a)-
  3*log(1-(pareto_type_2_cdf(t,0,lambda,a))^alpha)-
  ((pareto_type_2_cdf(t,0,lambda,a))^alpha/
  (1-(pareto_type_2_cdf(t,0,lambda,a))^alpha))^2+(beta-1)*
  log(1-exp(-((pareto_type_2_cdf(t,0,lambda,a))^alpha/
  (1-(pareto_type_2_cdf(t,0,lambda,a))^alpha))^2));
  return log_fl;
 }
 real gbxl_lccdf(real t, real alpha, real beta,real a, real lambda){
  real log_ccfl;
  log_ccfl=log(1-(1-exp(-((pareto_type_2_cdf(t,0,lambda,a))^alpha/
  (1-(pareto_type_2_cdf(t,0,lambda,a))^alpha))^2))^beta);
  return log_ccfl;
 }
 real surv_gbxl_lpdf(vector t, vector d, real alpha, real beta,
 real a, vector lambda){
  vector[num_elements(t)] llk_gbxl;
  real prob;
  for(i in 1:num_elements(t)){
   llk_gbxl[i]=log_mix(d[i],gbxl_lpdf(t[i]|alpha,beta,a,lambda[i]),
   gbxl_lccdf(t[i]|alpha,beta,a,lambda[i]));
  }
  prob=sum(llk_gbxl);
  return prob;
 }
}
data{
 int N;
 vector<lower=0>[N] y;
 vector<lower=0,upper=1>[N] censor;
 int M;
 matrix[N,M] x;
}
parameters{
```

```
 vector[M] b;
 real<lower=0> alpha;
 real<lower=0> beta;
 real<lower=0> a;
}
transformed parameters{
 vector[N] linpred;
 vector<lower=0>[N] lambda;
 linpred=x*b;
 for(i in 1:N){
  lambda[i]=exp(linpred[i]);
 }
}
model{
 //priors
 target+=cauchy_lpdf(alpha|0,25)- 1 * cauchy_lccdf(0|0,25);
 target+=cauchy_lpdf(beta|0,25)- 1 * cauchy_lccdf(0|0,25);
 target+=cauchy_lpdf(a|0,25)- 1 * cauchy_lccdf(0|0,25);
 target+=normal_lpdf(b|0,5);
 //liklihood
 target+=surv_gbxl_lpdf(y|censor,alpha,beta,a,lambda);
}
generated quantities{
 vector[N] log_lik;
 vector[N] yrepgbxl;
 for(n in 1:N) log_lik[n]=log_mix(censor[n],
 gbxl_lpdf(y[n]|alpha,beta,a,lambda[n]),
 gbxl_lccdf(y[n]|alpha,beta,a,lambda[n]));
 {real u;
  u=uniform_rng(0,1);
  for(n in 1:N) yrepgbxl[n]=lambda[n]*((1-
  (((-log(1-(1-u)^(1/beta)))^(1/2))/
  (1+(-log(1-(1-u)^(1/beta)))^(1/2)))^(1/alpha))^(-1/a)-1);
 }
}"
```

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 22, No. 1, 2024 (New Series), pp 263–277
http://www.ssca.org.in/journal



Asymptotic Confidence Interval Approach to Estimate the Portion of Area Under Multi-Class ROC Curve

Arunima S. Kannan and R. Vishnu Vardhan
Department of Statistics, Pondicherry University, Puducherry

Received: 26 October 2022; Revised: 03 September 2023; Accepted: 06 September 2023

Abstract

The area under the curve (AUC) gives an overall summary measure of the performance of the Receiver Operating Characteristic (ROC) curve. AUC summarizes the entire area under the curve. Sometimes clinical studies need to focus on the area with low FPR and high TPR rates. To find the area of portion of an ROC curve, partial AUC (pAUC) came into use. The seminal works on estimating the pAUC was in the framework of Bi-normal ROC curve. However, in a real-life scenario, we may come across non-normality, and the data may be of multi-class. In such cases the existing methodology of binormal ROC curve will not be of use and this creates the need to bring out a new methodology for estimating the pAUC under non-normal data. In this paper we made an attempt to address the above point and derived the expressions for the pAUC of multi-class ROC curve. Further estimating the partial AUC has been carried out by means of asymptotic confidence intervals of the false positive rates. Adding to this the constraint on TPR has been considered to elicit the focused area of the ROC curve and termed it as two-way pAUC (TpAUC). Good amount of simulations and two real datasets have been considered for necessary illustrations.

Key words: AUC; Exponential; Multi-class; pAUC; ROC; TpAUC.

AMS Subject Classifications: 62P10

1. Introduction

The Receiver Operating Characteristic (ROC) curve is the popular classification tool to evaluate the performance of a diagnostic test/marker. ROC curve was first used for signal detection during World War II (Peterson *et al.*, 1954; Tanner and Swets, 1954). ROC curve analysis was introduced into diagnostic medicine by Lusted (1971), and its applications can be seen in several clinical domains that rely extensively on screening and diagnostic procedures, laboratory testing, epidemiology, radiology, etc.(Obuchowski, 2003).

ROC curve is generated using the coordinate pairs, namely the false positive rates (FPRs) and true positive rates (TPRs), which are usually referred as intrinsic measures. Out of these thresholds, one has to choose a threshold that can give better accuracy with reasonable values of false positives and true positives. AUC is the summary measure of the ROC

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

curve, which is used to determine the performance/accuracy of a diagnostic test/marker. AUC has a theoretical value lies between 0 and 1. An AUC of 0.5 indicates that the classification is random, the accuracy of a diagnostic test or procedure will increase as the value of AUC gets closer to 1.

Let us assume that H denote the population 1 with distribution function F and D be the population 2 with distribution function G; then the ROC form is given as

$$y(t) = G(F_0^{-1}(t)), 0 \leq t \leq 1 \tag{1}$$

where, $G(x) = \int_x^\infty g(x)dx$ and $F_0(y) = \int_y^\infty f(y)dy$; *g(x)* and *f(y)* is the density functions of H and D populations respectively. The probability of detecting/ identifying a subject with condition is called as the $TPR = P(S > t/D)$, and the probability of classifying a subject is $FPR = P(S > t/H)$. Here S denotes the data value or score observed from a subject and t is the threshold. Each data point in the ROC serves as a threshold point, with which one can calculate the TPRs and FPRs. Yet there is a need to determine the optimal threshold among the set of all possible thresholds, which is done by using Youden's J index. The optimal threshold is determined by taking the maximum value of Youden's J index from a vector of values obtained from (2). Now, the FPR and TPR values corresponding to this optimal threshold have to be considered as the optimal FPR and TPR.

$$J = maximum\,(TPR(t) - FPR(t)) \tag{2}$$

which is the maximum distance between the curve and the chance line. With this optimal threshold, the subjects will be classified with the atmost accuracy and can also be used to assign the status of unspecified subjects. When J = 1, the test is perfect, meaning there are no false positives or negatives.
The AUC of an ROC curve is defined as

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

AUC is defined as the average TPR value for all possible TNR (1-FPR) values, which will consider the entire area under the curve (Obuchowski, 2003; Zhou *et al.*, 2009; McClish, 1989; Obuchowski and Bullen, 2018). Analyzing the entire ROC curve involves both strict and lax thresholds; hence, considering a portion of ROC curve will be more meaningful in some instances, and such portion is named as partial AUC (pAUC) (see Figure 2). From Figure 1, we can see that the lax threshold provides high TPRs and high FPRs, which are not a region of interest for clinical studies. Since most clinical studies involve living subjects, the FPR must be reasonably low. However, the strict threshold provides low FPR and TPR values, which also does not indicate better classification. Hence, this generates the need to speak about the portion of the ROC curve above the strict thresholds and below the lax thresholds, which is shown in Figure 2 for some arbitrary values of c_1 and c_2 from the set of FPRs. pAUC is now employed in numerous medical applications and has gained popularity, particularly in screening research (Ricamato and Tortorella, 2011).

The pAUC consider the area of the ROC space where data have been observed or that correspond to clinically significant TPR or TNR values. For example, only the lower tail of the ROC curve is of interest for cancer screening because the FPR must be minimal to

be acceptable (Zhang *et al.*, 2018). Baker and Pinsky (2001) also pointed out that low FPR needs to be maintained in cancer screening studies, which is important because it will avoid costly biopsies. In such cases, analyzing a restricted portion will be more meaningful than analyzing the entire area of the ROC curve. Seminal work on pAUC was done by McClish (1989), where a method of analyzing the portion of the ROC curve was proposed and also gave a transformation to obtain the standardized pAUC value. Thompson and Zucchini (1989) introduced the method to estimate the partial area under the binormal ROC curve over any specified region of interest. Later, Jiang *et al.* (1996) adopted the methodology of McClish's work and extended it to describe the partial area index for highly sensitive diagnostic test. Hillis and Metz (2012) derived analytic expressions to estimate the pAUC under the assumption of a latent binormal model.

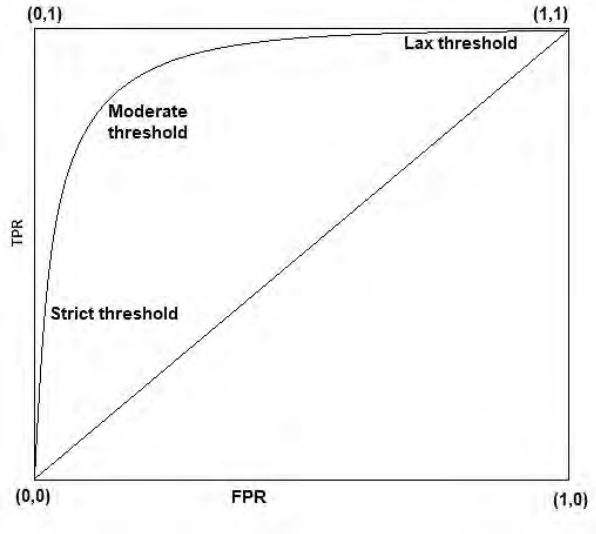


Figure 1: ROC curve depicting strict, moderate and lax thresholds

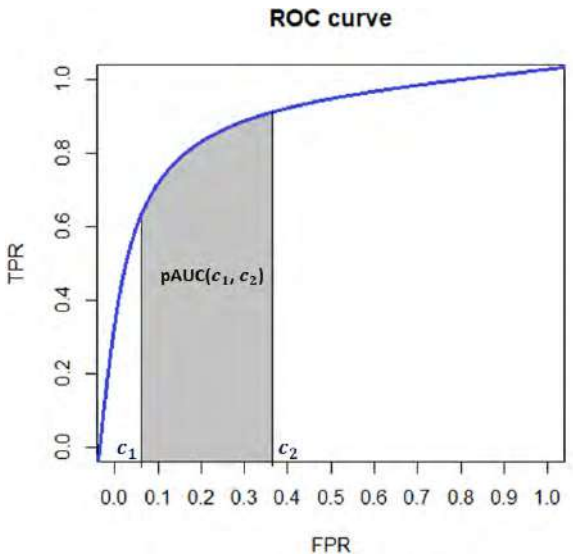


Figure 2: A typical plot of pAUC between a fixed range of FPR

In practice, diagnostic tests with high FPR lead to enormous economic expenses

because a significant fraction of healthy individuals would use up the limited supply of medical treatments. Furthermore, when diagnosing a fatal disease, failure to correctly identify severely ill patients (poor TPR) will result in severe ethical ramifications. Therefore, in such cases, it is necessary to simultaneously maintain FPR and TPR at low and high levels, respectively. So here, we have introduced a method to estimate the partial area by considering the constraints on both FPR and TPR simultaneously, which is termed as Two-Way pAUC (TpAUC). A diagrammatic representation of TpAUC is given in Figure 3 which considered the area of the ROC curve with FPR range (c_1,c_2) and minimum TPR of d_0, it can be denoted as TpAUC(c_1,c_2,d_0).

In literature, the works mainly focus on estimating the partial area of the binormal ROC curve, which can only be used when the data consists of two classes and follows normality. However, in a real-life scenario, we may come across non-normality, and the data may be of multi-class. In such cases the existing methodology of binormal ROC curve will not be of use and this creates the need to bring out a new methodology for estimating the pAUC under non-normal data. Estimation of AUC under the multi-class classification where data tend to follow normal distribution was addressed by Gönen (2013); Cheam and McNicholas (2016); Siva and Vishnu (2022). Recently Arunima and Vishnu (2022) proposed gamma mixture ROC curve to classify the multi-class data where the population follows gamma distributions, in which the gamma variate is transformed into normal by using the Wilson-Hilferty transformation. In this paper we have considered one of the well-known life time distribution; the exponential distribution and the partial area estimation of multi-class exponential ROC is discussed in detail. The study is supported with simulated and real datasets. Before we detail out the proposed methodology, a gentle introduction on Multi-class Exponential ROC Curve is given. Thereafter, along with the proposed methodology, the numerical illustrations are discussed in subsequent sections.

1.1. Multi-class exponential ROC curve

Let us assume that population 1, H $\sim exp(\theta_0)$ and population 2 has two sub populations namely D_1 and D_2 such that, $D_1 \sim exp(\theta_1)$ and $D_2 \sim exp(\theta_2)$. Then the expressions for intrinsic measures of mixture Exponential ROC (mEROC) are defined Arunima and Vishnu (2023) as below.
FPR of the mEROC (mFPR) is given as

$$mFPR = \pi_1 FPR_1 + \pi_2 FPR_2$$

where

$$FPR_1 = x(t_1) = P(S > t_1 \mid H) = e^{-\theta_0 t_1} \tag{3}$$

$$FPR_2 = x(t_2) = P(S > t_2 \mid D_1) = e^{-\theta_1 t_2} \tag{4}$$

where π_is are mixing proportions/weights; t_1 and t_2 are the respective threshold values for the classification of (H, D_1) and (D_1, D_2) respectively. From (3) and (4) we can write t_1 and t_2 as

$$t_1 = -\frac{log(x(t_1))}{\theta_0} \quad ; \quad t_2 = -\frac{log(x(t_2))}{\theta_1} \tag{5}$$

TPR of mEROC (mTPR) is given as

$$mTPR = \pi_1 TPR_1 + \pi_2 TPR_2$$

where

$$TPR_1 = y(t_1) = P(S > t_1 \mid D_1) = e^{-\theta_1 t_1} \tag{6}$$

$$TPR_2 = y(t_2) = P(S > t_2 \mid D_2) = e^{-\theta_2 t_2} \tag{7}$$

substituting (5) in (6) and (7) we will get the mEROC curve which be written as,

$$mEROC = \pi_1 x(t_1)^{\beta_1} + \pi_2 x(t_2)^{\beta_2}$$

where $\beta_1 = \frac{\theta_1}{\theta_0}$ and $\beta_2 = \frac{\theta_2}{\theta_1}$.
accuracy can be expressed notationally as

$$mAUC = \int_0^1 ROC(t)\, dt = \pi_1 \frac{\theta_0}{\theta_0 + \theta_1} + \pi_2 \frac{\theta_1}{\theta_1 + \theta_2}$$

2. Proposed methodology - partial area of mEROC curve

Let c_1 and c_2 denote any two arbitrary FPR values, then pAUC for mEROC can be defined as

$$mA_{(c_1,c_2)} = \pi_1 A_{1_{(c_1,c_2)}} + \pi_2 A_{2_{(c_1,c_2)}}$$

where $A_{1(c_1,c_2)}$ and $A_{2(c_1,c_2)}$ are the partial areas of H & D_1 and D_1 & D_2 respectively. which is defined as

$$\begin{aligned} A_{1_{(c_1,c_2)}} &= \int_{c_1}^{c_2} TPR_1(t)\, FPR_1'(t)\ dt = \int_{c_1}^{c_2} e^{-\theta_1 t}\, e^{-\theta_0 t}\, (-\theta_0)\, dt \\ &= \frac{\theta_0}{\theta_1 + \theta_0} \left[e^{-(\theta_0+\theta_1)c_2} - e^{-(\theta_0+\theta_1)c_1} \right] \end{aligned}$$

and

$$\begin{aligned} A_{2_{(c_1,c_2)}} &= \int_{c_1}^{c_2} TPR_2(t)\, FPR_2'(t)\ dt = \int_{c_1}^{c_2} e^{-\theta_2 t}\, e^{-\theta_1 t}\, (-\theta_1)\, dt \\ &= \frac{\theta_1}{\theta_2 + \theta_1} \left[e^{-(\theta_1+\theta_2)c_2} - e^{-(\theta_1+\theta_2)c_1} \right] \end{aligned}$$

The area (A+B) in Figure 3 indicates the pAUC between the FPR range c_1 and c_2. To this, one more additional constraint is added in the form of d_0. The area generated between c_1, c_2 and d_0 is termed as the two-way pAUC and denoted by $mTpAUC_{(c_1,c_2,d_0)}$. Here c_2 is the upper limit of FPR; d_0 is the lower limit of TPR, c_1 is the corresponding FPR value at d_0. The area encapsulated between the triplet combination (c_1, c_2, d_0) is indicated as B in Figure 3 and it can be obtained as

$$\begin{aligned} mTpAUC_{(c_1,c_2,d_0)} &= Area\ (A + B) - Area\ A \\ &= mA_{(c_1,c_2)} - [c_2 - c_1]d_0 \end{aligned}$$

In the above expression the constants c_1, c_2 and d_0 are to be estimated. To do so we need to employ two ways; one is arbitrarily choosing the values and the other is to estimate them using the method of asymptotic confidence interval. Below we describe the way of determining the c_1, c_2 and d_0.

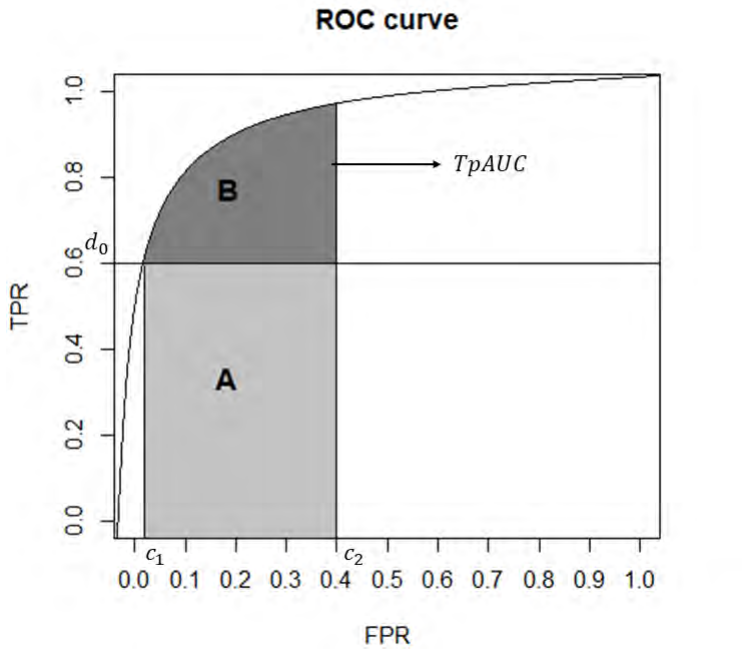


Figure 3: Depicting the area of A and B

Method I

In this method the c_2 and d_0 will be chosen arbitrarily and c_1 can be obtained from d_0. Instead of choosing arbitrarily one can impute using the knowledge from previous studies. In general clinicians prefer to have reasonably law FPR and moderate/high TPR.

Method II

Here we introduce the asymptotic confidence interval approach to define c_2 and d_0. For which the asymptotic confidence intervals are defined and respective variances for mFPR and mTPR are derived. The upper limit of mFPR will be taken as c_2 and lower limit of mTPR will be the d_0; the corresponding mFPR value is taken as c_1.

Asymptotic confidence intervals for mTPR and mFPR

The $100(1-\alpha)\%$ asymptotic confidence interval for mTPR is

$$mTPR \pm Z_{1-(\frac{\alpha}{2})}\sqrt{Var(mTPR)}$$

where $Z_{1-\frac{\alpha}{2}}$ is the $1-\frac{\alpha}{2}$ standard normal percentile and by delta method (Miller Jr, 1981), we can obtain variance of $mTPR$.

$$Var(mTPR) = \pi_1 Var(TPR_1) + \pi_2 Var(TPR_2)$$

where

$$Var(TPR_1) = \left(\frac{\partial TPR_1}{\partial \theta_0}\right)^2 Var(\theta_0) + \left(\frac{\partial TPR_1}{\partial \theta_1}\right)^2 Var(\theta_1)$$
$$= \left\{ \frac{e^{-\theta_1 \frac{(ln(\theta_0)-ln(\theta_1))}{\theta_0-\theta_1}} \left[\theta_1 \left(\theta_0 - \theta_1 - \theta_0(ln(\theta_0) + ln(\theta_1))\right)\right]}{\theta_0(\theta_0-\theta_1)^2} \right\}^2 \frac{{\theta_0}^2}{n_0}$$
$$+ \left\{ \frac{e^{-\theta_1 \frac{(ln(\theta_0)-ln(\theta_1))}{\theta_0-\theta_1}} \left[\theta_1 - \theta_0 ln(\theta_0) - \theta_0 + \theta_0 ln(\theta_0)\right]}{(\theta_0-\theta_1)^2} \right\}^2 \frac{{\theta_1}^2}{n_1}$$

$$Var(TPR_2) = \left(\frac{\partial TPR_2}{\partial \theta_1}\right)^2 Var(\theta_1) + \left(\frac{\partial TPR_2}{\partial \theta_2}\right)^2 Var(\theta_2)$$
$$= \left\{ \frac{e^{-\theta_2 \frac{(ln(\theta_1)-ln(\theta_2))}{\theta_1-\theta_2}} \left[\theta_2 \left(\theta_1 - \theta_2 - \theta_1(ln(\theta_1) + ln(\theta_2))\right)\right]}{\theta_1(\theta_1-\theta_2)^2} \right\}^2 \frac{{\theta_1}^2}{n_0}$$
$$+ \left\{ \frac{e^{-\theta_2 \frac{(ln(\theta_1)-ln(\theta_2))}{\theta_1-\theta_2}} \left[\theta_2 - \theta_1 ln(\theta_1) - \theta_1 + \theta_1 ln(\theta_1)\right]}{(\theta_1-\theta_2)^2} \right\}^2 \frac{{\theta_2}^2}{n_1}$$

For the method II we choose d_0 as lower limit of mTPR

$$d_0 = mTPR - Z_{1-(\frac{\alpha}{2})}\sqrt{Var(mTPR)} \tag{8}$$

then c_1 will be the corresponding FPR of the d_0.

Similarly, $100(1-\alpha)\%$ asymptotic confidence interval for $mFPR$ is,

$$mFPR \pm Z_{1-(\frac{\alpha}{2})}\sqrt{Var(mFPR)}$$

$$Var(mFPR) = \pi_1 Var(FPR_1) + \pi_2 Var(FPR_2)$$

where

$$Var(FPR_1) = \left(\frac{\partial FPR_1}{\partial \theta_0}\right)^2 Var(\theta_0) + \left(\frac{\partial FPR_1}{\partial \theta_1}\right)^2 Var(\theta_1)$$
$$= \left\{ \frac{e^{-\theta_0 \frac{(ln(\theta_0)-ln(\theta_1))}{\theta_0-\theta_1}} \left[\theta_0 - \theta_1 ln(\theta_0) - \theta_1 + \theta_1 ln(\theta_1)\right]}{(\theta_0-\theta_1)^2} \right\}^2 \frac{{\theta_0}^2}{n_0}$$
$$+ \left\{ \frac{e^{-\theta_0 \frac{(ln(\theta_0)-ln(\theta_1))}{\theta_0-\theta_1}} \left[\theta_0 \left(\theta_1 - \theta_0 + \theta_1 ln(\theta_0) + \theta_1 ln(\theta_1)\right)\right]}{\theta_1(\theta_0-\theta_1)^2} \right\}^2 \frac{{\theta_1}^2}{n_1}$$

and

$$Var(FPR_2) = \left(\frac{\partial FPR_2}{\partial \theta_1}\right)^2 Var(\theta_1) + \left(\frac{\partial FPR_2}{\partial \theta_2}\right)^2 Var(\theta_2)$$
$$= \left\{\frac{e^{-\theta_1 \frac{(ln(\theta_1)-ln(\theta_2))}{\theta_1-\theta_2}}\left[\theta_1 - \theta_2 ln(\theta_1) - \theta_2 + \theta_2 ln(\theta_2)\right]}{(\theta_1-\theta_2)^2}\right\}^2 \frac{{\theta_1}^2}{n_1}$$
$$+ \left\{\frac{e^{-\theta_1 \frac{(ln(\theta_1)-ln(\theta_2))}{\theta_1-\theta_2}}\left[\theta_1\left(\theta_2 - \theta_1 + \theta_2 ln(\theta_1) + \theta_2 ln(\theta_2)\right)\right]}{\theta_2(\theta_1-\theta_2)^2}\right\}^2 \frac{{\theta_2}^2}{n_2}$$

then the value of c_2 will be

$$c_2 = mFPR + Z_{1-\frac{\alpha}{2}}\sqrt{Var(mFPR)} \tag{9}$$

3. Numerical illustrations

For illustrating the proposed work both simulated and real datasets are considered and the results are tabulated accordingly.

3.1. Simulated datasets

Exponential random samples of size $n = (25, 50, 100, 200)$ are generated with different parameter combinations. The moderate and better classification scenarios will be demonstrated using the parameter combinations given in Table 1. Methods I and II are used to estimate the partial area on the samples that were generated.

For illustration purpose, we have chosen different parameter combinations of θ_0 and θ_1 and calculated the partial area by taking $c_2 = 0.5$ and $d_0 = 0.65$ (one may choose any values for c_2 and d_0 according to prior knowledge on the study). The results for each parameter combination with respective sample sizes are reported in Tables 1 and 2 and respective ROC curves are given in Figure 4. From Table 1, consider n=100 in set A, the overall $\widehat{mAUC}$ is observed to be 86.71%, with true positives about 76.61% and false positives of 13.40%, which indicates comparatively a better accuracy. In similar lines, in set B for n=100, the overall $\widehat{mAUC}$ is observed to be 68.27%, with true positives about 51.19% and false positives of 27.84%, which indicates a moderate accuracy. Table 2 gives the results pertaining to partial area for method I, and we can observe that, let say for n=100 for set A the values for d_0 and c_2 are chosen as 0.65 and 0.5 respectively, $\hat{c}_1$ is the corresponding mFPR at d_0 which is about 0.0816 provides mpAUC ($\widehat{mA}$) and $\widehat{mTpAUC}$ about 0.3786 and 0.1066 respectively. And for n=100 in set B, the corresponding mFPR at d_0 which is about 0.3911 provides mpAUC ($\widehat{mA}$) and $\widehat{mTpAUC}$ about 0.1496 and 0.0788 respectively. Here we can see that within a fixed range of mFPR and mTPR the $\widehat{mAUC}$, $\widehat{mA}_{(c_1,c_2)}$ and $\widehat{mTpAUC}_{(c_1,c_2,d_0)}$ are proportional to each other. For better understanding the $\widehat{mTpAUC}$ for n=100 of sets A and B using method II are depicted in Figures 5.

Table 1: ROC curve estimates of simulated dataset

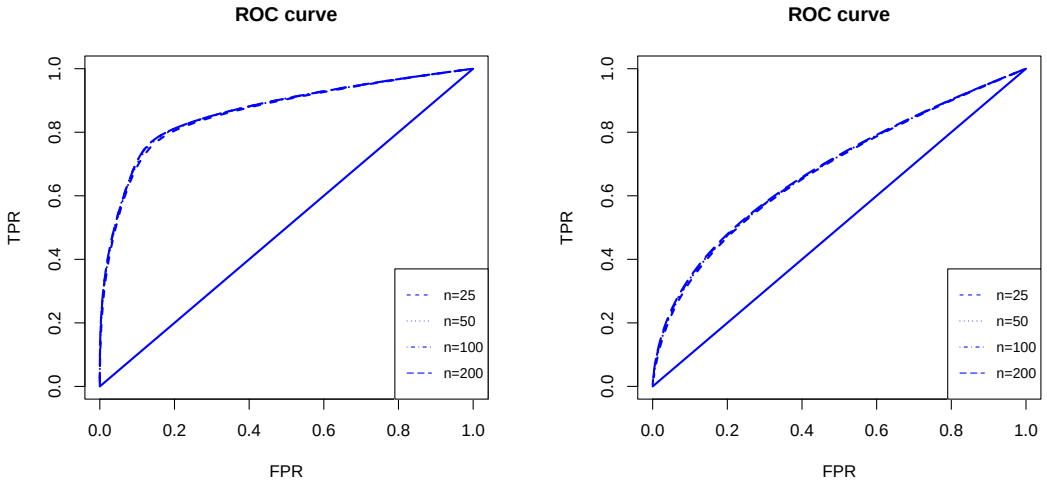
n	$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{\theta}_0$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{t}_1$	$\hat{t}_2$	$\hat{J}$	$\widehat{mFPR}$	$\widehat{mTPR}$	$\widehat{mAUC}$
Set A: $\hat{\theta}_0$=0.99; $\hat{\theta}_1$=0.3; $\hat{\theta}_2$=0.01											
25	0.5015	0.4985	1.0351	0.3136	0.0105	1.7019	11.5858	0.6277	0.1349	0.7625	0.8650
50	0.5024	0.4976	1.0151	0.3066	0.0102	1.7132	11.6439	0.6304	0.1335	0.7639	0.8662
100	0.4983	0.5017	0.9976	0.3035	0.0101	1.7264	11.6779	0.6325	0.1340	0.7664	0.8671
200	0.4986	0.5014	0.9908	0.3003	0.0101	1.7346	11.7392	0.6330	0.1342	0.7672	0.8674
Set B: $\hat{\theta}_0$=0.99; $\hat{\theta}_1$=0.6 $\hat{\theta}_2$=0.2											
25	0.5266	0.4734	1.0329	0.6287	0.2095	1.2612	2.6946	0.2485	0.2743	0.5228	0.6794
50	0.5143	0.4857	1.0122	0.6099	0.2033	1.2754	2.7382	0.2420	0.2792	0.5212	0.6831
100	0.5220	0.4780	0.9960	0.6066	0.2011	1.2823	2.7404	0.2415	0.2784	0.5199	0.6827
200	0.5109	0.4891	0.9940	0.6023	0.1997	1.2836	2.7512	0.2398	0.2819	0.5218	0.6851

Table 2: Partial area estimates of simulated datasets using method I

d_0	$\hat{c}_1$	c_2	$\widehat{mA}_{(c_1,c_2)}$	$\widehat{mTpAUC}_{(c_1,c_2,d_0)}$
Set A: $\hat{\theta}_0$=0.99; $\hat{\theta}_1$=0.3; $\hat{\theta}_2$=0.01				
0.65	0.08364	0.5	0.40231	0.13168
0.65	0.07992	0.5	0.39148	0.11842
0.65	0.08156	0.5	0.37855	0.10657
0.65	0.07934	0.5	0.37575	0.10233
Set B: $\hat{\theta}_0$=0.99; $\hat{\theta}_1$=0.6 $\hat{\theta}_2$=0.2				
0.65	0.39608	0.5	0.16230	0.09475
0.65	0.39208	0.5	0.15241	0.08226
0.65	0.39106	0.5	0.14953	0.07872
0.65	0.39220	0.5	0.14706	0.07699

Table 3: Partial area estimates of simulated datasets using method II

$Var(\widehat{mFPR})$	$Var(\widehat{mTPR})$	$\hat{d}_0$	$\hat{c}_1$	$\hat{c}_2$	$\widehat{mA}_{(c_1,c_2)}$	$\widehat{mTpAUC}_{(c_1,c_2,d_0)}$
Set A: $\hat{\theta}_0$=0.99; $\hat{\theta}_1$=0.3; $\hat{\theta}_2$=0.01						
0.00179	0.00206	0.68016	0.09353	0.21719	0.09823	0.01412
0.00096	0.00094	0.70578	0.10191	0.19527	0.07232	0.00642
0.00048	0.00046	0.72367	0.10988	0.17725	0.05133	0.00258
0.00025	0.00025	0.73621	0.11354	0.16539	0.03902	0.00084
Set B: $\hat{\theta}_0$=0.99; $\hat{\theta}_1$=0.6 $\hat{\theta}_2$=0.2						
0.00314	0.00302	0.41063	0.15767	0.35576	0.24532	0.16397
0.00118	0.00152	0.45334	0.17771	0.30952	0.15247	0.09272
0.00075	0.00116	0.46747	0.18847	0.29418	0.11795	0.06854
0.00046	0.00070	0.47924	0.20080	0.28167	0.08965	0.05089



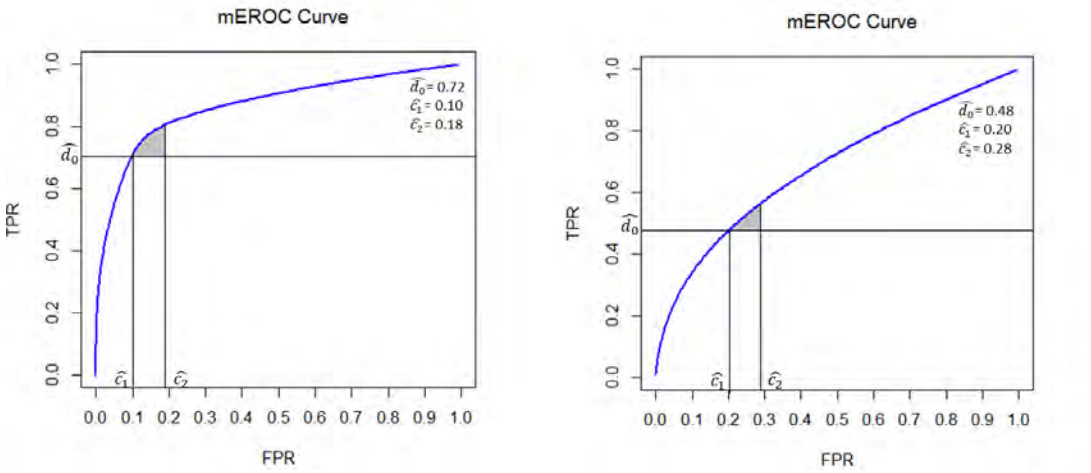
(a) Set A **(b) Set B**

Figure 4: ROC curve for simulated datasets

Coming to method II, from Table 3, we can observe that for n=100 in set A, by using the equations (8) and (9) the obtained value are $\hat{d}_0 = 0.7237$, $\hat{c}_2 = 0.1773$ and the respective $\widehat{mFPR}$ at $\hat{d}_0$ is $\hat{c}_1 = 0.1099$, and provides $\widehat{mA}_{(c_1,c_2)}$ and $\widehat{mTpAUC}_{(c_1,c_2,d_0)}$ of about 0.05133 and 0.00258 respectively. Similarly for n=100 in set B, the obtained value of $\hat{d}_0$, $\hat{c}_1$ and $\hat{c}_2$ are 0.46747, 0.18847 and 0.29418 which provides $\widehat{mA}_{(c_1,c_2)}$ and $\widehat{mTpAUC}_{(c_1,c_2,d_0)}$ of about 0.1179 and 0.06854 respectively.

3.2. Real datasets

To demonstrate the proposed methodology two real datasets are considered and their results are tabulated accordingly with respective ROC curves in Figure 7.



(a) Set A **(b) Set B**

Figure 5: TpAUC for set A and B for n=100

Data 1: Irradiated mice data

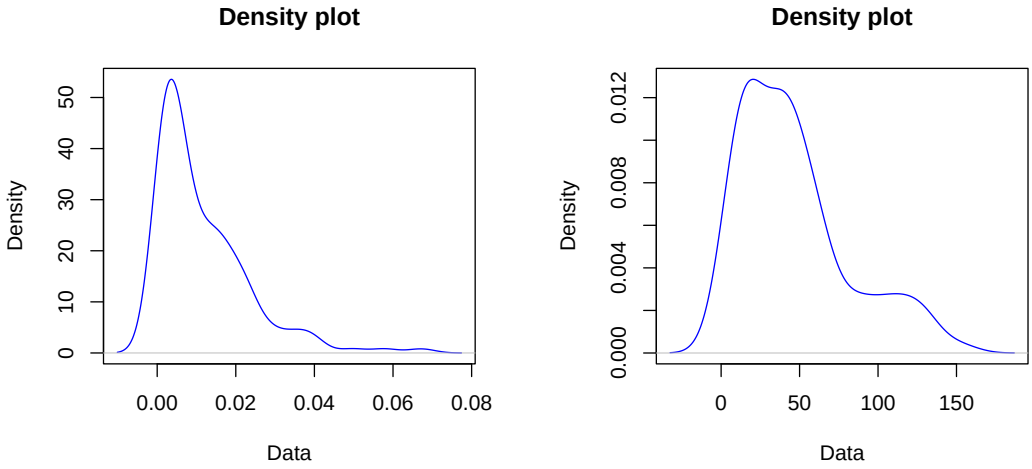
Irradiated mice data from Elandt-Johnson and Johnson (1980) is considered; the variable of interest is the time at the death of 99 mice. The p-value of Kolmogorov-Smirnov test for exponential distribution is 0.43 (test statistic (D)=0.113) which indicates that the data follows exponential distribution.

The density plot of the irradiated mice data is given in the Figure 6(a), and is very clear that there exists multi-modality which indicates the presence of sub-populations, i.e., the data is of multi-class. By using EM algorithm we identified that there are three classes and the estimated the parameters of the respective populations are $\hat{\theta}_0 = 0.3954$, $\hat{\theta}_1 = 0.0279$ and $\hat{\theta}_2 = 0.0201$. The proposed methodology is used to classify the data and the results are tabulated in the Tables 4 and 5.

It is observed that with thresholds $\hat{t}_1 = 12.8556$ and $\hat{t}_2 = 417.2446$, the overall $\widehat{mAUC}$ is about 0.7267, this indicates that the mEROC curve has accuracy about 72%, with false positives of 24% and true positives of 69%. This means to that a subject can be classified in the following manner.

$$\text{It is classified as} = \begin{cases} P_1, \text{ if } S \leq 12.8556 \\ P_2, \text{ if } 12.8556 < S \leq 417.2446 \\ P_3, \text{ if } S > 417.2446 \end{cases}$$

where P_1, P_2 and P_3 are the three respective classes. For method I, the d_0 and c_2 is taken as 0.6 and 0.4 respectively, and $\hat{c}_1$ which is the corresponding $\widehat{mFPR}$ at d_0 is 0.206488. Altogether, method I results $\widehat{mA}$ and $\widehat{mTpAUC}$ as 0.1873 and 0.0711 respectively. Coming to method II, by using equations (8) and (9) the obtained value for $\hat{d}_0$ and $\hat{c}_2$ are about 0.6948 and 0.2837, $\hat{c}_1$ which is the corresponding $\widehat{mFPR}$ at $\hat{d}_0$ is obtained as 0.2516. By method results the $\widehat{mA}$ and $\widehat{mTpAUC}$ are 0.0381 and 0.0158 respectively. Since, the difference between c_2 and c_1 is too small, the TpAUC portion on the mEROC curve is difficult to depict. However, the TpAUC is shown for breast cancer data (Figure 8).



(a) Mice data **(b) Breast Cancer data**

Figure 6: Density plots of real datasets

Data 2: Breast Cancer data

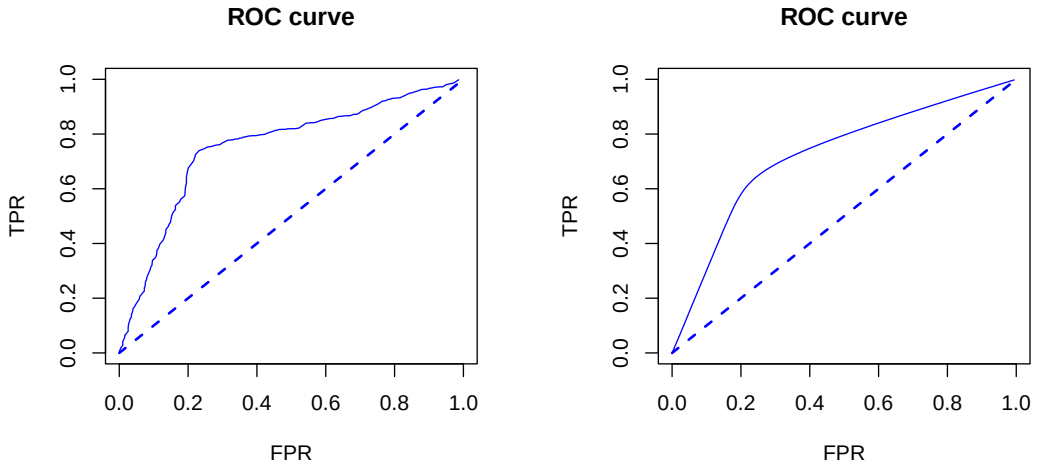
The real dataset represent the survival times of 121 patients with *breast cancer* obtained from a large hospital in a period from 1929 to 1938 (Lee and Wang, 2003). The p-value of K-S test for exponential distribution is 0.06024 (test statistics (D)=0.12031) which indicates that the data follows exponential distribution. The estimation is done by using both the methods and respective results are shown in Tables 4 and 5.

The density plot of the breast cancer data is given in Figure 6 (b), and is very clear that there exists multi-modality which indicates the presence of sub-populations. By using EM algorithm we identified that there are three classes and the estimated the parameters of the respective populations are $\hat{\theta}_0 = 0.4010$, $\hat{\theta}_1 = 0.0280$ and $\hat{\theta}_2 = 0.0202$. The optimal thresholds, $\hat{t}_1$ and $\hat{t}_2$ are 7.7311 and 44.6911, which gives accuracy about 75.26% with false positive rates about 20.63% and true positives of 63.93%. This means to that a subject can be classified in the following manner

$$\text{It is classified as} = \begin{cases} P_1, \text{ if } S \leq 7.7311 \\ P_2, \text{ if } 7.7311 < S \leq 44.6911 \\ P_3, \text{ if } S > 44.6911 \end{cases}$$

where P_1, P_2 and P_3 are the three respective classes with low, medium and high survival rate respectively.

For method I, the $\hat{d}_0$ and $\hat{c}_2$ are taken as 0.6 and 0.5 and the $\hat{c}_1$ corresponding to $\hat{d}_0$ is obtained as 0.1986, which results $\widehat{mA}$ and $\widehat{mTpAUC}$ about 0.1935 and 0.01266. By method II the values obtained for $\hat{d}_0$ and $\hat{c}_2$ are about 0.6123 and 0.2837, the corresponding $\hat{c}_1$ to $\hat{d}_0$ is 0.2142, altogether results $\widehat{mA}$ and $\widehat{mTpAUC}$ of 0.0509 and 0.0084 respectively. The TpAUC for breast cancer data is depicted in Figure 8.



(a) Mice data **(b) Breast Cancer data**

Figure 7: ROC curves for real datasets

Table 4: ROC curve measures of real datasets

$\hat{\pi}_1$	$\hat{\pi}_2$	$\hat{t}_1$	$\hat{t}_2$	$\widehat{mFPR}$	$\widehat{mTPR}$	$Var(\widehat{mFPR})$	$Var(\widehat{mTPR})$	$\hat{J}$	$\widehat{mAUC}$
Mice data									
0.4998	0.5002	12.8556	417.2446	0.2435	0.6973	0.000333	0.000308	0.45378	0.726739
Breast Cancer data									
0.4999	0.5001	7.7311	44.6911	0.20627	0.6393	0.000371	0.000286	0.432985	0.752647

Table 5: Partial area estimates of real datasets

Dataset	Method	$\hat{d}_0$	$\hat{c}_1$	$\hat{c}_2$	$\widehat{mA}_{(c_1,c_2)}$	$\widehat{mTpAUC}_{(c_1,c_2,d_0)}$
Mice	I	0.6	0.206488	0.4	0.187254	0.071147
	II	0.69483	0.251559	0.283712	0.038135	0.015794
Breast Cancer	I	0.6	0.19857	0.5	0.193517	0.012659
	II	0.61229	0.214157	0.283665	0.050896	0.008337

4. Summary

In this work, we made an attempt to explain the need and importance of analyzing a portion of the ROC curve for multi-class non-normal data. Methodological descriptions are given in detail for one-way and two-way pAUC. Expressions for mpAUC and mTpAUC are also derived, and the terms d_0 and c_2 involved in these expressions are obtained using asymptotic confidence intervals of mFPR and mTPR. Two real datasets and considerable simulations are used to demonstrate the proposed work. From the results it is observed that for a multi-class ROC curve, whose area is maximum (minimum), the areas within the mFPR range and mTpAUC will also have a larger (smaller) portion in the entire area.

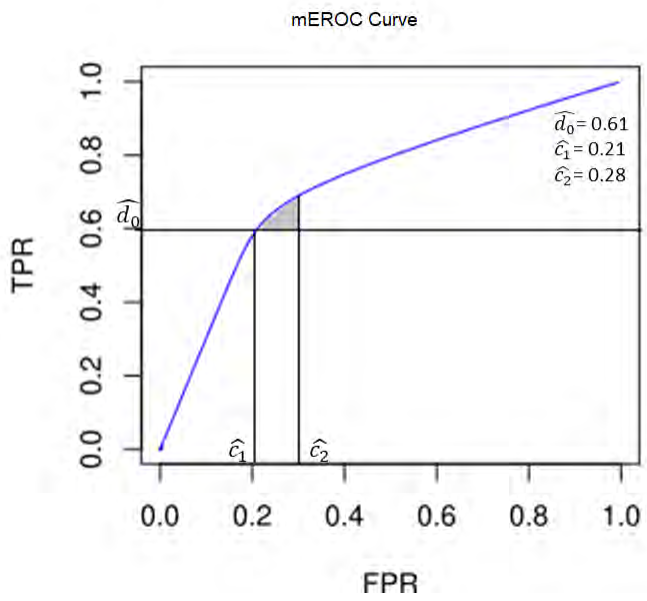


Figure 8: Two way pAUC for Breast Cancer data

Acknowledgements

The authors indeed grateful to the Editors for their guidance and counsel. The authors are grateful to the reviewer for valuable comments and suggestions of generously listing many useful references.

References

Arunima, S. K. and Vishnu, R. V. (2022). Estimation of area under the roc curve in the framework of gamma mixtures. *Communications in Statistics: Case Studies, Data Analysis and Applications*, **8**, 1–14.

Arunima, S. K. and Vishnu, R. V. (2023). Estimation of area under the multi-class roc for non-normal data. *Statistics and Applications*, **21**, 113–121.

Baker, S. G. and Pinsky, P. F. (2001). A proposed design and analysis for comparing digital and analog mammography: special receiver operating characteristic methods for cancer screening. *Journal of the American Statistical Association*, **96**, 421–428.

Cheam, A. S. and McNicholas, P. D. (2016). Modelling receiver operating characteristic curves using gaussian mixtures. *Computational Statistics and Data Analysis*, **93**, 192–208.

Elandt-Johnson, R. C. and Johnson, N. L. (1980). *Survival models and data analysis*. John Wiley and Sons.

Gönen, M. (2013). Mixtures of receiver operating characteristic curves. *Academic Radiology*, **20**, 831–837.

Hillis, S. L. and Metz, C. E. (2012). An analytic expression for the binormal partial area under the roc curve. *Academic Radiology*, **19**, 1491–1498.

Jiang, Y., Metz, C. E., and Nishikawa, R. M. (1996). A receiver operating characteristic partial area index for highly sensitive diagnostic tests. *Radiology*, **201**, 745–750.

Lee, E. T. and Wang, J. (2003). *Statistical Methods for Survival Data Analysis*, volume 476. John Wiley and Sons.

Lusted, L. B. (1971). Signal detectability and medical decision-making: Signal detectability studies help radiologists evaluate equipment systems and performance of assistants. *Science*, **171**, 1217–1219.

McClish, D. K. (1989). Analyzing a portion of the roc curve. *Medical Decision Making*, **9**, 190–195.

Obuchowski, N. A. (2003). Receiver operating characteristic curves and their use in radiology. *Radiology*, **229**, 3–8.

Obuchowski, N. A. and Bullen, J. A. (2018). Receiver operating characteristic (roc) curves: review of methods with applications in diagnostic medicine. *Physics in Medicine and Biology*, **63**, 1–17.

Peterson, W., Birdsall, T., and Fox, W. (1954). The theory of signal detectability. *Transactions of the IRE Professional Group on Information Theory*, **4**, 171–212.

Ricamato, M. T. and Tortorella, F. (2011). Partial auc maximization in a linear combination of dichotomizers. *Pattern Recognition*, **44**, 2669–2677.

Siva, G. and Vishnu, V. R. (2022). Multi-class classification using mixtures of univariate and multivariate roc curves. *Journal of Biostatistics and Epidemiology*, **8**, 209–233.

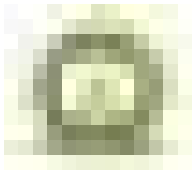
Tanner, J. W. P. and Swets, J. A. (1954). A decision-making theory of visual detection. *Psychological Review*, **61**, 401.

Thompson, M. L. and Zucchini, W. (1989). On the statistical analysis of roc curves. *Statistics in Medicine*, **8**, 1277–1290.

Zhang, L., Jie, Z., Xu, S., Zhang, L., and Guo, X. (2018). Use of receiver operating characteristic (roc) curve analysis for tyrer-cuzick and gail in breast cancer screening in jiangxi province, china. *Medical Science Monitor: International Medical Journal of Experimental and Clinical Research*, **24**, 5528.

Zhou, X.-H., McClish, D. K., and Obuchowski, N. A. (2009). *Statistical Methods in Diagnostic Medicine*. John Wiley and Sons.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 22, No. 1, 2024 (New Series), pp 279–288
http://www.ssca.org.in/journal



On the Robustness of LSD Layouts in the Presence of Neighbor Effects

Sobita Sapam[1] **and Bikas K Sinha**[2]
[1]*Manipur University, Canchipur, Imphal, Manipur, India*
[2]*Indian Statistical Institute, Kolkata (Retired Professor), India*
Received: 04 May 2023; Revised: 05 September 2023; Accepted: 11 September 2023

Abstract

In this paper we study the status of four non isomorphic Latin Square Designs (LSDs) of order four while finding out the optimal covariate matrices underlying the LSDs with and without Neighbor Effects (NEs). In these LSDs we consider the four sided NEs viz., left-sided, right-sided, top-sided and bottom-sided in the presence of covariates' effects. We utilize a circular model as was introduced by Kunert. Without NEs each of the four LSDs has six optimal covariate matrices whereas in the presence of the four-sided NEs the results are not as expected, for all the four LSDs.

Key words: Non isomorphic LSDs; Optimal designs for covariates effects; Neighbor effects; Circular models.

AMS Subject Classifications: 62K10

1. Introduction

As we know, there is a long history of use of ANCOVA Models for effective data analysis involving standard and non-standard experimental designs. Troya (1982a, 1982b) introduced the concept of Optimal Covariates Designs and presented optimality results in the context of CRDs. Inspired by Troya's formulation of optimality problems involving covariates effects, Das *et al.* (2003) got interested in this area of research and provided some combinatorial solutions. That was a modest beginning and much of the contents of the Monograph on Optimal Covariate Designs by Das *et al.* (2015) were motivated and inspired by 2003 paper. Prominent contributors in this area of research found their place and citations in the list of references of the monograph. This fascinating topic still holds rich rewards for serious researchers.

The first author (Sapam) got interested in this area of research and the recent works by Sapam *et al.* (2021) hold the key references for this paper. Optimal Covariate Designs (OCDs) are the designs which provide optimal or most efficient estimation of the covariates' effects in terms of the parameters in an assumed linear model. Lopes Troya (1982a, 1982b), Das *et al.* (2003), Shah and Sinha (1989), Dutta *et al.* (2014) are some of the related

Corresponding Author: Sobita Sapam
Email: sobita1@yahoo.com

references on the OCDs. Sapam *et al.* (2021) focused on OCDs incorporating the neighbor effects in four directions viz., left-sided, right-sided, top-sided and bottom sided in the assumed linear model in different RBD set ups. Sinha and Dutta (2017) worked on three different seasons of LSDs of order four without any NEs. We consider the four non-isomorphic LSDs across the three seasons, as considered in Sinha and Dutta (2017), with and without NEs. We crosscheck the earlier results and provide a few optimal matrices in the presence of the NEs. The notions of NEs are widely studied in the literature; some relevant references are Bailey (2003), Jaggi *et al.* (2007), Jaggi *et al.* (2018), Varghese *et al.* (2014), Sapam *et al.* (2019a, 2019b).

As the readers can realize, this area of research blends (block) designs [such as CRDs, RBDs, LSDs, GLSDs, BIBDs, *etc.*) and neighbour effects (introduced through what are known as circular models) and on the top, there are combinatorial arrangements of $(+1/-1)$'s. Most of the reference papers bear testimony to the authors' interest in these areas. An interested reader will benefit by reading Das *et al.* (2003) before proceeding to venture in complicated set-ups. Not to obscure the essential steps of reasoning and understanding, we describe the linear model in simple terms with quantitative covariates and with/without neighbour effects. Since we will be primarily dealing with LSDs in this paper, we restrict to the LSD of order 4 shown in Table 1. There are altogether $4 \times 4 = 16$ plots and we have an LSD laid out there. We describe the linear model for some special plots - covering all diverse positions of the treatments - with/without neighbour effects of the treatments [in all the four positions].

Confining to LSD - S-1, wrt the treatment 1 in the first row and first column, the model specifications is as follows.

y(1,1;1)= $\mu+\rho_1+\gamma_1+\ \tau_1\ +e_{11}$
y(1,1;1)= $\mu+\rho_1+\gamma_1+\ \tau_1\ +LN4+RN2+TN4+BN2+e_{11}$

The parameters involved in the model are obvious. In linear model with four-sided neighbor effects, we have inserted LN, RN, TN and BN effects on the (Direct) Effects ('s) of the treatments.

In order that the readers can accompany and comprehend the thought process we refer to Das *et al.* (2003) wherein the conditions for existence of OCDs has been explicitly laid down.

2. LSDs of order 4 with covariates without neighbor effects

Taking the four treatments 1, 2, 3, 4 let us perform the complete enumeration principle to obtain all the possible forms of Standard Latin Squares. The following four non-isomorphic Standard Latin Square designs viz., S-1, S-2, S-3, S-4 are the only possible LSDs. The matrix X is the general form of covariate matrix wrt the four LSDs.

There should be three conditions for all optimal X- matrices wrt each of the above LSDs, viz., S-1, S-2, S-3 and S-4 without any neighbor effects, the elements of X-matrices being (+1/-1):
(i) Row totals of the optimal X - matrices = 0
(ii) Column totals of the optimal X - matrices = 0

(iii) Each treatment totals of the optimal X - matrices = 0

Table 1: Four non-isomorphic Standard Latin Square designs

LSDs				
S-1	1	2	3	4
	2	1	4	3
	3	4	1	2
	4	3	2	1
S-2	1	2	3	4
	2	3	4	1
	3	4	1	2
	4	1	2	3
S-3	1	2	3	4
	2	1	4	3
	3	4	2	1
	4	3	1	2
S-4	1	2	3	4
	2	4	1	3
	3	1	4	2
	4	3	2	1

Table 2: General X-matrix

a	b	c	d
e	f	g	h
i	j	k	l
m	n	o	p

Sinha and Dutta (2017) studied the LSDs of order 4 in 3 Different SEASONS, viz., S-1, S-2, S-3 and worked out forms of some optimal covariate matrices in the absence of neighbor effects. Below we are showing six optimal covariate matrices for each of S-1, S-2, S-3 and for another additional design S-4 as well. We denote the six optimal X-matrices by S-1C1 to S-1C6 wrt S-1 design, with the same notation S-2C1 to S-2C6 wrt S-2 design and so on.

3. LSDs S-1, S-2, S-3 and S-4 with covariates in the presence of neighbor effects

The following conditions should hold in the presence of the four-sided neighbor effects viz., Left Neighbor (LN), Right Neighbor (RN), Top Neighbor (TN) and Bottom Neighbor (BN), involving the elements (+1/-1) of each of the X-matrices for the LSDs.

(i) row sum of the optimal X - matrices = 0
(ii) column sum of the optimal X - matrices = 0
(iii) each treatment sum of the optimal X - matrices = 0
(iv) sum of covariate-values of the optimal X - matrices corresponding to the LN of each

Table 3: Six optimal covariates matrix of S-1

	1	-1	1	-1
	1	-1	1	-1
S-1C1	-1	1	-1	1
	-1	1	-1	1
	1	-1	-1	1
	-1	1	1	-1
S-1C2	1	-1	-1	1
	-1	1	1	-1
	1	1	-1	-1
	-1	-1	1	1
S-1C3	1	1	-1	-1
	-1	-1	1	1
	1	-1	1	-1
	-1	1	-1	1
S-1C4	-1	1	-1	1
	1	-1	1	-1
	1	-1	1	-1
	1	-1	1	-1
S-1C5	-1	1	-1	1
	-1	1	-1	1
	1	1	-1	-1
	-1	-1	1	1
S-1C6	1	1	-1	-1
	-1	-1	1	1

treatment = 0
(v) sum of covariate-values of the optimal X - matrices corresponding to the RN of each treatment =0
(vi) sum of covariate-values of the optimal X - matrices corresponding to the TN of each treatment =0
(vii) sum of covariate-values of the optimal X - matrices corresponding to the BN of each treatment = 0.

When we consider the four-sided neighbor effects for each of the designs S-1, S-2, S-3 and S-4, the above $6 \times 4 = 24$ covariate matrices [optimal in the absence of neighbor effects (NEs)] do not all satisfy all the properties listed in (i)-(vii). The designs S-1 and S-2 has each six optimal covariate matrices in the presence of four-sided NEs, viz., S-1C1, S-1C2, S-1C3, S-1C4, S-1C5, S-1C6 and S-2C1, S-2C2, S-2C3, S-2C4, S-2C5 and S-2C6, satisfying all the conditions (i)- (vii) mentioned above for being optimal X -matrices in the presence of all the four-sided NEs. On the other hand, the designs S-3 and S-4 there is not even a single optimal X-matrix in the presence of four sided neighbor effects.

Table 4: Six optimal covariates matrix of S-2

	1	1	-1	-1
	-1	-1	1	1
S-2C1	1	1	-1	-1
	-1	-1	1	1
	1	-1	1	-1
	-1	1	-1	1
S-2C2	-1	1	-1	1
	1	-1	1	-1
	1	-1	-1	1
	-1	1	1	-1
S-2C3	1	-1	-1	1
	-1	1	1	-1
	1	-1	1	-1
	1	-1	1	-1
S-2C4	-1	1	-1	1
	-1	1	-1	1
	1	1	-1	-1
	-1	1	1	-1
S-2C5	-1	-1	1	1
	1	-1	-1	1
	1	-1	-1	1
	1	1	-1	-1
S-2C6	-1	1	1	-1
	-1	-1	1	1

4. Existence and non-existence of optimal X-matrices with/without NEs: Status of the LSDs S-1 and S-4

Consider the LSD S-1 in the presence of four sided NEs and examine all the eight combinations corresponding to the choices of (b,e,f), setting a=1. If there exists a solution satisfying the above conditions (i)- (vii) with the solution space [1,-1], an optimal X- matrix will be available in the presence of four-sided neighbor effects.

Case1: b=e=f=1: no solution,
Case 2: b= -1, e=f=1: no solution,
Case 3: b=f= -1, e=1: two solutions viz., S-1C1 & S-1C5,
Case 4: e= -1, b=f=1: no solution,
Case 5: f= -1, b=e=1: no solution,
Case 6: b=e= -1, f=1: two solutions, viz., S-1C2 & S-1C4,
Case 7: e=f= -1, b=1: one solution, viz., S-1C3 & S-1C6,
Case 8: b=e=f= -1: no solution.

Therefore, total number of optimal X- matrices obtained wrt S-1 is six [viz., S-1C1, S-1C2, S-1C3, S-1C4, S-1C5, S-1C6] in the presence of four-sided NEs.

Table 5: Six optimal covariates matrix of S-3

	1	1	-1	-1
	-1	-1	1	1
S-3C1	1	1	-1	-1
	-1	-1	1	1
	1	1	-1	-1
	-1	-1	1	1
S-3C2	-1	-1	1	1
	1	1	-1	-1
	1	-1	-1	1
	-1	1	1	-1
S-3C3	1	-1	1	-1
	-1	1	-1	1
	1	-1	1	-1
	-1	1	-1	1
S-3C4	-1	1	1	-1
	1	-1	-1	1
	1	-1	1	-1
	1	-1	1	-1
S-3C5	-1	1	-1	1
	-1	1	-1	1
	1	-1	-1	1
	1	-1	-1	1
S-3C6	-1	1	1	-1
	-1	1	1	-1

Next, consider LSD S-4 in the presence of four sided NEs. If there exists a solution satisfying the above conditions (i)- (vii) with the solution space [1,-1], an optimal X- matrix exists. WOLG, using the notations of the general covariate matrix given above, we set, a=1 and examine all the eight combinations corresponding to the choices of (b, e, f). The following are the cases:
Case1: b=e=f=1; there is no solution [since, treatment 2 sum cannot be zero]
Case 2: b=e=1, f=-1; there is no solution [since, 2nd column sum cannot be zero]
Case 3: b=f=1, e= -1; there is no solution, [since, LN of Tr. 1 sum cannot be zero]
Case 4: b=1, e=f= -1; there is no solution, [since, LN of Tr. 1 sum cannot be zero]
Case 5: b= -1, e=f=1; there is no solution, [since, LN of Tr. 1 sum cannot be zero]
Case 6: b= -1, e= -1, f=1; there is no solution, [here two subcases arise: in one case TN of Tr1 sum is not equal to zero and in another subcase LN of Tr. 1 sum is not equal to zero]
Case 7: b= -1, e=1, f= -1; there is no solution, [since, LN of Tr. 2 sum is not equal to zero]
Case 8: b= e= f= -1; there is no solution [since, 2nd column sum cannot be zero].

This shows that there is not even a single optimal X- matrix in the presence of four sided NEs for the LSD S-4. Further, consider the LSD S-4 without NEs. If there exists a solution satisfying the above conditions (i)- (iii) of section 2 with the solution space [1,-1], an

Table 6: Six optimal covariates matrix of S-4

	1	1	-1	-1
	-1	1	-1	1
S-4C1	1	-1	1	-1
	-1	-1	1	1
	1	1	-1	-1
	-1	-1	1	1
S-4C2	-1	-1	1	1
	1	1	-1	-1
	1	-1	1	-1
	1	1	-1	-1
S-4C3	-1	-1	1	1
	-1	1	-1	1
	1	-1	-1	1
	-1	1	1	-1
S-4C4	1	-1	-1	1
	-1	1	1	-1
	1	-1	1	-1
	-1	1	-1	1
S-4C5	-1	1	-1	1
	1	-1	1	-1
	1	-1	-1	1
	1	-1	-1	1
S-4C6	-1	1	1	-1
	-1	1	1	-1

optimal X- matrix exists. WOLG, using the notations of the general covariate matrix given above, we set, a=1 and examine all the eight combinations corresponding to the choices of (b, e, f).
Case1: b=e=f=1; there is no solution [since, treatment 2 total cannot be zero].
Case 2: b=e=1, f=-1; there is no solution [since, second column total cannot be zero].
Case 3: b=f=1, e= -1; there is one solution, viz., S-4C1
Case 4: b=1, e=f= -1; there is one solution, viz., S-4C2
Case 5: b= -1, e=f=1; there is one solution, viz., S-4C3
Case 6: b= -1, e= -1, f=1; there are two solutions, viz., S-4C4 and S-4C5
Case 7: b= -1, e=1, f= -1; there is one solution, viz., S-4C6
Case 8: b= e= f= -1; there is no solution [since, 1st column sum cannot be zero]
These eight cases show the existence of six optimal X-matrices in the absence of NEs wrt S-4 design.

5. Concluding remarks

In the study of four non isomorphic LSDs of order four with and without NEs we can summarize that in the absence of NEs, for each of the designs S-1, S-2, S-3 and S-4 of LSD

of order 4, we can find out all the possible (six) optimal X-matrices. On the other hand, in all the four LSDs, these optimal matrices fail to be optimal when we incorporate the four sided NEs. Only for the designs S-1 and S-2 all the six optimal X-matrices continue to be so even in the presence of NEs. The other two LSDs S-3 and S-4 has no X-matrix. Now we can sum up in the following table as Annexure I and II, the reasons of disqualification and their corresponding optimal X-matrices with respect to the design S-4 in the presence of neighbor effects.

Acknowledgements

The first author acknowledges financial support from Women Scientist Project No. **DST/WOS-A/PM-27/2021, DST, New Delhi**. She also thanks her Mentor Professor KK Singh Meitei for providing all facilities towards successfully pursuing her research in the broad area of DoE and also for arranging academic visits of Professor Sinha to the Manipur University for collaborative research.

References

Bailey, R. A. (2003). Designs for one sided neighbor effects. *Journal of Indian Society of Agricultural Statistics*, **56**, 302-314.

Das, K., Mandal, N. K., and Sinha, B. K. (2003). Optimal experimental designs for models with covariates. *Journal of Statistical Planning and Inference*, **115**, 273-285.

Das, P., Dutta. G., Mandal, N. K., and Sinha, B. K. (2015). *Optimal Covariate Designs*. Springer Verlag Text Book Series.

Dutta, G., Das, P., and Mandal, N. K. (2014). D-optimal designs for covariate models. *Communication in Statistics - Theory and Methods*, **43**, 165-174.

Jaggi, S., Varghese, C., and Gupta, V. K. (2007). Optimal circular block designs for competition effects. *Journal of Applied Statistics*, **34**, 577-584.

Jaggi, S., Pateria, D. K., Varghese, C., Varghese, E., and Bhowmik, A. (2018). A note on Circular neighbor balanced designs. *Communication in Statistics - Simulation and Computation*, **47**, 2896-2905.

Sapam, S., Meitei, K. K. S., and Sinha, B. K. (2021). Randomized block designs, balanced incomplete block designs and latin square designs with neighbor effects in the presence of covariates. *Statistics and Applications* (New Series), **19**, 19 - 31. Special Volume in Memory of Late Aloke Dey.

Sapam, S., Mandal, N. K., and Sinha, B. K. (2019a). Latin square designs with neighbor effects -part II. *Communication in Statistics - Theory and Methods*, https://doi.org/10.1080/03610926.2019.1702694.

Sapam, S., Mandal, N. K., and Sinha, B. K. (2019b). Latin square designs with neighbor effects. *Journal of Indian Society of Agricultural Statistics*, **73**, 91-98.

Shah, K. R. and Sinha, B. K. (1989). *Theory of Optimal Designs*. Lecture notes in Statistics. Series 54, Springer, New York.

Sinha, B. K. and Dutta, G. (2017). Groups of Latin square designs in agricultural experiments with covariate, *RASHI*, **2**, 01-05.

Troya, L. J. (1982a). Optimal designs for covariates models. *Journal of Statistical Planning and Inference*, **6**, 373-419.

Troya, L. J (1982b). Cyclic designs for covariates models. *Journal of Statistical Planning and Inference*, **7**, 49-75.

Varghese, E., Jaggi, S., and Varghese, C. (2014). Neighbor balanced-row column designs. *Communication in Statistics - Theory and Methods*, **43**, 1261-1276.

ANNEXURE I

Table 7: Reasons for disqualification of the X-matrices in the presence of NE wrt S-4

Sl no.	X- matrices*	Reasons for disqualification in the presence of NEs
1	1 1 -1 -1 -1 1 -1 1 1 -1 1 -1 -1 -1 1 1	LN=RN=2 for Treatment 1
2	1 1 -1 -1 -1 -1 1 1 -1 -1 1 1 1 1 -1 -1	LN= - 4 and RN= 4 for Treatment 1
3	1 -1 1 -1 1 1 -1 -1 -1 -1 1 1 -1 1 -1 1	LN= -2 and RN =2 for Treatment 1
4	1 -1 -1 1 -1 1 1 -1 1 -1 -1 1 -1 1 1 -1	LN = 4 and RN = -2 for Treatment 1
5	1 -1 1 -1 -1 1 -1 1 -1 1 -1 1 1 -1 1 -1	TN of Tr1 = 4 and BN = - 4
6	1 -1 -1 1 1 -1 -1 1 -1 1 1 -1 -1 1 1 -1	LN= 4 and RN = 4 wrt Tr. 2

X- Matrices* are without NEs

ANNEXURE II

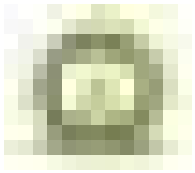
The design S-4 in the presence of four sided the neighbor effects the above six covariate matrices wrt S-4, we can see that there is no optimal matrix. The reasons are shown as below.

Table 8: Reasons for disqualification of the X-matrices in the presence of four-sided NEs wrt S-4C1 to S-4C6

Sl no.	Treatments	LN effects	RN effects
1	1	-1+1+1+1= 2	1+1+1-1 = 2
S-4C1	2	1+1+1-1 = 2	-1+1+1+1= 2
	3	1-1-1-1 = -2	-1-1-1+1= -2
	4	-1-1-1+1 =-2	1-1-1-1 = -2
2	1	-1-1-1-1 = -4	1+1+1+1 = 4
S-4C2	2	1+1+1+1 = 4	-1 -1 -1 -1 = -4
	3	1+1+1+1 = 4	-1 -1 -1 -1 = -4
	4	-1 -1 -1 -1 = -4	1+1+1+1 = 4
3	1	-1 + 1 - 1 - 1 = -2	-1 - 1 + 1 - 1 = -2
S-4C3	2	1 - 1 + 1 + 1 = 2	1 + 1 - 1 + 1 = 2
	3	-1 - 1 + 1 - 1 = -2	-1 + 1 - 1 - 1 = -2
	4	1 + 1 - 1 + 1 = 2	1 - 1 + 1 + 1 = 2
4	1	1+1+1+1 =4	-1 -1 -1 -1 =4
S-4C4	2	1 -1 - 1 +1 =0	-1 +1 +1 -1 =0
	3	-1 +1 +1-1 =0	1 -1 -1 +1 =0
	4	-1 -1 -1 -1 = -4	1+ 1+1+1 = 4
5	1	1 -1 -1 +1 =0	-1 +1 +1 -1 = 0
S-4C6	2	1+1 +1+1= 4	-1 -1 -1 -1 = - 4
	3	-1 -1 -1 -1 = - 4	1+ 1+1+1 = 4
	4	-1 +1 +1 -1 = 0	1 - 1 -1 +1 = 0
	Treatments	TN effects	BN effects
6	1	1+1+1+1 = 4	-1 -1 -1 -1 = - 4
S-4C5**	2	-1+1+1 -1 = 0	1 - 1 - 1 +1 = 0
	3	1 - 1 -1 + 1 = 0	-1 +1 +1 -1 = 0
	4	-1 -1 -1 -1 = - 4	1 +1 + 1+1 =4

In the case of S-4C5** conditions for both LN and RN effects for each treatment are satisfied whereas for Top Neighbor (TN) and Bottom Neighbor (BN) effects conditions are not satisfied, hence we take up only TN and BN effects in the Sl no. 6.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 22, No. 1, 2024 (New Series), pp 289–301
http://www.ssca.org.in/journal



Statistical Analysis on Optimal Lockdown Schedule by Developing a Multivariate Prediction model SEIRDVIm

Subhasree Bhattacharjee[1], Kunal Das[2], Sahil Zaman[2], Arindam Sadhu[3] and Bikramjit Sarkar[4]

[1]*Computer Application Department*
Narula Institute of Technology, Kolkata, West Bengal, India
[2]*Computer Science Department*
Acharya Prafulla Chandra College, New Barrackpur, Kolkata-700131, India
[3]*ECE department, Greater Kolkata College of Engineering and Management*
Dudhnai, Ramnagar, Baruipur, West Bengal 743387, India
[4]*Computer Science and Engineering Department, JIS College of Engineering, Nadia, West Bengal, India.*

Received: 21 July 2023; Revised: 21 September 2023; Accepted: 09 October 2023

Abstract

Considering the high infection rate and bed scarcity in hospitals amidst the COVID-19 pandemic it is necessary to find out an optimal lockdown schedule for minimizing infection rate as well as maintaining economic sustainability. This paper proposes an effective compartmental model SEIRDVI$_{\mathrm{m}}$ and yields an optimal lockdown schedule using classical and quantum knapsack algorithms. When the available bed count falls below a certain threshold, the city goes into lockdown mode, and vice versa. The R^2 value of SEIRDVIm is 0.8797 and the Mean Squared Error (RMSE) is 34.59. The proposed model yields better results compared to the classical SEIR model. Variation of infected with vaccination rate and effectiveness of vaccination is demonstrated. Using 10 predictors it is found that for 60 days, quantum-assisted lockdown yields a death toll of 15062 compared to 20123 in classical knapsack induced lockdown.

Key words: SEIRDVIm model; Death rate; Knapsack problem; Lockdown schedule; Mean Squared Error; R-squared (R^2).

AMS Subject Classifications: 62K05, 05B05

1. Introduction

In December 2019 the outbreak of the novel severe acute respiratory syndrome Coronavirus called SARS-CoV-2 started locally in Wuhan, China, and rapidly spread all over the world. As reported 65.8 lakh deaths all over the world on 23rd October 2022. For deciding public policy several epidemic models have been used by the nation during the past few

Corresponding Author: Kunal Das
Email: kunal@apccollege.ac.in

years, see Ferguson *et al.* (2020). It is important to understand the impact of precautionary measures and medical intervention on multiple variants. The impact of effective vaccination in to fight against COVID-19 is tremendous. Vaccination production and proper distribution are important. Future prediction on pandemic significantly dominates vaccine distribution. For studying the effect of vaccination, additional compartments have been added to the existing models to analyze the effectiveness of vaccination. Matrajt *et al.* (2021) studied the effectiveness of vaccination for allocating vaccines properly. In the past also several analyses have been done on vaccination at the time of previous outbreaks, see Feng *et al.* (2011), Scherer and McLean (2002) and Chowell *et al.* (2019). For studying the spread of a disease in a population, SIR-based epidemic models are widely used, see Cooper *et al.* (2020), Kuhl and Kuhl (2021), and others. An extended SEIR-based model to predict the future trend of COVID-19 has been proposed by Lal *et al.* (2021). The main framework of the study by Davies *et al.* (2020) is the transmission of disease using age-based modeling. In research it is discussed the population behavior a level of caution and sense of safety while considering vaccine efficacy, see Usherwood *et al.* (2021).

This paper proposes an effective compartmental model SEIRDVI_m and yields an optimal lockdown schedule using classical and quantum knapsack algorithms. When the available bed count falls below a certain threshold, the city goes into lockdown mode, and vice versa. The novel contributions of this research article are as follows:

Proposed a new compartmental model SEIRDVI_m for designing an optimal lockdown schedule using the quantum knapsack algorithm by maximizing the objective function, available bed capacity and minimizing the death and then converting the objective function into an energy function using binary quadratic model (bqm). Then minimize the same by D-Wave Quantum Annealer.

This paper is represented as follows. Section 1 illustrates the introduction. The newly proposed model SEIRDVIm is discussed in Section 2. Section 3 highlights the result of the evolution of the proposed model with lockdown optimization using classical and quantum knapsack. Section 4 provides the discussion of the work and at last section 5 concludes the paper.

2. Methods

The proposed SEIRDVI_m model in Figure 1 divides the population into susceptible (S), exposed (E), infected incompletely vaccinated (I_{iv}), infected completely vaccinated (I_{cv}), recovered (R), vaccinated (V), Immunized (I_m), and deceased (D).

SEIRDVI_m model is described by eight linear differential equations. Variation of eight compartments S, E, I_{iv}, I_{cv}, R, V, I_m, and D with time (t) are depicted in equations 1 to 8. The assumptions of the model are:

I. The population is fixed.

II. After being completely vaccinated, a person can become infected with a lower rate of infection.

III. After complete and successful vaccination, immunity may be gained at rate η.

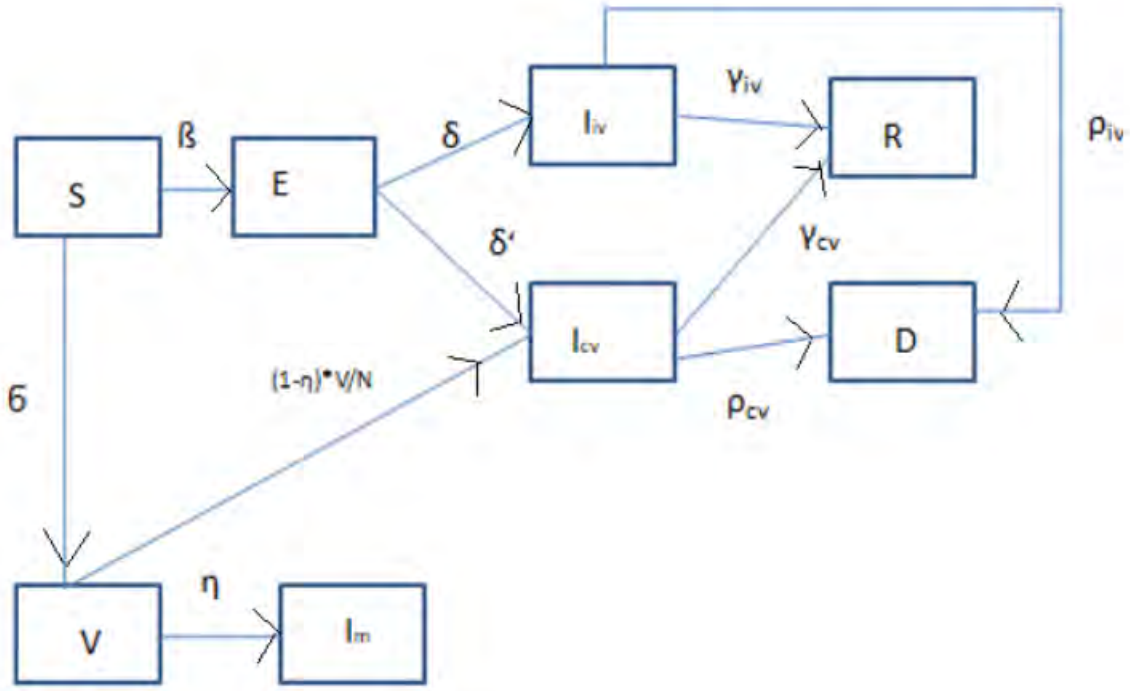


Figure 1: SEIRDVI$_m$ model

The differential equations 1 to 8 of the model are given below:

$$\frac{dS}{dt} = -\beta S\frac{(I_{IV} + I_{CV})}{N} - \sigma S + \mu I_m \tag{1}$$

$$\frac{dE}{dt} = -\beta S\frac{(I_{IV} + I_{CV})}{N} - \delta E + \delta' E \tag{2}$$

$$\frac{dI_{IV}}{dt} = -\delta E - (1-\alpha)\gamma_{IV} I_{IV} - \alpha\rho_{IV} I_{IV} \tag{3}$$

$$\frac{dI_{CV}}{dt} = -\delta' E - (1-\alpha)\gamma_{CV} I_{CV} - \alpha\rho_{CV} I_{CV} + (1-\eta)\frac{V}{N} \tag{4}$$

$$\frac{dR}{dt} = (1-\alpha)\gamma_{CV} I_{CV} + (1-\alpha)\gamma_{IV} I_{IV} \tag{5}$$

$$\frac{dD}{dt} = \alpha\rho_{CV} I_{CV} + \alpha\rho_{IV} I_{IV} \tag{6}$$

$$\frac{dv}{dt} = \sigma\frac{S}{N} - \frac{V}{N} \tag{7}$$

$$\frac{dI_m}{dt} = \eta\frac{V}{N} - \mu I_m \tag{8}$$

The parameters of the equations are described in Table 1, see Lobinska *et al.* (2022) and Rella *et al.* (2021).

Table 1: Parameters of the model

Parameter	Value
Transmission of disease β	{0.0155, 0.18}
Infection rate δ	1.1
Infection rate after vaccination δ'	0.5
Death rate after incomplete vaccination ρ_{IV}	0.2
Death rate after complete vaccination ρ_{CV}	0.01
Recovery rate after incomplete vaccination γ_{IV}	0.076
Recovery rate after complete vaccination γ_{CV}	0.79
Fatality rate α	0.05
Vaccination rate σ	{0.3, 0.8}
Vaccine effectiveness η	{0.2, 0.7}

3. Results

The newly proposed SEIRDVI_m model is used to run for 51 days. The time frame is divided into intervals of five days. SEIRDVI_m model is used to run for each interval of time for each of the five cities. SEIRDVI_m model in Figure 2 depicts the variation of incompletely vaccinated people with vaccination rate and effectiveness of vaccination. The vaccination rate has varied from 0.3 to 0.8 with effectiveness 0.2 to 0.7. From Figure 3, the Variation of infection with completely vaccinated with vaccination rate and effectiveness is seen. Figure 4 exhibits the variation of death with a product of vaccination rate and effectiveness. As the product increases total death count decreases. Figure 5 depicts the comparison of the total infected in the simulation result and the actual data value. The registry data of the United States is collected in the period of 1st March 2020 to 23rd March 2020, see Liu *et al.* (2021) and Alamo *et al.* (2020). The data set is used for validation of the model.

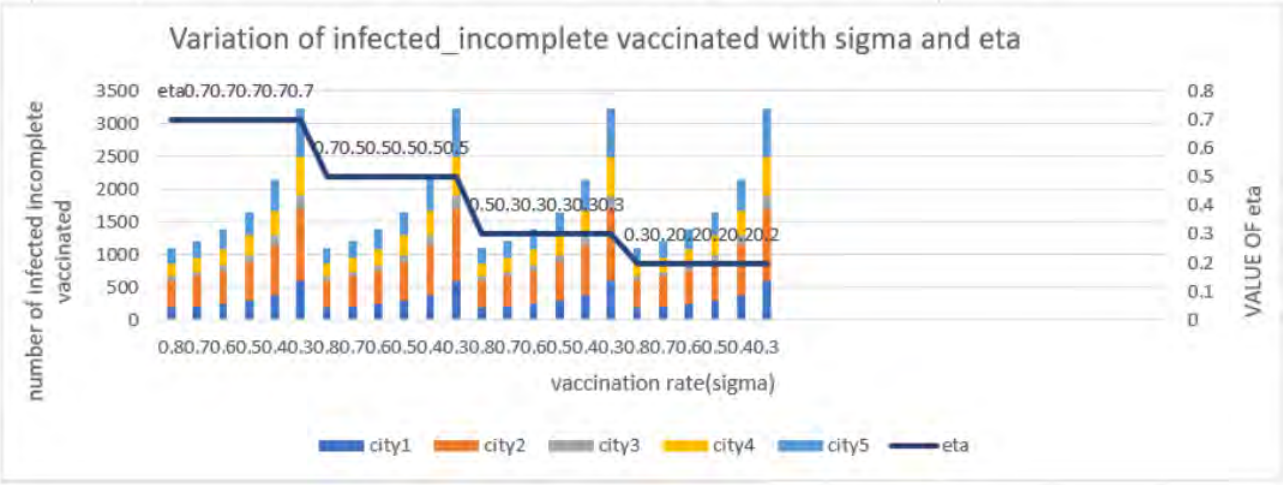


Figure 2: Variation of infected incompletely vaccinated with vaccination rate and effectiveness

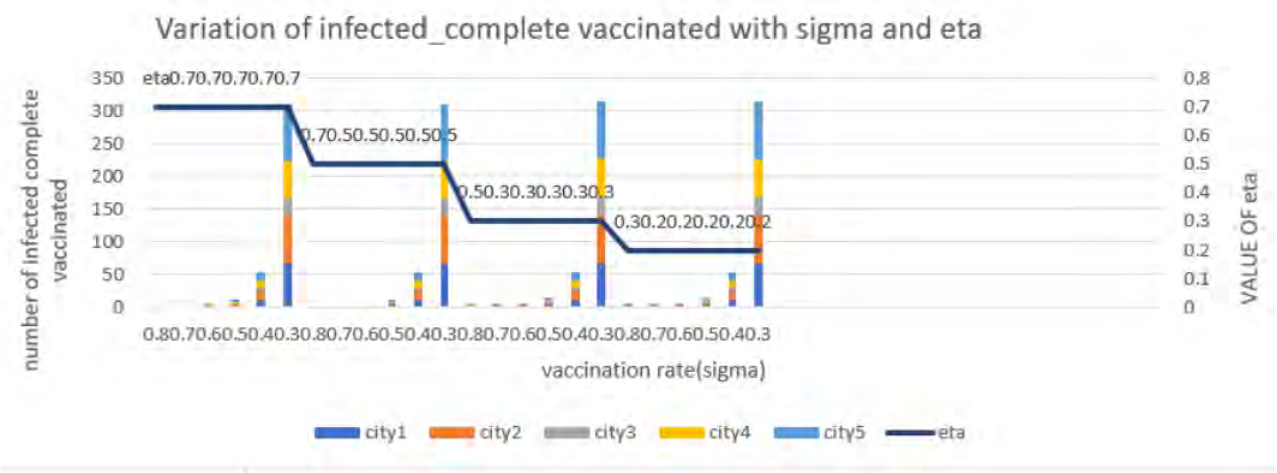


Figure 3: Variation of infection with completely vaccinated with vaccination rate and effectiveness

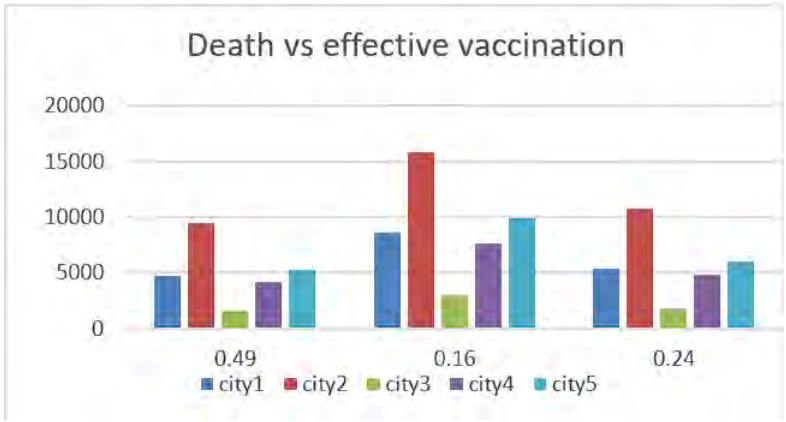


Figure 4: Variation of Death with effective vaccination

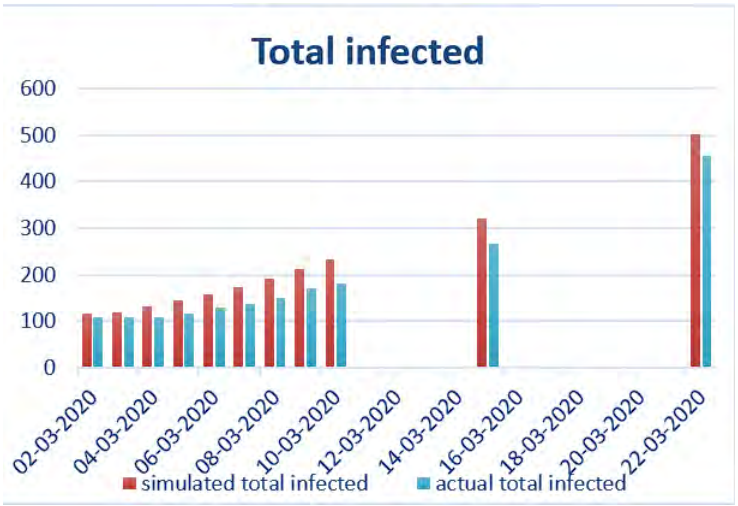


Figure 5: Comparison on total infected with actual data and simulated results

In order to analyze the model, R-squared (R^2) and Mean Squared Error (RMSE) are used for comparison. R-squared (R^2) is a statistical measure of how close the simulation result matches the actual data. The higher the value of R-squared (R^2), the better the model fits the actual data. Equations 9 and 10 describe R-squared (R^2) and Mean Squared Error (RMSE).

$$R^2 = \frac{\sum(simulated\ result - actual\ value)^2}{\sum(actual\ value - Mean\ value)^2} = 0.8797 \tag{9}$$

The RMSE value calculates the error between the simulated result value and the real data. The more the RMSE value closes to 0, the better the result, see Lucas (2014).

$$RMSE = \frac{\sqrt{\sum_{i=1}^{N}(simulated\ value\ i - actual\ value\ i)^2}}{N} = 34.59 \tag{10}$$

Table 2: Comparison of Models

Parameter	Classical SEIR, Liu *et al.* (2021)	Proposed Model SEIRDVIm
R^2	0.60624	0.8979
RMSE	4132.2348	34.59

3.1. Model 1: Lockdown using classical knapsack

Lockdown state is represented by 0 and open state is represented by 1. Figure 6 depicts the scenario when the city is in open or closed states.

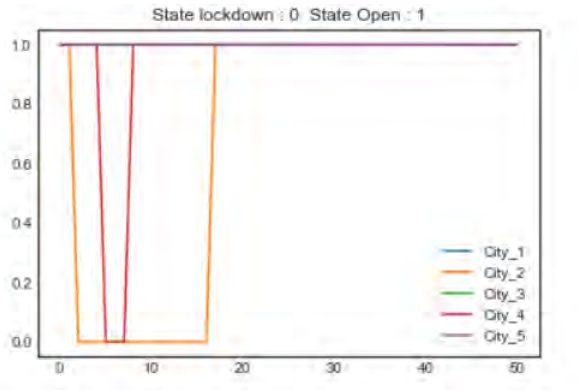


Figure 6: Lockdown in cities

3.1.1. Lockdown and open state for five cities

We are using the same parameters as in the SEIRDVI_m model with no lockdown scenario as in Table 1. The classical knapsack algorithm is applied every five days to obtain the optimal lockdown schedule. We consider bed capacity as cost and the number of infected as weight. The knapsack will contain only the open cities. The cities are selected in such a way that the number of available beds is maximized and death will be minimized. Figure 7

describes the classical knapsack-imposed lockdown schedule. The variation of bed capacity for each city with the number of days in lockdown is portrayed in Figure 8.

Day City	0	5	10	15	20	25	30	35	40	45	50	55	60
City_1	O	O	O	O	O	O	O	O	O	O	O	O	O
City_2	O	X	X	X	X	O	O	O	O	O	O	O	O
City_3	O	X	X	O	O	O	O	O	O	O	O	O	O
City_4	O	X	X	X	O	X	O	O	O	O	O	O	O
City_5	O	X	X	X	X	X	O	O	O	O	O	O	O

Figure 7: Lockdown schedule using classical knapsack

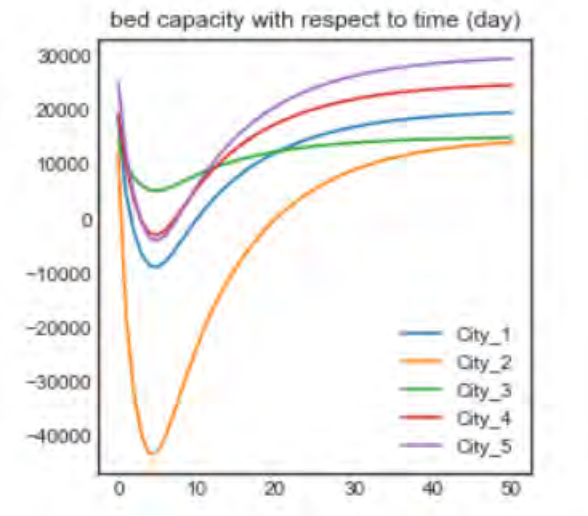


Figure 8: Variation of available bed capacities for five cities with number of days in lockdown

It is reflected in the result that for classical knapsack-imposed lockdown, the total death is 20123 after 60 days, where 356169 have been recovered and 7060317 have been vaccinated.No real data is used to derive the lockdown schedule.

3.2. Model 2: Lockdown using quantum knapsack

For deriving an optimal lockdown schedule in quantum, we need to transfer the objective, i.e. maximizing available bed capacity and minimizing the death into an energy function using a binary quadratic model (bqm). Then we minimize the energy function by D-Wave Quantum Annealer. Lucas (2014) described the quantum algorithm for the knapsack problem. The Quantum algorithm for the knapsack problem is built by using the algorithm Q-Knapsack ($city_{index}$, $city_{GDP}$, $city_{infected}$, $city_{bedCapacity}$) where $city_{GDP}$ represents the GDP

of each city, $city_{infected}$ is the number of infected in the city, and $city_{bedCapacity}$ represents the hospital bed capacity of each city. See Annexure for algorithm 1 of quantum knapsack algorithm for generating binary quadratic model, from which lockdown schedule is obtained based on closed and open cities sample set.

For deriving optimal an lockdown schedule using quantum knapsack we are using the same parameters as described in Table 1. Figure 9 depicts the lockdown schedule as time is divided into five-days intervals.

Day City	0	5	10	15	20	25	30	35	40	45	50	55	60
City_1	O	X	X	X	O	X	X	X	X	X	X	X	O
City_2	X	X	X	X	X	X	X	X	X	X	X	X	O
City_3	X	X	X	X	O	X	O	O	X	O	O	X	X
City_4	X	X	X	O	X	O	O	O	X	X	X	X	O
City_5	X	X	X	X	X	O	X	X	O	O	X	O	X

Figure 9: Quantum imposed lockdown schedule

Using the same rule, we are putting the cities in the knapsack such that available bed increases and death decreases. The cities that are not in knapsack need to be in lockdown. Algorithm 2 describes the quantum algorithm for lockdown. Figure 10 shows the variation of bed capacity for each city with a number of days in lockdown.

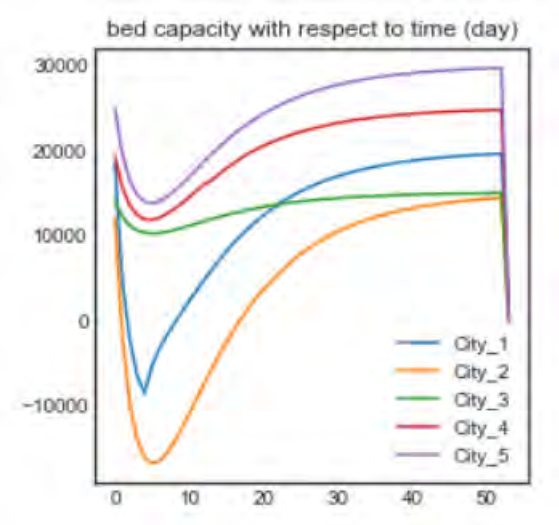


Figure 10: variation of bed capacity with number of days in lockdown

It is reflected in the result that for quantum-imposed lockdown the total number of deaths is 15062 after 60 days, where 192804 have been recovered and 7230041 have been vaccinated. Figure 11 depicts the comparison of infected who are incompletely vaccinated by

classical knapsack-imposed lockdown and Quantum knapsack-imposed lockdown. Algorithm 2 describes the algorithm of quantum-imposed lockdown, see Annexure for Algorithm 2.

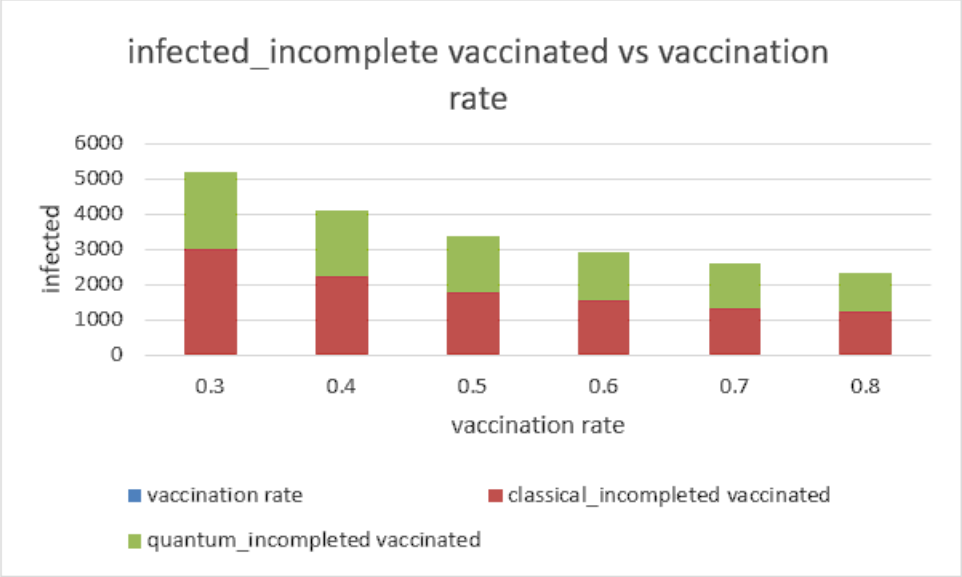


Figure 11: Comparison of infected who are incompletely vaccinated by classical knapsack-imposed lockdown and quantum knapsack-imposed lockdown

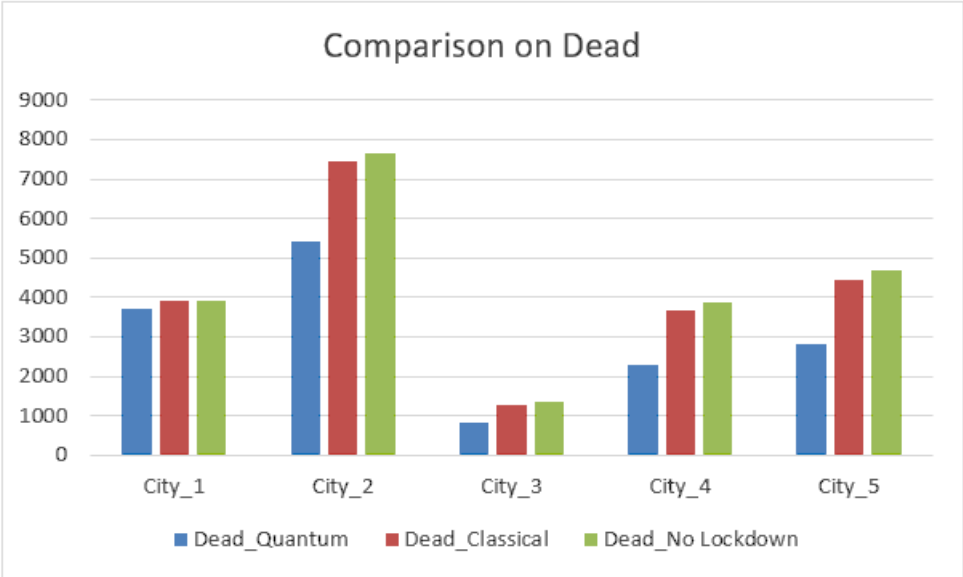


Figure 12: Comparison of dead by no lockdown, classical knapsack-imposed lockdown and quantum knapsack-imposed lockdown

In Figure 12, a comparison of the death count is done between no lockdown, classical knapsack-assisted lockdown, and quantum knapsack-assisted lockdown algorithm. It is reflected from the result that quantum causes 15062 total deaths whereas in classical knapsack death count is 20123.

4. Discussion

This paper fulfills the objective by validating the proposed model with real data. The model fits better and the root mean square error concerning actual data is lesser compared to the classical SEIR model. The model shows the effectiveness of vaccination by showing the variation of infected and death with vaccination rate. It is observed that number of infected is much lesser in complete vaccination compared to incomplete vaccination. An optimal lockdown schedule is derived by applying a quantum knapsack algorithm and it is found that compared with classical knapsack-based lockdown quantum assisted lockdown results in lesser death. For 60 days, quantum-assisted lockdown yields a death toll of 15062 compared to 20123 in classical knapsack-induced lockdown. Compared to classical, quantum knapsack implements a lockdown schedule more efficiently so that the number of infections decreases resulting increase in available bed capacity and thus number of deaths. Because of this, the death toll of quantum-assisted method is much smaller compared to classical Knapsack algorithm. However, the paper has the limitations that the exact date of obtaining predictor values is not known. Despite this limitation, the SEIRDVI_m model can predict the possible infected and death as well as help to decide on lockdown.

5. Conclusion

In this paper, our objective is to propose an effective compartmental model SEIRDVI_m considering complete and partially vaccinated populations with immunized as a separate compartment. The model yields better results compared to the classical SEIR model in terms of R^2 and RMSE values. This model yields an optimal lockdown schedule using classical and quantum knapsack algorithms. It is reflected in the result that for 60 days, quantum-based lockdown resulted death toll of 15062 compared to 20123 in classical knapsack-induced lockdown.

Acknowledgement

The authors are thankful to Dark Star Quantum Lab Inc. for funding to access D-Wave. Authors are indeed grateful to the Editors for their guidance and counsel. Authors are very grateful to the reviewer for valuable comments and suggestions.

References

Alamo, T., Reina, D. G., Mammarella, M., and Abella, A. (2020). Covid-19: Open-data resources for monitoring, modeling, and forecasting the epidemic. *Electronics*, **9**, 827.

Chowell, G., Tariq, A., and Kiskowski, M. (2019). Vaccination strategies to control ebola epidemics in the context of variable household inaccessibility levels. *PLOS Neglected Tropical Diseases*, **13**, e0007814.

Cooper, I., Mondal, A., and Antonopoulos, C. G. (2020). A sir model assumption for the spread of covid-19 in different communities. *Chaos, Solitons & Fractals*, **139**, 110057.

Davies, N. G., Kucharski, A. J., Eggo, R. M., Gimma, A., Edmunds, W. J., Jombart, T., O'Reilly, K., Endo, A., Hellewell, J., Nightingale, E. S., et al. (2020). Effects of non-pharmaceutical interventions on covid-19 cases, deaths, and demand for hospital services in the uk: a modelling study. *The Lancet Public Health*, **5**, e375–e385.

Feng, Z., Towers, S., and Yang, Y. (2011). Modeling the effects of vaccination and treatment on pandemic influenza. *The AAPS Journal*, **13**, 427–437.

Ferguson, N., Laydon, D., Nedjati-Gilani, G., Imai, N., Ainslie, K., Baguelin, M., Bhatia, S., Boonyasiri, A., Cucunubá, Z., Cuomo-Dannenburg, G., et al. (2020). Report 9: Impact of non-pharmaceutical interventions (npis) to reduce covid19 mortality and healthcare demand. *Imperial College London*, **10**, 491–497.

Kuhl, E. and Kuhl, E. (2021). The classical sir model. *Computational Epidemiology: Data-Driven Modeling of COVID-19*, **1**, 41–59.

Lal, R., Huang, W., and Li, Z. (2021). An application of the ensemble kalman filter in epidemiological modelling. *Plos One*, **16**, e0256227.

Liu, X.-X., Fong, S. J., Dey, N., Crespo, R. G., and Herrera-Viedma, E. (2021). A new seaird pandemic prediction model with clinical and epidemiological data analysis on covid-19 outbreak. *Applied Intelligence*, **51**, 4162–4198.

Lobinska, G., Pauzner, A., Traulsen, A., Pilpel, Y., and Nowak, M. A. (2022). Evolution of resistance to covid-19 vaccination with dynamic social distancing. *Nature Human Behaviour*, **6**, 193–206.

Lucas, A. (2014). Ising formulations of many np problems. *Frontiers in Physics*, **2**, 5.

Matrajt, L., Eaton, J., Leung, T., and Brown, E. R. (2021). Vaccine optimization for covid-19: Who to vaccinate first? *Science Advances*, **7**, eabf1374.

Rella, S. A., Kulikova, Y. A., Dermitzakis, E. T., and Kondrashov, F. A. (2021). Rates of sars-cov-2 transmission and vaccination impact the fate of vaccine-resistant strains. *Scientific Reports*, **11**, 15729.

Scherer, A. and McLean, A. (2002). Mathematical models of vaccination. *British Medical Bulletin*, **62**, 187–199.

Usherwood, T., LaJoie, Z., and Srivastava, V. (2021). A model and predictions for covid-19 considering population behavior and vaccination. *Scientific Reports*, **11**, 12051.

ANNEXURE

Algorithm 1: An algorithm to obtain Binary Quadratic Model for Quantum Knapsack

Require: $city_{index}$, $city_{GDP}$, $city_{infected}$, $city_{bedCapacity}$

bqm : Binary Quadratic Model
$lagrange \leftarrow \max(city_{value})$
$x_{size} \leftarrow length(city_{infected})$
$y_{index_{max}}$: maximum index in y
for $k \leftarrow 1$, x_{size} **do**
    bqm.setLinear($city_{index_k}, lagrange * (city_{infected_k})^2 - city_{GDP_k}$))
end for

for $i \leftarrow 1$, x_{size} **do**
    for $j \leftarrow i + 1$, x_{size} **do**
        bqm.setQuadratic[$city_{index_i}, city_{index_j}$] $\leftarrow 2(lagrange * city_{infected_i} * city_{infected_j})$
    end for
end for

for $k \leftarrow 1$, $y_{index_{max}}$ **do**
    bqm.setLinear('y' + string(k), $lagrange * (y_k)^2$)
end for

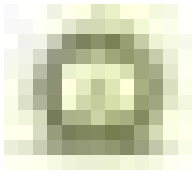
for $i \leftarrow 1$, $y_{index_{max}}$ **do**
    for $j \leftarrow i + 1$, $y_{index_{max}}$ **do**
        bqm.setQuadratic[y_i, y_j] $\leftarrow 2 * lagrange * y_i * y_j$
    end for
end for

for $i \leftarrow 1$, x_{size} **do**
    for $j \leftarrow i + 1$, $y_{index_{max}}$ **do**
        bqm.setQuadratic[$city_{index_i}, y_j$] $\leftarrow -2 * lagrange * city_{infected_i} * y_j$
    end for
end for

Algorithm 2: An algorithm to lockdown city based on Binary Quadratic Model from Quantum Knapsack

Require: *bqm* : Binary Quadratic Model for Quantum Knapsack
Get OpenCity & ClosedCity sampleset from *bqm* based on bed Capacity threshold.
for each city **do**
if $city_{index}$ is in OpenCities **then** LockdownList $\leftarrow 1$
else LockdownList $\leftarrow 0$
end if
end for
return LockdownList

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 22, No. 1, 2024 (New Series), pp 303–319
http://www.ssca.org.in/journal



Cost & Profit Analysis of Two-Dimensional State M/M/2 Queueing Model with Multiple Vacation, Feedback, Catastrophes and Balking

Sharvan Kumar and Indra
Department of Statistics & O. R, Kurukshetra University, Kurukshetra-136119

Received: 19 June 2023; Revised: 05 October 2023; Accepted: 11 October 2023

Abstract

A time-dependent solution of the two-dimensional state M/M/2 queueing system with multiple vacation, feedback, catastrophes and balking is obtained in this study. Inter-arrival and service times follow an exponential distribution with parameters λ and μ respectively. Both the servers go on vacation with probability one when there are no units in the system. All the units are ejected from the system when catastrophes occur and the system becomes temporarily unavailable. The system reactivates when new units arrive. Occurrence of catastrophes follow Poisson distribution with rate ξ. The units come and wait in the queue for service; the served units either leave the system or rejoin immediately at the early end of the queue to receive satisfactory service, known as feedback. Laplace transform approach has been used to find the time-dependent solution. The efficiency of a queuing system has been verified by evaluating some key measures along with "total expected cost" and "total expected profit". Numerical analyses have been done by using Maple software.

Key words: Time-dependent solution; Two-dimensional state model; Balking; Catastrophes; Feedback; Multiple vacation.

AMS Subject Classifications: 62K05, 05B05

1. Introduction

In the present study, the two-dimensional state model has been used to simplify the complicated transient analysis of some queueing problems. This model is used to study the queueing system more categorically for arrivals and departures. The idea of two-dimensional state for the M/M/1 queue was first given by Pegden and Rosenshine (1982). After that, two-dimensional state model has received considerable attention by many researchers to analyse various queuing systems.

Various studies have been conducted to evaluate different performance measures to verify the robustness of the system in which a server takes a break for a random period of time *i.e.* vacation. When the server returns from a vacation and finds the queue empty, it

Corresponding Author: Indra
Email: indra@kuk.ac.in

immediately goes on another vacation and if it finds at least one waiting unit, then it will commence service according to the prevailing service policy, *i.e.* multiple vacation. Different queueing systems with multiple vacation have been extensively investigated and effectively used in several fields including industries, computer & communication systems, telecommunication systems *etc.* Different types of vacation policies are available in literature such as single vacation, multiple vacation and working vacations. Researches on vacation models have grown tremendously in the last several years. Cooper (1970) was the first to study the vacation model and determined the mean waiting time for a unit arrive at a queue served in cyclic order. Doshi (1986) and Ke *et al.* (2010) have done outstanding researches on queueing system with vacations and released some excellent surveys. Xu and Zhang (2006) considered the Markovian multi-server queue with a single vacation (e, d)-policy. They also formulated the system as a quasi-birth-and-death process and computed the various stationary performance measures. Altman and Yechiali (2006) studied the customer's impatience in queues with server vacations. Kalidass *et al.* (2014) obtained the time-dependent solution of a single server queue with multiple vacations. Ammar (2015) analysed M/M/1 queue with impatient units and multiple vacations. Sharma and Indra (2020) investigated the dynamic aspects of a two-dimensional state single server Markovian queueing system with multiple vacations and reneging.

Also, units may be served repeatedly for many reasons, *e.g.* when a unit is unsatisfied with a service, the unit may try for a satisfactory service. For example, we visit to the online shopping store and order a full-sleeve jacket but when we receive the order it turn out to be half-sleeve jacket. Since we are unsatisfied with the service, so we go for a return policy or exchange policy provided by the shopping store and to receive satisfactory service. Many researchers have been attracted to the study of queues with feedback as large number of applications have been found in many areas including production systems, post offices, supermarkets, hospital management, financial sectors, ticket offices, grocery stores, ATMs and so forth. Takacs (1963) determined the distribution of the queue size and the first two moments of the distribution for a queue with feedback. D'Avignon and Disney (1976) studied the non-Markovian queue with a state-dependent feedback mechanism. Disney *et al.* (1980) investigated a number of random processes that occur in queues with instantaneous Bernoulli feedback. Choudhury and Paul (2005) derived two phases of heterogeneous services with Bernoulli feedback systems. Chowdhury and Indra (2020) analysed two-node tandem queue with feedback.

Queueing systems with catastrophes are getting a lot of attention nowadays and may be used to solve a wide range of real-world problems. Catastrophes may occur at any time, resulting in the loss of all the units and the deactivation of the service centre, because they are totally unpredictable in nature. Such types of queues with catastrophes play an important role in computer programs, telecommunication, ticket counter *etc.* For example, virus or hacker attacking a computer system or program causing the system fail or become idle. Chao (1995) obtained steady-state probability of the queue size and a product form solution of a queueing network system with catastrophes. Krishna Kumar *et al.* (2007) obtained time-dependent solution for M/M/1 queueing system with catastrophes. Kalidass *et al.* (2012) derived explicit closed form analytical expressions for the time-dependent probabilities of the system size. Dharmaraja and Kumar (2015) studied Markovian queueing system with heterogeneous servers and catastrophes. Chakravarthy (2017) studied delayed catastrophic model in steady state using the matrix analytic method. Suranga Sampath and

Liu (2018) studied an M/M/1 queue with reneging, catastrophes, server failures and repairs using modified Bessel function, Laplace transform and probability generating function approach. de Oliveira Souza and Rodriguez (2021) worked on fractional M/M/1 queue model with catastrophes.

Queues with balking have a wide range of practical applications in everyday life. Balking occurs if units avoid joining the queue, when they perceive the queue to be too long. Long queues at cash counters, ticket booths, banks, barber shops, grocery stores, toll plaza *etc.* Kumar *et al.* (1993) obtained time-dependent solution of an M/M/1 queue with balking. Zhang *et al.* (2005) analysed the M/M/1/N queueing system with balking, reneging, and server vacation. Sharma and Kumar (2012) used a single-server Markovian feedback queuing system with balking.

With above concepts in mind, we analyse a two-dimensional state M/M/2 queueing model with multiple vacation, feedback, catastrophes and balking.

Out of the many physical situations, one can be in the post office, where an unit arrives to receive the service and is unsatisfied by the service, then it re-joins at the early end of the queue to receive satisfactory service; may be considered as feedback unit. On arrival, if the unit finds a long queue and decides not to join; may be considered as balking unit and if the computer system fails due to virus or any other reason; may be considered as occurrence of catastrophes. After service completion, the server may take a break, when he finds an empty queue.

The paper has been structured as follows. In Section 2, the model assumptions, notations and description are given. In Section 3 the differential-difference equations to find out the time-dependent solution are given and Section 4 describes important performance measures. Section 5 investigates the total expected cost function and total expected profit function for the given queueing system. In Section 6, we present the numerical results in the form of tables and graphs to illustrate the impact of various factors on performance measures. The last Section contains discussion on the findings and suggestions for future work.

2. Model assumptions, notations and description

- Arrivals follow Poisson distribution with parameter λ.
- There are two homogeneous servers and the service times at each server follow an exponential distribution with parameter μ.
- The vacation time of the server follows an exponential distribution with parameter w.
- After completion of the service, the dissatisfied units rejoin at the early end of the queue to receive service with probability q.
- On arrival a unit either decides to join the queue with probability β or not to join the queue with probability 1-β.
- Occurrence of catastrophes follows Poisson distribution with parameter ξ.

- Various stochastic processes involved in the system are statistically independent of each other.

Initially, the system starts with zero units and the server is on vacation, *i.e.*

$$P_{0,0,V}(0) = 1 \quad , \quad P_{0,0,B}(0) = 0 \tag{1}$$

$$\delta_{i,j} = \begin{cases} 1 & ; for\ i = j \\ 0 & ; for\ i \neq j \end{cases} \tag{2}$$

$$\sum_{i}^{j} = 0 \quad for \quad j < i$$

The two-dimensional state model

$P_{i,j,V}(t)$=The probability that there are exactly i arrivals and j departures by time t and the server is on vacation.

$P_{i,j,B}(t)$=The probability that there are exactly i arrivals and j departures by time t and the server is busy in relation to the queue.

$P_{i,j}(t)$=The probability that there are exactly i arrivals and j departures by time t.

3. The differential-difference equations for the queueing model under study

$$\frac{d}{dt}P_{i,i,V}(t) = -\lambda\beta P_{i,i,V}(t) + q\mu P_{i,i-1,B}(t)(1-\delta_{i,0}) + \xi(1 - P_{i,i,V}(t)) \quad i \geq 0 \tag{3}$$

$$\frac{d}{dt}P_{i+1,i,B}(t) = -(\lambda\beta + q\mu + \xi)P_{i+1,i,B}(t) + 2q\mu P_{i+1,i-1,B}(t)(1-\delta_{i,0}) + wP_{i+1,i,V}(t) \quad i \geq 0 \tag{4}$$

$$\frac{d}{dt}P_{i,j,V}(t) = -(\lambda\beta + w + \xi)P_{i,j,V}(t) + \lambda\beta P_{i-1,j,V}(t) \quad i > j \geq 0 \tag{5}$$

$$\frac{d}{dt}P_{i,j,B}(t) = -(\lambda\beta + 2q\mu + \xi)P_{i,j,B}(t) + \lambda\beta P_{i-1,j,B}(1-\delta_{i-1,j})(t) + 2q\mu P_{i,j-1,B}(t)(1-\delta_{j,0})$$
$$+wP_{i,j,V}(t) \quad i > j+1 \tag{6}$$

The preceding equations (3) to (6) are solved by taking the Laplace transforms together with initial conditions:

$$\bar{P}_{0,0,V}(s) = \frac{\xi + s}{s(s + \lambda\beta + \xi)} \tag{7}$$

$$\bar{P}_{i,0,V}(s) = \frac{(\lambda\beta)^i(\xi + s)}{s(s + \lambda\beta + \xi)(s + \lambda\beta + w + \xi)^i} \quad i > 0 \tag{8}$$

$$\bar{P}_{i,i,V}(s) = \frac{q\mu}{s + \lambda\beta + \xi}P_{j,j-1,B}(s) \quad i > 0 \tag{9}$$

$$\bar{P}_{i,0,B}(s) = \frac{w(\lambda\beta)^i(\xi + s)}{s(s + \lambda\beta + \xi)(s + \lambda\beta + w + \xi)(s + \lambda\beta + q\mu + \xi)(s + \lambda\beta + 2q\mu + \xi)^{i-1}} +$$

$$+w(\lambda\beta)^i \sum_{m=1}^{i-1} \frac{1}{s(s+\lambda\beta+\xi)(s+\lambda\beta+w+\xi)^{m+1}(s+\lambda\beta+q\mu+\xi)(s+\lambda\beta+2q\mu+\xi)^{i-m}} \quad i \geq 1 \tag{10}$$

$$\bar{P}_{i+1,i,B}(s) = \frac{2q\mu}{s+\lambda\beta+q\mu+\xi} P_{i+1,i-1,B}(s) + \frac{q\mu w\lambda\beta}{(s+\lambda\beta+\xi)(s+\lambda\beta+w+\xi)(s+\lambda\beta+q\mu+\xi)} P_{i,i-1,B}(s) \quad i > 0 \tag{11}$$

$$\bar{P}_{i,j,V}(s) = \frac{(q\mu)}{(s+\lambda\beta+w+\xi)} \frac{(\lambda\beta)^{i-j}}{(s+\lambda\beta+w+\xi)^{i-j}} P_{j,j-1,B}(s) \quad i > j \geq 1 \tag{12}$$

$$\bar{P}_{i,j,B}(s) = \frac{\lambda\beta}{s+\lambda\beta+2q\mu+\xi} P_{i-1,j,B}(s) + \frac{2q\mu}{s+\lambda\beta+2q\mu+\xi} P_{i,j-1,B}(s) + \frac{q\mu}{s+\lambda\beta+\xi} \frac{w}{s+\lambda\beta+2q\mu+\xi} \frac{(\lambda\beta)^{i-j}}{(s+\lambda\beta+w+\xi)^{i-j}} P_{j,j-1,B}(s) \quad i > j+1, j > 0 \tag{13}$$

It is seen that

$$\sum_{i=0}^{\infty} \sum_{j=0}^{i} [\bar{P}_{i,j,V}(s) + \bar{P}_{i,j,B}(s)(1-\delta_{i,j})] = \frac{1}{s} \tag{14}$$

and hence

$$\sum_{i=0}^{\infty} \sum_{j=0}^{i} [\bar{P}_{i,j,V}(t) + \bar{P}_{i,j,B}(t)(1-\delta_{i,j})] = 1 \tag{15}$$

a verification.

4. Performance measures

(a) The Laplace transform of $P_{i.}(t)$ the probability that exactly i units arrive by time t; when initially there are no units in the system is given by

$$\bar{P}_{i.}(s) = \sum_{j=0}^{i} [\bar{P}_{i,j,V}(s) + \bar{P}_{i,j,B}(s)(1-\delta_{i,j})] = \sum_{j=0}^{i} \bar{P}_{i,j}(s) = \frac{(\lambda\beta)^i}{(s+\lambda\beta)^{i+1}} \tag{16}$$

and its inverse Laplace transform is

$$P_{i.}(t) = \frac{e^{-\lambda\beta t}(\lambda\beta t)^i}{i!} \tag{17}$$

The arrivals follow a Poisson distribution as the probability of the total number of arrivals is not affected by vacation time of the server.

(b) $P_{.j}(t)$ is the probability that exactly j units have been served by time t. In terms of $P_{i,j}(t)$ we have

$$P_{.j}(t) = \sum_{i=j}^{\infty} P_{i,j}(t) \tag{18}$$

(c)The probability of exactly n units in the system at time t, denoted by $P_n(t)$, can be expressed in terms $P_{ij}(t)$ as

$$P_n(t) = \sum_{j=0}^{\infty} P_{j+n,j}(t) \tag{19}$$

(d) The Laplace transform of mean number of arrivals by time t is

$$\sum_{i=0}^{\infty} i\bar{P}_{i.}(s) = \frac{\lambda\beta}{s^2} \tag{20}$$

and inverse of the mean number of arrivals by time t is

$$\sum_{i=0}^{\infty} i\ P_{i.}(t) = \lambda t \tag{21}$$

(e) The mean number of units in the queue is calculated as follows

$$Q_L(t) = \sum_{n=0}^{\infty} nP_V(t) + \sum_{n=2}^{\infty}(n-2)P_B(t) \tag{22}$$

where $n = i - j$.

5. Cost function and profit function

For the given queueing system, the following notations have been used to represent various costs to find out the total expected cost and total expected profit per unit time
Let
C_H: Cost per unit time for unit in the queue.
C_B: Cost per unit time for a busy server.
C_μ: Cost per service per unit time.
C_V: Cost per unit time when the server is on vacation.
$C_{\mu-q}$: Cost per unit time when a unit rejoins at the early end of the queue as a feedback unit.
If I is the total expected amount of income generated by delivering a service per unit time then
a) Total expected cost per unit at time t is given by

$$TC(t) = C_H * Q_L(t) + C_B * P_B(t) + C_V * P_V(t) + \mu * (C_\mu + C_{\mu-q}) \tag{23}$$

b) Total expected income per unit at time t is given by

$$TE_I(t) = I * \mu * (1 - P_V(t)) = I * \mu * P_B(t) \tag{24}$$

c) Total expected profit per unit at time t is given by

$$TE_P(t) = TE_I(t) - TC(t) \tag{25}$$

6. Numerical results

6.1. Numerical validity check

1. For the state when the server is on vacation

$$P_V(t) = \sum_{j=0}^{i} P_{i,j,V}(t) \tag{26}$$

2. For the state when the server is busy in relation to the queue

$$P_B(t) = \sum_{j=0}^{i-1} P_{i,j,B}(t) \tag{27}$$

3. The probability $P_{i.}(t)$ that exactly i units arrive by time t is

$$P_{i.}(t) = \sum_{j=0}^{i} P_{i,j}(t) = \sum_{j=0}^{i} P_{i,j,V}(t) + \sum_{j=0}^{i-1} P_{i,j,B}(t) \tag{28}$$

4. A numerical validity check of inversion of $\bar{P}_{i,j}(s)$ is based on the relationship

$$P_r\{i \quad arrivals \quad in \quad (0,t)\} = \frac{e^{-(\lambda\beta t)} * (\lambda\beta t)^i}{i!} = \sum_{j=0}^{i} P_{i,j}(t) = P_{i.}(t) \tag{29}$$

The probabilities of this model shown in last column of Table 1 given below are consistent to the last column of "Pegden and Rosenshine (1982)"

Table 1: Numerical validity check of inversion $\bar{P}_{i,j}(s)$

λ	μ	t	i	w	q	ξ	β	$\frac{e^{-(\lambda t)} * (\lambda t)^i}{i!}$	$\sum_{j=0}^{i} P_{i,j,V}(t)$	$\sum_{j=0}^{i-1} P_{i,j,B}(t)$	$\sum_{j=0}^{i} P_{i,j}(t)$
1	2	3	1	1	1	0	1	0.149361	0.12688	0.02247	0.14936
1	2	3	3	1	1	0	1	0.224041	0.14971	0.07433	0.22404
1	2	3	5	1	1	0	1	0.100818	0.05262	0.04818	0.10081
2	2	3	1	1	1	0	1	0.014871	0.01263	0.00223	0.01487
2	2	3	3	1	1	0	1	0.089234	0.05962	0.02960	0.08923
2	2	3	5	1	1	0	1	0.160622	0.08384	0.07677	0.16062
1	2	4	1	1	1	0	1	0.073261	0.06443	0.00882	0.07326
1	2	4	3	1	1	0	1	0.195366	0.14187	0.05349	0.19536
1	2	4	5	1	1	0	1	0.156292	0.09401	0.06227	0.15629
2	2	4	1	1	1	0	1	0.002682	0.00236	0.00032	0.00268
2	2	4	3	1	1	0	1	0.028625	0.02078	0.00783	0.02862
2	2	4	5	1	1	0	1	0.091602	0.05510	0.03650	0.09160
2	4	4	5	1	1	0	1	0.091602	0.07219	0.01940	0.09160
1	2	4	4	1	1	0	1	0.195366	0.12931	0.06605	0.19536
1	2	3	6	1	1	0	1	0.050409	0.02299	0.02741	0.05040

Table 2: Probabilities of exactly n units in the system at time t

n	$t=1$	$t=2$	$t=3$	t=4	$t=5$
0	0.3064059	0.2756176	0.2194939	0.1313295	0.0614414
1	0.3512870	0.3112144	0.2199253	0.1139011	0.0462964
2	0.1935174	0.1626266	0.0998428	0.0447517	0.0158779
3	0.0917319	0.0804006	0.0421561	0.0161968	0.0049851
4	0.0368324	0.0375051	0.0167183	0.0053778	0.0014074
5	0.0123808	0.0160725	0.0061926	0.0016003	0.0003411
6	0.0033284	0.0059066	0.0020629	0.0004081	0.0000646

Table 3: Probabilities of exactly j departures by time t

j	$t=1$	$t=3$	$t=5$	t=7	$t=10$
0	0.67968674	0.099509402	0.007046060	0.000420927	0.00016181
1	0.17711300	0.081454841	0.009095394	0.000581746	0.00005037
2	0.09298400	0.117972870	0.018124103	0.001391221	0.00005942
3	0.03598100	0.140843910	0.031453844	0.003012981	0.00008380
4	0.01081200	0.140978350	0.047484726	0.005811690	0.00014277
5	0.00259600	0.119261779	0.061887004	0.009825172	0.00026424
6	0.00050430	0.084578352	0.068296881	0.014198767	0.00046744

6.2. Sensitivity analysis

This part focuses on the impact of the arrival rate (λ), service rate (μ), vacation rate (w), catastrophes rate (ξ), feedback probability (q) and balking probability (1-β) on the probability when the server is on vacation ($P_V(t)$), probability when the server is busy ($P_B(t)$), expected queue length ($Q_L(t)$), total expected cost ($TC(t)$), total expected income ($TE_I(t)$) and total expected profit ($TE_P(t)$) at time t. To calculate the numerical results for the sensitivity of the queueing system one parameter varied while keeping all the other parameters fixed.

Impact of arrival rate: We examine the behaviour of the queueing system using measures of effectiveness along with cost and profit analysis by varying λ with time t, while keeping all other parameters fixed; μ=5, w=2, β=0.5, ξ=0.0001, q=0.7, C_H=10, C_B=8, C_V=5, C_μ=4, $C_{\mu-q}$=2, I=100 and N=8. In Table 4, we observe that as the value of λ increases with time t, $P_B(t)$, $Q_L(t)$, $TC(t)$, $TE_I(t)$ and $TE_P(t)$ increases but $P_V(t)$ decreases.

Table 4: Measures of effectiveness versus λ

t	λ	$P_V(t)$	$P_B(t)$	$Q_L(t)$	$TC(t)$	$TE_I(t)$	$TE_P(t)$
1	1.00	0.905411	0.094588	0.2173509	37.457268	47.2940	9.8367320
2		0.880912	0.119086	0.2429674	37.786922	59.5430	21.756078
3		0.878559	0.121412	0.2454781	37.818872	60.7060	22.887128
4		0.878264	0.121497	0.2455127	37.818423	60.7485	22.930077
5		0.877657	0.121202	0.2447627	37.805528	60.6010	22.795472
1	1.25	0.884191	0.115808	0.2722840	38.070259	57.9040	19.833741
2		0.855421	0.144572	0.3036798	38.470479	72.2860	33.815521
3		0.852827	0.147024	0.3064271	38.504598	73.5120	35.007402
4		0.852077	0.146782	0.3056222	38.490863	73.3910	34.900137
5		0.849575	0.145499	0.3023700	38.435567	72.7495	34.313933
1	1.50	0.863798	0.136201	0.3275511	38.684109	68.1005	29.416391
2		0.831169	0.168802	0.3646441	39.152702	84.4010	45.248298
3		0.828243	0.171206	0.3671780	39.182643	85.6030	46.420357
4		0.826167	0.170030	0.3638632	39.129707	85.0150	45.885293
5		0.818924	0.166270	0.3543160	38.967940	83.1350	44.167060

Impact of service rate: The behaviour of the queueing system using measures of

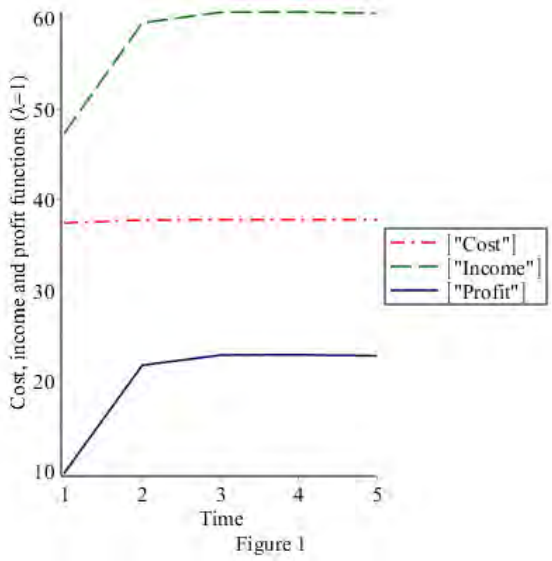


Figure 1

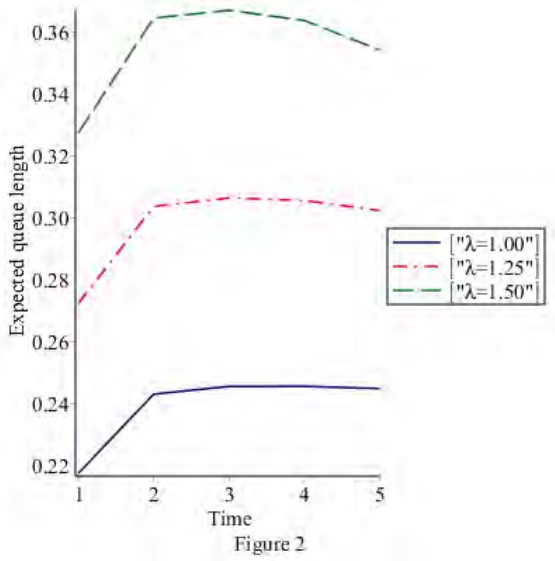


Figure 2

Figure 1: Shows the variation of cost, income and profit at an arrival rate λ=1.00 with time t while keeping the other parameters fixed (μ=5, w=2, ξ=0.0001, q=0.7, β=0.5)

Figure 2: Shows the variation of $Q_L(t)$ with time t by varying arrival rate λ(=1.00, 1.25, 1.50) while keeping the other parameters fixed (μ=5, w=2, ξ=0.0001, q=0.7, β=0.5)

effectiveness along with cost and profit analysis by varying μ with time t, while keeping all other parameters fixed; λ=1, w=2, β=0.5, ξ=0.0001, q=0.7, C_H=10, C_B=8, C_V=5, C_μ=4, $C_{\mu-q}$=2, I=100 and N=8. In Table 5, we observe that as the value of μ increases with time t, $P_B(t)$, $Q_L(t)$, $TC(t)$, $TE_I(t)$ and $TE_P(t)$ increases but $P_V(t)$ decreases.

Table 5: Measures of effectiveness versus μ

t	μ	$P_V(t)$	$P_B(t)$	$Q_L(t)$	$TC(t)$	$TE_I(t)$	$TE_P(t)$
1	3.75	0.885251	0.114748	0.220270	30.046939	43.03050	12.983561
2		0.845732	0.154266	0.247027	30.433058	57.84975	27.416692
3		0.840988	0.158983	0.248985	30.466654	59.61862	29.151971
4		0.840598	0.159163	0.248765	30.463944	59.68612	29.222181
5		0.840100	0.158760	0.247921	30.433790	58.78500	28.351210
1	4.25	0.894093	0.105906	0.218863	33.006343	45.01005	12.003707
2		0.861897	0.138101	0.244849	33.362783	58.69292	25.330142
3		0.858481	0.141490	0.247023	33.394555	60.13325	26.738695
4		0.858160	0.141601	0.246935	33.392958	60.18042	26.787467
5		0.857609	0.141250	0.246144	33.371485	59.60625	26.234765
1	4.75	0.901872	0.098127	0.217789	35.972266	46.61032	10.638059
2		0.875156	0.124842	0.243461	36.309126	59.29995	22.990824
3		0.872529	0.127443	0.245866	36.340849	60.53542	24.194576
4		0.872229	0.127532	0.245867	36.340071	60.57770	24.237629
5		0.871638	0.127221	0.245106	36.319018	59.95497	23.635957

Impact of vacation rate: We observe that the behaviour of the queueing system

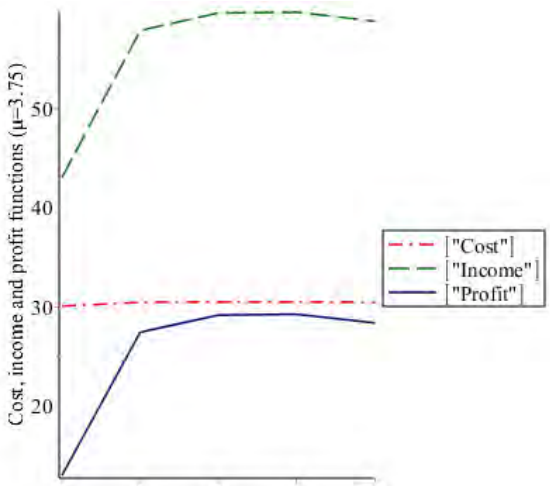


Figure 3

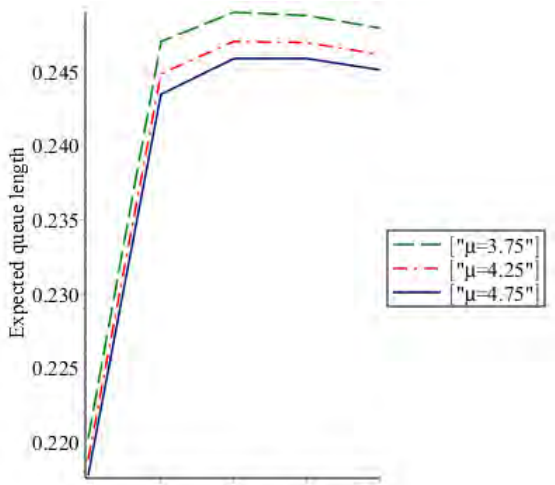


Figure 4

Figure 3: Shows the variation of cost, income and profit at a service rate μ=3.75 with time t while keeping the other parameters fixed (λ=1, w=2, ξ=0.0001, q=0.7, β=0.5)

Figure 4: Shows the variation of $Q_L(t)$ with time t by varying service rate μ(=3.75, 4.25, 4.75) while keeping the other parameters fixed (λ=1, w=2, ξ=0.0001, q=0.7, β=0.5)

using measures of effectiveness along with cost and profit analysis by varying w with time t, while keeping all other parameters fixed; λ=1, μ=5, q=0.7, β=0.5, ξ=0.0001, C_H=10, C_B=8, C_V=5, C_μ=4, $C_{\mu-q}$=2, I=100 and N=8. In Table 6 , we observe that as the value of w increases with time t, $P_B(t)$, $Q_L(t)$, $TC(t)$, $TE_I(t)$ and $TE_P(t)$ increases but $P_V(t)$ decreases.

Impact of catastrophes rate: We see that the behaviour of the queueing system using measures of effectiveness, along with cost and profit analysis by varying ξ with time t, while keeping all other parameters fixed; λ=1, μ=5, w=2, q=0.7, β=0.5, C_H=10, C_B=8, C_V=5 and C_μ=4, $C_{\mu-q}$=2, I=100, N=8. In Table 7, we observe that as the value of ξ increases with time t, $P_B(t)$, $Q_L(t)$, $TC(t)$, $TE_I(t)$ and $TE_P(t)$ increases but $P_V(t)$ decreases.

Impact of feedback probability: We observe that the behaviour of the queueing system using measures of effectiveness along with cost and profit analysis by varying q with time t, while keeping all other parameters fixed; λ=1, μ=5, w=2, β=0.5, ξ=0.0001, C_H=10, C_B=8, C_V=5, C_μ=4, $C_{\mu-q}$=2, I=100 and N=8. In Table 8, we observe that as the value of q increases with time t, $P_B(t)$, $Q_L(t)$, $TC(t)$, $TE_I(t)$ and $TE_P(t)$ increases but $P_V(t)$ decreases.

Impact of joining probability: We observe that the behaviour of the queueing system using measures of effectiveness along with cost and profit analysis by varying β with time t, while keeping all other parameters fixed; λ=1, μ=5, w=2, q=0.7, ξ=0.0001, C_H=10, C_B=8, C_V=5, C_μ=4, $C_{\mu-q}$=2, I=100 and N=8. In Table 9, we observe that as the value of β increases with time t, $P_B(t)$, $Q_L(t)$, $TC(t)$, $TE_I(t)$ and $TE_P(t)$ increases but $P_V(t)$ decreases.

Table 6: Measures of effectiveness versus w

t	w	$P_V(t)$	$P_B(t)$	$Q_L(t)$	$TC(t)$	$TE_I(t)$	$TE_P(t)$
1	2.00	0.905411	0.094588	0.217350	37.457259	47.2940	9.8367410
2		0.880912	0.119086	0.242967	37.786918	59.5430	21.756082
3		0.878559	0.121412	0.245478	37.818871	60.7060	22.887129
4		0.878264	0.121497	0.245512	37.818416	60.7485	22.930084
5		0.877657	0.120202	0.244762	37.797521	60.1010	22.303479
1	2.25	0.900868	0.099131	0.200253	37.299918	49.5655	12.265582
2		0.878961	0.121037	0.218097	37.544071	60.5185	22.974429
3		0.877366	0.122605	0.219252	37.560190	61.3025	23.742310
4		0.877175	0.122587	0.219121	37.557781	61.2935	23.735719
5		0.876583	0.121276	0.218425	37.537373	60.6380	23.100627
1	2.50	0.897077	0.102922	0.185305	37.161811	51.4610	14.299189
2		0.877506	0.122492	0.197743	37.344896	61.2460	23.901104
3		0.876420	0.123552	0.198212	37.352636	61.7760	24.423364
4		0.876276	0.123486	0.198033	37.349598	61.7430	24.393402
5		0.875691	0.122168	0.197394	37.329739	61.0840	23.754261

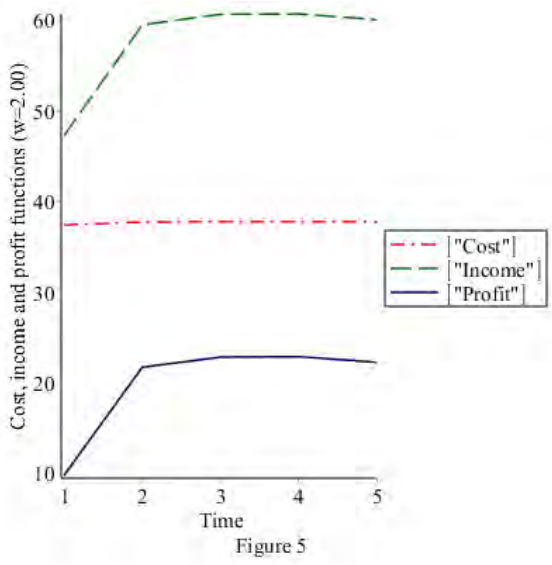


Figure 5

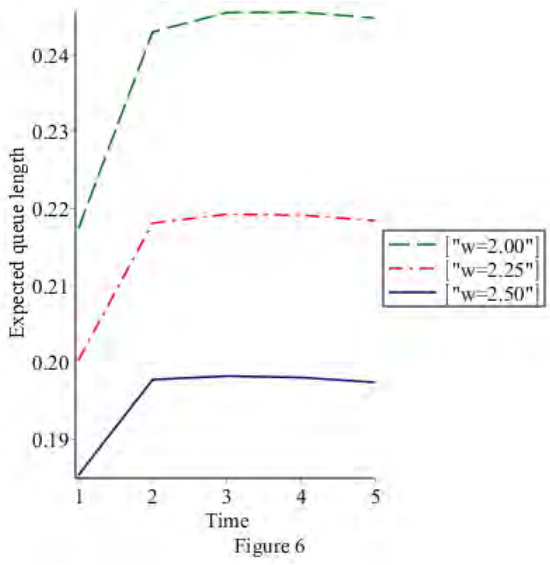


Figure 6

Figure 5: Shows the variation of cost, income and profit at a vacation rate w=2.00 with time t while keeping the other parameters fixed (λ=1, μ=5, ξ=0.0001, q=0.7, β=0.5)

Figure 6: Shows the variation of $Q_L(t)$ with time t by varying vacation rate w(=2.00, 2.25, 2.50) while keeping the other parameters fixed (λ=1, μ=5, ξ=0.0001, q=0.7, β=0.5)

7. Discussion

Figure 1 shows the variation of cost, income and profit with time t by keeping λ constant (=1.00). The value of cost, income and profit increases with increase in t upto t(=3.00, 4.00, 4.00) respectively then decreases slightly. The variation in queue length with time t is represented in figure 2 by varying the arrival rate λ(=1.00, 1.25, 1.50). Queue

Table 7: Measures of effectiveness versus ξ

t	ξ	$P_V(t)$	$P_B(t)$	$Q_L(t)$	$TC(t)$	$TE_I(t)$	$TE_P(t)$
1	0.0001	0.905411	0.094588	0.217350	37.457259	47.2940	9.8367410
2		0.880912	0.119086	0.242967	37.786918	59.5430	21.756082
3		0.878559	0.121412	0.245478	37.818871	60.7060	22.887129
4		0.878264	0.121497	0.245512	37.818416	60.7485	22.930084
5		0.877657	0.121202	0.244762	37.797521	60.1010	22.303479
1	0.0002	0.905415	0.094584	0.217343	37.457177	47.2920	9.8348230
2		0.880920	0.119078	0.242956	37.786784	59.5390	21.752216
3		0.878567	0.121404	0.245466	37.818727	60.7020	22.883273
4		0.878273	0.121489	0.245501	37.818287	60.7445	22.926213
5		0.877666	0.120194	0.244751	37.797392	60.0970	22.299608
1	0.0003	0.905420	0.094579	0.217336	37.457092	47.2895	9.8324080
2		0.880928	0.119070	0.242945	37.786650	59.5350	21.748350
3		0.878576	0.121395	0.245455	37.818590	60.6975	22.878910
4		0.878282	0.121480	0.245489	37.818140	60.7400	22.921860
5		0.877675	0.120185	0.244740	37.797255	60.0925	22.295245

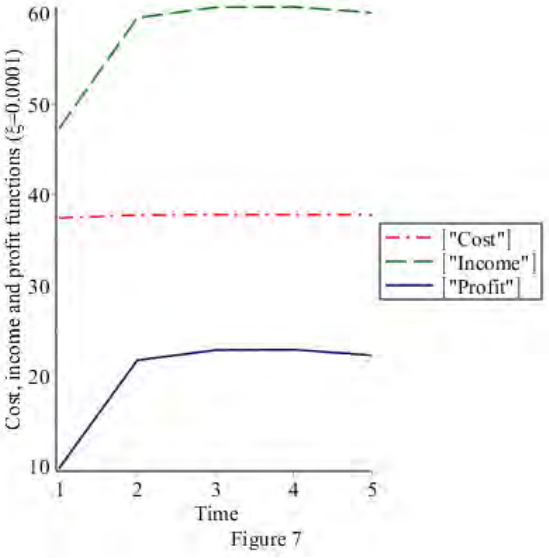


Figure 7

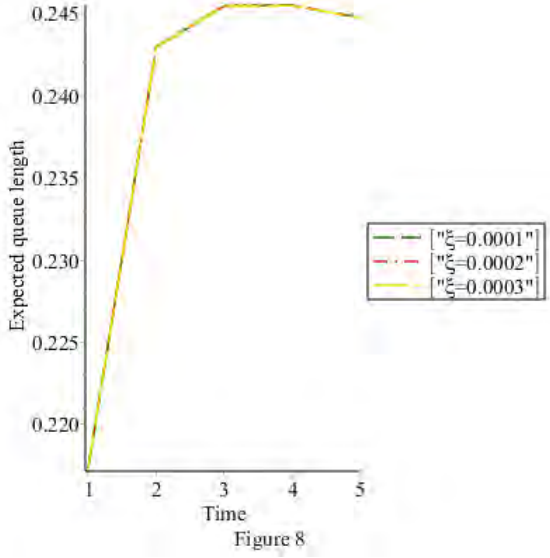


Figure 8

Figure 7: Shows the variation of cost, income and profit at a catastrophes rate ξ=0.0001 with time t while keeping the other parameters fixed (λ=1, μ=5, w=2, q=0.7, β=0.5)

Figure 8: Shows the variation of $Q_L(t)$ with time t by varying catastrophes rate ξ(=0.0001, 0.0002, 0.0003) while keeping the other parameters fixed (λ=1, μ=5, w=2, q=0.7, β=0.5)

length values increases with increase in time up to t(=4.00, 3.00, 3.00) also then decreases slightly. Hence we get the optimal value of t=1 when λ=1.00 and t=3 when λ=1.50 for minimum cost and maximum profit respectively.

Figure 3 shows the variation of cost, income and profit with time t by keeping μ constant (=3.75). The value of cost, income and profit increases with increase in t upto

Table 8: Measures of effectiveness versus q

t	q	$P_V(t)$	$P_B(t)$	$Q_L(t)$	$TC(t)$	$TE_I(t)$	$TE_P(t)$
1	0.55	0.888542	0.111457	0.2197238	37.531604	55.7285	18.196896
2		0.851890	0.148108	0.2461336	37.905650	74.0540	36.148350
3		0.847705	0.152266	0.2481621	37.938274	76.1330	38.194726
4		0.847350	0.152412	0.2479947	37.935993	76.2060	38.270007
5		0.846830	0.151029	0.2471729	37.914111	75.5145	37.600389
1	0.65	0.900287	0.099712	0.2179962	37.479093	49.8560	12.376907
2		0.872523	0.127475	0.2437086	37.819501	63.7375	25.917999
3		0.869757	0.130214	0.2460660	37.851157	65.1070	27.255843
4		0.869455	0.130306	0.2460511	37.850234	65.1530	27.302766
5		0.868872	0.128987	0.2452851	37.829107	64.4935	26.664393
1	0.75	0.910104	0.089895	0.2168165	37.437845	44.9475	7.5096550
2		0.888300	0.111698	0.2424285	37.759369	55.8490	18.089631
3		0.886247	0.113724	0.2450750	37.791777	56.8620	19.070223
4		0.885957	0.113804	0.2451474	37.791691	56.9020	19.110309
5		0.885330	0.112529	0.2444091	37.770973	56.2645	18.493527

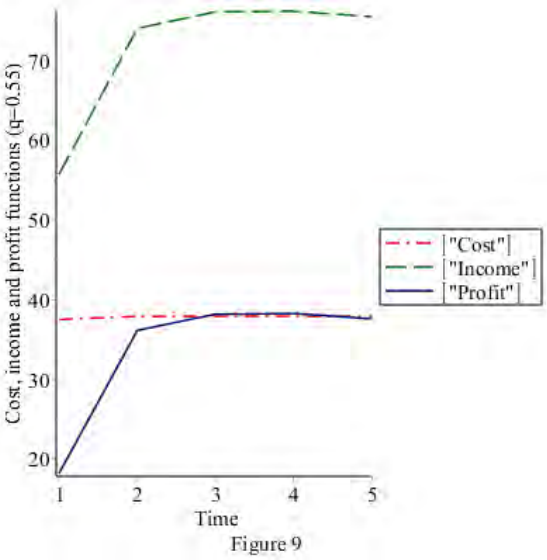


Figure 9

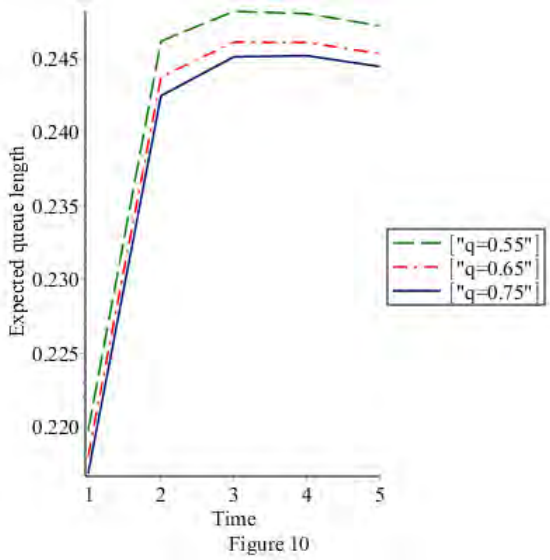


Figure 10

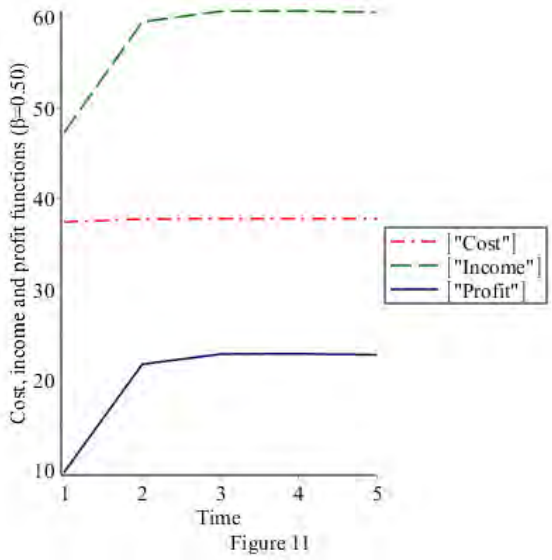
Figure 9: Shows the variation of cost, income and profit at a feedback probability q=0.55 with time t while keeping the other parameters fixed (λ=1, μ=5, w=2, ξ=0.0001, β=0.5)

Figure 10: Shows the variation of $Q_L(t)$ with time t by varying feedback probability q(=0.55, 0.65, 0.75) while keeping the other parameters fixed (λ=1, μ=5, w=2, ξ=0.0001, β=0.5)

t(=3.00, 4.00, 4.00) then decreases slightly. The variation in queue length with time t is represented in figure 4 by varying the service rate μ(=3.75, 4.25, 4.75). Queue length values increases with increase in time up to t(=3.00, 3.00, 4.00) also then decreases slightly. Hence we get the optimal value of t=1 when μ=3.75 and t=4 when μ=3.75 for minimum cost and maximum profit respectively.

Table 9: Measures of effectiveness versus β

t	β	$P_V(t)$	$P_B(t)$	$Q_L(t)$	$TC(t)$	$TE_I(t)$	$TE_P(t)$
1	0.50	0.905411	0.094588	0.2173509	37.457268	47.2940	9.8367320
2		0.880912	0.119086	0.2429674	37.786922	59.5430	21.756078
3		0.878559	0.121412	0.2454781	37.818872	60.7060	22.887128
4		0.878264	0.121497	0.2455127	37.818423	60.7485	22.930077
5		0.877657	0.121202	0.2447627	37.805528	60.6010	22.795472
1	0.60	0.888367	0.111632	0.2612720	37.947611	55.8160	17.868389
2		0.860414	0.139580	0.2915200	38.333910	69.7900	31.456090
3		0.857871	0.142019	0.2942445	38.367952	71.0095	32.641548
4		0.857261	0.141876	0.2937015	38.358328	70.9380	32.579672
5		0.855314	0.140883	0.2911838	38.315472	70.4415	32.126028
1	0.70	0.871860	0.128139	0.3054014	38.438426	64.0695	25.631074
2		0.840733	0.159250	0.3402231	38.879896	79.6250	40.745104
3		0.837961	0.161701	0.3429228	38.912641	80.8505	41.937859
4		0.836564	0.161003	0.3408982	38.879826	80.5015	41.621674
5		0.831663	0.158467	0.3344632	38.770683	79.2335	40.462817



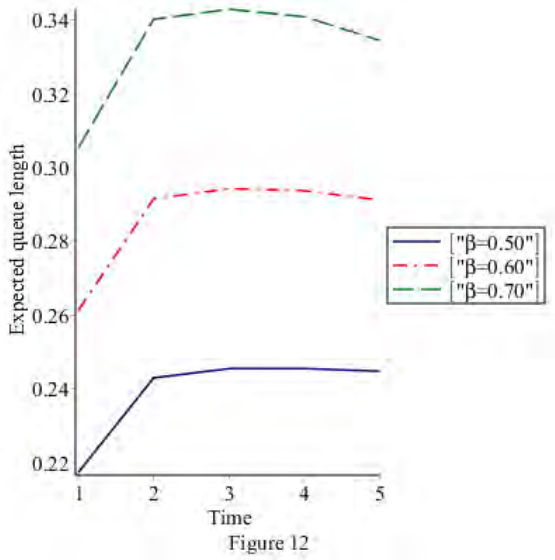


Figure 11: Shows the variation of cost, income and profit at a joining probability β=0.50 with time t while keeping the other parameters fixed (λ=1, μ=5, w=2, ξ=0.0001, q=0.7)

Figure 12: Shows the variation of $Q_L(t)$ with time t by varying joining probability β(=0.50, 0.60, 0.70) while keeping the other parameters fixed (λ=1, μ=5, w=2, ξ=0.0001, q=0.7)

Figure 5 shows the variation of cost, income and profit with time t by keeping w constant (=2.00). The value of cost, income and profit increases with increase in t upto t(=3.00, 4.00, 4.00) then decreases slightly. The variation in queue length with time t is represented in figure 6 by varying the vacation rate w(=2.00, 2.25, 2.50). Queue length values increases with increase in time up to t(=4.00, 3.00, 3.00) also then decreases slightly. Hence we get the optimal value of t=1 when w=2.50 and t=3 when w=2.50 for minimum

cost and maximum profit respectively.

Figure 7 shows the variation of cost, income and profit with time t by keeping ξ constant (=0.0001). The value of cost, income and profit increases with increase in t upto t(=3.00, 4.00, 4.00) then decreases slightly. The variation in queue length with time t is represented in figure 8 by varying the catastrophes rate ξ(=0.0001, 0.0002, 0.0003). Queue length values increases with increase in time up to t(=4.00) then decreases slightly. Hence we get the optimal value of t=1 when ξ=0.0003 and t=4 when ξ=0.0001 for minimum cost and maximum profit respectively. Finally, the variation in rate of catastrophes shows the minor effect on cost and profit.

Figure 9 shows the variation of cost, income and profit with time t by keeping q constant (=0.55). The value of cost, income and profit increases with increase in t upto t(=3.00, 4.00, 4.00) then decreases slightly. The variation in queue length with time t is represented in figure 10 by varying the feedback probability q(=0.55, 0.65, 0.75). Queue length values increases with increase in time up to t(=3.00, 4.00, 4.00) also then decreases slightly. Hence we get the optimal value of t=1 when q=0.75 and t=4 when q=0.55 for minimum cost and maximum profit respectively.

Figure 11 shows the variation of cost, income and profit with time t by keeping β constant (=0.50). The value of cost, income and profit increases with increase in t upto t(=3.00, 4.00, 4.00) then decreases slightly. The variation in queue length with time t is represented in figure 12 by varying the joining probability β(=0.50, 0.60, 0.70). Queue length values increases with increase in time up to t(=4.00, 3.00, 3.00) also then decreases slightly. Hence we get the optimal value of t=1 when β=0.50 and t=3 when β=0.70 for minimum cost and maximum profit respectively.

8. Conclusions and future work

The time-dependent solution for the M/M/2 queueing system with multiple vacation, feedback, catastrophes and balking has been obtained using a two-dimensional state model. Based on various performance measures, total expected cost and total expected profit, the best optimal value is at t=1 when service rate=3.75 and t=3 when arrival rate=1.50 for minimum cost and maximum profit respectively. Some key measures give a greater understanding of system model behaviour. This model finds its application in post office, computer networks, supermarkets, hospital administrations, financial sector and many others.
As part of future study, this model may be examined further for Non-Markovian queues, bulk queues, tandem queues *etc.*

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

Altman, E. and Yechiali, U. (2006). Analysis of customers' impatience in queues with server vacations. *Queueing Systems*, **52**, 261–279.

Ammar, S. I. (2015). Transient analysis of an m/m/1 queue with impatient behavior and multiple vacations. *Applied Mathematics and Computation*, **260**, 97–105.

Chakravarthy, S. R. (2017). A catastrophic queueing model with delayed action. *Applied Mathematical Modelling*, **46**, 631–649.

Chao, X. (1995). A queueing network model with catastrophes and product form solution. *Operations Research Letters*, **18**, 75–79.

Choudhury, G. and Paul, M. (2005). A two phase queueing system with bernoulli feedback. *International Journal of Information and Management Sciences*, **16**, 35.

Chowdhury, A. R. and Indra (2020). Prediction of two-node tandem queue with feedback having state and time dependent service rates. In *Journal of Physics: Conference Series*, volume 1531, page 012063. IOP Publishing.

Cooper, R. B. (1970). Queues served in cyclic order: Waiting times. *Bell System Technical Journal*, **49**, 399–413.

de Oliveira Souza, M. and Rodriguez, P. M. (2021). On a fractional queueing model with catastrophes. *Applied Mathematics and Computation*, **410**, 126468.

Dharmaraja, S. and Kumar, R. (2015). Transient solution of a markovian queuing model with heterogeneous servers and catastrophes. *Opsearch*, **52**, 810–826.

Disney, R. L., McNickle, D. C., and Simon, B. (1980). The m/g/1 queue with instantaneous bernoulli feedback. *Naval Research Logistics Quarterly*, **27**, 635–644.

Doshi, B. T. (1986). Queueing systems with vacations—a survey. *Queueing Systems*, **1**, 29–66.

D'Avignon, G. and Disney, R. (1976). Single-server queues with state-dependent feedback. *INFOR: Information Systems and Operational Research*, **14**, 71–85.

Kalidass, K., Gnanaraj, J., Gopinath, S., and Kasturi, R. (2014). Transient analysis of an m/m/1 queue with a repairable server and multiple vacations. *International Journal of Mathematics in Operational Research*, **6**, 193–216.

Kalidass, K., Gopinath, S., Gnanaraj, J., and Ramanath, K. (2012). Time dependent analysis of an m/m/1/n queue with catastrophes and a repairable server. *Opsearch*, **49**, 39–61.

Ke, J.-C., Wu, C.-H., and Zhang, Z. G. (2010). Recent developments in vacation queueing models: a short survey. *International Journal of Operations Research*, **7**, 3–8.

Krishna Kumar, B., Krishnamoorthy, A., Pavai Madheswari, S., and Sadiq Basha, S. (2007). Transient analysis of a single server queue with catastrophes, failures and repairs. *Queueing Systems*, **56**, 133–141.

Kumar, B. K., Parthasarathy, P., and Sharafali, M. (1993). Transient solution of an m/m/1 queue with balking. *Queueing Systems*, **13**, 441–448.

Pegden, C. D. and Rosenshine, M. (1982). Some new results for the m/m/1 queue. *Management Science*, **28**, 821–828.

Sharma, R. and Indra (2020). Dynamic aspect of two dimensional single server markovian queueing model with multiple vacations and reneging. In *Journal of Physics: Conference Series*, volume 1531, page 012060. IOP Publishing.

Sharma, S. K. and Kumar, R. (2012). A markovian feedback queue with retention of reneged customers and balking. *Advanced Modeling and Optimization*, **14**, 681–688.

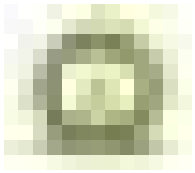
Suranga Sampath, M. and Liu, J. (2018). Transient analysis of an m/m/1 queue with reneging, catastrophes, server failures and repairs. *Bulletin of the Iranian Mathematical Society*, **44**, 585–603.

Takacs, L. (1963). A single-server queue with feedback. *Bell system Technical journal*, **42**, 505–519.

Xu, X. and Zhang, Z. G. (2006). Analysis of multi-server queue with a single vacation (e, d)-policy. *Performance Evaluation*, **63**, 825–838.

Zhang, Y., Yue, D., and Yue, W. (2005). Analysis of an m/m/1/n queue with balking, reneging and server vacations. In *Proceedings of the 5th International Symposium on OR and its Applications*, pages 37–47.

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 22, No. 1, 2024 (New Series), pp 321–341
http://www.ssca.org.in/journal



Singh Maddala Dagum Distribution with Application to Income Data

Ashlin Varkey and Haritha N. Haridas
Department of Statistics, Farook College (Autonomous), Kozhikode, University of Calicut, Kerala, India 673632
Received: 15 September 2023; Revised: 08 October 2023; Accepted: 12 October 2023

Abstract

This article introduces the Singh Maddala Dagum distribution as the sum of the quantile functions of the Singh-Maddala and Dagum distributions. The distributional properties, income inequality measures, and poverty measures of this distribution are derived. Poverty measures such as the poverty gap ratio and the Foster-Greer-Thorbecke measure were converted to quantile forms. The least squares method is used to estimate the parameters of the proposed distribution, and the model is applied to two real datasets.

Key words: Singh-Maddala distribution; Dagum distribution; Quantile function; Income inequality measures; Poverty measures.

AMS Subject Classifications: 33B15, 33B20, 60E05, 62G07, 62P20

1. Introduction

The two equivalent techniques for modeling and analyzing statistical data are by using the distribution functions and quantile functions. The quantile function for a real-valued and continuous random variable X with distribution function $F(x)$ is given as

$$Q(u) = F^{-1}(u) = \inf\{x : F(x) \geq u\}, \; 0 \leq u \leq 1.$$

Even though Galton (1875) first proposed the formal concept of quantiles, the work of Hastings *et al.* (1947) provided a notable advancement in depicting quantile functions to represent distributions. Parzen's (1979) paper and Tukey's (1977) research on exploratory data analysis stimulated the development of the quantile functions as a vital tool in statistical analysis instead of the distribution functions.

The quantile function holds a number of characteristics that the distribution function does not have. In particular, two quantile functions added together and two positive quantile functions multiplied together are again quantile functions. Also, $\frac{1}{Q(1-u)}$ is the quantile function of $\frac{1}{X}$, if $Q(u)$ is the quantile function of X. For a comprehensive review of this concept, one can refer to Nair *et al.* (2013), Gilchrist (2000), Sankaran and Dileep Kumar (2018), and the references therein.

Corresponding Author: Ashlin Varkey
Email: ashlinvarkey@gmail.com

Tarsitano (2004) used a general form of the Tukey lambda family of distributions proposed by Ramberg and Schmeiser (1972), to provide a good start for quantile-based income modeling. However, the model put forth by Tarsitano (2004) is not valid throughout the parametric space. To solve this issue, Haritha *et al.* (2007) utilized the four-parameter generalized lambda distribution proposed by Freimer *et al.* (1988) for income modeling. Later, using the quantile function method, the Zenga measure and other measures of income inequality were examined by Sreelakshmi and Nair (2014).

The objective of this paper is to introduce a new quantile function that is useful for the analysis of income data. Since Singh-Maddala (SM) and Dagum distributions are adaptable and frequently used in income modeling, we propose the Singh Maddala Dagum (SMD) distribution derived from the sum of the quantile functions of the two models.

Singh and Maddala introduced the SM distribution in 1975 and refined it in 1976, has received special attention among income distributions. The SM distribution is a special case of the generalized beta 2 (GB2) distribution and is known as Burr XII or simply Burr distribution. For a detailed study on the SM distribution, one could refer to Kleiber and Kotz (2003),Shahzad and Asghar (2013b), and Kumar (2017). The distribution and quantile functions of the SM distribution are given by

$$G(x) = 1 - \left[1 + \left(\frac{x}{b}\right)^{a}\right]^{-q}, \quad x > 0, \tag{1}$$

and

$$Q_1(u) = b\left[(1-u)^{-\frac{1}{q}} - 1\right]^{\frac{1}{a}}, \quad 0 < u < 1, \tag{2}$$

where all three parameters a, b, q are positive.

Dagum distribution proposed by Dagum (1977) is also a special case of GB2 distribution and is known as Burr III distribution. Dagum distribution has numerous applications in the fields of reliability, meteorology, quality control, insurance, business failure data, and income modeling. A detailed discussion of the Dagum distribution can be found in Kleiber and Kotz (2003) and Shahzad and Asghar (2013a). Using the SM and Dagum distributions Saulo *et al.* (2023) proposed parametric quantile regressions. The distribution and quantile functions of the Dagum distribution are given by

$$H(x) = \left[1 + \left(\frac{x}{b}\right)^{-a}\right]^{-p}, \quad x > 0, \tag{3}$$

and

$$Q_2(u) = b\left[u^{-\frac{1}{p}} - 1\right]^{-\frac{1}{a}}, \quad 0 < u < 1, \tag{4}$$

where all three parameters a, b, p are positive.

The remaining portion of the article is structured as follows. We define SMD distribution and its basic aspects in Section 2. Section 3 deals with some popular distributions that belong to the proposed class or that result from pertinent transformations on the proposed quantile function. Section 4 covers the distributional properties, such as skewness, kurtosis, L-moments, order statistics, etc. Section 5 discusses the major income inequalities and poverty measures of the proposed class. The inference method and its application to real data are carried out in Section 6. Overall findings from the study are given in the final Section 7.

2. Singh Maddala Dagum (SMD) quantile function

If X and Y are two non-negative random variables with quantile functions $Q_1(u)$ and $Q_2(u)$ respectively. Then

$$Q(u) = Q_1(u) + Q_2(u),$$

is again a quantile function. Likewise, the sum of two quantile density functions results in a quantile density function. Now we define a new quantile function

$$Q(u) = b\left[\left((1-u)^{-\frac{1}{q}} - 1\right)^{\frac{1}{a}} + \left(u^{-\frac{1}{p}} - 1\right)^{-\frac{1}{a}}\right],\ 0 < u < 1,\ a,\ b,\ p,\ q > 0, \tag{5}$$

which is the sum of quantile functions in (2) and (4). The proposed class of distribution is known as SMD distribution and its support is $(0, \infty)$. The quantile density function of the SMD distribution is

$$\begin{aligned} q(u) &= \frac{dQ(u)}{du} \\ &= b\left[\frac{(1-u)^{-\frac{1}{q}-1}\left((1-u)^{-\frac{1}{q}} - 1\right)^{\frac{1}{a}-1}}{aq} + \frac{u^{-\frac{1}{p}-1}\left(u^{-\frac{1}{p}} - 1\right)^{-\frac{1}{a}-1}}{ap}\right]. \end{aligned}$$

The density and distribution functions are not available in closed form for the family of distributions given in (5). However, these can be computed by numerical inversion of the quantile function. In terms of the distribution function, the density function $f(x)$ of the proposed class can be written as

$$f(x) = \frac{1}{b}\left[\frac{apq\ \ F(x)^{\frac{1}{p}+1}(1-F(x))^{\frac{1}{q}+1}}{pF(x)^{\frac{1}{p}+1}[(1-F(x))^{-\frac{1}{q}} - 1]^{\frac{1}{a}-1} + q(1-F(x))^{\frac{1}{q}+1}(F(x)^{-\frac{1}{p}} - 1)^{-\frac{1}{a}-1}}\right]. \tag{6}$$

The density function is plotted for various parameter combinations and is given in Figure 1. For various parameter values, it can be seen that the family includes decreasing, unimodal, positive, and negatively skewed models.

3. Members of the family

We can obtain several popular distributions from the suggested model (5) for various parameter values and by utilizing some transformations given in Gilchrist (2000).

Case 1. $b > 0, q > 0, a = 1$ and $p \to 0$

The quantile function of the suggested class tends to the Lomax distribution and is given as

$$Q(u) = b\left[(1-u)^{-\frac{1}{q}} - 1\right]. \tag{7}$$

Case 2. $b > 0, a > 0, q = 1$ and $p \to 0$

The quantile function of the suggested class tends to the Fisk distribution and is given as

$$Q(u) = b\left[(1-u)^{-1} - 1\right]^{\frac{1}{a}}. \tag{8}$$

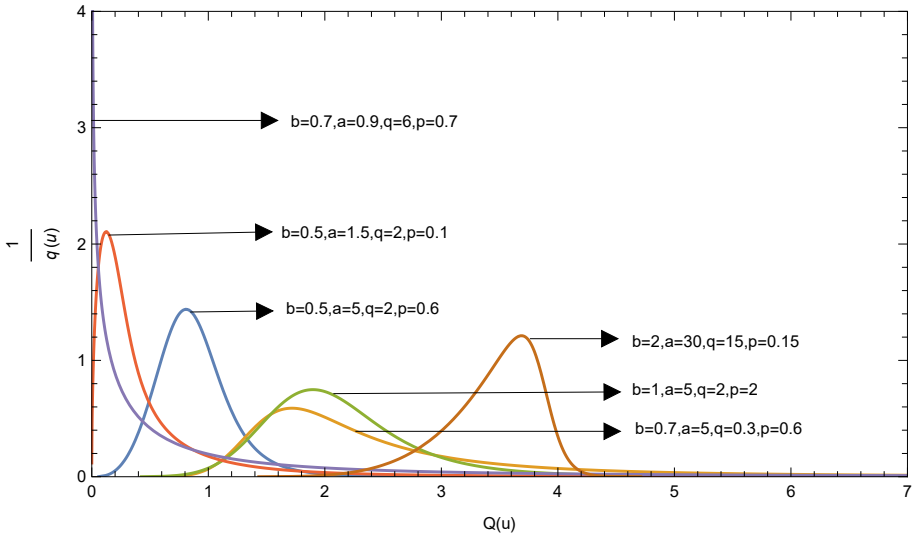


Figure 1: Plots of density function for different values of parameters

Case 3. $b > 0, a = q$ and $p \to 0$
The quantile function of the suggested class tends to the Paralogistic distribution and is given as

$$Q(u) = b\left[(1-u)^{-\frac{1}{a}} - 1\right]^{\frac{1}{a}}. \tag{9}$$

On applying reciprocal transformation on (9), we get the inverse Paralogistic distribution with quantile function

$$Q(u) = \frac{1}{Q(1-u)} = k\left(u^{-\frac{1}{a}} - 1\right)^{-\frac{1}{a}},$$

where $k = \frac{1}{b}$ and a are the parameters. Further details on paralogistic and inverse paralogistic distributions can be found in Klugman *et al.* (2019).

The following theorems give the relationships between the random variables representing the SM, SMD, and Dagum distributions.

Theorem 1: If $V \sim SM(a, b, q)$ then the random variable,
$U = V + b\left\{\left[1 - \left(1 + \left(\frac{V}{b}\right)^a\right)^{-q}\right]^{-\frac{1}{p}} - 1\right\}^{-\frac{1}{a}}$ has $SMD(a, b, p, q)$ distribution.

Proof:

Let S and R represent two random variables with distribution functions $F_S(x)$ and $F_R(x)$ and quantile functions $Q_S(u)$ and $Q_R(u)$ respectively. Assume $Q^*(u) = Q_S(u) + Q_R(u)$, then the random variable that corresponds to the quantile function $Q^*(u)$ is $S + Q_R(F_S(S))$ or $R + Q_S(F_R(R))$ (Sankaran *et al.*, 2016).

Let $V \sim SM(a, b, q)$ and $W \sim Dagum(a, b, p)$; then $V + Q_W(F_V(V))$ has $SMD(a, b, p, q)$ distribution by above result.
We have, $Q_W(u) = b\left(u^{-\frac{1}{p}} - 1\right)^{-\frac{1}{a}}$ and $F_V(V) = 1 - \left[1 + \left(\frac{V}{b}\right)^a\right]^{-q}$

Therefore, $V + Q_W(F_V(V)) = V + b\left\{\left[1 - \left(1 + \left(\frac{V}{b}\right)^a\right)^{-q}\right]^{-\frac{1}{p}} - 1\right\}^{-\frac{1}{a}}$ has $SMD(a, b, p, q)$ distribution. □

Theorem 2: If $W \sim Dagum(a, b, p)$, then the random variable,
$U = W + b\left\{\left[1 - \left(1 + \left(\frac{W}{b}\right)^{-a}\right)^{-p}\right]^{-\frac{1}{q}} - 1\right\}^{\frac{1}{a}}$ has $SMD(a, b, p, q)$ distribution.

Proof: The proof is omitted since it is similar to that of Theorem 1. □

4. Distributional characteristics

The use of quantile functions reduces the effort needed to describe a distribution through its moments. Hence it is common in statistical analysis to use quantile-based measurements of distributional features like location, dispersion, skewness, and kurtosis. These measurements can be used to estimate the model's parameters by matching population characteristics with corresponding sample characteristics.

4.1. Measures of location, spread and shape

The r^{th} order traditional moment is given as

$$E(X^r) = \int_0^1 (Q(u))^r du.$$

In particular, the mean of the SMD distribution is

$$\mu = b\left[\frac{\Gamma\left(1 + \frac{1}{a}\right)\Gamma\left(q - \frac{1}{a}\right)}{\Gamma(q)} + \frac{\Gamma\left(p + \frac{1}{a}\right)\Gamma\left(1 - \frac{1}{a}\right)}{\Gamma(p)}\right].$$

For the model given in (5), the median (M) is

$$\begin{aligned} M &= Q(0.5) \\ &= b\left[\left(2^{\frac{1}{q}} - 1\right)^{\frac{1}{a}} + \left(2^{\frac{1}{p}} - 1\right)^{-\frac{1}{a}}\right]. \end{aligned} \tag{10}$$

The interquartile range (IQR) is

$$\begin{aligned} IQR &= Q(0.75) - Q(0.25) \\ &= b\Big\{\left[(0.25)^{-\frac{1}{q}} - 1\right]^{\frac{1}{a}} - \left[(0.75)^{-\frac{1}{q}} - 1\right]^{\frac{1}{a}} \\ &+ \left[(0.75)^{-\frac{1}{p}} - 1\right]^{-\frac{1}{a}} - \left[(0.25)^{-\frac{1}{p}} - 1\right]^{-\frac{1}{a}}\Big\}. \end{aligned} \tag{11}$$

Galton's skewness (S) and Moors kurtosis (T) measures are given in (12) and (13) respectively.

$$S = \frac{Q(0.25) + Q(0.75) - 2M}{IQR} = \frac{S_1 + S_2}{\left[(0.25)^{-\frac{1}{q}} - 1\right]^{\frac{1}{a}} - \left[(0.75)^{-\frac{1}{q}} - 1\right]^{\frac{1}{a}} + \left[(0.75)^{-\frac{1}{p}} - 1\right]^{-\frac{1}{a}} - \left[(0.25)^{-\frac{1}{p}} - 1\right]^{-\frac{1}{a}}}, \quad (12)$$

where $S_1 = \left[(0.25)^{-\frac{1}{q}} - 1\right]^{\frac{1}{a}} + \left[(0.75)^{-\frac{1}{q}} - 1\right]^{\frac{1}{a}} - 2\left[2^{\frac{1}{q}} - 1\right]^{\frac{1}{a}}$,
and $S_2 = \left[(0.25)^{-\frac{1}{p}} - 1\right]^{-\frac{1}{a}} + \left[(0.75)^{-\frac{1}{p}} - 1\right]^{-\frac{1}{a}} - 2\left[2^{\frac{1}{p}} - 1\right]^{-\frac{1}{a}}$.

$$T = \frac{Q(0.875) - Q(0.625) + Q(0.375) - Q(0.125)}{IQR} = \frac{T_1 + T_2}{\left[(0.25)^{-\frac{1}{q}} - 1\right]^{\frac{1}{a}} - \left[(0.75)^{-\frac{1}{q}} - 1\right]^{\frac{1}{a}} + \left[(0.75)^{-\frac{1}{p}} - 1\right]^{-\frac{1}{a}} - \left[(0.25)^{-\frac{1}{p}} - 1\right]^{-\frac{1}{a}}}, \quad (13)$$

where $T_1 = \left[(0.125)^{-\frac{1}{q}} - 1\right]^{\frac{1}{a}} - \left[(0.375)^{-\frac{1}{q}} - 1\right]^{\frac{1}{a}} + \left[(0.625)^{-\frac{1}{q}} - 1\right]^{\frac{1}{a}} - \left[(0.875)^{-\frac{1}{q}} - 1\right]^{\frac{1}{a}}$,
and $T_2 = \left[(0.875)^{-\frac{1}{p}} - 1\right]^{-\frac{1}{a}} - \left[(0.625)^{-\frac{1}{p}} - 1\right]^{-\frac{1}{a}} + \left[(0.375)^{-\frac{1}{p}} - 1\right]^{-\frac{1}{a}} - \left[(0.125)^{-\frac{1}{p}} - 1\right]^{-\frac{1}{a}}$.

4.2. L-moments

The L-moments are alternatives to the classical moments and are the expected values of linear functions of order statistics. The work on order statistics by Sillitto (1969) and Greenwood *et al.* (1979) laid the foundation for L-moments, but Hosking (1990) developed a comprehensive theory on L-moments. These moments are resistant to outliers and typically have reduced sample variances. Like classical moments, L-moments can be used to identify distributions, summarise measures of probability distributions, and fit models to data. The r^{th} L-moment is represented as

$$L_r = \int_0^1 \sum_{k=0}^{r-1} (-1)^{r-1-k} \binom{r-1}{k} \binom{r-1+k}{k} u^k Q(u) du.$$

The first four L-moments of SMD distributions are

$$L_1 = b\left[A_1 O_1 + A_2 R_1\right],$$

$$L_2 = b\left[A_1 (O_1 - O_2) - A_2 (R_1 - R_2)\right],$$

$$L_3 = b\left[A_1 (O_1 - 3O_2 + 2O_3) + A_2 (R_1 - 3R_2 + 2R_3)\right],$$

$$L_4 = b\left[A_1 (O_1 - 6O_2 + 10O_3 - 5O_4) - A_2 (R_1 - 6R_2 + 10R_3 - 5R_4)\right],$$

where $A_1 = \Gamma\left(1+\frac{1}{a}\right)$, $A_2 = \Gamma\left(1-\frac{1}{a}\right)$, $O_i = \frac{\Gamma\left(iq-\frac{1}{a}\right)}{\Gamma(iq)}$, $R_i = \frac{\Gamma\left(ip+\frac{1}{a}\right)}{\Gamma(ip)}$ and $i = 1, 2, 3, 4$.
The L-coefficient of variation (τ_2), which is an alternative to the coefficient of variation based on traditional moments is

$$\begin{aligned}\tau_2 &= \frac{L2}{L1} \\ &= \frac{A_1(O_1 - O_2) - A_2(R_1 - R_2)}{A_1 O_1 + A_2 R_1}.\end{aligned} \tag{14}$$

The L-coefficient of skewness (τ_3) and L-coefficient kurtosis (τ_4) of the SMD distribution, is given in (15) and (16).

$$\begin{aligned}\tau_3 &= \frac{L3}{L2} \\ &= \frac{A_1(O_1 - 3O_2 + 2O_3) + A_2(R_1 - 3R_2 + 2R_3)}{A_1(O_1 - O_2) - A_2(R_1 - R_2)}.\end{aligned} \tag{15}$$

$$\begin{aligned}\tau_4 &= \frac{L4}{L2} \\ &= \frac{A_1(O_1 - 6O_2 + 10O_3 - 5O_4) - A_2(R_1 - 6R_2 + 10R_3 - 5R_4)}{A_1(O_1 - O_2) - A_2(R_1 - R_2)}.\end{aligned} \tag{16}$$

The plots of L-coefficients of skewness (τ_3) and kurtosis (τ_4) for different parameter values are given in Figures 2, 3 and 4. In Figure 2, the curve of τ_3 decreases with a for fixed value of q and p but the curve of τ_4 decreases with a for fixed value of q and p, when $p > 1$. In Figure 3, the curves of τ_3 and τ_4 increase with p for fixed values of a and q when $q \geq 1$. The curves of τ_3 and τ_4 for fixed values of a and p and for varying q are given in Figure 4.

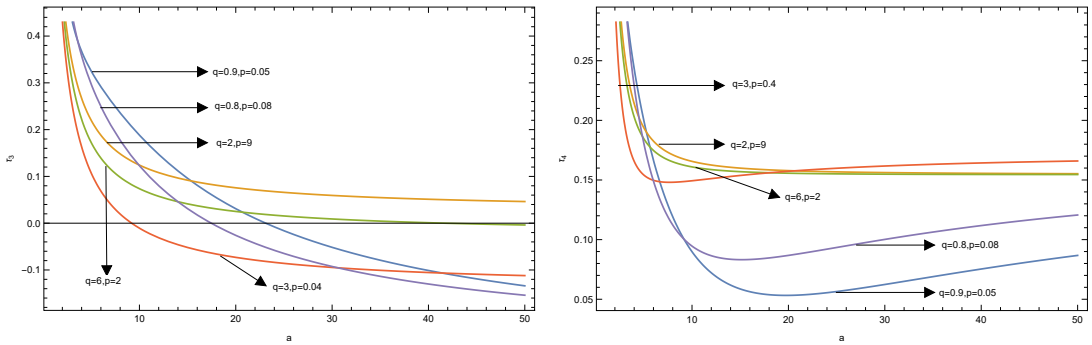


Figure 2: Plot of L-coefficients of skewness and kurtosis for particular values of q and p as a function of the parameter a

4.3. Order statistics

In a random sample of size n, let $X_{r:n}$ represent the r^{th} order statistic. Then, $X_{r:n}$ has density function $f_r(x)$ and is given as

$$f_r(x) = \frac{1}{\beta(r, n-r+1)} f(x) F(x)^{r-1} (1 - F(x))^{n-r}.$$

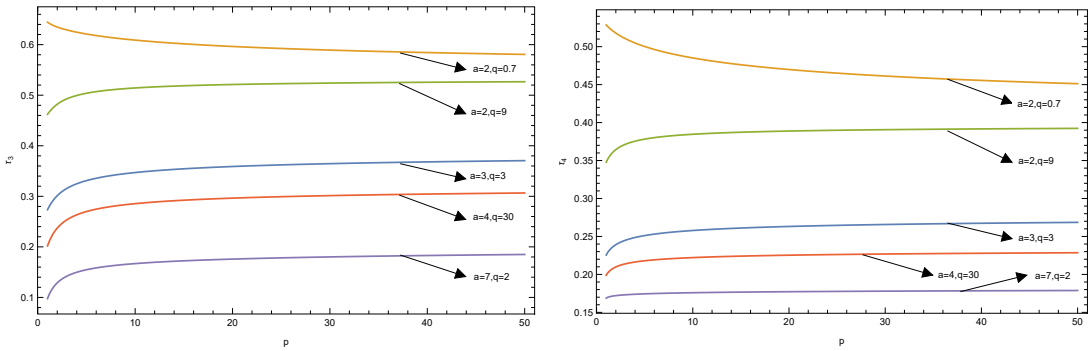


Figure 3: Plot of L-coefficients of skewness and kurtosis for particular values of a and q as a function of the parameter p

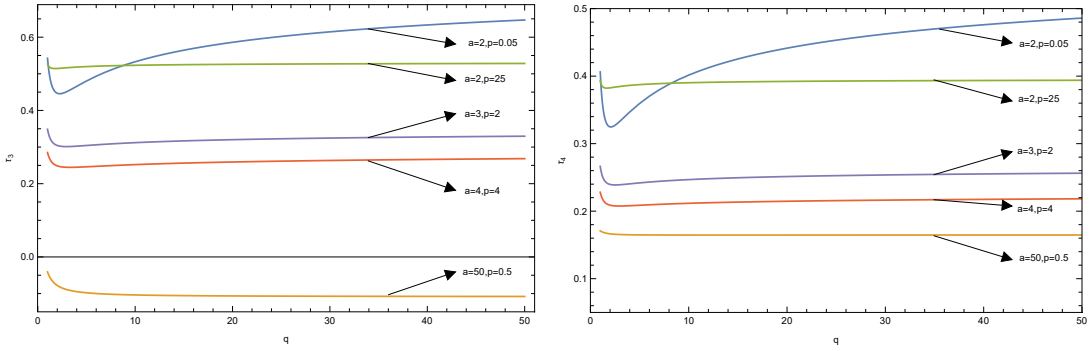


Figure 4: Plot of L-coefficients of skewness and kurtosis for particular values of a and p as a function of the parameter q

From (6) we get

$$f_r(x) = \frac{apq}{b\beta(r, n-r+1)} \frac{F(x)^{r+\frac{1}{p}}(1-F(x))^{n+\frac{1}{q}+1-r}}{pF(x)^{\frac{1}{p}+1}\left[(1-F(x))^{-\frac{1}{q}}-1\right]^{\frac{1}{a}-1} + q(1-F(x))^{\frac{1}{q}+1}\left(F(x)^{-\frac{1}{p}}-1\right)^{-\frac{1}{a}-1}}.$$

Thus

$$E(X_{r:n}) = \frac{apq}{b\beta(r, n-r+1)} \times \int_0^\infty \frac{xF(x)^{r+\frac{1}{p}}(1-F(x))^{n+\frac{1}{q}+1-r}}{pF(x)^{\frac{1}{p}+1}\left[(1-F(x))^{-\frac{1}{q}}-1\right]^{\frac{1}{a}-1} + q(1-F(x))^{\frac{1}{q}+1}\left(F(x)^{-\frac{1}{p}}-1\right)^{-\frac{1}{a}-1}} dx.$$

In quantile terms, the above expression can be written as

$$E(X_{r:n}) = \frac{apq}{b\beta(r, n-r+1)} \int_0^1 \frac{Q(u)\, u^{r+\frac{1}{p}}(1-u)^{n+\frac{1}{q}+1-r}}{p\, u^{\frac{1}{p}+1}\left((1-u)^{-\frac{1}{q}}-1\right)^{\frac{1}{a}-1} + q(1-u)^{\frac{1}{q}+1}\left(u^{-\frac{1}{p}}-1\right)^{-\frac{1}{a}-1}} du.$$

For SMD distribution the first-order statistic $X_{1:n}$ has a quantile function

$$\begin{aligned} Q_{(1)}(u) &= Q\left[1-(1-u)^{\frac{1}{n}}\right] \\ &= b\left\{\left[(1-u)^{-\frac{1}{nq}}-1\right]^{\frac{1}{a}}+\left[\left(1-(1-u)^{\frac{1}{n}}\right)^{-\frac{1}{p}}-1\right]^{-\frac{1}{a}}\right\}, \end{aligned} \tag{17}$$

and the n^{th} order statistic $X_{n:n}$ has the quantile function

$$\begin{aligned} Q_{(n)}(u) &= Q\left(u^{\frac{1}{n}}\right) \\ &= b\left\{\left[\left(1-u^{\frac{1}{n}}\right)^{-\frac{1}{q}}-1\right]^{\frac{1}{a}}+\left[u^{-\frac{1}{np}}-1\right]^{-\frac{1}{a}}\right\}. \end{aligned} \tag{18}$$

5. Income inequality and poverty measures

In statistical and economics literature, the study of income inequality and poverty measures are always popular and favorite subjects. A measure of income inequality is intended to give an index, that can reduce the differences in income that exist among the members of a group, whereas a poverty measure evaluates the severity of poverty experienced by those whose income is below a pre-determined poverty level.

5.1. Income inequality measures

The Lorenz curve proposed by Lorenz (1905) is a flexible tool for reporting and graphically depicting income inequality. When the income is arranged in increasing order of magnitude, the points $(u, L(u))$ define a Lorenz curve, where u denotes the cumulative frequency of income receiving units and $L(u)$ denotes the cumulative frequency of income. Gastwirth (1971) gave a general definition of Lorenz curve as

$$L(u) = \frac{1}{\mu}\int_0^u Q(p)dp,$$

where $\mu = \int_0^1 Q(p)dp$. For SMD distribution the Lorenz curve is

$$L(u) = \frac{q\beta_{1-(1-u)^{\frac{1}{q}}}\left(1+\frac{1}{a}, q-\frac{1}{a}\right)+p\beta_{u^{\frac{1}{p}}}\left(p+\frac{1}{a}, 1-\frac{1}{a}\right)}{q\beta\left(1+\frac{1}{a}, q-\frac{1}{a}\right)+p\beta\left(p+\frac{1}{a}, 1-\frac{1}{a}\right)}, \tag{19}$$

where $\beta_*(.,.)$, is an incomplete beta function.

The Gini index is a well known income inequality proposed by Gini (1914) and is defined as two times the area between the Lorenz curve and the egalitarian line. The Gini index for the class of distributions in (5) is

$$\begin{aligned} G &= 1-2\int_0^1 L(u)du \\ &= 1-2\left[\frac{q\beta\left(1+\frac{1}{a}, 2q-\frac{1}{a}\right)+p\beta\left(p+\frac{1}{a}, 1-\frac{1}{a}\right)-p\beta\left(2p+\frac{1}{a}, 1-\frac{1}{a}\right)}{q\beta\left(1+\frac{1}{a}, q-\frac{1}{a}\right)+p\beta\left(p+\frac{1}{a}, 1-\frac{1}{a}\right)}\right]. \end{aligned} \tag{20}$$

Pietra (1932) developed the Pietra index which measures the maximal vertical distance between the Lorenz curve and the line of equality. The Pietra index and relative mean deviation in quantile terms are

$$P = \frac{\vartheta_1}{2\mu},$$
$$\tau_2 = \frac{\vartheta_1}{\mu},$$

where $\vartheta_1 = \int_0^1 |Q(u) - Q(u_0)|du$ and $\mu = Q(u_0)$ for some $0 < u_0 < 1$. Further, by solving for u in the equation $\mu = Q(u)$, u_0 can be obtained, and μ represents the mean of the distribution.
Now, the Pietra index of the SMD distribution is given as

$$P = \frac{u_0 Q(u_0) - b\left[q\beta_{1-(1-u_0)^{\frac{1}{q}}}\left(1+\frac{1}{a}, q-\frac{1}{a}\right) + p\beta_{u_0^{\frac{1}{p}}}\left(p+\frac{1}{a}, 1-\frac{1}{a}\right)\right]}{\mu}. \tag{21}$$

The Bonferroni curve proposed by Bonferroni (1930) is used to quantify the variability in income distribution. For an absolutely continuous and non-negative random variable, the Bonferroni curve in quantile terms is given as

$$B_F(u) = \frac{L(u)}{u}$$
$$= \frac{1}{u\mu}\int_0^u Q(p)dp.$$

For SMD distribution the Bonferroni curve is

$$B_F(u) = \frac{q\beta_{1-(1-u)^{\frac{1}{q}}}\left(1+\frac{1}{a}, q-\frac{1}{a}\right) + p\beta_{u^{\frac{1}{p}}}\left(p+\frac{1}{a}, 1-\frac{1}{a}\right)}{u\left[q\beta\left(1+\frac{1}{a}, q-\frac{1}{a}\right) + p\beta\left(p+\frac{1}{a}, 1-\frac{1}{a}\right)\right]}. \tag{22}$$

A more realistic curve was introduced by Zenga (2007) based on the conditional expectation of the concerned distribution. The Zenga curve in quantile terms is

$$Z(u) = 1 - \frac{(1-u)}{u}\frac{\int_0^u Q(p)dp}{\int_u^1 Q(p)dp}.$$

For SMD distribution the Zenga curve is given as

$$Z(u) = \frac{q\,\mathfrak{Z}_1 + p\,\mathfrak{Z}_2}{q\,\mathfrak{Z}_3 + p\,\mathfrak{Z}_4}, \tag{23}$$

where

$$\mathfrak{Z}_1 = \left[\beta\left(1+\frac{1}{a}, q-\frac{1}{a}\right) - u^{-1}\beta_{1-(1-u)^{\frac{1}{q}}}\left(1+\frac{1}{a}, q-\frac{1}{a}\right)\right],$$
$$\mathfrak{Z}_2 = \left[\beta\left(p+\frac{1}{a}, 1-\frac{1}{a}\right) - u^{-1}\beta_{u^{\frac{1}{p}}}\left(p+\frac{1}{a}, 1-\frac{1}{a}\right)\right],$$
$$\mathfrak{Z}_3 = \left[\beta\left(1+\frac{1}{a}, q-\frac{1}{a}\right) - \beta_{1-(1-u)^{\frac{1}{q}}}\left(1+\frac{1}{a}, q-\frac{1}{a}\right)\right],$$
$$\mathfrak{Z}_4 = \left[\beta\left(p+\frac{1}{a}, 1-\frac{1}{a}\right) - \beta_{u^{\frac{1}{p}}}\left(p+\frac{1}{a}, 1-\frac{1}{a}\right)\right].$$

The Lorenz, Bonferroni, and Zenga curves of SMD distribution are given in Figure 5, 6, and 7 respectively.

The Frigyes measures developed by Éltetö and Frigyes (1968) have clear economic interpretations and are given as

$$\varphi = \frac{m}{m_1}, \psi = \frac{m_2}{m_1}, \omega = \frac{m_2}{m},$$

where $m = E(X)$, $m_1 = E(X|X < m)$, and $m_2 = E(X|X \geqslant m)$. The measure ψ can be considered as an inequality measure for the complete income distribution, whereas φ and ω denote the inequalities of the two respective portions of the distribution below and above the mean.

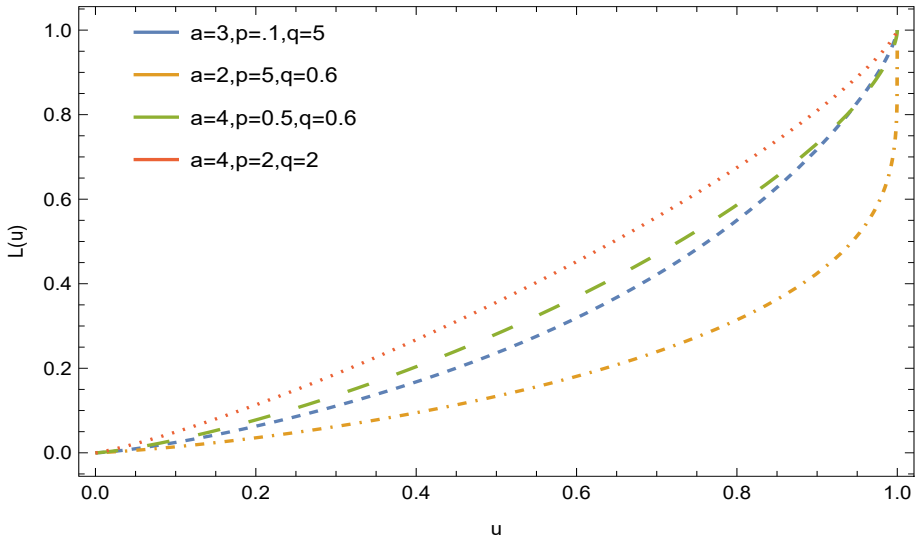


Figure 5: Graph of SMD Lorenz curve

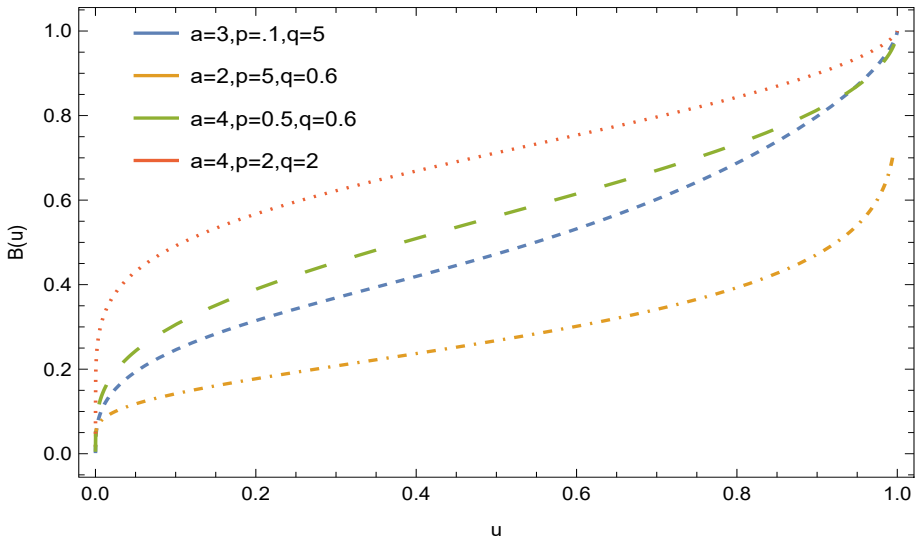


Figure 6: Graph of SMD Bonferroni curve

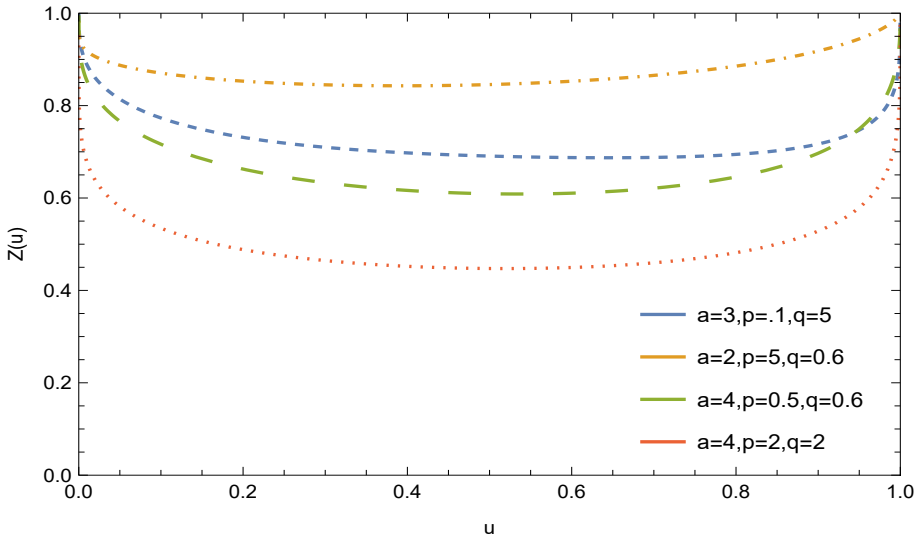


Figure 7: Graph of SMD Zenga curve

In quantile terms, these measures are given as

$$\varphi = \frac{u_0 Q(u_0)}{\int_0^{u_0} Q(u)du},$$
$$\psi = \frac{u_0}{1-u_0}\frac{\int_{u_0}^1 Q(u)du}{\int_0^{u_0} Q(u)du},$$
$$\omega = \frac{\int_{u_0}^1 Q(u)du}{(1-u_0)Q(u_0)}.$$

For SMD distribution these measures are

$$\varphi = \frac{u_0\left[\left((1-u_0)^{-\frac{1}{q}}-1\right)^{\frac{1}{a}} + \left(u_0^{-\frac{1}{p}}-1\right)^{-\frac{1}{a}}\right]}{q\beta_{1-(1-u_0)^{\frac{1}{q}}}\left(1+\frac{1}{a},q-\frac{1}{a}\right) + p\beta_{u_0^{\frac{1}{p}}}\left(p+\frac{1}{a},1-\frac{1}{a}\right)}, \tag{24}$$

$$\psi = \frac{u_0}{(1-u_0)}\left[\frac{q\beta\left(1+\frac{1}{a},q-\frac{1}{a}\right)+p\beta\left(p+\frac{1}{a},1-\frac{1}{a}\right)}{q\beta_{1-(1-u_0)^{\frac{1}{q}}}\left(1+\frac{1}{a},q-\frac{1}{a}\right)+p\beta_{u_0^{\frac{1}{p}}}\left(p+\frac{1}{a},1-\frac{1}{a}\right)}-1\right], \tag{25}$$

$$\omega = \frac{q\,\mathfrak{w}_1 + p\,\mathfrak{w}_2}{(1-u_0)\left[\left((1-u_0)^{-\frac{1}{q}}-1\right)^{\frac{1}{a}} + \left(u_0^{-\frac{1}{p}}-1\right)^{-\frac{1}{a}}\right]}. \tag{26}$$

where

$$\mathfrak{w}_1 = \left[\beta\left(1+\frac{1}{a},q-\frac{1}{a}\right) - \beta_{1-(1-u_0)^{\frac{1}{q}}}\left(1+\frac{1}{a},q-\frac{1}{a}\right)\right],$$
$$\mathfrak{w}_2 = \left[\beta\left(p+\frac{1}{a},1-\frac{1}{a}\right) - \beta_{u_0^{\frac{1}{p}}}\left(p+\frac{1}{a},1-\frac{1}{a}\right)\right].$$

5.2. Poverty measures

Measures of poverty are primarily used to track socioeconomic development and set goals for success or failure. Most of the poverty measurements can be stated as the average deprivation faced by the poor. If the function $D(z, y)$ describes the level of deprivation experienced by an individual whose income y is less than the poverty line z.
Hence

$$\begin{aligned} P &= E_y \left[D\left(z, y\right) I\left(y < z\right) \right] \\ &= \int_0^z D\left(z, y\right) f(y) dy, \end{aligned} \tag{27}$$

where $I\left(y < z\right)$ represents an indicator function which takes value 1 when $y < z$ and 0 otherwise, and $f(y)$ represents probability density function. For a detailed reading on poverty measures, one can refer to Kakwani (1980) and Morduch (2008). Chotikapanich *et al.* (2013) derived poverty measures from generalized beta distribution and examined how poverty has changed in south and southeast Asian nations.

The headcount ratio is the most basic and widely used measure of poverty, it represents the proportion of the population who are poor and is denoted by H. By definition

$$H = \frac{N_p}{N},$$

where N_p and N denotes the number of poor and total population respectively. That is, the head-count ratio ignores the severity of the deprivation experienced by the poor.

A number of alternatives to the head-count ratio have been proposed in order to establish a measure that takes into account both the proportion of poor as well as the intensity of poverty among those who are characterized as poor. The poverty gap ratio calculates the amount of money by which each person falls below the poverty line. It can be obtained from (27), by setting $D\left(z, y\right) = \left(\frac{z-y}{z}\right)$. Thus

$$\begin{aligned} PG &= \int_0^z D\left(z, y\right) f(y) dy \\ &= \int_0^z \left(\frac{z-y}{z}\right) f(y) dy. \end{aligned} \tag{28}$$

Using the transformation, $F(z) = u$ and $F(y) = p$, where $0 < u < 1$ and $0 < p < 1$ in (28), we get the poverty gap ratio in quantile form and is given as

$$PG = \int_0^u \left(\frac{Q(u) - Q(p)}{Q(u)}\right) dp. \tag{29}$$

The poverty gap ratio defined here is also known as the income gap ratio of the poor in Haritha *et al.* (2007). The poverty gap ratio can be written in terms of reversed mean residual quantile function as follows

$$PG = \frac{uR(u)}{Q(u)},$$

where $R(u) = u^{-1}\int_0^u (Q(u) - Q(p))\, dp$. For, SMD distribution the poverty gap ratio is

$$PG = u - \frac{q\beta_{1-(1-u)^{\frac{1}{q}}}\left(1+\frac{1}{a}, q-\frac{1}{a}\right) + p\beta_{u^{\frac{1}{p}}}\left(p+\frac{1}{a}, 1-\frac{1}{a}\right)}{\left((1-u)^{-\frac{1}{q}}-1\right)^{\frac{1}{a}} + \left(u^{-\frac{1}{p}}-1\right)^{-\frac{1}{a}}}. \tag{30}$$

The Foster-Greer-Thorbecke (FGT) measure proposed by Foster *et al.* (1984) generalizes the poverty gap ratio. Here, $D(z,y) = \left(\frac{z-y}{z}\right)^{\alpha}$ and the measure is

$$FGT(\alpha) = \int_0^z \left(\frac{z-y}{z}\right)^{\alpha} f(y)dy,$$

where $\alpha \geq 1$ is the inequality aversion parameter. The lower tail of the income distribution receives more emphasis as the value of α increases. When $\alpha = 1$, the FGT measure becomes equivalent to the poverty gap ratio. The quantile version of the FGT measure can be obtained by using the same transformation in the poverty gap ratio and is given in (31). Moreover, it does not have a closed form expression for the SMD distribution.

$$FGT(\alpha) = \int_0^u \left(\frac{Q(u) - Q(p)}{Q(u)}\right)^{\alpha} dp. \tag{31}$$

Watts (1968) introduced the first distribution-sensitive poverty index called Watt's index. This index satisfies the focus, monotonicity, and transfer axioms of poverty and in quantile terms, it is given as

$$W = \int_0^u \ln\left(\frac{Q(u)}{Q(p)}\right) dp. \tag{32}$$

Kakwani (1999) has proposed a measure that is closely related to the Watts index and is given by, $K^* = 1 - e^{-W}$. For SMD distribution these indices do not have simple algebraic expressions.

Sen (1976) put forward a measure that attempted to incorporate the effects of the number of poor, the severity of their poverty, and poverty distribution within the group. In quantile terms, it is

$$S = u\left(\frac{u\,\Delta\rho_1(u) + \rho_2(u)}{u\,\Delta\rho_1(u) + \rho_1(u)}\right),$$

where $\rho_1(u) = \frac{1}{u}\int_0^u Q(p)dp$, $\rho_2(u) = \frac{1}{u^2}\int_0^u (2p-u)\,Q(p)dp$, and $\Delta\,\rho_1$ denotes derivative of ρ_1 with respect to u. For SMD distribution $\rho_1(u)$ and $\rho_2(u)$ are given as

$$\rho_1(u) = \frac{b}{u}\left[q\beta_{1-(1-u)^{\frac{1}{q}}}\left(1+\frac{1}{a}, q-\frac{1}{a}\right) + p\beta_{u^{\frac{1}{p}}}\left(p+\frac{1}{a}, 1-\frac{1}{a}\right)\right],$$

$$\rho_2(u) = \frac{b}{u^2}\Bigg[(2-u)q\beta_{1-(1-u)^{\frac{1}{q}}}\left(1+\frac{1}{a}, q-\frac{1}{a}\right) - 2q\beta_{1-(1-u)^{\frac{1}{q}}}\left(1+\frac{1}{a}, 2q-\frac{1}{a}\right)$$
$$+ 2p\beta_{u^{\frac{1}{p}}}\left(2p+\frac{1}{a}, 1-\frac{1}{a}\right) - up\beta_{u^{\frac{1}{p}}}\left(p+\frac{1}{a}, 1-\frac{1}{a}\right)\Bigg].$$

The Gini index for the poor has the quantile form

$$\eta(u) = 1 - \frac{2}{\rho_1(u)} \int_0^u Q(p) \left(\frac{u-p}{u^2} \right) dp$$
$$= \frac{\rho_2(u)}{\rho_1(u)}. \tag{33}$$

Now, for SMD distribution the above index can be written as

$$\eta(u) = \frac{2}{u} \times \frac{A}{B} - 1, \tag{34}$$

where

$$A = q\beta_{1-(1-u)^{\frac{1}{q}}} \left(1 + \frac{1}{a}, q - \frac{1}{a}\right) - q\beta_{1-(1-u)^{\frac{1}{q}}} \left(1 + \frac{1}{a}, 2q - \frac{1}{a}\right) + p\beta_{u^{\frac{1}{p}}} \left(2p + \frac{1}{a}, 1 - \frac{1}{a}\right),$$
$$B = q\beta_{1-(1-u)^{\frac{1}{q}}} \left(1 + \frac{1}{a}, q - \frac{1}{a}\right) + p\beta_{u^{\frac{1}{p}}} \left(p + \frac{1}{a}, 1 - \frac{1}{a}\right).$$

6. Inference and applications

Here, we estimate the parameters of the family of distributions (5) and use real data sets to assess the model's effectiveness and applications.

The parameters of the distribution in a quantile setup can be estimated using a variety of methods. The L-moments method, the percentile approach, the minimum absolute deviation method, the least squares method, and the maximum likelihood method are frequently used techniques. We employ the method of least squares to estimate the parameters of the model (5). In order to estimate the generalized Tukey lambda distribution, Öztürk and Dale (1985) utilized this estimation method. Hankin and Lee (2006) also used this method to estimate the parameters of the Davies distribution. The least square estimation is illustrated as follows.

Let $X_{(i)}$ denote the i^{th} order statistic from a random sample of size n from SMD distribution, and $u_{(i)}$ be the i^{th} order statistic of the associated uniformly distributed random variable, $u = F(X)$. In the ideal situation, both the random variables $X_{(i)}$ and $Q(u_{(i)}, \hat{\delta})$ have the same distribution, where $\hat{\delta}$ is the estimator of the model's parameter vector. In this estimation technique, we estimate $\delta = (a, b, p, q)$ that minimizes $\zeta(\delta)$

$$\zeta(\delta) = \sum_{i=1}^{n} \left(X_{(i)} - Q(u_{(i)}, \delta) \right)^2.$$

6.1. Real data analysis

The applicability of the model (5) can be demonstrated with the aid of two real income datasets. The first data is taken from `https://www.bea.gov`, which deals with the per capita personal income of 46 counties in South Carolina State, 2018. Using midyear population estimates from the Census Bureau, per capita personal income was calculated. We use the least squares method discussed above to estimate the parameters. The estimate is based on the parameter value that minimizes the residual sum of squares and is obtained as

$$\hat{a} = 8.2443,\ \hat{b} = 10618.9,\ \hat{q} = 2.83288,\ \text{and}\ \hat{p} = 1856.39.$$

The Q-Q plot and the chi-square test are the two goodness-of-fit criteria used here to evaluate how well the model fits the data. The Q-Q plot given in Figure 8, shows that the fit is satisfactory. We conducted the chi-square goodness-of-fit test and obtained the test statistic value as 6.89418 with p-value 0.648136. Hence, the proposed model (5) fits the given dataset reasonably well. Since the quantile functions of the SM and Dagum distributions are added to obtain our model, we fitted the above data to these distributions, and the results are given in Table 1. Figure 9 illustrates the histogram of the data along with the density functions for the SM, SMD, and Dagum distributions. It is clear from the figure that the SMD distribution fits the dataset more accurately than the other two models.

Table 1: Parameter estimates, chi-square statistic, and p-value of SM and Dagum distributions for dataset 1

Distribution	Parameter estimates	Chi-square statistic	p-value
SM	a = 24.0223		
	b = 33492.8	8.66856	0.468414
	q = 0.31186		
Dagum	a = 10.477		
	b = 36651.4	7.74593	0.559939
	p = 1.27739		

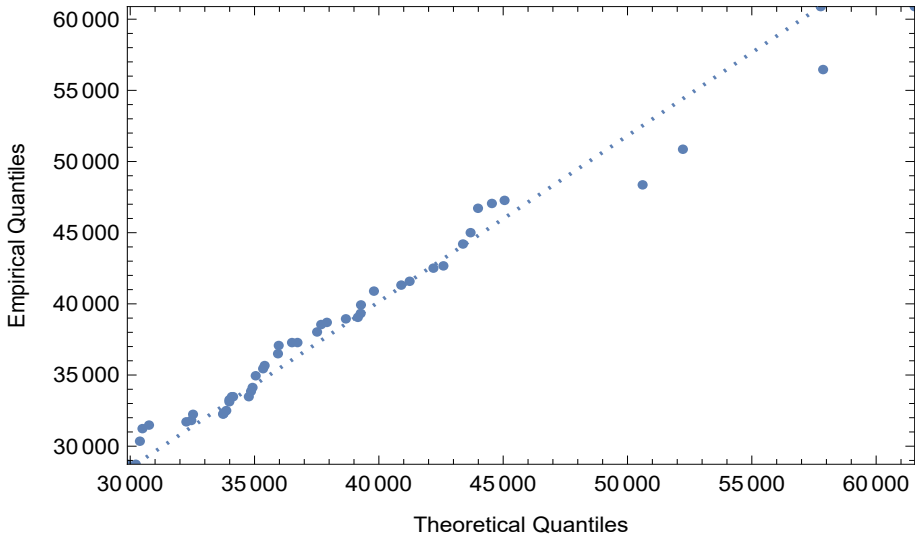


Figure 8: Q-Q plot for the per capita personal income of counties in South Carolina State in 2018

The second dataset is also taken from `https://www.bea.gov`, which deals with the per capita personal income of 120 counties in Kentucky State, 2020. The method of least squares is employed to estimate the parameters and is obtained as

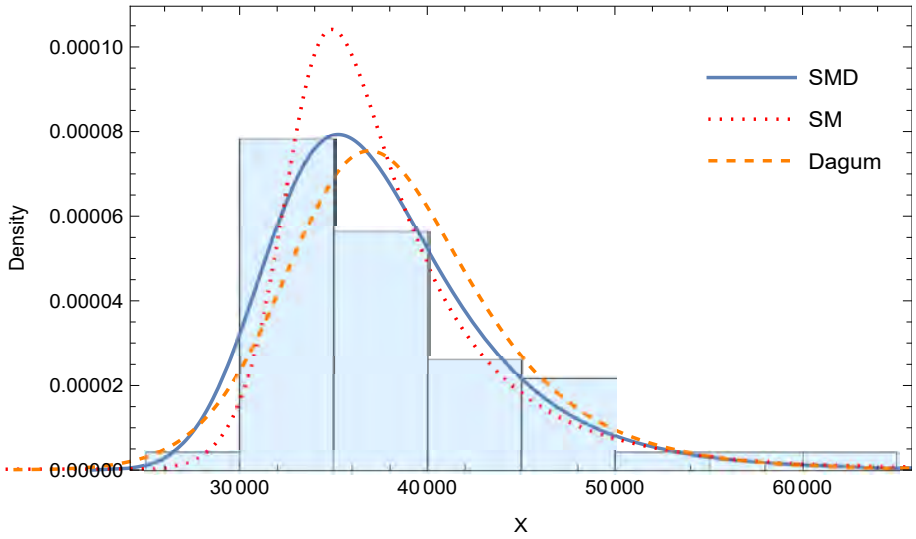


Figure 9: The densities of the SM, SMD, and Dagum distributions for the per capita personal income of counties in South Carolina State in 2018

$$\hat{a} = 9.46169,\ \hat{b} = 17513.9,\ \hat{q} = 3.20767,\ \text{and}\ \hat{p} = 27.4554.$$

Two goodness-of-fit methods are used to evaluate how well the model fits the data. The first one is the Q-Q plot in Figure 10, which shows that the suggested model is appropriate for the given data set. In addition, we perform the chi-square goodness-of-fit test and get test statistic value 4.35791 with p-value 0.986754. This indicates the fit of SMD distribution for the given data. The SM and Dagum distributions are also fitted to income data of Kentucky State, and the results are given in Table 2. The histogram of the data and the density functions for the SM, SMD, and Dagum distributions are shown in Figure 11. From the figures and the chi-square values the SMD model appears to be better than SM and Dagum distributions.

Table 2: Parameter estimates, chi-square statistic, and p-value of SM and Dagum distributions for dataset 2

Distribution	Parameter estimates	Chi-square statistic	p-value
SM	a = 16.1799		
	b = 39576.5	7.20778	0.891129
	q = 0.640004		
Dagum	a = 9.86976		
	b = 36637.5	4.49484	0.984699
	p = 2.34804		

7. Conclusion

In this article, we propose the quantile function known as SMD distribution by adding the quantile functions of the SM and Dagum distributions. Several popular distributions are

members of the proposed class of distributions. We studied the distributional properties, the major income inequality, and the poverty measures of the proposed class. We also derived the quantile version of poverty measures, such as the poverty gap ratio and the Foster-Greer-Thorbecke measure. The estimation of the parameters of the model was done using the method of least squares. The proposed class of distribution was used for the analysis of two real income data and it gives a better fit than SM and Dagum distributions.

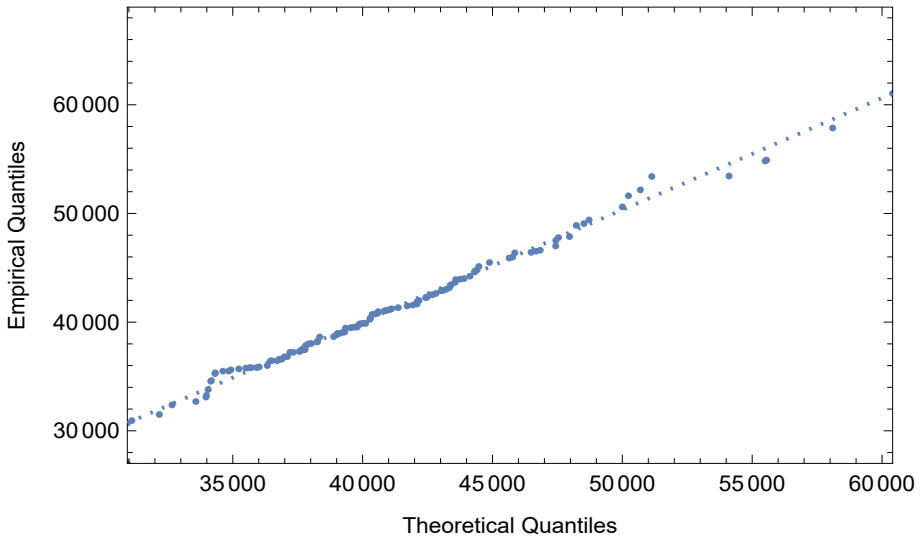


Figure 10: Q-Q plot for the per capita personal income of counties in Kentucky State in 2020

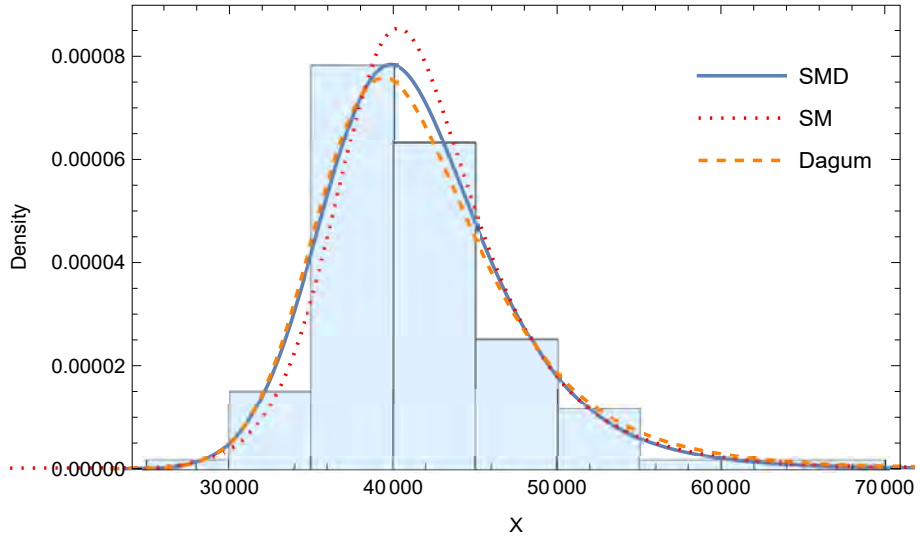


Figure 11: The densities of the SM, SMD, and Dagum distributions for the per capita personal income of counties in Kentucky State in 2020

Acknowledgements

We thank the referee and the editor for their valuable comments and suggestions. The first author is thankful to the Kerala State Council for Science, Technology and Environment

(KSCSTE) for the financial support. The authors are also thankful to the Department of Science & Technology (DST) for the technical support.

References

Bonferroni, C. E. (1930). *Elements of General Statistics*. Seeber, Florence.

Chotikapanich, D., Griffiths, W., Karunarathne, W., and Prasada Rao, D. (2013). Calculating poverty measures from the generalized beta income distribution. *Economic Record*, **89**, 48–66.

Dagum, C. (1977). A new model of personal income distribution: specification and estimation. *Économie Appliquée*, **30**, 413–437.

Éltetö, Ö. and Frigyes, E. (1968). New income inequality measures as efficient tools for causal analysis and planning. *Econometrica*, **36**, 383–396.

Foster, J., Greer, J., and Thorbecke, E. (1984). A class of decomposable poverty measures. *Econometrica*, **52**, 761–766.

Freimer, M. L., Kollia, G., Mudholkar, G. S., and Lin, C. T. (1988). A study of the generalized Tukey lambda family. *Communications in Statistics-Theory and Methods*, **17**, 3547–3567.

Galton, F. (1875). Statistics by intercomparison, with remarks on the law of frequency of error. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, **49**, 33–46.

Gastwirth, J. L. (1971). A general definition of the Lorenz curve. *Econometrica*, **39**, 1037–1039.

Gilchrist, W. (2000). *Statistical Modeling with Quantile Functions (1st ed.)*. Chapman and Hall/CRC, New York.

Gini, C. (1914). On the measurement of the concentration and variability of characters. *Proceedings of the Royal Veneto Institute of Sciences, Letters and Arts*, **73**, 1203–1248.

Greenwood, J. A., Landwehr, J. M., Matalas, N., and Wallis, J. (1979). Probability weighted moments: Definition and relation to parameters of several distributions expressable in inverse form. *Water Resources Research*, **15**, 1049–1054.

Hankin, R. K. S. and Lee, A. (2006). A new family of non-negative distributions. *Australian & New Zealand Journal of Statistics*, **48**, 67–78.

Haritha, N. H., Nair, K. R. M., and Nair, N. U. (2007). *Income modeling using quantile functions*. Doctoral dissertation, Cochin University of Science and Technology, Cochin.

Hastings, J. C., Mosteller, F., Tukey, J. W., and Winsor, C. P. (1947). Low moments for small samples: A comparative study of order statistics. *The Annals of Mathematical Statistics*, **18**, 413–426.

Hosking, J. R. M. (1990). L-moments: Analysis and estimation of distributions using linear combinations of order statistics. *Journal of the Royal Statistical Society. Series B (Methodological)*, **52**, 105–124.

Kakwani, N. (1980). *Income Inequality and Poverty: Methods of Estimation and Policy Applications*. Oxford University Press, Oxford.

Kakwani, N. (1999). *Inequality, Welfare And Poverty: Three Interrelated Phenomena*, pages 599–634. Springer Netherlands, Dordrecht.

Kleiber, C. and Kotz, S. (2003). *Statistical Size Distributions in Economics and Actuarial Sciences*. John Wiley Sons, Ltd, Hoboken, New Jersey.

Klugman, S. A., Panjer, H. H., and Willmot, G. E. (2019). *Loss Models From Data to Decisions*. John Wiley, New York, 5^{th} edition.

Kumar, D. (2017). The Singh–Maddala distribution: Properties and estimation. *International Journal of System Assurance Engineering and Management*, **8 (Suppl. 2)**, 1297–1311.

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

Morduch, J. (2008). Poverty measures. In *United Nations Handbook of Poverty Statistics*, chapter 3, pages 52–83. United Nations, New York.

Nair, N. U., Sankaran, P. G., and Balakrishnan, N. (2013). Quantile Function Models. In *Quantile-Based Reliability Analysis*, chapter 3, pages 59–103. Springer, New York.

Öztürk, A. and Dale, R. F. (1985). Least squares estimation of the parameters of the generalized lambda distribution. *Technometrics*, **27**, 81–84.

Parzen, E. (1979). Nonparametric statistical data modeling. *Journal of the American Statistical Association*, **74**, 105–121.

Pietra, G. (1932). New contributions to the methodology of variability and concentration indices. *Proceedings of the Royal Veneto Institute of Sciences, Letters and Arts*, **91**, 989–1008.

Ramberg, J. S. and Schmeiser, B. W. (1972). An approximate method for generating symmetric random variables. *Communications of the ACM*, **15**, 987–990.

Sankaran, P. G. and Dileep Kumar, M. (2018). Pareto-Weibull quantile function. *Journal of Applied Probability and Statistics*, **13**, 81–95.

Sankaran, P. G., Nair, N. U., and Midhu, N. N. (2016). A new quantile function with applications to reliability analysis. *Communications in Statistics - Simulation and Computation*, **45**, 566–582.

Saulo, H., Vila, R., Borges, G. V., Bourguignon, M., Leiva, V., and Marchant, C. (2023). Modeling income data via new parametric quantile regressions: Formulation, computational statistics, and application. *Mathematics*, **11**.

Sen, A. (1976). Poverty: An ordinal approach to measurement. *Econometrica*, **44**, 219–231.

Shahzad, M. and Asghar, Z. (2013a). Comparing TL-moments, L-moments and conventional moments of Dagum distribution by simulated data. *Revista Colombiana de Estadística*, **36**, 79–93.

Shahzad, M. and Asghar, Z. (2013b). Parameter estimation of Singh-Maddala distribution by moments. *International Journal of Advanced Statistics and Probability*, **1**, 121–131.

Sillitto, G. P. (1969). Derivation of approximants to the inverse distribution function of a continuous univariate population from the order statistics of a sample. *Biometrika*, **56**, 641–650.

Singh, S. K. and Maddala, G. S. (1975). A stochastic process for income distribution and tests for income distribution functions. In *ASA Proceedings of the Business and Economic Statistics Section*, pages 551–553.

Sreelakshmi, N. and Nair, K. R. M. (2014). *A quantile based analysis of income data*. Doctoral dissertation, Cochin University of Science and Technology, Cochin.

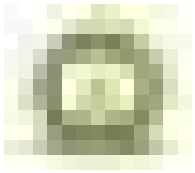
Tarsitano, A. (2004). Fitting the generalized lambda distribution to income data. In *COMPSTAT 2004 Symposium*, pages 1861–1867, New York. Springer.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Addsion Wesley, Massachusetts.

Watts, H. W. (1968). *An Economic Definition of Poverty*. Basic Books, New York.

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

Statistics and Applications {ISSN 2454-7395 (online)}
Volume 22, No. 1, 2024 (New Series), pp 343–374
http://www.ssca.org.in/journal



Bayesian Integration for Small Areas by Supplementing a Probability Sample with a Non-probability Sample

Balgobin Nandram[1] and J. N. K. Rao[2]

[1]*Department of Mathematical Sciences*
Worcester Polytechnic Institute, Worcester, MA 01609, United States of America
[2]*School of Mathematics and Statistics, Carleton University, Ottawa, K1S 5B6, Canada*

(This is a special invited paper on request from the Chair Editor.)

Received: 26 June 2023; Revised: 5 January 2024; Accepted: 8 January 2024

Abstract

We consider the problem of data integration in small area estimation, where a non-probability sample (nps) and a relatively small probability sample (ps) are available from each area. By definition, for the nps, there are no survey weights, but for the ps, there are survey weights. A recent method, based on a pseudo-likelihood, is used to estimate the survey weights in the nps, and thereafter assumed known. The key issue we address is that the nps, although much larger than the ps, can lead to a biased estimator of a finite population quantity of each area but with much smaller variance. We assume that there are common covariates and responses for everyone in the two samples, no covariates available for nonsampled units, and no overlaps of the two samples by area. In the data integration, we use the nps to construct a prior for the ps, and partial discounting of the nps is incorporated to avoid a dominance of the prior. Inverse probability weighting is used to assist Bayesian predictive inference via surrogate sampling of the finite population means and percentiles. The Gibbs sampler, with some collapsing to speed up convergence and to provide strong mixing, is carefully executed to fit the joint posterior density. In our illustrative example on body mass index, our data-integrated model is preferred over the ps only model and other competitors. The data-integrated model provides small area estimates, roughly similar to those of the ps only model, with larger precision.

Key words: Bayesian diagnostics; Finite population quantities; Gibbs sampler; Inverse probability weighting; Power prior; Surrogate samples.

1. Introduction

We assume that there are data from a number of small areas, and from each area we have a non-probability sample (nps,1) and a probability sample (ps, 2), the probability sample being much smaller than the non-probability sample. The problem is how to improve

Corresponding Author: Balgobin Nandram
Email: balnan@wpi.edu & balgobin.nandram@usda.gov

inference for each area based on the ps, but supplemented by the nps, and we do not want the nps to dominate the analysis. While the nps may be biased, the ps is considered unbiased when the survey weights are incorporated. In a similar manner, because of its size the nps will provide improved precision but it will provide biased estimates, which we do not want to happen. Probability sampling is the gold standard among all data collection procedures, but this is still problematic because nonresponse has become a serious concern. How can we provide small area estimates with relatively small bias, possibly closed to the ps, with better precision than only the ps can provide?

There are efforts to combine both probability and nonprobability samples to produce a single inference that compensates for the limitations of each process. Typically the nonprobability sample is relatively large, as in big data, but one needs to be careful with the bias it introduces into the final estimates. Meng (2018) argued that a small bias in big data can be catastrophic; see also Nandram and Rao (2021, 2023) for a review and an interpretation of Meng (2018) relevant to nonprobability sample. Perhaps if one uses only the covariates in the big data, there may not be significant bias, but it is a different issue if one wants to use the study variable from the big data as well.

Most of the work on nonprobability sampling has been in the non-Bayesian setting, mostly randomization-based analysis. For example, Elliott and Haviland (2007) evaluated a composite estimator to supplement a standard probability sample with a nonprobability sample. They showed that the estimator, based on a linear combination of both sample processes and a bias function, can produce estimates with a smaller mean squared error (MSE) relative to a probability-only sample. See Elliott and Valliant (2017) for an informative review of the design-based approach, where they discussed quasi-randomization.

Sakshaug *et al.* Blom (2019) and Wisnioski *et al.* (2020) introduced a Bayesian approach in which survey weights are incorporated as a covariate and there is no need to estimate the probabilities of the nps. The underlying idea is that probability and nonprobability samples can be integrated in a way that exploits their advantages to compensate for their weaknesses and improve estimation of model parameters. Salvatore *et al.* (2023) used a similar idea for binary data via logistic regression. Nandram and Rao (2021, 2023) showed how to combine a nps and a ps using a Bayesian model. They argued that the nps should be used to construct a prior, together with a discounting factor, and to obtain a prior for the hyper-parameters in the model, which begins with a weighted likelihood. As pointed out by both Sakshaug *et al.* (2019) and Wisnioski *et al.* (2020), it will be better to use a nonprobability sample to supplement a probability sample; see also Nandram and Rao (2021, 2023). Salvatore *et al.* (2023) also supported our idea.

Chen, Li and Wu (2020) used a ps and a nps to obtain survey weights in the nps in the design-based approach, where they made strong use of the Horvitz-Thompson and the Hajek estimators. There was no study variable in the ps and so this is really a very limited data integration problem. Actually their method cannot be extended to accommodate a study variable in the ps. Also, their method uses logistic regression to construct the propensity scores and then the survey (design) weights are obtained by taking reciprocals. This is ignorable selection. A summary of the approach of Chen, Li and Wu (2020) for propensity scores is given in Appendix A. For small area estimation, the computational procedure of Chen, Li and Wu (2020) is unstable, so we had to do this procedure for the entire ensemble

at once, not each area at a time.

As pointed out by a reviewer, we should present the reasons why we do not use non-ignorable selection. We used ignorable selection because within the framework of Chen, Li and Wu (2020), it is not possible to obtain the propensity scores unless you want to use a mass imputation to 'manufacture' the values of the study variable for the probability sample. This is not in the spirit of our paper because we have the study variable for both the non-probability sample and the probability sample; yet we use the method of Chen, Li and Wu (2020) to get the propensity scores. We realized that we could have used both samples to get the propensity scores in the non-probability samples, but it does not fit directly into the framework of Chen, Li and Wu (2020). Non-ignorable selection is defined as

$$f(I = k, y \mid \underline{x}) = P(I = k \mid y, \underline{x}) f(y \mid \underline{x}), k = 0, 1,$$

where I is the participation variable, y the study variable and $\underline{x}$ the covariates. We have ignorable selection if $P(I \mid y, \underline{x}) = P(I \mid \underline{x})$, which is simpler than non-ignorable selection. Clearly, non-ignorable selection is preferred but it leads to computational instability. See Nandram and Choi (2010) and Nandram (2022) for more discussions on non-ignorability. One difficulty is that one needs y to be strongly related to $\underline{x}$ and at the same time, both $\underline{x}$ and y are used as covariates in the participation model. New research is needed at least within the Bayesian paradigm; see Marella (2023) for recent work on nonignorability, not within the Bayesian paradigm though. Data integration can be discussed without mentioning how the data are selected; see Salvatore *et al.* (2023) for binary data and others.

It is worth noting that all the above mentioned work do not consider data integration for small areas. Beaumont (2020) argued that it is sensible to use a non-probability sample to supplement a probability sample in small area estimation; see also Beaumont and Rao (2021). For one thing, small sample sizes within small areas do not lead to adequate precision. The small area model will include random effects as an attempt to discriminate the areas. These works use the area-level Fay-Herriot model. However, there is virtually no work using the unit level model like that of Battese, Harter and Fuller (1988) for integration of a non-probability sample and a probability sample partly because it is a less practical to get unit-level data in both the nps and the ps, but it is possible. Again see Nandram and Rao (2021, 2023).

Rao (2020) stated that a non-probability sample can be used to construct covariates for probability samples in small area estimation. The use of area level big data as additional predictors in the area level model has the potential of providing good predictors for modeling. He mentioned four applications that have used big data covariates in an area level model; see Marchetti *et al.* (2015), Porter *et al.* (2014), Schmid *et al.* (2017) and Muhyi *et al.* (2019) for the four applications. Rao (2020) also cited applications where unit level models are used; see Chambers *et al.* (2019). Again, if one wants to use both the study variable and the covariates from the big data, one might need the unknown selection probabilities. However, one does not really need to estimate the selection probabilities because one can use structural (measurement error) models; see discussions in the concluding remarks and Berg *et al.* (2021). One drawback of structural models is that there will be non-identifiable parameters which will create difficulties in model fitting, especially if Markov chain Monte Carlo methods must be used.

In our paper, we actually used a power prior to discount the non-probability sample,

which we treat as historical data to construct a prior distribution for the parameters of the probability sample. The parameters in the two models are basically the same, and their priors come from the non-probability sample. In general, if you start with a density, $g(y \mid \underline{\theta})$, we can penalize it by using

$$f(y \mid \underline{\theta}, a) = \frac{\{g(y \mid \underline{\theta})\}^a}{\int \{g(y \mid \underline{\theta})\}^a dy}, 0 \leq a \leq 1.$$

So we actually use $f(y \mid \underline{\theta}, a)$ for the non-probability sample and $g(y \mid \underline{\theta})$ for the probability sample. For example, if $a = 1$, there will be no discounting, and if $a = 0$, the non-probability sample will not be used. Details of the power prior in data integration is reviewed in Nandram and Rao (2021, 2022); see Ibrahim and Chen (2000) and Ibrahim *et al.* (2015) for a review and many applications of the power prior in more general settings.

Small area estimation (SAE) is an important problem facing many government agencies. They want to do estimation for each area, but for most small areas the direct estimates are unreliable. Then, pooling of the data over the entire ensemble is required to get reliable estimates for each area. While the SAE problem is difficult in its own right, there is additional complexity to integrate the non-probability sample and the probability sample.

To focus our development, we study body mass index (BMI) as the variable of interest with covariates, age, race and sex, from eight counties in California, based on a probability sample. The covariates, responses (BMI) and survey weights are all known. We construct a small-area example out of these data with two samples from each of the eight counties (about 80% for nps and 20% for the ps). To form a practical example, we discarded the weights from the nps and they are assumed unknown. The population size of each county is roughly the sum of the survey weights in the ps. Here, the covariates, responses and survey weights in the nps are respectively $(\underline{x}_{1ij}, y_{1ij}, w_{1ij}), i = 1, \ldots, \ell, j = 1, \ldots, n_{1i}$, and the covariates, responses and survey weights of the ps are $(\underline{x}_{2ij}, y_{2ij}, w_{2ij}), i = 1, \ldots, \ell, j = 1, \ldots, n_{2i}$; the survey weights w_{1ij} are unknown in the nps.

The small area model has the following features.

a. The two sets of covariates are commensurate (*i.e.*, the same covariates are measured in the non-probability sample and the probability sample; or at least only a common set of covariates will be used).

b. Pooling will take place using a common set of regression coefficients and variance components over all areas in the two samples. The nps is essentially used to construct a prior for the hyperparameters and this prior is discounted using a power prior.

c. Within an area, the random effects are the same in the model that links the non-probability sample and the probability sample.

d. It is possible to have some areas with only a probability sample, and some areas with only a non-probability sample, but there must be a common set. This can be done within our approach, but we will not pursue this issue further in this paper.

Finally, a reviewer asked why there is a need for super-population models. Clearly, it will be better to do data integration without specifying the super-population model. Most

Bayesian methods used the super-population model; Wang *et al.* (2018) is an exception and it uses an approximate Bayesian method. In fact, they used the sampling distribution of a summary statistic to derive the posterior distribution of the parameters of interest, but this is not quite within the Bayesian paradigm. However, there is a need to robustify both models for the study variable and the participation variable. We have indicated how to do so in the concluding remarks, and this is an on-going activity. In our on-going research work, by using a double mass imputation (Kim, *et al.* 2021, Chen, *et al.* 2022), we can avoid a participation model but we do need a robust model for the study variable if we use a Bayesian method. One of the authors gave a couple of talks on this topic already.

This paper has five sections, including this one, and it is an extension of Nandram and Rao (2021, 2023) to cover Bayesian data integration for small areas. In Section 2, we review the single area model of Nandram and Rao (2021, 2023). In Section 3, we discuss small area estimation using a unit-level model, show how to operationalize the proposed model to provide fast computation for a large number of areas, and describe how to estimate finite population percentiles. In Section 4, we provide the analysis of the numerical example as we described above. Section 5 provides some concluding remarks and extensions. The appendices provide technical details on propensity scores, computation for the small area model, Bayesian model diagnostics, and the ps only model.

2. Review of the single area model

In this paper, we extend the single area model of Nandram and Rao (2021, 2023) to accommodate data integration for small areas. Therefore, it is pertinent for us to describe the single area model to motivate the small area model.

We have two samples from a single area, which are the nps (1) and the ps (2). We have $(W_{ti}, \underline{x}_{ti}, y_{ti}), i = 1, \ldots, n_t, t = 1, 2$, where W_{1i} are unknown, but W_{2i} are assumed known. We plan to construct a prior for the regression coefficients and the variance parameters using a discount factor (power prior) to help mitigate the nps from dominating the ps. (Throughout, as covariates are assumed fixed, conditioning on them will be omitted.)

For the nps, propensity scores, assumed strictly positive, are estimated using logistic regression (Chen, Li and Wu, 2020; see Appendix A of the current paper for a review), so for the nps probability enters through quasi-randomization (*e.g.*, Elliott and Valliant, 2017). The method of CLW is used to estimate the propensity scores, π_{1i}, and the weights of the nps are $W_{1i} = N\frac{1/\pi_i}{\sum_{j=1}^{n_1} 1/\pi_j}, i = 1, \ldots, n_1$, where N is the population size, and the Horvitz-Thompson estimator of N is $\sum_{i=1}^{n_2} W_{2i}$. This assumes ignorability in which given the covariates, the study variable and the participation variable are independent and it also assumes that the propensity scores depend only on the covariates, which is not unreasonable; see Nandram (2022) for a discussion about nonignorability. These estimated weights, W_{1i}, are assumed known throughout our work. In our models, associated with weighted likelihood, we use normalized densities with adjusted weights to get a more appropriate measure of variability. The adjusted weights are

$$w_{ti} = \hat{n}_t W_{ti} / \sum_{j=1}^{n_t} W_{tj}, \hat{n}_t = \left(\sum_{j=1}^{n_t} W_{tj}\right)^2 / \sum_{j=1}^{n_t} W_{tj}^2, i = 1, \ldots, n_t, t = 1, 2,$$

where $\hat{n}_t$ is the effective sample size; see Potthof *et al.* (1992).

The population model, which we assume holds, is

$$y_i \mid \underline{\beta}, \sigma^2 \overset{ind}{\sim} \text{Normal}(\underline{x}_i'\underline{\beta}, \sigma^2), i = 1, \ldots, N.$$

A finite population quantity (mean or percentile) can be estimated using surrogate sampling (Nandram 2007). That is, the entire population is sampled given $(\underline{\beta}, \sigma^2)$. However, the question is how to get samples of $(\underline{\beta}, \sigma^2)$, and this is where most of the work is needed. We need to adjust the population model to accommodate the two samples, in which the nps is penalized using a power prior; see Nandram and Rao (2021, 2023) for a quick review of the power prior and how it is used in our work.

The model that combines the two samples, in which the nps is used to supplement the ps is

$$y_{ti} \mid \underline{\beta}, \sigma^2 \overset{ind}{\sim} \text{Normal}\left(\underline{x}_i'\underline{\beta}, \frac{\sigma^2}{a_t w_{ti}}\right),$$

$$\pi(\underline{\beta}, \sigma^2, a) \propto 1/\sigma^2, a_2 = 1, 0 < a_1 = a < 1, i = 1, \ldots, n_t, t = 1, 2,$$

where a is the discounting factor with a uniform prior and w_{ti} are adjusted weights. The joint posterior density of $(\underline{\beta}, \sigma^2, a)$ has been shown to be proper and it can be fit using a grid sample (the posterior density of a is non-standard); see Nandram and Rao (2021, 2023) for details.

Nandram and Rao (2021, 2023) obtained Bayesian predictive inference for the finite population mean using

$$\pi(\bar{Y} \mid \underline{y}_1, \underline{y}_2) = \int f(\bar{Y} \mid \underline{\beta}, \sigma^2)\pi(\underline{\beta}, \sigma^2 \mid \underline{y}_1, \underline{y}_2)d\underline{\beta}d\sigma^2,$$

where $\underline{y}_1$ and $\underline{y}_2$ are the two samples. Note that $f(\bar{Y} \mid \underline{\beta}, \sigma^2)$ does not depend on $(\underline{y}_1, \underline{y}_2)$, unlike standard Bayesian predictive inference, a feature of surrogate sampling; see Nandram (2007). Note that

$$\bar{Y} \mid \underline{\beta}, \sigma^2 \sim \text{Normal}\left(\underline{\bar{X}}'\underline{\beta}, \frac{\sigma^2}{N}\right),$$

where we use the Horvitz-Thompson estimator of the finite population mean vector covariate, $\underline{\bar{X}}$, which is $\frac{1}{N}\sum_{i=1}^{n_2} W_{2i}\underline{x}_{2i}$; this is actually the Hajek estimator because N is assumed unknown.

Inference about a finite population percentile is a related, but different, problem. This is discussed in Section 3. Inference about the finite population percentiles is also a problem in our study on body mass index (*e.g.*, the 85^{th} percentile is a measure of overweight).

3. A small area model for data integration

We show how to extend the model of Nandram and Rao (2021) to accommodate a number of areas. This uses an extended version of the unit-level model of Battese, Harter and Fuller (BHF, 1988). See also Toto and Nandram (2011) and Molina, Nandram and Rao (2014) for the Bayesian version of the BHF model.

We assume there are ℓ areas and within the i^{th} area, there are a non-probability sample of size n_{1i} and a probability sample of size n_{2i} where "1" and "2" respectively refer to the non-probability sample and the probability sample, maintaining the notation in the single area example, and the population size is N_i. [Note that the nps and the ps of each area come from the same distinct sub-population; so there is single subscript in N_i.] For $i = 1, \ldots, \ell$, the covariates are $(\underline{x}_{sij}, j = 1, \ldots, n_{sij}, s = 1, 2)$, but the covariates are unobserved for the nonsampled units, and the responses are $y_{sij}, j = 1, \ldots, n_{si}$. There are also survey weights for the probability sample, denoted by W_{2i} (known). There are no survey weights for the non-probability sample and these are estimated using the method of Chen, Li and Wu (2020); again see Appendix A. The population size for the i^{th} area is estimated by $N_i = \sum_{j=1}^{n_{2i}} W_{2ij}, i = 1, \ldots, \ell$. Bayesian predictive inference is required for the finite population area means,

$$\bar{Y}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} y_{ij}, i = 1, \ldots, \ell,$$

based on the non-probability samples and probability samples, where y_{ij} are the unknown population values. Of course, the model permits the use of the non-probability sample, as we have seen for the single sample model. That is, there is pooling across areas and within areas from both the non-probability sample and the probability sample.

As we have stated, the discounting factors will only be included for the nps, which will be used to construct the prior (the nps is viewed as historical data) and the ps will be used as the actual data. For generality, these discounting factors depend on areas. That is, for $s = 1$ (nps), $a_{si} = a_i, i = 1, \ldots, \ell$ (allowing discounting) and for $s = 2$ (ps), $a_{si} = 1, i = 1, \ldots, \ell$ (no discounting).

3.1. Proposed small area model

Our model for the two samples over the areas is

$$y_{sij} \mid \nu_i, \underline{\beta} \overset{ind}{\sim} \text{Normal}(\underline{x}_{sij}\underline{\beta} + \nu_i, \frac{\sigma^2}{a_{si} w_{sij}}), j = 1, \ldots, n_{si}, s = 1, 2,$$

where w_{sij} are the adjusted weights within areas. The weights for the nps are obtained using the method of Chen, Li and Wu (2020) over the entire ensemble (assumed known henceforth), and then the weights for both the nps and ps are used to provide the adjusted weights, as was done in the single area example. The fact that we are assuming the estimated weights are known is an important caveat of our work, and this is on-going research activity. A priori, for the random effects, we assume that

$$\nu_i \mid \rho, \sigma^2 \overset{ind}{\sim} \text{Normal}(0, \frac{\rho}{1-\rho}\sigma^2), i = 1, \ldots, \ell,$$

and for the hyperparameters, we assume

$$\pi(\underline{\beta}, \sigma^2, \rho) \propto \frac{1}{\sigma^2}, 0 < \rho < 1.$$

Again note that these are two BHF models, one for the non-probability samples and the other for the probability samples. But they are connected because they have the same parameters

(except the nps has the discounting factors), and this is how we link the nps, ps and the small areas.

For the discounting factors $0 \le a_i \le 1$, we will assume that for $i = 1, \ldots, \ell$,

$$a_i \mid \theta, \gamma \overset{ind}{\sim} \text{Beta}\left\{\theta\frac{1-\gamma}{\gamma}, (1-\theta)\frac{1-\gamma}{\gamma}\right\}, 0 < \theta, \gamma < 1.$$

We need to specify the priors for θ and γ. We make a modest assumption that the distribution of each a_i is log-concave, and a sufficient condition for this to happen is that $\theta\frac{1-\gamma}{\gamma} > 1$ and $(1-\theta)\frac{1-\gamma}{\gamma} > 1$. (A log-concave density has very nice properties, specifically its moment generating function exists.) This means that $0 < \gamma < \frac{1}{3}, \frac{\gamma}{1-\gamma} < \theta < \frac{1-2\gamma}{1-\gamma}$. Therefore, the prior for $(\underline{a}, \theta, \gamma, \rho)$ is

$$\pi(\underline{a}, \theta, \gamma, \rho) = \left\{\prod_{i=1}^{\ell} \frac{a_i^{\theta\frac{1-\gamma}{\gamma}-1}(1-a_i)^{(1-\theta)\frac{1-\gamma}{\gamma}-1}}{B\{\theta\frac{1-\gamma}{\gamma}, (1-\theta)\frac{1-\gamma}{\gamma}\}}\right\}, 0 < \gamma < \frac{1}{3}, \frac{\gamma}{1-\gamma} < \theta < \frac{1-2\gamma}{1-\gamma}, 0 < \rho < 1.$$

Note that this model holds for the entire population with $w_{sij} \equiv 1$.

Using Bayes' theorem, letting $\underline{y}$ (both nps and ps) denote the vector of all observations, the joint posterior density is

$$\pi(\underline{\nu}, \underline{a}, \underline{\beta}, \sigma^2, \rho, \theta, \gamma \mid \underline{y}) \propto$$

$$\frac{1}{\sigma^2}\prod_{i=1}^{\ell}\left\{\left[\prod_{j=1}^{n_{1i}}\sqrt{\frac{a_i w_{1ij}}{2\pi\sigma^2}}e^{-\frac{a_i w_{1ij}}{2\sigma^2}(y_{1ij}-\underline{x}'_{1ij}\underline{\beta}-\nu_i)^2}\prod_{j=1}^{n_{2i}}\sqrt{\frac{w_{2ij}}{2\pi\sigma^2}}e^{-\frac{w_{2ij}}{2\sigma^2}(y_{2ij}-\underline{x}'_{2ij}\underline{\beta}-\nu_i)^2}\right]\right.$$
$$\left.\times\sqrt{\frac{1-\rho}{2\pi\rho\sigma^2}}e^{-\frac{1-\rho}{2\rho\sigma^2}\nu_i^2}\ \frac{a_i^{\theta\frac{1-\gamma}{\gamma}-1}(1-a_i)^{(1-\theta)\frac{1-\gamma}{\gamma}-1}}{B\{\theta\frac{1-\gamma}{\gamma}, (1-\theta)\frac{1-\gamma}{\gamma}\}}\right\}. \tag{1}$$

Letting $\Omega_1 = (\underline{a}, \theta, \gamma, \rho)$ and $\Omega_2 = (\underline{\nu}, \underline{\beta}, \sigma^2)$, to fit the posterior density in (1), we will first integrate out Ω_2 and sample the posterior density of $\Omega_1 \mid \underline{y}$ using the Gibbs sampler; see Appendix B. Then, we can sample $\Omega_2 \mid \Omega_1, \underline{y}$ using the composition method via

$$\pi(\Omega_2 \mid \Omega_1, \underline{y}) = \pi_1(\sigma^2 \mid \Omega_1, \underline{y})\pi_2(\underline{\beta} \mid \sigma^2, \Omega_1, \underline{y})\pi_3(\underline{\nu} \mid \underline{\beta}, \sigma^2, \Omega_1, \underline{y}),$$

where $\pi_1(\sigma^2 \mid \Omega_1, \underline{y})$, $\pi_2(\underline{\beta} \mid \sigma^2, \Omega_1, \underline{y})$ and $\pi_3(\underline{\nu} \mid \underline{\beta}, \sigma^2, \Omega_1, \underline{y})$ are all in standard forms, inverse gamma, p-variate normal and independent normals respectively; see Appendix B. This strategy provides a more efficient computational algorithm (better convergence and mixing of the Gibbs sampler).

Bayesian predictive inference is required for $\bar{Y}_i = \frac{1}{N_i}\sum_{i=1}^{N_i} y_{ij}$, where y_{ij} are the population values (unknown). As the sample values, y_{sij}, are corrupted because of the survey weights, we cannot use them. So we use surrogate sampling; in principle the entire population is drawn, not the values for the individual units though. Therefore,

$$\bar{Y}_i \mid \nu_i, \underline{\beta}, \sigma^2 \overset{ind}{\sim} \text{Normal}\left(\bar{\underline{X}}'_i\underline{\beta} + \nu_i, \frac{\sigma^2}{N_i}\right), i = 1, \ldots, \ell,$$

where $\underline{\bar{X}}_i = \frac{1}{N_i}\sum_{i=1}^{N_i} \underline{x}_{2ij}$ and N_i are assumed unknown. We use the Horvitz-Thompson estimators $\underline{\bar{x}}_{2i} = \frac{\sum_{j\in S_{2i}} w_{2ij}\underline{x}_{2ij}}{\sum_{j\in S_{2i}} w_{2ij}}$ and $\sum_{j\in S_{2i}} w_{2ij}$ to estimate $\underline{\bar{X}}_{2i}$ and N_i respectively (inverse probability weighted estimators - IPW), where S_{2i} is the set of units in the i^{th} area of the ps. Then,

$$\bar{Y}_i \mid \nu_i, \underline{\beta}, \sigma^2 \overset{ind}{\sim} \text{Normal}\left(\underline{\bar{x}}_{2i}'\underline{\beta} + \nu_i, \frac{\sigma^2}{\sum_{j\in S_{2i}} w_{2ij}}\right), i = 1, \ldots, \ell. \tag{2}$$

Once we have drawn $(\underline{\nu}, \underline{\beta}, \sigma^2)$ using the Gibbs sampler, we simply draw the Y_i from (2). According to the model, all the sampled data are used in the predictive inference.

Observe that $\mathrm{E}(\bar{Y}_i \mid \nu_i, \underline{\beta}, \sigma^2, \rho) = \underline{\bar{x}}_{2i}'\underline{\beta} + \lambda_i(\bar{\bar{y}}_i - \underline{\bar{\bar{x}}}_i{}'\underline{\beta})$, where

$$\lambda_i = \frac{\rho\sum_{s=1}^{2}\sum_{j=1}^{n_{si}} a_{si}w_{sij}}{\rho\sum_{s=1}^{2}\sum_{j=1}^{n_{si}} a_{si}w_{sij} + (1-\rho)}, \phi_{sij} = \frac{a_{si}w_{sij}}{\sum_{s=1}^{2}\sum_{j=1}^{n_{si}} a_{si}w_{sij}},$$

$$\bar{\bar{y}}_i = \sum_{s=1}^{2}\sum_{j=1}^{n_{si}} \phi_{sij}y_{sij}, \quad \underline{\bar{\bar{x}}}_i = \sum_{s=1}^{2}\sum_{j=1}^{n_{si}} \phi_{sij}\underline{x}_{sij};$$

see Appendix B for definitions. Then,

$$\mathrm{E}(\bar{Y}_i \mid \underline{\beta}, \sigma^2, \rho, \underline{y}) = \lambda_i\bar{\bar{y}}_i + (1-\lambda_i)\underline{\bar{\bar{x}}}_i{}'\underline{\beta} + (\underline{\bar{x}}_{2i} - \underline{\bar{\bar{x}}}_i)'\underline{\beta}$$

and

$$\mathrm{Var}(\bar{Y}_i \mid \underline{\beta}, \sigma^2, \rho, \underline{y}) = \left\{\frac{1}{\sum_{j=1}^{n_{2i}} w_{2ij}} + \frac{\rho}{\rho\sum_{s=1}^{2}\sum_{j=1}^{n_{si}} a_{si}w_{sij} + (1-\rho)}\right\}\sigma^2.$$

These can be used to form Rao-Blackwellized density estimators for $\bar{Y}_i$.

More importantly, we can study the behavior of $\mathrm{E}(\bar{Y}_i \mid \underline{\beta}, \sigma^2, \rho, \underline{y})$ and $\mathrm{Var}(\bar{Y}_i \mid \underline{\beta}, \sigma^2, \rho, \underline{y})$ to see the importance of ρ. As $\rho \to 0$, $\lambda_i \to 0$,

$$\mathrm{E}(\bar{Y}_i \mid \underline{\beta}, \sigma^2, \rho, \underline{y}) \to \underline{x}_{2i}'\underline{\beta}$$

and

$$\mathrm{Var}(\bar{Y}_i \mid \underline{\beta}, \sigma^2, \rho, \underline{y}) \to \frac{\sigma^2}{\sum_{j=1}^{n_{2i}} w_{2ij}}.$$

That is, the non-probability sample does not play a major role. As $\rho \to 1$, $\lambda_i \to 1$,

$$\mathrm{E}(\bar{Y}_i \mid \underline{\beta}, \sigma^2, \rho, \underline{y}) \to \underline{x}_{2i}'\underline{\beta} + (\bar{\bar{y}}_i - \underline{\bar{\bar{x}}}'\underline{\beta})$$

and

$$\mathrm{Var}(\bar{Y}_i \mid \underline{\beta}, \sigma^2, \rho, \underline{y}) \to \left\{\frac{1}{a\sum_{j=1}^{n_{1i}} w_{1ij} + \sum_{j=1}^{n_{2i}} w_{2ij}} + \frac{1}{\sum_{j=1}^{n_{2i}} w_{2ij}}\right\}\sigma^2.$$

Both samples are important.

3.2. Operationalizing the small area model

Apart from the exchangeable assumption on the a_i, the current small area model is essentially robust with respect to the a_i. But with a large number of areas, it will be too slow to sample all the a_i using the grid method. One possibility is to smooth out the a_i in an attempt to operationalize the algorithm.

We can assume that the a_i are "proportional" to the sample sizes or better yet to their logarithms. This will also eliminate the exchangeability assumption. Therefore, one possibility is to take

$$a_i = \frac{e^{\gamma_0+\gamma_1 \log(n_i)}}{1+e^{\gamma_0+\gamma_1 \log(n_i)}}, i = 1, \ldots, \ell,$$

where for the i^{th} area, n_i is the sample size of the nonprobability sample or the total sample size. We are assuming here that $-\infty < \gamma_0 < \infty, 0 < \gamma_1 < \infty$.

Then, clearly

$$a_i = \frac{\alpha_0 n_i^{\gamma_1}}{1+\alpha_0 n_i^{\gamma_1}}, \alpha_0 = e^{\gamma_0}, i = 1, \ldots, \ell.$$

Now, letting $\alpha_0 = \frac{\phi_0}{1-\phi_0}$ and $\alpha_1 = \frac{\phi_1}{1-\phi_1}$, we have

$$a_i = \frac{\phi_0 n_i^{\frac{\phi_1}{1-\phi_1}}}{1-\phi_0+\phi_0 n_i^{\frac{\phi_1}{1-\phi_1}}}, i = 1, \ldots, \ell, \tag{3}$$

where $0 < \phi_0, \phi_1 < 1$. Note that if $\phi_1 = 0$, then $a_i = \phi_0$ and there will be no dependence on the n_i. Now, simply substitute the a_i in (3) into the SAE model and use the prior

$$\phi_0, \phi_1 \overset{ind}{\sim} \text{Uniform}(0,1).$$

This reduces the number of parameters for this part of the model from $\ell+2$ to just two and actually the two parameters, θ and γ, are now eliminated or replaced by ϕ_0 and ϕ_1. So if ℓ is large, not just 8, there will be large gains in computational time. This is how the procedure is operationalized.

3.3. Percentiles

As we consider each area individually, we can drop the subscript, i, to get the population model

$$y_j \mid \underline{\beta}, \nu, \sigma^2 \overset{ind}{\sim} \text{Normal}(\underline{x}_j'\underline{\beta}+\nu, \sigma^2), j = 1, \ldots, N.$$

We recall that the nonsampled covariates are unknown. In principle, if we can get the nonsampled covariates, then, given $\underline{\beta}, \nu, \sigma^2$, we can sample $y_j, j = 1, \ldots, N$. Then, for $0 < \gamma < 1$, the $[\gamma N]$ percentile is $Y_{[\gamma N]}$, an order statistic. But this procedure is prohibitively expensive because the nonsampled covariates are unknown and N is large.

However, it is possible to obtain finite population percentiles (needed for BMI data) without the nonsampled covariates. For BMI, the 85^{th} and 95^{th} percentiles respectively measure overweight and obsesity. First, note that

$$P(Y_j < t_j \mid \nu, \underline{\beta}, \sigma^2) = \Phi\left\{\frac{t_j - \underline{x}_j'\underline{\beta}-\nu}{\sigma}\right\},$$

where $\Phi(\cdot)$ is the standard normal cdf. Therefore, with $0 < \gamma < 1$, the $100(1-\gamma)^{th}$ percentile of Y_j is $t_j = \underline{x}_j'\underline{\beta} + \nu + \sigma\Phi^{-1}(\gamma)$, Then, for the h^{th} iterate from the Gibbs sampler, the $100(1-\gamma)^{th}$ percentile of Y_j is

$$t_j^{(h)} = \underline{x}_j'\underline{\beta}^{(h)} + \nu^{(h)} + \sigma^{(h)}\Phi^{-1}(\gamma),$$

and the $100(1-\gamma)^{th}$ finite population percentile is $\frac{\sum_{j=1}^{n_2} W_{2j}t_j^{(h)}}{\sum_{j=1}^{n_2} W_{2j}}$. Some improvements can be made; actually such improvements are not necessary because N is very large, and like the finite population mean, the variance is approximately zero.

Walker (1968) showed that the sample γ-quantile, $Y_{([N\gamma])} \sim aN\{\epsilon_\gamma, \frac{\gamma(1-\gamma)}{Nf^2(\epsilon_\gamma)}\}$, where ϵ_γ is the γ^{th} quantile of the population, $f(\cdot)$, which is assumed to be continuous with $f(\epsilon_\gamma) > 0$ and $F(\epsilon_\gamma) = \gamma$ uniquely. Here, we simply take $\epsilon_\gamma = \frac{\sum_{j=1}^{n_{2i}} W_{2ij}t_{ij}^{(h)}}{\sum_{j=1}^{n_{2i}} W_{2ij}}$ and because the variance is $o(\frac{1}{N})$ and N is very large, essentially $Y_{([N\gamma])}$ is a point mass at ϵ_γ. A similar result holds for $\bar{Y}_i$.

One question is how to define $f(y)$. We write $y_j \mid \underline{\beta}, \nu, \sigma^2 \overset{ind}{\sim} \text{Normal}(\underline{x}_j'\underline{\beta} + \nu, \sigma^2), j = 1, \ldots, N$. Then, we replace $\underline{x}_j, j = 1, \ldots, N$, by the weighted average, $\underline{d} = \frac{\sum_{j=1}^{n_2} W_{2j}\underline{x}_j}{\sum_{j=1}^{n_2} W_{2j}}$, to get $y_j \mid \underline{\beta}, \sigma^2 \overset{ind}{\sim} \text{Normal}(\underline{d}'\underline{\beta} + \nu, \sigma^2), j = 1, \ldots, N$. Finally, $f(\epsilon_\gamma) = \frac{1}{\sigma}\phi(\frac{\epsilon_\gamma - \underline{d}'\underline{\beta} - \nu}{\sigma})$, where $\phi(\cdot)$ is the standard normal density.

4. Numerical example on small area estimation

We use the BMI data from the eight counties of California to construct a practical example; see Nandram and Choi (2010) for design issues in the National Health and Nutrition Examination Survey (NHANES III). We use Bayesian model diagnostics to compare all the models. Then, we compare our selected model with data integration and the ps only model via Bayesian predictive inference of the finite population mean and the 85^{th} finite population percentile.

But, first we discuss the performance of the Gibbs sampler for the model with discounting (the other models are similar). The entire computation consists of three parts (a) constructing the unknown survey weights for the nps, (b) fitting the individual area model, and (c) fitting of the small area model. The entire computation took nearly 40 minutes with (c) taking almost all the time. We started the Gibbs sampler arbitrarily by taking the a_i to be the corresponding posterior means from the individual area model, set $\rho = .5$, its mid range, and as the mid point of the interval $(\frac{\gamma}{1-\gamma}, \frac{1-2\gamma}{1-\gamma})$ is .5, set $\theta = .5$ and and $\gamma = 1/6$, its mid range. We ran $21,000$ iterates, used 1000 as a "burn in" and systematically selected every twentieth to get a 'random' sample of $M = 1,000$. We also performed the diagnostic procedures for the Gibbs sampler. The auto-correlations are not significant, the trace plots show no trend, Geweke tests of stationarity are all passed and the effective sample size are all satisfactory, mostly near to 1000. Table 1 has the p-values and the effective sample sizes. The fact that the effective sample size (ESS) is about 550, not 1000, for θ and γ is not a problem because θ and γ are hyperparameters of the a_i, which perform well.

In Table 2 we present diagnostic measures to compare the small area models. These are the negative log pseudo marginal likelihood (LPML), the deviance information criterion (DIC), the Bayesian predictive p-value (BPP), a divergence measure (DM) and the posterior root mean square error (PRMSE); see Appendix C for a review of the definitions of these measures. Smaller values of all quantities, except BPP, show better fit; values of BPP in $(.05, .95)$ show good fitting models.

All measures show that the model without discounting is not competitive, and DM and PRMSE show that the PS only model is not competitive, leaving us with two models, discounting and logit. In terms of PRMSE, the model with discounting is approximately 10% better than the logit model, which is not robust because it assumes linearity between the discounting factors and log sample sizes, thereby making the model with discounting the best. Also, the posterior standard deviations of the finite population means of the different areas under the model with discounting are at least as similar to those from the other models, better than the ps only model.

Table 1: Gibbs sampler diagnostics for the model with discounting using the BMI data of the eight counties

Parameter	n_1	n_2	Pval	ESS
a_1	140	24	0.804	1000
a_2	138	38	0.750	1000
a_3	667	128	0.395	1000
a_4	133	29	0.709	1000
a_5	96	29	0.813	1000
a_6	119	22	0.144	1000
a_7	100	28	0.332	884
a_8	137	39	0.447	1000
ρ	-	-	0.465	1000
θ	-	-	0.886	541
γ	-	-	0.473	545

NOTE: Pval is the p-value of the Geweke test and ESS is the effective sample size of the Gibbs sampler

Table 2: Comparison of five models using BMI data of eight counties

Model	LPML	DIC	BPP	DM	PRMSE
Discounting	977.491	1946.369	0.553	2.626	1.606
	(0.8)	(1.3)	(-)	(-2.0)	(-52.4)
Logit	975.866	1943.152	0.528	2.623	1.783
	(0.7)	(1.1)	(-)	(-2.2)	(-47.1)
No discounting	1235.930	2472.066	1.000	2.616	1.718
	(27.5)	(28.6)	(-)	(-2.5)	(-49.1)
No nps weights	978.573	1948.031	.541	2.597	1.521
	(0.9)	(1.3)	(-)	(-3.1)	(-54.9)
PS only	969.371	1922.219	0.493	2.682	3.373

NOTE: For PRMSE, the true value is taken to be the weighted average of all BMI values. The model with discounting is the one described, the logit model regresses the a_i on the logarithm of sample sizes, and the model without discounting has all a_i set to unity. The measures are calculated for PS data only. Gibbs sampling is needed for the models with discounting. Wang *et al.* (2011) has the divergence measure (DM). The parenthesis $(\cdot)$ shows the percent each measure is larger than the one for the ps. The model with discounting has PRMSE 9.9% smaller than the logit model.

Table 3 has posterior inference about the discounting factors. There are some discrimination among the small areas as the a_i range from .066 to .141. The posterior standard deviations are small making the CVs standing between .102 and .160 and so the inference is very precise and reliable. Consequently, the 95% HPDIs are reasonably tight. Therefore, as there is much discounting, the a_i are playing a consequential role in this application. Nandram and Rao (2021, 2023) gave interpretations of the discounting factor for a single area.

For comparisons, we use the following idea in Tables 4 & 5. For two standard deviations, a, b, assuming independence, $\max(a, b) \leq \sqrt{(a^2 + b^2)} \leq a + b$. That is, assuming independence of two sources, the standard deviation of the difference is at least the larger one.

In Table 4, we compare inference about the finite population means using integrated data and the probability sample only (ps only model). Note that the data from the nps are not used in the ps model only; see Appendix D for a discussion of the ps only model. As expected, there are large gains in precision over the ps only model when the model with discounting is used. Three of the PMs under the model with discounting are smaller than the corresponding ones under ps only model. Therefore, there is possibly some selection bias in the model with discounting. The 95% HPDIs for the nps have considerable overlaps on

Table 3: Posterior summaries of the discounting factors for BMI data of eight counties

County	n_1	n_2	PM	PSD	NSE	CV	95% HPDI
1	140	24	0.130	0.019	0.001	0.147	(0.097, 0.171)
2	138	38	0.066	0.010	0.000	0.160	(0.043, 0.085)
3	667	128	0.111	0.011	0.000	0.102	(0.091, 0.132)
4	133	29	0.112	0.017	0.001	0.149	(0.081, 0.146)
5	96	29	0.095	0.015	0.001	0.158	(0.069, 0.130)
6	119	22	0.141	0.022	0.001	0.155	(0.101, 0.184)
7	100	28	0.101	0.016	0.001	0.160	(0.071, 0.131)
8	137	39	0.099	0.015	0.000	0.148	(0.071, 0.126)

NOTE: The discounting factors, a_i, are small.

the right to those of the ps. Therefore, there is not much difference between the two models in terms of PMs.

In Table 5, we compare inference about the 85^{th} percentile of the finite population using the model with discounting and the probability sample only. Again, as expected, there are large gains in precision when the model with discounting is used. Three of the PMs under the model with discounting are smaller than the corresponding ones under ps model only. For each area, the intervals under the nps overlap considerably on the right of those for the ps. Therefore, again there is possibly some selection bias in the model with discounting. There are similar results for the 90^{th} and 95^{th} percentiles (not shown) with much larger variability, of course.

In Figures 1 & 2 we show plots of the posterior densities of the finite population means by county. For all counties, the model with discounting gives more precise estimates than the ps only model, and the plots overlap with various degrees, with the plot of the nps to the right of the ps, indicating some degree of selection bias remaining; five counties (2, 3, 4, 5, 7), appear similar. There appears to be no differences in sample size except for county 3 with very large county size (667, 128).

Table 4: Comparison of the nps model (with discounting) and the ps only model via posterior summaries of the finite population mean of the eight counties using the BMI data

County	n_1	n_2	Model	PM	PSD	NSE	CV	95% HPDI
1	140	24	nps	26.193*	0.293	0.009	0.011	(25.559, 26.704)
1	140	24	ps	25.229	0.450	0.014	0.018	(24.437 26.199)
2	138	38	nps	27.483*	0.299	0.010	0.011	(26.908, 28.052)
2	138	38	ps	27.100	0.363	0.010	0.013	(26.393 27.740)
3	667	128	nps	26.931*	0.149	0.006	0.005	(26.642, 27.219)
3	667	128	ps	26.769	0.222	0.006	0.008	(26.346 27.204)
4	133	29	nps	26.299	0.364	0.010	0.014	(25.593, 26.951)
4	133	29	ps	26.481	0.878	0.026	0.033	(24.535 28.090)
5	96	29	nps	27.017	0.355	0.011	0.013	(26.290, 27.652)
5	96	29	ps	27.416	0.521	0.017	0.019	(26.356 28.339)
6	119	22	nps	26.352*	0.299	0.008	0.011	(25.841, 26.954)
6	119	22	ps	25.102	0.469	0.013	0.019	(24.100 25.939)
7	100	28	nps	26.845*	0.305	0.010	0.011	(26.253, 27.389)
7	100	28	ps	26.467	0.416	0.014	0.016	(25.720 27.297)
8	137	39	nps	27.350	0.295	0.012	0.011	(26.789, 27.930)
8	137	39	ps	28.406	0.457	0.013	0.016	(27.530 29.276)

NOTE: Posterior inference is based on 1000 iterates that provide posterior mean, PM, posterior standard deviation, PSD, numerical standard error, NSE, coefficient of variation, CV, and 95% highest posterior density interval, HPDI. The PMs of the model with data discounting are larger than those under the PS only model by $3.8, 1.4, .6, -.7, -1.5, 5.0, 1.4, -3.7$ percent. PMs are larger for counties marked $(*)$.

Table 5: Comparison of the nps model (with discounting) and the ps only model via posterior summaries of the finite population 85^{th} percentile of the eight counties using the BMI data

County	n_1	n_2	Model	PM	PSD	NSE	CV	95% HPDI
1	140	24	nps	27.762*	0.311	0.011	0.011	(27.126, 28.309)
1	140	24	ps	26.724	0.455	0.015	0.017	(25.856, 27.649)
2	138	38	nps	29.036*	0.318	0.010	0.011	(28.392, 29.625)
2	138	38	ps	28.574	0.376	0.012	0.013	(27.846, 29.290)
3	667	128	nps	28.490*	0.169	0.006	0.006	(28.128, 28.790)
3	667	128	ps	28.255	0.235	0.006	0.008	(27.836, 28.774)
4	133	29	nps	27.859	0.378	0.011	0.014	(27.105, 28.553)
4	133	29	ps	27.955	0.827	0.022	0.030	(26.149, 29.415)
5	96	29	nps	28.580	0.351	0.012	0.012	(27.924, 29.302)
5	96	29	ps	28.908	0.505	0.018	0.017	(27.867, 29.300)
6	119	22	nps	27.932*	0.332	0.011	0.011	(27.268, 28.553)
6	119	22	ps	26.600	0.475	0.014	0.018	(25.700, 27.540)
7	100	28	nps	28.409*	0.323	0.012	0.011	(27.786, 29.013)
7	100	28	ps	27.934	0.429	0.011	0.015	(27.089, 28.756)
8	137	39	nps	28.905	0.297	0.009	0.010	(28.352, 29.494)
8	137	39	ps	29.913	0.422	0.015	0.014	(29.091, 30.726)

NOTE: Posterior inference is based on 1000 iterates that provide PM, posterior mean, PSD, posterior standard deviation, W, width of the 95% HPD interval and CV, coefficient of variation. PMs are larger for counties marked $(*)$.

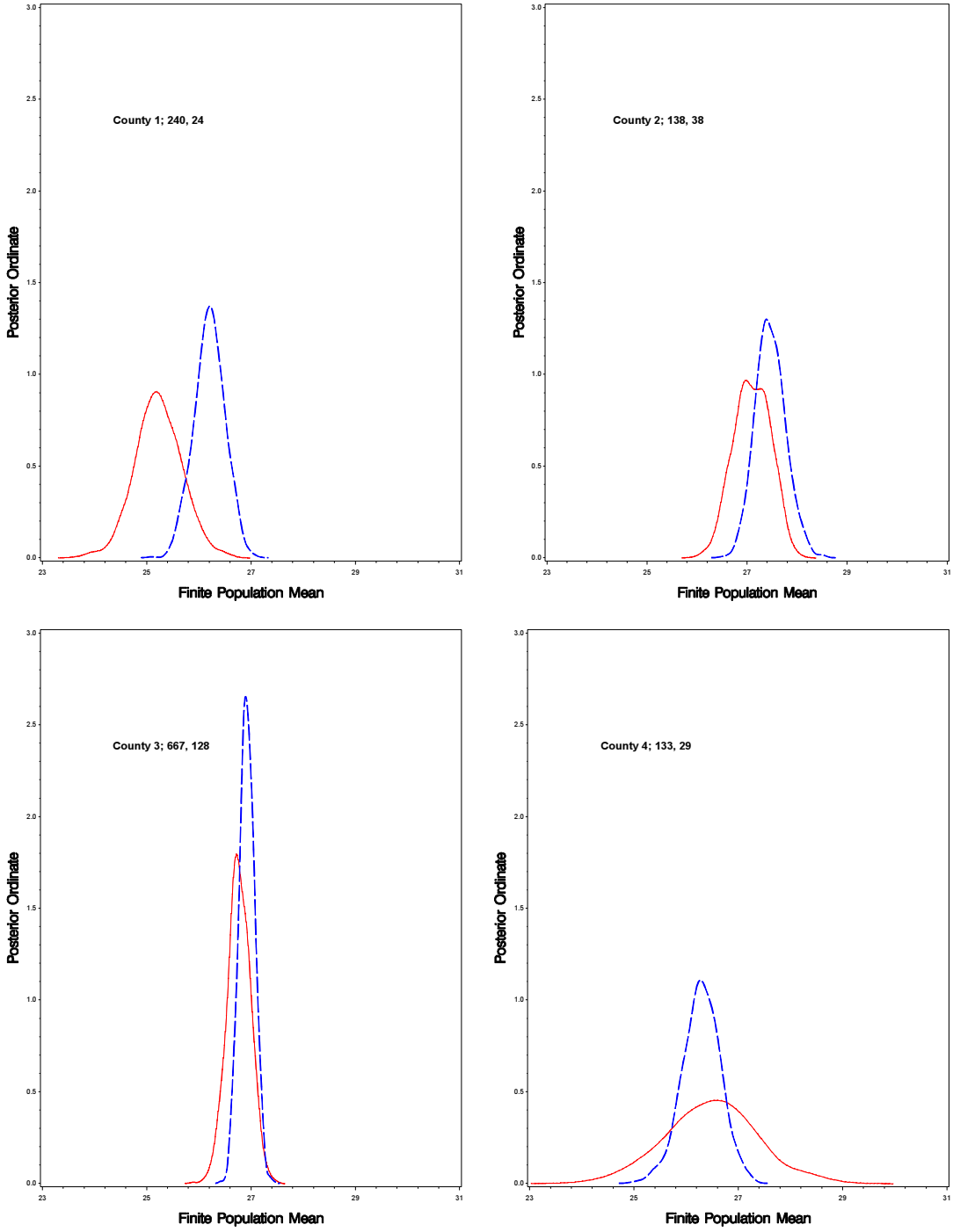


Figure 1: Comparison for the posterior distributions of the finite population mean for nps and ps models by county (dashed: discounting model; solid: ps only model)

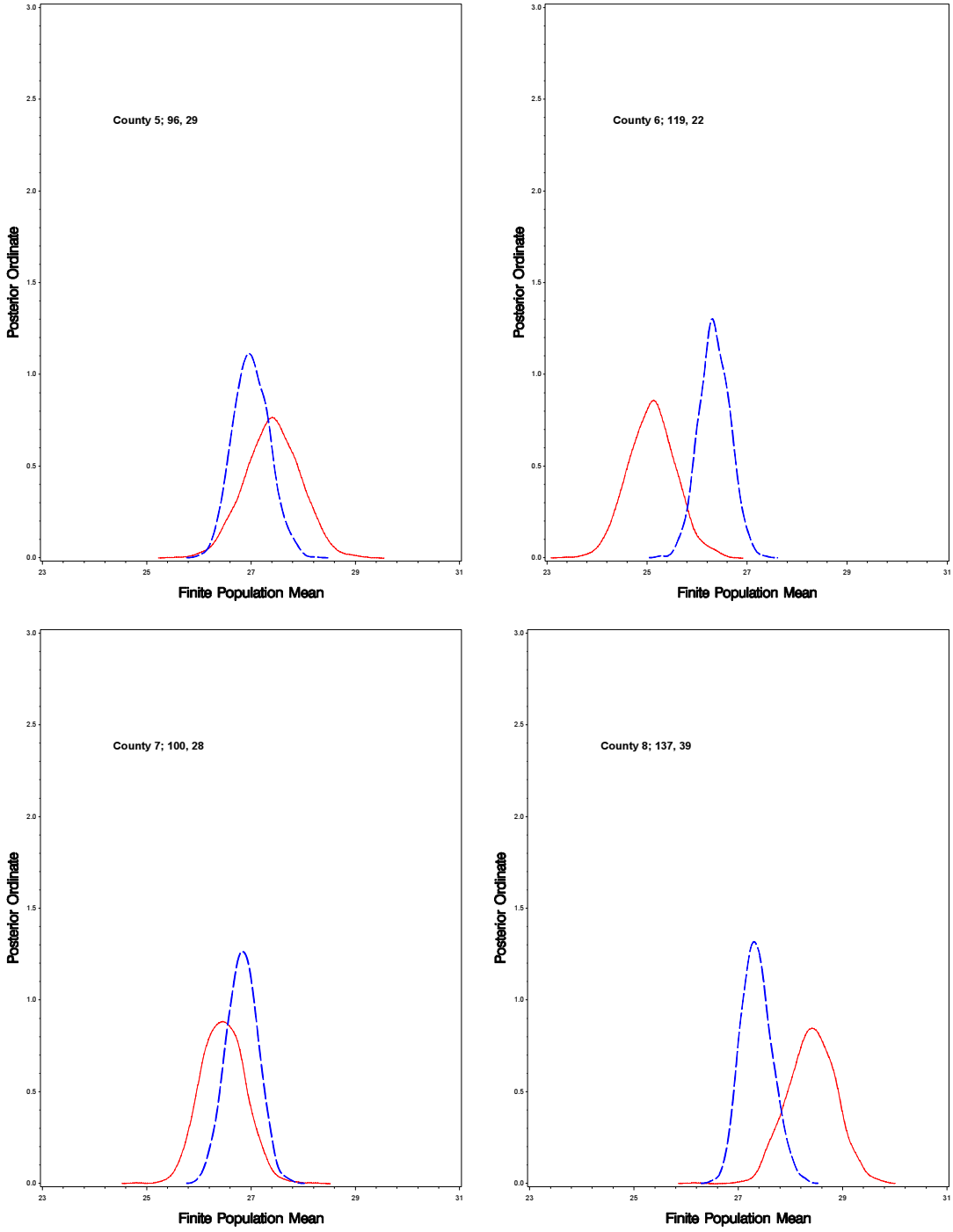


Figure 2: Comparison for the posterior distributions of the finite population mean for nps and ps models by county (dashed: discounting model; solid: ps only model)

5. Concluding remarks

This section has two subsections. The first subsection is a summary of the paper with general comments and the second subsection is on robustification of the models for the study variable and the participation variable.

5.1. Summary and comments

In our illustrative example on body mass index, our data-integrated model with discounting is preferred over the ps only model and other competitors. The logit data-integrated model is a strong competitor. The data-integrated model provides small area estimates, roughly similar to those of the ps only model, with larger precision. It is difficult to remove all biases completely. We outline some important problems we are currently working on, particularly how the assumptions on the participation variable and the study variable can be relaxed.

We have shown how to extend our approach to cover small area estimation. We have done so for the unit-level small area model (a bit less practical); this is an extension of Nandram and Rao (2021, 2023) to cover small areas. We have extended Toto and Nandram (2010) or Molina, Nandram and Rao (2014), who provided a Bayesian approach, to solve the problem without combining a nps and a ps. However, our work here was motivated by Beaumont (2020), Rao (2020) and Beaumont and Rao (2020) but these authors provided limited discussion of unit-level models; Beaumont and Rao (2020) showed how to use the area-level Fay-Herriot model to improve inference for the small areas in the ps, covariates being drawn from the nps (Big Data).

The assumption of normality on the BMI data is perhaps not a very good one because the BMI data are skewed (true for most continuous survey data) and discrete; see Yin and Nandram (2020 a,b) on how the Dirichlet process is used for BMI data without data integration. Also, more robust methods on propensity scores are needed. Stick-breaking priors can be used to provide more robust models, but these models are difficult to fit when all uncertainty is taken into account and this is on-going work; see Ishwaran and James (2001). It is also possible to use BART in data integration (*e.g.*, Rafei, *et al.* 2021). But BART is not a fully Bayesian procedure because it double-uses the data, it suffers from overshrinkage, and there is no underlying theory of BART (just a machine learning algorithm like random forest); see Hill, Linero and Murray (2020) for more detailed discussions and criticisms about BART. Yet, one does not need to express a relation between study variable and covariates; see Lockwood (2023, PhD Dissertation) for an important advance.

It is possible to avoid estimation of the survey weights of the non-probability sample by using a structural (measurement error) model; see Berg *et al.* (2021) for a start. We have been doing similar work at National Agriculural Statistics Service, USDA. For the nps (1), we consider

$$y_{1ij} \overset{ind}{\sim} \text{Normal}\left\{\gamma_0 + \gamma_1(\underline{x}'_{1ij}\underline{\beta} + \nu_i), \frac{\sigma^2}{a_i}\right\}, j = 1, \ldots, n_{1i}, i = 1, \ldots, \ell,$$

and for the ps (2),

$$y_{2ij} \overset{ind}{\sim} \text{Normal}\left\{\underline{x}'_{2ij}\underline{\beta} + \nu_i, \frac{\sigma^2}{w_{2ij}}\right\}, j = 1, \ldots, n_{2i}, i = 1, \ldots, \ell.$$

Here, γ_0 and γ_1 are weakly identified and can lead to poor performance of a Gibbs sampler. One can define the true values of y_{2ij} as $\theta_{ij} = \underline{x}'_{2ij}\underline{\beta} + \nu_i$. We do not need to estimate nps weights. Note again that n_{1i} is much larger than n_{2i}, and a discount factor is used to increase variability and help avoiding the nps to dominate the ps. Note that the parameters, $\underline{\beta}$, σ^2 and ν_i are the same in both the nps and the ps. Finally, a standard assumption on the area random effects is

$$\nu_i \mid \rho, \sigma^2 \overset{ind}{\sim} \text{Normal}\left\{0, \frac{\rho}{1-\rho}\sigma^2\right\}, i = 1, \ldots, \ell.$$

Of course, this can be overcome using the Pitman-Yor stick breaking procedure. Because of non-identifiability issues, we will assume that γ_0 and γ_1 are independent with

$$\gamma_0 \sim \text{Uniform}(c_1, c_2), \gamma_1 \sim \text{Uniform}(d_1, d_2),$$

where (c_1, c_2) and (d_1, d_2) are to be specified using exploratory data analysis. This can be done by fitting $\bar{y}_{1i} = \gamma_0 + \gamma_1 \bar{y}_{2i} + e_i, i = 1, \ldots, \ell$, and using the bootstrap distributions of the least squares estimators of γ_0 and γ_1 to get their ranges. For the a_i, we will assume the same prior as before, and we also assume that

$$\pi(\underline{\beta}, \sigma^2, \rho) \propto \frac{1}{\sigma^2}.$$

Also, as before prediction is done by using

$$y_{ij} \mid \nu_i, \underline{\beta}, \sigma^2 \overset{ind}{\sim} \text{Normal}(\underline{x}'_{ij}\underline{\beta} + \nu_i, \sigma^2), j = 1, \ldots, N_i, i = 1, \ldots, \ell,$$

and the prediction procedure is similar to the one done earlier. For

$$\bar{Y}_i \mid \nu_i, \underline{\beta}, \sigma^2 \overset{ind}{\sim} \text{Normal}(\underline{\bar{X}}'_i\underline{\beta} + \nu_i, \frac{\sigma^2}{N_i}), i = 1, \ldots, \ell,$$

where $\underline{\bar{X}}_i = \frac{\sum_{j=1}^{N_i} \underline{x}_{ij}}{N_i}$ is unknown and N_i may also be unknown. Design-based estimators of N_i and $\underline{\bar{X}}_i$ are respectively $N_i = \sum_{i=1}^{n_{2i}} W_{2ij}$ and $\underline{\bar{X}}_i = \frac{\sum_{j=1}^{n_{2i}} W_{2ij}\underline{x}_{2ij}}{N_i}$ (Hajek or Horvitz-Thompson). Inference for finite population percentiles is also possible.

5.2. Robustification

Looking towards double robustness as in non-Bayesian methods, we can use a mixture model for the study variable and a t-link for the participation variable of any number of areas within the Bayesian paradigm.

5.2.1. Robustification of the model of the study variable

For the study variable, we use a three-component mixture model. For the non-probability sample,

$$f(y_{1ij} \mid \nu_i, \underline{\beta}, p, q, \rho, \gamma) = (1-p-q)\text{Normal}_{y_{1ij}}(\underline{x}'_{1ij}\underline{\beta} + \nu_i, \frac{\rho\gamma\sigma^2}{aw_{1ij}})$$

$$+p\text{Normal}_{y_{1ij}}(\underline{x}'_{1ij}\underline{\beta} + \nu_i, \frac{\gamma\sigma^2}{aw_{1ij}}) + q\text{Normal}_{y_{1ij}}(\underline{x}'_{1ij}\underline{\beta} + \nu_i, \frac{\sigma^2}{aw_{1ij}}), i = 1, \ldots, n_{1i}.$$

and, for the probability sample, we have

$$f(y_{2ij} \mid \nu_i, \underline{\beta}, p, q, \rho, \gamma) = (1-p-q)\text{Normal}_{y_{2ij}}(\underline{x}'_{2ij}\underline{\beta} + \nu_i, \frac{\rho\gamma\sigma^2}{w_{2i}})$$

$$+p\text{Normal}_{y_{2ij}}(\underline{x}'_{2ij}\underline{\beta} + \nu_i, \frac{\gamma\sigma^2}{w_{2ij}}) + q\text{Normal}_{y_{2ij}}(\underline{x}'_{2ij}\underline{\beta} + \nu_i, \frac{\sigma^2}{w_{2ij}}), i = 1, \ldots, n_{2i}, i = 1, \ldots, \ell.$$

Finally,

$$\nu_i \mid \psi, \sigma^2 \overset{ind}{\sim} \text{Normal}(0, \frac{\psi}{1-\psi}\sigma^2), i = 1, \ldots, \ell.$$

It is also sensible to use the constraint $p > q$ and $0 < p, q, p+q, \rho, \gamma < 1$. In each case, the first component corresponds to ordinary observations, the second component corresponds to mild outliers and the third component to severe outliers. See Chakraborty, Datta, and Mandal (2019) for the much simpler two-component mixture model. There is on-going work on this topic.

5.2.2. Robustification of the model of the participation variable

We consider the following mixture model for the selection indicators, $r_i, i = 1, \ldots, N$, and we consider one large area (all areas combined). We make the robust assumption,

$$r_i \mid T = g, \underline{\theta} \overset{ind}{\sim} \text{Bernoulli}\{\mathcal{T}_{a_g}(\underline{z}'_i\underline{\theta})\}, i = 1, \ldots, N,$$

$$P(T = g \mid \lambda_g) = \lambda_g, g = 1, \ldots, G,$$

where $(a_g, \lambda_g), g = 1, \ldots, G$, and G are to be specified. We define the propensity scores as

$$\pi_i = \sum_{g=1}^{G} \lambda_g \mathcal{T}_{a_g}(\underline{z}'_i\underline{\theta}), i = 1, \ldots, N.$$

We can now develop a pseudo-density for each g and average all the pseudo-densities over g. Specifically, we have the mixture pseudo-density,

$$P(\underline{r} \mid \underline{z}, \underline{\theta}) = \sum_{g=1}^{G} \lambda_g \prod_{i=1}^{n_1} \left\{ \frac{\mathcal{T}_{a_g}(\underline{z}'_{1i}\underline{\theta})}{1 - \mathcal{T}_{a_g}(\underline{z}'_{1i}\underline{\theta})} \right\} \prod_{i=1}^{n_2} \left\{1 - \mathcal{T}_{a_g}(\underline{z}'_{2i}\underline{\theta})\right\}^{W_{2i}}, \tag{4}$$

where $\mathcal{T}_{a_g}, g = 1, \ldots, G$, is the Student's t cdf on a_g degrees of freedom. The estimated propensity scores we need are then

$$\hat{\pi}_i = \sum_{g=1}^{G} \lambda_g \mathcal{T}_{a_g}(\underline{z}'_{1i}\hat{\underline{\theta}}), i = 1, \ldots, n_1,$$

where $\hat{\underline{\theta}} = E(\underline{\theta} \mid \underline{r})$; it is possible to use other summaries as well (*e.g.*, the posterior median or the posterior mode).

This is a generalization of the logistic regression model, and it covers many cases (Cauchy, logistic and normal). It is well-known that when the Student's t density and/or the logistic distribution are appropriately rescaled, a plot of the quantiles of the Student's t density on roughly 8 degrees of freedom versus the quantiles of the logistic distribution is almost a 45^o straight line through the origin. Here $\lambda_g, g = 1, \ldots, G$, are specified weights at degrees of freedom $a_g, g = 1, \ldots, G$, and to look at variation around the logistic distribution, we can place more probability at $a_g = 8$. For example, we have used $a_g = 1, 4, 8, 13, 20, 30, 40, 50$ for $G = 8$, $a_g = 40, 50$ will be close to a standard normal density, and $\lambda_g = .125, .125, .25, .125, .125, .125, .080, .045$. There is on-going work on this topic.

Acknowledgments

Balgobin Nandram was supported by a grant from the Simons Foundation (#353953, Balgobin Nandram) and J. N. K. Rao was supported by a research grant from the Natural Sciences and Engineering Research Council of Canada. Balgobin Nandram presented invited talks on this topic at the SAE2022 conference, University of Maryland, and at Banaras Hindu University, India, February 2023. The reviewers helped to improve the presentation.

APPENDIX A: Propensity scores

Let $\underline{x_i}, i = 1, \ldots, N$, denote the covariates; these are observed in the ps and the nps, but they are not observed for the rest of the population. Again, for the nps, we have $\underline{x}_{1i}, i = 1, \ldots, n_1$, and for the ps, we have $\underline{x}_{2i}, i = 1, \ldots, n_2$. Chen, Li and Wu (2020) has a method to get the propensity scores for the nps, and therefore the survey weights, which they defined as the reciprocals of the propensity scores. They assume that the propensity scores can be modeled parametrically using

$$\pi_i = P(r_i = 1 \mid \underline{x}_i) = \pi(\underline{x}_i; \underline{\theta}),$$

with independence over i, where $\underline{\theta}$ are to be estimated. Here $r_i = 1$ for the ps or nps; $r_i = 0$ for the nonsamples. Then, the likelihood function is

$$\ell(\underline{\theta}) = \prod_{i=1}^{N} \{\pi(\underline{x}_i; \underline{\theta})\}^{r_i} \{1 - \pi(\underline{x}_i; \underline{\theta})\}^{1-r_i}.$$

The propensity scores are obtained in two steps.

First, they wrote the log-likelihood as

$$\ell^*(\underline{\theta}) = \sum_{i=1}^{n_1} \log\left\{\frac{\pi(\underline{x}_{1i};\underline{\theta})}{1-\pi(\underline{x}_{1i};\underline{\theta})}\right\} + \sum_{i=1}^{N} \log\{1-\pi(\underline{x}_i;\underline{\theta})\}.$$

Second, they used the pseudo-log-likelihood by replacing the second term by the Horvitz-Thompson estimator since the nonsample $\underline{x}_i$ are unknown, as

$$\ell^*(\underline{\theta}) = \sum_{i=1}^{n_1} \log\left\{\frac{\pi(\underline{x}_{1i};\underline{\theta})}{1-\pi(\underline{x}_{1i};\underline{\theta})}\right\} + \sum_{i=1}^{n_2} W_{2i} \log\{1-\pi(\underline{x}_{2i};\underline{\theta})\},$$

which can now be maximized for $\hat{\underline{\theta}}$. The propensity scores for the nps are then $\pi(\underline{x}_{1i};\hat{\underline{\theta}}), i = 1,\ldots,n_1$. Henceforth, they specialize to logistic regression.

One caveat is that the propensity scores are not really selection probabilities (*i.e.*, quasi-randomization). This is true because the propensity scores must be obtained for the entire population (*i.e.*, all N units) and then calibrated to the nps sample size. Only in this case, quasi-randomization makes any sense at all. This is still an open problem. Also, they assumed ignorability (given the covariates, the participation variable is independent of the study variable), but see Nandram (2022) for nonignorability. Chen, Li and Wu (2020) did not assume non-ignorability because they assumed that the study variable is missing in the probability sample; they need to mass impute the the missing values, but this is not in the spirit of their work.

APPENDIX B: Computation for the small area model

We discuss how to fit the proposed model. Recall $\Omega_1 = (\underline{a}, \theta, \gamma, \rho)$ and $\Omega_2 = (\underline{\nu}, \underline{\beta}, \sigma^2)$. Our strategy is to integrate out Ω_2 from $\pi(\Omega_1, \Omega_2 \mid \underline{y})$ to get $\pi(\Omega_1 \mid \underline{y})$ and then sample $\pi(\Omega_1 \mid \underline{y})$ using the Griddy-Gibbs sampler (Ritter and Tanner, 1992).

For convenience, we will keep $a_{si}, s = 1,2, i = 1,\ldots,\ell$, free in $(0,1)$ and sometimes $a_{1i} = a_i$ and $a_{2i} = 1$, $i = 1,\ldots,\ell$. Then, letting $n = \sum_{s=1}^{2}\sum_{i=1}^{\ell} n_{si}$, the total number of observations,

$$\pi(\Omega_1, \Omega_2 \mid \underline{y}) \propto \pi(\Omega_1)\left(\prod_{i=1}^{\ell}\sqrt{a_i}\right) \times$$

$$\left(\frac{1}{\sigma^2}\right)^{\frac{n+\ell}{2}+1}\left(\frac{1-\rho}{\rho}\right)^{\ell/2}\prod_{i=1}^{\ell}\left[e^{-\frac{1}{2\rho\sigma^2}\left\{\rho\sum_{s=1}^{2}\sum_{j=1}^{n_{si}} a_{si}w_{sij}(y_{sij}-\underline{x}'_{sij}\underline{\beta}-\nu_i)^2+(1-\rho)\nu_i^2\right\}}\right]. \tag{B.1}$$

We will integrate out Ω_2. Momentarily, we will drop $\pi(\Omega_1)$, but we will retain $\prod_{i=1}^{\ell}\sqrt{a_i}$.

Define the following quantities,

$$\lambda_i = \frac{\rho\sum_{s=1}^{2}\sum_{j=1}^{n_{si}} a_{si}w_{sij}}{\rho\sum_{s=1}^{2}\sum_{j=1}^{n_{si}} a_{si}w_{sij} + (1-\rho)}, \quad \phi_{sij} = \frac{a_{si}w_{sij}}{\sum_{s=1}^{2}\sum_{j=1}^{n_{si}} a_{si}w_{sij}},$$

$$\bar{\bar{y}}_i = \sum_{s=1}^{2}\sum_{j=1}^{n_{si}} \phi_{sij}y_{sij}, \quad \underline{\bar{\bar{x}}}_i = \sum_{s=1}^{2}\sum_{j=1}^{n_{si}} \phi_{sij}\underline{x}_{sij},$$

$$\tilde{y}_{sij} = y_{sij} - \bar{\bar{y}}_i, \quad \tilde{\underline{x}}_{sij} = \underline{x}_{sij} - \bar{\bar{\underline{x}}}_i.$$

Note that while the λ_i are functions of ρ, but the ϕ_{sij}, $\bar{\bar{y}}_i$ and $\bar{\bar{\underline{x}}}_i$ are not functions of ρ.

We can now rewrite the exponent in (B.1),

$$\exp\left\{-\frac{1}{2\sigma^2}\left\{\sum_{s=1}^{2}\sum_{j=1}^{n_{si}} a_{si}w_{sij}(y_{sij} - \underline{x}'_{sij}\underline{\beta} - \nu_i)^2 + \frac{1-\rho}{\rho}\nu_i^2\right\}\right\},$$

as

$$\exp\left\{-\frac{1}{2\sigma^2}\left\{\sum_{s=1}^{2}\sum_{j=1}^{n_{si}} a_{si}w_{sij}(\tilde{y}_{sij} - \tilde{\underline{x}}'_{sij}\underline{\beta})^2 + \frac{1-\rho}{\rho}(\sum_{s=1}^{2}\sum_{j=1}^{n_{si}} a_{si}w_{sij})(\bar{\bar{y}}_i - \bar{\bar{\underline{x}}}'\underline{\beta} - \nu_i)^2\right\}\right\}.$$

Then, it is easy to show that

$$\nu_i \mid \underline{\beta}, \sigma^2, \rho, \underline{y} \overset{ind}{\sim} \text{Normal}\{\hat{\nu}_i, \frac{\rho}{1-\rho}\sigma^2(1-\lambda_i)\}, i = 1, \ldots, \ell,$$

where $\hat{\nu}_i = \lambda_i(\bar{\bar{y}}_i - \bar{\bar{\underline{x}}}'_i\underline{\beta})$. This is a standard form in small area estimation and it combines both the probability sample and the non-probability sample over all areas; note the common $\underline{\beta}$ and σ^2.

Then, integrating out the ν_i from (B.1), we have

$$\pi(\underline{\beta}, \sigma^2, \rho \mid \underline{y}) \propto \left(\frac{1}{\sigma^2}\right)^{\frac{n}{2}+1} \prod_{i=1}^{\ell}\sqrt{a_i(1-\lambda_i)}$$

$$\times \prod_{i=1}^{\ell}\left[\exp\left\{-\frac{1}{2\sigma^2}\left\{\sum_{s=1}^{2}\sum_{j=1}^{n_{si}} a_{si}w_{sij}(\tilde{y}_{sij} - \tilde{\underline{x}}'_{sij}\underline{\beta})^2 + P_i(\bar{\bar{y}}_i - \bar{\bar{\underline{x}}}'_i\underline{\beta})^2\right\}\right\}\right], \tag{B.2}$$

where

$$P_i = \left(\sum_{s=1}^{2}\sum_{j=1}^{n_{si}} a_{si}w_{sij}\right)(1-\lambda_i)^2 + \frac{1-\rho}{\rho}\lambda_i^2, i = 1, \ldots, \ell.$$

Then,

$$\underline{\beta} \mid \sigma^2, \rho, \underline{y} \sim \text{Normal}\{\hat{\underline{\beta}}, \sigma^2\Delta\},$$

where

$$\Delta = \left\{\sum_{i=1}^{\ell}\sum_{s=1}^{2}\sum_{j=1}^{n_{si}} a_{si}w_{sij}\tilde{\underline{x}}_{sij}\tilde{x}'_{sij} + \sum_{i=1}^{\ell} P_i\bar{\bar{\underline{x}}}_i\bar{\bar{\underline{x}}}'_i\right\}^{-1}$$

and

$$\hat{\underline{\beta}} = \left\{\sum_{i=1}^{\ell}\sum_{s=1}^{2}\sum_{j=1}^{n_{si}} a_{si}w_{sij}\tilde{\underline{x}}_{sij}\tilde{x}'_{sij} + \sum_{i=1}^{\ell} P_i\bar{\bar{\underline{x}}}_i\bar{\bar{\underline{x}}}'_i\right\}^{-1}\left\{\sum_{i=1}^{\ell}\sum_{s=1}^{2}\sum_{j=1}^{n_{si}} a_{si}w_{sij}\tilde{\underline{x}}_{sij}\tilde{y}_{sij} + \sum_{i=1}^{\ell} P_i\bar{\bar{\underline{x}}}_i\bar{\bar{y}}_i\right\}.$$

Then integrating $\underline{\beta}$ from (B.2), we have

$$\pi(\sigma^2, \rho \mid \underline{y}) \propto \left(\frac{1}{\sigma^2}\right)^{\frac{n-p}{2}+1} \mid \Delta \mid^{1/2} \prod_{i=1}^{\ell}\sqrt{a_i(1-\lambda_i)}$$

$$\times e^{-\frac{1}{2\sigma^2}\left\{\sum_{i=1}^{\ell}\sum_{s=1}^{2}\sum_{j=1}^{n_{si}} a_{si}w_{sij}\{\tilde{y}_{sij}-\underline{\tilde{x}}'_{sij}\underline{\hat{\beta}}\}^2+\sum_{i=1}^{\ell} P_i(\bar{\tilde{y}}_i-\underline{\bar{\tilde{x}}}'_i\underline{\hat{\beta}})^2\right\}}. \tag{B.3}$$

Therefore,

$$\sigma^2 \mid \rho, \underline{y} \sim \text{InvGam}\left\{\frac{n-p}{2}, \frac{Q}{2}\right\}, \tag{B.4}$$

where

$$Q = \sum_{i=1}^{\ell}\sum_{s=1}^{2}\sum_{j=1}^{n_{si}} a_{si}w_{sij}\{\tilde{y}_{sij} - \underline{\tilde{x}}'_{sij}\underline{\hat{\beta}}\}^2 + \sum_{i=1}^{\ell} P_i(\bar{\tilde{y}}_i - \underline{\bar{\tilde{x}}}'_i\underline{\hat{\beta}})^2.$$

Integrating out σ^2 from (B.3), we have

$$\pi(\rho \mid \underline{y}) \propto \frac{\mid \Delta \mid^{1/2} \prod_{i=1}^{\ell}\sqrt{a_i(1-\lambda_i)}}{Q^{(n-p)/2}}, 0 \le \rho \le 1. \tag{B.5}$$

Actually $\pi(\rho \mid \Omega_1, \underline{y})$ is defined for all values of ρ in $[0,1]$ because the P_i and λ_i are well defined for all values of ρ in $[0,1]$. Note that the a_i are constants (given) above, specifically they are constants in (B.5).

Bringing back $\pi(\Omega_1)$ into the picture, we have

$$\pi(\Omega_1 \mid \underline{y}) \propto \pi(\Omega_1)\pi(\rho \mid \underline{y}),$$

and therefore,

$$\pi(\Omega_1 \mid \underline{y}) \propto \frac{\mid \Delta \mid^{1/2} \prod_{i=1}^{\ell}\sqrt{a_i(1-\lambda_i)}}{Q^{(n-p)/2}} \left\{\prod_{i=1}^{\ell} \frac{a_i^{\theta(\frac{1-\gamma}{\gamma})-1}(1-a_i)^{(1-\theta)(\frac{1-\gamma}{\gamma})-1}}{B\{\theta(\frac{1-\gamma}{\gamma}), (1-\theta)(\frac{1-\gamma}{\gamma})\}}\right\}, \tag{B.6}$$

$\frac{\gamma}{1-\gamma} \le \theta \le \frac{1-2\gamma}{1-\gamma}, 0 < \gamma < 1/3, 0 \le \rho \le 1$. It is worth noting that the a_i are not independent; Δ and Q contain all the a_i, which is contained by λ_i also.

In (B.6), $\pi(\Omega_1 \mid \underline{y})$ is well defined for all values of $\underline{a}, \theta, \gamma, \rho$ because $0 < a_i < 1, i = 1, \ldots, \ell$, $0 < \rho < 1$, $0 < \gamma < \frac{1}{3}$, $\frac{\gamma}{1-\gamma} \le \theta \le \frac{1-2\gamma}{1-\gamma}$. Therefore, it follows that the joint posterior density $\pi(\Omega_1, \Omega_2 \mid \underline{y})$ is proper. Next, we present the rather obvious conditional posterior densities (CPDs) necessary to run the Gibbs sampler.

First, we consider the CPD of the $a_i, i = 1, \ldots, \ell$. Letting $\underline{a}_{(i)} = (a_1, \ldots, a_{i-1}, a_{i+1}, \ldots, a_\ell)', i = 1, \ldots, \ell$ (a_i is eliminated), then for $0 < a_i < 1$,

$$\pi(a_i \mid \underline{a}_{(i)}, \rho, \theta, \gamma, \underline{y}) \propto \frac{\mid \Delta \mid^{1/2} \prod_{i=1}^{\ell}\sqrt{a_i(1-\lambda_i)}}{Q^{(n-p)/2}} \left\{\prod_{i=1}^{\ell} a_i^{\theta(\frac{1-\gamma}{\gamma})-1}(1-a_i)^{(1-\theta)(\frac{1-\gamma}{\gamma})-1}\right\}. \tag{B.7}$$

Second, the CPD of ρ is

$$\pi(\rho \mid \underline{a}, \theta, \gamma, \underline{y}) \propto \frac{\mid \Delta \mid^{1/2} \prod_{i=1}^{\ell}\sqrt{(1-\lambda_i)}}{Q^{(n-p)/2}}, 0 < \rho < 1. \tag{B.8}$$

Third, the joint CPD of (θ, γ) is

$$\pi(\theta, \gamma \mid \underline{a}, \rho, \underline{y}) \propto \left\{ \prod_{i=1}^{\ell} \frac{a_i^{\theta(\frac{1-\gamma}{\gamma})-1}(1-a_i)^{(1-\theta)(\frac{1-\gamma}{\gamma})-1}}{B\{\theta(\frac{1-\gamma}{\gamma}), (1-\theta)(\frac{1-\gamma}{\gamma})\}} \right\}, \frac{\gamma}{1-\gamma} \leq \theta \leq \frac{1-2\gamma}{1-\gamma}, 0 < \gamma < 1/3. \tag{B.9}$$

$\frac{\gamma}{1-\gamma} \leq \theta \leq \frac{1-2\gamma}{1-\gamma}, 0 < \gamma < 1/3$. The CPD of θ or γ is easy to write down.

We note that all the CPDs are nonstandard, but all the parameters lie in $(0, 1)$, so we have used a grid method, with 100 grid points, to sample each of the CPDs. The number grid points can be reduced for the a_i perhaps to 50 or so, but we need the number grid points to be around 100 for (ρ, θ, γ); hyperparameters are more difficult to sample. We have done this, and we have reduced the entire computation time from roughly 40 minutes to 20 minutes with little change in the results.

APPENDIX C: Bayesian model diagnostics and measures

We test concordance of the ps (2) part of the model,

$$y_{2ij} \mid \nu_i, \underline{\beta}, \sigma^2 \overset{ind}{\sim} \text{Normal}(\underline{x}'_{2ij}\underline{\beta} + \nu_i, \frac{\sigma^2}{W_{2ij}}), j = 1, \ldots, n_{2i}, i = 1, \ldots, \ell,$$

with the observed data of the ps (2). It is not sensible to study concordance with the observed data of the nps (1) because they are biased. The posterior density of $(\underline{\nu}, \underline{\beta}, \sigma^2)$ comes from their respective models. We describe five Bayesian measures, which are the negative log-pseudo marginal likelihood (LPML), the deviance information criterion (DIC), the Bayesian predictive p-value (BPP), the divergence measure (DM) and the posterior root mean squared error (PRMSE); LPML and DM are based on Bayesian cross-validation.

First, the conditional posterior ordinate is $CPO_{ij} = f(y_{2ij} \mid \underline{y}_{(2ij)})$, where $\underline{y}_{(2ij)}$ is the vector of all values excluding $y_{(2ij)}$. Let $\underline{\Omega} = (\underline{\nu}', \underline{\beta}', \sigma^2)'$ and $\underline{\Omega}^{(h)}$ denote the h^{th} iterate from the Gibbs sampler of the appropriate parameters. Then, CPO_{ij} is usually estimated by

$$\widehat{CPO}_{ij} = \left[\frac{1}{M} \sum_{h=1}^{M} \frac{1}{f(y_{2ij} \mid \underline{\Omega}^{(h)})} \right]^{-1},$$

the harmonic mean, and $LPML = \sum_{i=1}^{\ell} \sum_{j=1}^{n_{2i}} \log(\widehat{CPO}_{ij})$.

Second, letting $\hat{\underline{\Omega}}$ denote the posterior mean of $\underline{\Omega}$, the DIC is

$$DIC = 2\hat{D}(\underline{y}) - D(\underline{y} \mid \hat{\underline{\Omega}}),$$

where $D(\underline{y} \mid \underline{\Omega}) = -2\log\{f(\underline{y} \mid \underline{\Omega})\}$ and $\hat{D}(\underline{y}) = \frac{1}{M}\sum_{h=1}^{M} D(\underline{y} \mid \underline{\Omega}^{(h)})$.

Third, letting T_2 denote a test (discrepancy) function, the BPP is

$$P(T_2^{rep} > T_2^{obs} \mid \underline{y}^{obs}),$$

where we have used

$$T_2 = \sum_{i=1}^{\ell} \sum_{j=1}^{n_{2i}} W_{2ij} \frac{(y_{2ij} - \underline{x}'_{2ij}\underline{\beta} - \nu_i)^2}{\sigma^2}.$$

Fourth, the divergence measure is

$$DM = \frac{1}{\sum_{i=1}^{\ell} n_{2i}} \sum_{i=1}^{\ell} \sum_{j=1}^{n_i} |y_{2ij} - E(y_{2ij}|\underline{y}_{(2ij)})|;$$

see Wang *et al.* (2012).

Fifth, letting $T = \sum_{s=1}^{2} \sum_{i=1}^{\ell} \sum_{j=1}^{n_{si}} W_{sij} y_{sij} / \sum_{s=1}^{2} \sum_{i=1}^{\ell} \sum_{j=1}^{n_{si}} W_{sij}$,

$$PRMSE = \sqrt{\sum_{i=1}^{\ell} \{(PM_{2i} - T)^2 + PSD_{2i}^2\}},$$

where $PM_{2i} = \mathrm{E}(\bar{Y}_{2i}|\underline{y}_1, \underline{y}_2)$ and $PSD_{2i}^2 = \mathrm{Var}(\bar{Y}_{2i}|\underline{y}_1, \underline{y}_2)$.

APPENDIX D: Adding survey weights into the bayesian BHF model

We describe how to fit the ps only model. This is essentially adding survey weights to the BHF model.

The population model is

$$y_{ij} \mid \nu_i, \underline{\beta}, \rho \overset{ind}{\sim} \text{Normal}\{\underline{x}'_{ij}\underline{\beta} + \nu_i, (1-\rho)\sigma^2\}, j = 1, \ldots, N_i,$$

where $\underline{x}_{ij}$ has p components, including an intercept, and

$$\nu_i \mid \sigma^2, \rho \overset{ind}{\sim} \text{Normal}(0, \rho\sigma^2), i = 1, \ldots, \ell.$$

The reparameterization with respect to ρ is similar, but slightly different, to the one we have used before. The correlation of the values within an area is ρ, and the model is defined for all values of ρ in $[0, 1]$. Let $\bar{Y}_i = \frac{1}{N_i} \sum_{j=1}^{N_i} y_{ij}$, the finite population mean of the i^{th} area, and let $\underline{\bar{X}}_i$ denote the finite population mean covariate vector.

Therefore,

$$\bar{Y}_i \mid \nu_i, \underline{\beta}, \rho \overset{ind}{\sim} \text{Normal}\{\underline{\bar{X}}'_i\underline{\beta} + \nu_i, \frac{(1-\rho)\sigma^2}{N_i}\}.$$

Then, integrating out the ν_i, we have

$$\bar{Y}_i \mid \underline{\beta}, \sigma^2, \rho \overset{ind}{\sim} \text{Normal}\{\underline{\bar{X}}'_i\underline{\beta}, \rho\sigma^2 + \frac{(1-\rho)\sigma^2}{N_i}\}.$$

So that σ^2 has a direct impact in prediction even when the N_i are very large, and ρ plays an important role here. This is different from the case when there is just a single sample (*i.e.*, no random effects), where for large N_i, the variance is approximately 0.

The sample model is

$$y_{ij} \mid \nu_i, \underline{\beta}, \rho \overset{ind}{\sim} \text{Normal}\{\underline{x}'_{ij}\underline{\beta} + \nu_i, \frac{(1-\rho)\sigma^2}{w_{ij}}\}, j = 1, \ldots, n_i,$$

$$\nu_i \mid \sigma^2, \rho \overset{ind}{\sim} \text{Normal}(0, \rho\sigma^2), i = 1, \ldots, \ell,$$

$$\pi(\underline{\beta}, \sigma^2, \rho) \propto \frac{1}{\sigma^2}.$$

Letting $n = \sum_{i=1}^{\ell} n_i$, the total number of observations over the ℓ small areas, the joint posterior density is

$$\pi(\underline{\nu}, \underline{\beta}, \sigma^2, \rho \mid \underline{y}) \propto \frac{1}{(\sigma^2)^{(n+\ell)/2+1}} \frac{1}{(\rho)^{\ell/2}} \frac{1}{(1-\rho)^{n/2}}$$

$$\times \exp\left[-\frac{1}{2\rho(1-\rho)\sigma^2}\left\{\rho\sum_{j=1}^{n_i} w_{ij}(y_{ij} - \underline{x}'_{ij}\underline{\beta} - \nu_i)^2 + (1-\rho)\nu_i^2\right\}\right]. \tag{D.1}$$

We will decompose $\pi(\underline{\nu}, \underline{\beta}, \sigma^2, \rho \mid \underline{y})$ as

$$\pi(\underline{\nu}, \underline{\beta}, \sigma^2, \rho \mid \underline{y}) = \pi_1(\underline{\nu} \mid \underline{\beta}, \sigma^2, \rho, \underline{y})\pi_2(\underline{\beta} \mid \sigma^2, \rho, \underline{y})\pi_3(\sigma^2 \mid \rho, \underline{y})\pi_4(\rho \mid \underline{y}),$$

where $\pi_1(\underline{\nu} \mid \underline{\beta}, \sigma^2, \rho, \underline{y})$, $\pi_2(\underline{\beta} \mid \sigma^2, \rho, \underline{y})$, $\pi_3(\sigma^2 \mid \rho, \underline{y})$, except $\pi_4(\rho \mid \underline{y})$, are all in standard forms. Next, we will demonstrate this decomposition, and at the same time, we will prove propriety of the joint posterior density.

For $i = 1, \ldots, \ell$, let $\bar{\underline{x}}_i = \frac{\sum_{j=1}^{n_i} w_{ij}\underline{x}_{ij}}{\sum_{j=1}^{n_i} w_{ij}}$, $\bar{y}_i = \frac{\sum_{j=1}^{n_i} w_{ij}y_{ij}}{\sum_{j=1}^{n_i} w_{ij}}$, and $\lambda_i = \frac{\rho\sum_{j=1}^{n_i} w_{ij}}{\rho\sum_{j=1}^{n_i} w_{ij}+1-\rho}$. Note that the λ_i are not functions of σ^2. Then, it is easy to show that

$$\nu_i \mid \underline{\beta}, \sigma^2, \rho, \underline{y} \overset{ind}{\sim} \text{Normal}\{\hat{\nu}_i, (1-\lambda_i)\rho\sigma^2\}, i = 1, \ldots, \ell,$$

where $\hat{\nu}_i = \lambda_i(\bar{y}_i - \bar{\underline{x}}'_i\underline{\beta})$.

Let $t_{ij} = y_{ij} - \lambda_i\bar{y}_i$ and $\underline{d}_{ij} = \underline{x}_{ij} - \lambda_i\bar{\underline{x}}_i, i = 1, \ldots, \ell$. Then, integrating ν_i from (D.1), we have

$$\pi(\underline{\beta}, \sigma^2, \rho \mid \underline{y}) \propto \frac{1}{(\sigma^2)^{n/2+1}} \frac{1}{(1-\rho)^{n/2}} \prod_{i=1}^{\ell} \sqrt{1-\lambda_i}$$

$$\times \exp\left[-\frac{1}{2\rho(1-\rho)\sigma^2}\left\{\rho\sum_{i=1}^{\ell}\sum_{j=1}^{n_i} w_{ij}(t_{ij} - \underline{d}'_{ij}\underline{\beta})^2 + (1-\rho)\sum_{i=1}^{\ell}\lambda_i^2(\bar{y}_i - \bar{\underline{x}}'_i\underline{\beta})^2\right\}\right]. \tag{D.2}$$

Now, let

$$\hat{\underline{\beta}} = \Delta\left\{\rho\sum_{i=1}^{\ell}\sum_{j=1}^{n_i} w_{ij}\underline{d}_{ij}t_{ij} + (1-\rho)\sum_{i=1}^{\ell}\lambda_i^2\bar{\underline{x}}_i y_i\right\},$$

where

$$\Delta^{-1} = \rho\sum_{i=1}^{\ell}\sum_{j=1}^{n_i} w_{ij}\underline{d}_{ij}\underline{d}'_{ij} + (1-\rho)\sum_{i=1}^{\ell}\lambda_i^2\bar{\underline{x}}_i\bar{\underline{x}}'_i.$$

Note that $\hat{\beta}$ does not depend on σ^2. Then, it is easy to show that

$$\underline{\beta} \mid \sigma^2, \rho, \underline{y} \sim \text{Normal}(\underline{\hat{\beta}}, \rho(1-\rho)\sigma^2\Delta).$$

Now, integrating $\underline{\beta}$ from (D.2), we have

$$\pi(\sigma^2, \rho \mid \underline{y}) \propto \frac{1}{(\sigma^2)^{(n-p)/2+1}} \frac{|\Delta|}{(1-\rho)^{(n-p)/2}} \rho^{p/2} \prod_{i=1}^{\ell} \sqrt{1-\lambda_i}$$

$$\times \exp\left[-\frac{1}{2\rho(1-\rho)\sigma^2}\left\{\rho \sum_{i=1}^{\ell}\sum_{j=1}^{n_i} w_{ij}(t_{ij} - \underline{d}'_{ij}\underline{\hat{\beta}})^2 + (1-\rho)\sum_{i=1}^{\ell} \lambda_i^2(\bar{y}_i - \underline{\bar{x}}'_i\underline{\hat{\beta}})^2\right\}\right]. \quad \text{(D.3)}$$

Finally, it follows easily that

$$\sigma^2 \mid \rho, \underline{y} \sim \text{InvGam}\left\{\frac{n-p}{2}, \frac{\rho\sum_{i=1}^{\ell}\sum_{j=1}^{n_i} w_{ij}(t_{ij} - \underline{d}'_{ij}\underline{\hat{\beta}})^2 + (1-\rho)\sum_{i=1}^{\ell}\lambda_i^2(\bar{y}_i - \underline{\bar{x}}'_i\underline{\hat{\beta}})^2}{2\rho(1-\rho)}\right\}$$

and integrating σ^2 from (D.3), we have

$$\pi(\rho \mid \underline{y}) \propto \prod_{i=1}^{\ell} \frac{1}{(\rho\sum_{j=1}^{n_i} w_{ij} + 1 - \rho)^{1/2}}$$

$$\times \frac{\rho^{n/2}(1-\rho)^{\ell/2}|\Delta|^{1/2}}{\{\rho\sum_{i=1}^{\ell}\sum_{j=1}^{n_i} w_{ij}(t_{ij} - \underline{d}'_{ij}\underline{\hat{\beta}})^2 + (1-\rho)\sum_{i=1}^{\ell}\lambda_i^2(\bar{y}_i - \underline{\bar{x}}'_i\underline{\hat{\beta}})^2\}^{(n-p)/2}}, 0 < \rho < 1. \quad \text{(D.4)}$$

Note that $\pi(\rho \mid \underline{y})$ is defined for all values of $\rho \in [0,1]$; we only need Δ to be well defined, and this is true because $\sum_{i=1}^{\ell}\sum_{j=1}^{n_i} \underline{d}_{ij}\underline{d}'_{ij}$ is full rank for all values of ρ (*i.e.*, the matrix $(\underline{x}'_{ij})$ is full rank provided that it has at least p linearly independent rows). Of course, $n > p$ as in standard regression problems. Therefore, the joint posterior density is proper.

References

Berg, E., Im, J., Zhu, Z., Lewis-Beck, C., and Li, J. (2021). Integration of statistical and administrative agricultural data from Namibia. *Statistical Journal of the IAOS*, **37**, 557–578, DOI: 10.3233/SJI-200634.

Battese, G. E., Harter, R., and Fuller, W. A. (1988). An error-components model for prediction of county crop Areas using survey and satellite data. *Journal of the American Statistical Association*, **83**, 28-36.

Beaumont, J-F. (2020). Are probability surveys bound to disappear for the production of official statistics? *Survey Methodology*, **46**, 1-28.

Beaumont, J-F. and Rao, J. N. K. (2021). Pitfalls of making inferences from non-probability samples: Can data integration through probability samples provide remedies? *The Survey Statistician*, **83**, 11-22.

Chakraborty, A., Datta, G. S., and Mandal, A. (2019). A robust hierarchical bayes small area estimation for nested error linear regression model. *International Statistical Reviews*, **87**, S1, S158-S156, DOI: 10.1111/insr.12283.

Chambers, R. L., Fabrizi, E., and Salvati, N. (2019). Small area estimation with linked data. *arXiv:1904.0036v1 [Stat.ME], 31 March 2019*, pg. 1-29.

Chen, Y., Li, P., and Wu, C. (2020). Doubly robust inference with nonprobability survey samples. *Journal of the American Statistical Association*, **115**, 2011-2021, DOI: 10.1080/01621459.2019.1677241.

Chen, S., Yang, S., and Kim, J. K. (2022). Nonparametric mass imputation for data integration. *Journal of Survey Statistics and Methodology.* **10**, 1-24, DOI: 10.1093/jssam/smaa036.

Elliott, M. N. and A. Haviland (2007). Use of a web-based convenience sample to supplement a probability sample. *Survey Methodology*, **33**, 211–215.

Elliott, M. R. and Valliant, R. (2017). Inference for nonprobability samples. *Statistical Science*, **32**, 249–264, DOI: 10.1214/16-STS598.

Hill, J., Linero, A., and Murray, J. (2020). Bayesian additive regression trees: A review and look forward. *Annual Review of Statistics and Its Applications*, **7**, 251-278, DOI: 10.1146/annurev-statistics-031219-041110.

Ibrahim, J. G. and Chen, M-H. (2000). Power prior distributions for regression models. *Statistical Science*, **15**, 46-60, DOI: 10.1214/ss/1009212673.

Ibrahim, J. G., Chen, M-H., Gwon, Y., and Chen, F. (2015). The power prior: Theory and applications. *Statistics in Medicine*, **34**, 3724-3749, DOI: 10.1002/sim.6728.

Ishwaran, H. and James, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association*, **96**, 161-173.

Kim, J. K., Park, S., Chen, Y., and Wu, C. (2021). Combining Non-probability and Probability Survey Samples Through Mass Imputation. *Journal of the Royal Statistical Society*, Series A, **184**, 941-963.

Lockwood, A. (2023). *Bayesian predictive inference for a study variable without specifying a link to the Covariates.* PhD Dissertation, Department of Mathematical Sciences, Worcester Polytechnic Institute, pg. 1-110.

Marchetti, S., Giusti, C., Pratesi, M., Salvati, N., Giannotti, F., Pedreschi, D., Rjnzivillo, S., Pappalardo, L., and Gabrjelli, L. (2015). Small area model-based estimators using big data sources. *Journal of Official Statistics*, **31**, 263-281.

Marella, D. (2023). Adjusting for selection bias in non-probability samples by empirical likelihood approach. *Journal of Official Statistics*, **39**, 2023, 151-172, DOI: 10.2478/JOS-2023-0008.

Meng, X-L (2018). Statistical paradises and paradoxes in big data (I): Law of large populations, big data paradox, and the 2016 US presidential election. *The Annals of Applied Statistics*, **12**, 685–726, DOI: 10.1214/18-AOAS1161SF.

Molina, I., Nandram, B., and Rao, J. N. K., (2014). Small area estimation of general parameters with application to poverty indicators: A hierarchical bayes approach. *The Annals of Applied Statistics*, **8**, 852-885, DOI: 10.1214/13-AOAS702.

Muhyi, F. A., Sartono, B., Sulvianti, I. D., and Kurnia, A. (2019). Twitter utilization in application of small area estimation to estimate electability of candidate central java governor. *IOP Conference Series in Earth and Environmental Science*, **299** 012033, 1-10.

Nandram, B. (2022). A Bayesian assessment of non-ignorable selection of a non-probability Sample. *Indian Bayesians' News Letter, Invited Paper*, 1-16.

Nandram, B. (2007). Bayesian predictive inference under informative sampling via surrogate samples. In *Bayesian Statistics and Its Applications, Eds. S.K. Upadhyay, Umesh Singh and Dipak K. Dey*, Anamaya, New Delhi, Chapter **25**, 356-374.

Nandram, B. and Choi, J. W. (2010). A Bayesian analysis of body mass index data from small domains under nonignorable nonresponse and selection. *Journal of the American Statistical Association*, **105**, 120-135.

Nandram, B., Choi, J. W., and Liu, Y. (2021). Integration of nonprobability and probability samples via survey weights. *International Journal of Statistics and Probability*, **10**, 4-17, DOI: 10.5539/ijsp.v10n6p5.

Nandram, B. and Rao, J. N. K (2021). A bayesian approach for integrating a small probability sample with a nonprobability sample. *Proceedings of the American Statistical Association, Survey Research Methods Section*, 1568-1603.

Nandram, B. and Rao, J. N. K (2023). Bayesian predictive inference when integrating a nonprobability sample and a probability sample. *arXiv:2305.08997v1 [Stat.ME], 15 May 2023*, pg. 1-35.

Nandram, B. (2023). Overcoming challenges associated with early bayesian state estimation of planted acres in the United States. Special Proceedings: Society of Statistics, Computing and Applications, ISBN #: 978-81-950383-2-9, 25th Annual Conference, 15-17 February 2023; pp 51-78.

Porter, A. T., Holan, S. H., Wikle, C. K., and Cressie, N. (2014). Spatial Fay-Herriot model for small area estimation with functional covariates. *Spatial Statistics*, **10**, 27-42.

Potthoff, R. F., Woodbury, M. A., and Manton, K. G. (1992). "Equivalent sample size" and "equivalent degrees of freedom" refinements for inference using survey weights under superpopulation models. *Journal of the American Statistical Association*, **87**, 383-396.

Rafei, A., Flannagan,C. A. C., West, B. T., and Elliott, M. R. (2022). Robust bayesian inference for big data: Combining sensor-based records with traditional survey. *arxiv:2101.07456V2 [Stat.ME], 26 March 2022*, pp. 1-58.

Rao, J. N. K. (2020). On making valid inferences by integrating data from surveys and other sources. *Sankhya*, Series B, 3-33, DOI: 10.1007/s13571-020-00227-w.

Ritter, C. and Tanner, M. A. (1992). Facilitating the Gibbs sampler: The Gibbs stopper and the Griddy-Gibbs sampler. *Journal of the American Statistical Association*, **87**, 861-868, DOI: 10.2307/2290225.

Sakshaug, J. W., Wisniowski, A., Ruiz, D. A. P., and Blom, A. G. (2019). Supplementing small probability samples with nonprobability samples: A bayesian approach. *Journal of Official Statistics*, **35**, 653-681, DOI: 10.1093/jssam/smad041.

Salvatore, C., Biffignandi, S., Sakshaug, J. W., Wisniowski, A., and Struminskaya, B. (2023). Bayesian integration of probability and non-probability samples for logistic regression. *Journal of Survey Statistics and Methodology*, 00,1-35, DOI: 10.1093/jssam/smad041.

Schmid, T., Bruckschen, F., Salvati, N., and Zbiranski, T. (2017). Constructing sociodemographic indicators for national statistical institutes by using mobile phone data: Estimating literacy rates in Senegal. *Journal of the Royal Statistical Society*, Series A, **180**, 1163-1190, DOI: 10.1111/rssa.12305.

Toto, M. C. S. and Nandram, B. (2010). A bayesian predictive inference for small area means incorporating covariates and sampling weights. *Journal of Statistical Planning and Inference*, **140**, 2963-2979, DOI: 10.1016/j.jspi.2010.03.043.

Yin, J. and Nandram, B. (2020a). A bayesian small area model with dirichlet processes on responses. *Statistics in Transition, New Series*, **21**, 1-19, DOI: 10.21307/stattrans-2020-041.

Yin, J. and Nandram, B. (2020b). A nonparametric bayesian analysis of response data with gaps, outliers and ties. *Statistics and Applications, New Series*, **18**, 121-141, ISSN 2452-7395(online).

Walker, A. M. (1968). A note on the asymptotic distribution of the sample quantiles. *Journal of the Royal Statistical Society*, Series B, **30**, 570-575.

Wang, J. C., Scott, H. H., Nandram, B., Barboza, W., Toto, C., and Anderson, E. (2012). A bayesian approach to estimating agricultural yield based on multiple repeated surveys. *Journal of Agricultural, Biological, and Environmental Statistics*, **17**, 84-106, D0I: 10.107/513253-011-0067-5.

Wang, Z., Kim, J. K., and Yang, S. (2018). Approximate bayesian inference under informative sampling. *Biometrika*, **105**, 91-10, DOI: 10.1093/biomet/asx073.

Wisniowski, A., Sakshaug, J. W., Ruiz, D. A. P., and Blom, A. G. (2020). Integrating probability and nonprobability samples for survey inference. *Journal of Survey Statistics and Methodology*, **8**, 120-147, DOI: 10.1093/jssam/smz051.

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